Statistical tests for analysing particle trajectories : application to intracellular imaging

Vincent Briane

ANNÉE 2017

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Traitement du signal et Télécommunications

Ecole doctorale MATHSTIC

présentée par

Vincent Briane

préparée au centre Inria Rennes - Bretagne Atlantique

Tests statistiques pour l'analyse de trajectoires de particules : Application à l'imagerie intracellulaire.

Thèse soutenue à Rennes
le 20 décembre 2017

devant le jury composé de :

Xavier DESCOMBES
Directeur de recherche, INRIA Sophia
rapporteur

Agnès DESOLNEUX
Directrice de recherche, CMLA-ENS Cachan
rapporteure

Delphine BLANKE
Professeur des universités, Université d'Avignon
examinatrice

Maxime DAHAN
Directeur de recherche, Institut Curie
examinateur

Valentin PATILEA
Professeur des universités, ENSAI
examinateur

Myriam VIMOND
Maître de conférence, ENSAI
co-directrice de thèse

Charles KERVRANN
Directeur de recherche, INRIA Rennes
directeur de thèse

Remerciements

Avant toute chose, je tiens à remercier mes deux directeurs de thèse Charles Kervrann et Myriam Vimond. Charles Kervrann m'a donné l'opportunité de réaliser cette thèse en statistiques appliquées à l'imagerie intracellulaire, un domaine alors nouveau pour moi que j'ai très vite apprécié. Grâce à son dynamisme, j'ai pu présenter tout au long de ma thèse mes travaux dans différentes conférences nationales et internationales, des expériences très enrichissantes à tout point de vue. Myriam Vimond, quant à elle, m'a permis de développer mes propres idées tout en gardant une rigueur mathématique irréprochable. Avec son aide, j'ai approfondi mon bagage en statistiques et acquis une plus grande rigueur dans la modélisation mathématique.

Je souhaite aussi remercier Valentin Patilea qui a rendu possible la collaboration entre l'Inria et l'Ensai-Crest, collaboration sans laquelle cette thèse n'aurait pas vu le jour. J'en profite aussi pour remercier Marian Hristache qui m'a permis d'assurer mes travaux dirigés à l'Ensai en toute sérénité, toujours prêt à m'aider en cas de problème.

Je remercie maintenant l'ensemble des équipes Serpico et Fluminance, et spécialement Huguette, toujours à l'écoute pour résoudre d'épineux problèmes logistiques. Je remercie aussi les postdocs et doctorants avec qui j'ai partagé de bons moments et de (longues?) pauses cafés tout au long de mon doctorat. Un merci tout particulier à Antoine, Cédric, Abed, Sandeep, Pranav, Benoît, Yvan, Pierre et Emmanuel.

Je remercie enfin mes amis: les éternels du lycée, Quentin, Lucie, Coralie, Thibault et Maxime; François pour des vacances hivernales 2016 aussi mémorables que nécessaires; Abdou, Lauric, Erwann et Thibault, mes compagnons de l'Ensai; tout le club du stade rennais athlétisme, avec une mention spéciale pour Antoine, avec qui j'ai pu me vider la tête tous les mardis et jeudis soirs à l'entraînement (pour la reremplir inexorablement le lendemain!)

Enfin, je remercie de tout cœur mon père, ma mère, ma soeur et mon frère.

Contents

List of Figures

List of Tables

Résumé

L'objet de cette thèse est l'étude quantitative du mouvement des particules intra-cellulaires (e.g., biomolécules). La cellule est un environnement complexe composé d'une multitude de structures inter-connectées. Ces structures échangent de la matière organique directement via le cytosol ou par l'intermédiaire de filaments comme les microtubules, les filaments d'actine ou encore les filaments intermédiaires. La dynamique des particules échangées détermine l'organisation et les fonctions cellulaires [Bressloff, 2014, Chapter 9]. Ainsi, l'estimation du mouvement des particules au sein de la cellule est d'un intérêt majeur en biologie cellulaire puisqu'elle permet de quantifier précisément les interactions entre les différents composants de la cellule.

Plusieurs techniques de microscopie quantitative permettent d'analyser le mouvement des particules intra-cellulaires. Les plus populaires sont le Recouvrement de Fluorescence Après Photoblanchiment (FRAP), la Spectroscopie à Corrélation de Fluorescence (FCS) et enfin le suivi de particules individuelles "Single Particle Tracking" (SPT). Les techniques FRAP et FCS reposent sur l'analyse moyenne d'un grand nombre de particules tandis que la méthode SPT permet de révéler des comportements individuels à une résolution moléculaire. Dans cette thèse, nous étudierons le mouvement individuel de particules correspondant à des biomolécules dans le contexte biophysique. Nous adoptons une approche lagrangienne puisque nous analysons la trajectoire d'une même particule au cours du temps. Ce concept est à distinguer du paradigme eulérien qui étudie le mouvement local moyen des particules dans une région au cours du temps.

Nous modélisons les trajectoires des particules avec des processus stochastiques puisque le milieu intra-cellulaire est soumis à de nombreux aléas. Les diffusions – processus à trajectoires continues– permettent de modéliser un large panel de mouvements intra-cellulaires. Les diffusions sont très souvent étudiées en biophysique [Qian et al., 1991, Saxton and Jacobson, 1997]. Dans la littérature, on distingue ainsi quatre principaux types de diffusion: le mouvement brownien, la super-diffusion, la diffusion confinée et la diffusion anomale. Des trajectoires représentatives de ces quatre groupes de diffusion sont illustrées sur la Figure 0.1. Ces différents types de mouvement correspondent à des scénarios biologiques distincts. Le déplacement d'une particule évoluant sans contrainte dans le cytosol est modélisé par un mouvement brownien; la particule ne se déplace pas dans une direction précise et atteint sa destination en un temps long en moyenne. Les particules (appelées dans ce contexte "cargos"), sont propulsées par des moteurs moléculaires le long des microtubules et des filaments d'actine qui constituent le cytosquelette de la cellule. Leurs mouvements sont alors modélisées par des

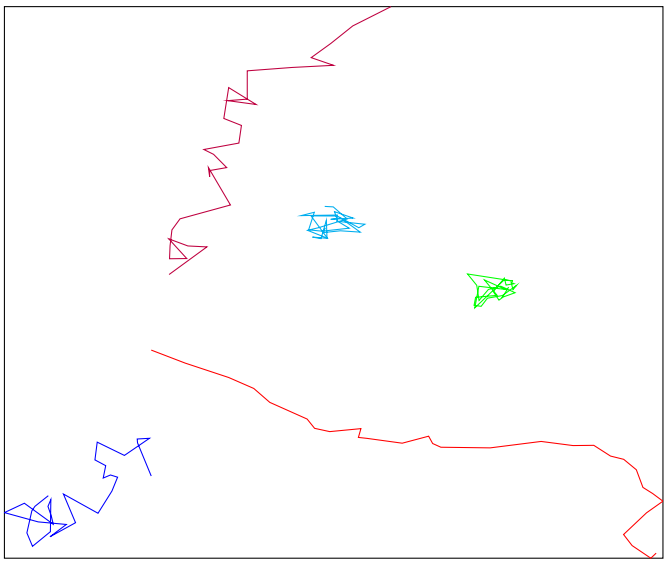


Figure 0.1: Trajectoires simulées représentant les différents types de diffusion. La trajectoire bleue est brownienne; la trajectoire violette est générée par un mouvement brownien avec dérive constante (4.2.3) et illustre la super-diffusion; la trajectoire rouge est issue d'un mouvement brownien fractionnaire de paramètre $\mathfrak{h} > 1/2$ (2.5.7) et correspond à une super-diffusion; la trajectoire cyan simulée avec un processus d'Ornstein-Uhlenbeck (3.3.14) est un exemple de diffusion confinée; la trajectoire verte issue d'un mouvement brownien fractionnaire de paramètre $\mathfrak{h} < 1/2$ (2.5.7) est associée à une diffusion anomale. Les paramètres des processus ci-dessus sont reportés dans le Tableau (5.1)

super-diffusions.

La diffusion confinée [Metzler and Klafter, 2000, Hoze et al., 2012] correspond à la situation où les particules sont bloquées dans des microdomaines. Quand une particule se fraye un chemin dans un milieu encombré, son mouvement est modélisé par une diffusion de type anomale [Saxton, 1994, Berry and Chaté, 2014]. Par la suite, nous ne ferons pas de distinction entre diffusion anomale et diffusion confinée. Nous les rassemblons sous une terminologie unique dite "sous-diffusion". En effet, dans cette thèse, nous nous plaçons dans un contexte non-paramétrique dans lequel la distinction entre diffusion confinée et anomale ne peut être réalisée pour des trajectoires courtes. Un exemple de notre classification selon les trois groupes de diffusions considérés est présenté sur la Figure 0.2.

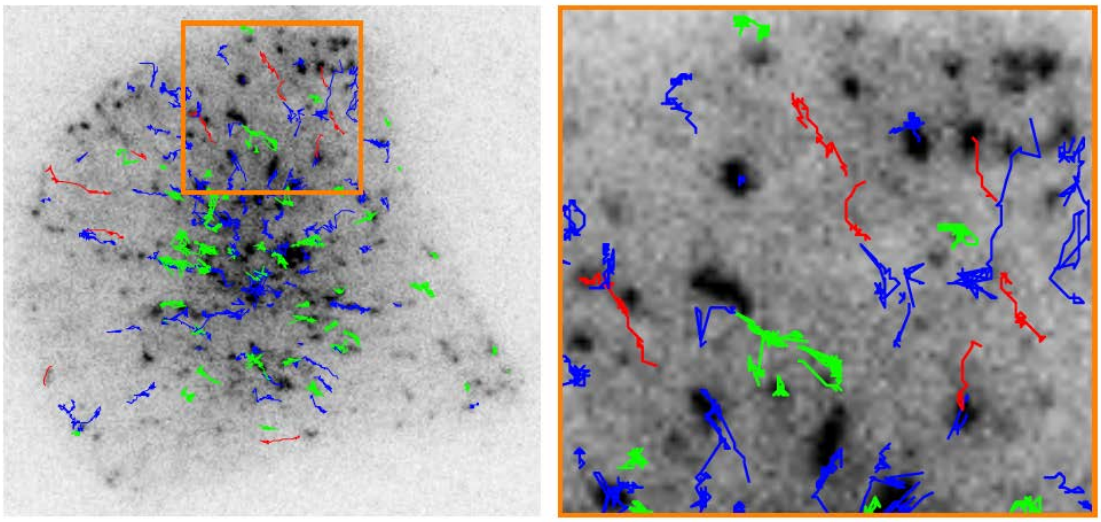


Figure 0.2: Classification de trajectoires 2D de la protéine Rab11a. La séquence est obtenue par microscopie TIRF (en collaboration avec l' UMR 144 CNRS Institut Curie PICT IBiSA). Nous avons utilisé notre test à trois décisions au niveau $\alpha = 5\%$ (voir Chapitre 4). Les trajectoires browniennes sont labellisées en bleu, les trajectoires sous-diffusives en vert et les super-diffusives en rouge.

Problématique

Déplacement carré moyen

Dans la littérature biophysique, les différents types de diffusion sont caractérisés par le déplacement carré moyen "Mean Square Displacement" (MSD) [Qian et al., 1991]. Soit $(X_t)_{t>0}$ le processus décrivant la trajectoire d'une particule. La courbe MSD est une fonction du temps définie comme suit,

$$\mathrm{MSD}(t) = \mathbb{E}\left(\|X_{t+t_0} - X_{t_0}\|^2\right), \tag{0.0.1}$$

où $\|\cdot\|$ est la norme euclidienne et $\mathbb{E}(\cdot)$ est l'espérance sur l'espace probabilisé. La fonction MSD du mouvement brownien est linéaire ($\mathrm{MSD}(t) \propto t$). Cette propriété remarquable explique la popularité du MSD. Une représentation schématique des courbes MSD de la sous-diffusion et de la super-diffusion est donnée sur la Figure 0.3. En pratique, nous observons les positions successives d'une même particule $X_{t_0}, X_{t_1}, \ldots, X_{t_n}$ en 2 ou 3 dimensions à intervalles de temps réguliers, c'est-à-dire que l'on a $t_{i+1} - t_i = \Delta$. On estime la fonction MSD au temps $j\Delta$, où le décalage j est un entier, par:

$$\widehat{\mathrm{MSD}}(j\Delta) = \frac{1}{n-j+1}\sum_{k=0}^{n-j}\|X_{t_{k+j}} - X_{t_k}\|^2. \tag{0.0.2}$$

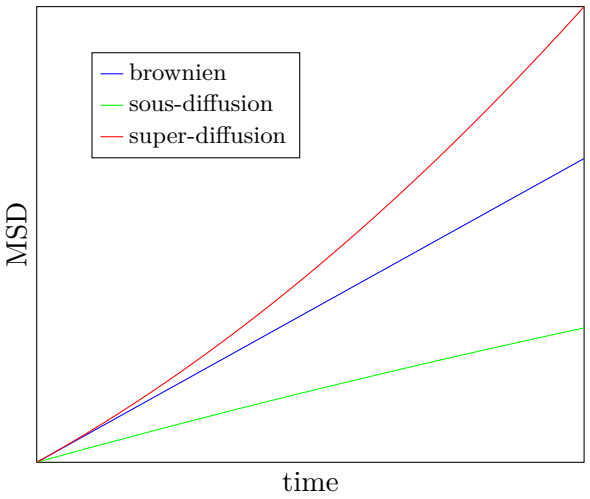


Figure 0.3: Schéma des courbes MSD typiques des différents types de diffusion.

Les estimations successives (0.0.2) du MSD à différents décalages j constituent une estimation de la fonction MSD (0.0.1). Une méthode simple, largement utilisée en biophysique, consiste à ajuster la courbe obtenue à la fonction $t \to t^{\beta}$. Ainsi, Feder et al. [1996] déclare que la trajectoire est sous-diffusive si $\beta < 0.9$, super-diffusive si $\beta > 1.1$ et brownienne si $0.9 < \beta < 1.1$ (voir Figure 0.4). Si $\beta < 0.1$, la trajectoire correspond à un mouvement nul.

Limites

Le critère du MSD connaît néammoins certaines limites. Tout d'abord, en tant que moment d'ordre deux, il ne permet pas de caractériser complètement la dynamique d'une trajectoire. C'est la raison pour laquelle Gal et al. [2013] présentent d'autres statistiques qui peuvent être associées au MSD pour analyser des trajectoires.

Le second problème majeur du MSD est lié à son estimation. La variance de l'estimateur (0.0.1) augmente avec le décalage j. Ce problème est illustré sur la Figure 0.4 dans le cas de trajectoires browniennes. Les résultats exposés suggèrent que la classification de Feder et al. [1996], basée sur le paramètre β, tend à classer une trajectoire brownienne dans le groupe de la sous-diffusion ou de la super-diffusion engendrant une erreur de classification. De plus, la variance du MSD est aussi affectée à des décalages j plus petits par la présence de bruit due à l'erreur de localisation. Pour tenir compte de ces erreurs, Michalet and Berglund [2012] détaillent une méthode itérative pour déterminer le nombre optimal de décalages j à utiliser en présence de bruit et rendre ainsi l'estimation plus robuste.

D'autres méthodes exploitant la fonction MSD ont été proposées ces dernières années.

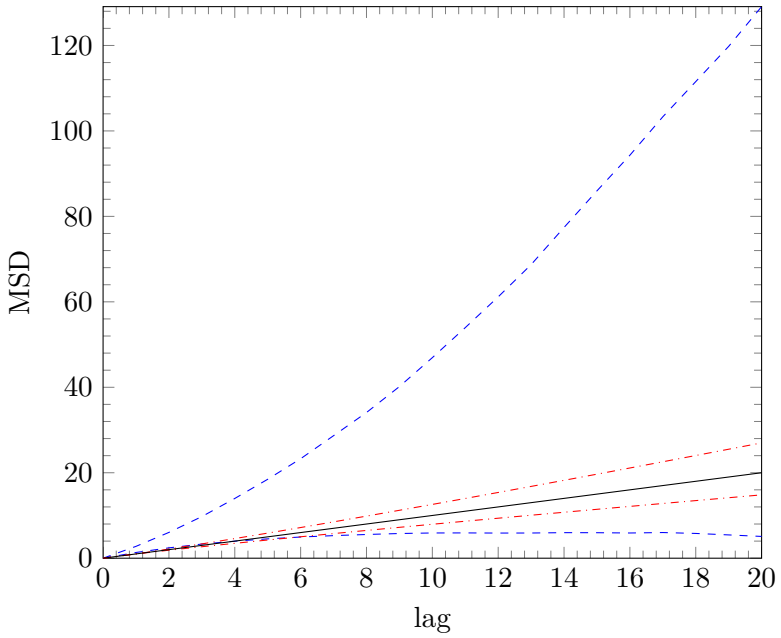


Figure 0.4: Règle de classification pour détecter les différents modes de diffusion à partir du MSD. Les courbes rouges sont les courbes MSD limites définies par Feder et al. [1996] $t \to t^{\beta}$, $\beta = 0.9$ et 1.1.. Les courbes bleues constituent un intervalle ponctuel de probabilité 95% associé au MSD empirique calculé sur des trajectoires browniennes de taille $n = 30$. Les courbes correspondent aux quantiles d'ordre 2.5% et 97.5% de (0.0.2) et sont calculées par simulation de Monte Carlo à partir de 10 001 trajectoires browniennes de taille $n = 30$.

Lund et al. [2014] proposent un arbre de décision pour sélectionner le meilleur modèle de mouvement qui combine le MSD, le critère d'information bayésien (BIC) et le rayon de gyration. Lysy et al. [2016] présentent une inférence basée sur un calcul de vraisemblance pour distinguer deux modèles de sous-diffusion : le mouvement brownien fractionnaire contre un processus solution de l'équation de Langevin généralisée. Les auteurs considèrent un modèle bayésien pour estimer les paramètres de diffusion et recourent au facteur de Bayes pour comparer les modèles.

Afin de réduire la variabilité de l'estimateur du MSD, certains auteurs calculent les courbes MSD sur un ensemble de trajectoires indépendentes plutôt que sur une seule. Ces trajectoires peuvent avoir des longueurs différentes mais leurs dynamiques sont supposées être identiques. Par exemple, Pisarev et al. [2015] considèrent un estimateur des moindres carrés pondéré pour β, la pondération étant calculée à partir de la variance du MSD. Les auteurs procèdent ensuite à une sélection de modèles basée sur le critère d'Akaike modifié. Monnier et al. [2012] proposent une approche bayésienne pour calculer les probabilités relatives d'un ensemble de modèles de mouvement. En général,

l'estimation du mouvement par la moyennisation des dynamiques de plusieurs particules conduit à une simplification des processus biologiques sous-jacents [Gal et al., 2013].

Contributions

Dans cette thèse, nous proposons une nouvelle statistique T_n qui s'affranchit de certaines limitations du MSD. La statistique T_n est définie comme une normalisation de la plus grande distance parcourue par la particule depuis son point de départ. Nous interprétons cette mesure comme suit:

1. si T_n est faible, la particule est restée proche de sa position initiale indiquant qu'elle est potentiellement bloquée dans un microdomaine ou freinée par des obstacles (sous-diffusion);

2. si la valeur de T_n est grande, la particule s'est beaucoup éloignée de sa position initiale, probablement propulsée par un moteur moléculaire (super-diffusion).

Dans un premier temps, nous utilisons cette statistique pour classer la trajectoire associée à une particule dans un des trois groupes de diffusion. Ensuite, nous proposons un algorithme basé sur T_n pour détecter les temps de rupture le long d'une trajectoire. Enfin, nous appliquons notre méthode de classification dans le cadre d'une analyse spatiale des mouvements intracellulaires.

Classification de trajectoires

Nous voulons associer les trajectoires observées à un des trois groupes de diffusion: mouvement brownien, sous-diffusion et super-diffusion. Pour cela, nous avons développé un test à trois décisions [Shaffer, 1980] dont la statistique de test est T_n. L'hypothèse nulle indique que la trajectoire observée est générée par un mouvement brownien tandis que les deux alternatives correspondent à la sous-diffusion et à la super-diffusion. Dans cette thèse, nous étudions le comportement asymptotique de notre test sous l'hypothèse nulle et en considérant quatre modèles paramétriques associés la super-diffusion et à la sous-diffusion. Ces modèles paramétriques sont communément utilisés dans la littérature biophysique. Enfin, nous avons proposé une procédure de tests multiples pour tester simultanément une collection de trajectoires indépendantes observées dans une seule cellule. Cette procédure est une extension de la méthode de Benjamini and Hochberg [1995] aux tests à trois décisions. Cette procédure permet de contrôler le taux de fausses découvertes "False Discovery Rate" (FDR).

Détection de changement de mouvement le long d'une trajectoire

Lorsqu'on observe une trajectoire longue (plus de 100 points), il est possible que la particule en question ait changé de dynamique au cours du temps. Dans une telle situation,

imposer un modèle unique de déplacement peut conduire à des interprétations fausses. Nous avons donc développé un algorithme pour détecter les temps de rupture, c'est-à-dire les instants qui correspondent à des transitions entre deux dynamiques différentes.

De nombreuses méthodes de détection de rupture existent, développées dans différents contextes. Tout d'abord, Page [1954] a introduit le populaire CUSUM test pour détecter des changements dans l'évolution d'un paramètre θ (par exemple la moyenne), en supposant que $t \to \theta(t)$ est une fonction constante par morceaux au cours du temps. Plus récemment, Spokoiny [2009] a proposé une procédure basée sur des hypothèses paramétriques locales pour détecter les ruptures dans des séries temporelles non stationnaires. Les détections de rupture pour des processus de diffusion sont aussi étudiées: Pollak and Siegmund [1985] estiment le changement de la dérive ("drift") dans le cas du mouvement brownien avec dérive; une autre approche est celle des modèles de Markov cachés [Rabiner and Juang, 1986]. En biophysique, Monnier et al. [2015] supposent que les trajectoires observées peuvent être modélisées par un mélange de K mouvements browniens avec des dérives et des coefficients de diffusion différents. Les états cachés correspondent aux différentes valeurs des paramètres de ces mouvements browniens avec dérive.

Dans notre cas, nous considérons que les changements de dynamique correspondent à des changements de diffusion (brownien, sous-diffusion et super-diffusion). Nous souhaitons aussi détecter les temps de rupture dans un contexte non-paramétrique. A notre connaissance, il n'existe pas de méthodes pour détecter les temps de rupture associés à un changement de type de diffusion (brownien, sous-diffusion et super-diffusion). Par conséquent, nous avons mis au point un algorithme séquentiel basé sur la statistique T_n, calculée sur des sous-trajectoires de la trajectoire initiale et avons adapté le schéma séquentiel proposé par Cao and Wu [2015]. Cet algorithme permet de contrôler la probabilité de détecter un faux point de rupture lorsque la trajectoire est brownienne du début à la fin.

Analyse spatiale

La régulation des processus cellulaires, comme la transmission synaptique, repose sur des interactions moléculaires (liaison et dissociation à des ligands). Les molécules intervenant dans ces processus sont d'abord capturées et confinées dans des microdomaines où s'effectuent des liaisons chimiques. Hoze et al. [2012] modélisent ces microdomaines par des puits de potentiel. Les auteurs utilisent une approche eulérienne basée sur une estimation non-paramétrique du paramètre de dérive du processus de diffusion. Cette approche suppose l'observation d'un grand nombre de particules dans le domaine spatial étudié. Cette situation ne coïncide pas toujours avec la réalité expérimentale. De plus, à cause de son caractère eulérien, l'algorithme de Hoze et al. [2012] n'est pas en mesure de détecter un mélange de plusieurs modèles de diffusion observé localement dans une région. Au contraire, il va estimer un mouvement moyen qui ne correspond pas au pro-

cessus biologique sous-jacent.

Nous proposons une procédure alternative pour détecter ces zones. Nous avons recours à un algorithme de clustering couplé avec notre procédure de test pour détecter les régions avec une grande concentration de particules sous-diffusives. Nous choisissons l'algorithme DBSCAN [Ester et al., 1996] pour sa capacité à distinguer les vrais clusters dans un environnement bruité. D'autres techniques de clustering sont capables d'identifier des clusters en présence de bruit. Par exemple, Cao et al. [2007] proposent une autre approche basée sur la méthode *a-contrario*. Nous ne considérons pas ce type d'algorithme ici par souci de simplicité même si leur intérêt est indéniable dans notre contexte. Nous évaluons notre méthode sur des données artificielles générées par le logiciel FLUOSIM développé par M. Lagardere et O.Thoumine (Institut Interdisciplinaire de Neuroscience (IINS), Université de Bordeaux 2). Ce logiciel est spécialement conçu pour simuler les dynamiques moléculaires dans le contexte de l'imagerie par fluorescence.

Organisation de la thèse

Cette thèse comporte trois parties. Dans la Partie I nous définissons la statistique T_n ainsi que les procédures de test à trois décisions pour classer les trajectoires des particules selon trois groupes. Dans la partie II, nous décrivons un algorithme séquentiel pour détecter les changements de dynamique d'une même particule au cours du temps. Enfin, dans la Partie III, nous estimons les microdomaines qui piègent les particules. L'organisation de la thèse est présentée de manière synthétique ci-dessous.

Chapter 2 Nous introduisons le concept probabiliste de diffusion présenté dans Karlin [1981] et Klebaner et al. [2012]. Nous définissons d'abord la notion générale de processus stochastique. Ensuite, nous présentons le mouvement brownien à partir duquel nous construisons les processus de diffusion. Finalement, nous exposons une extension du mouvement brownien à savoir le mouvement brownien fractionnaire.

Chapter 3 Nous expliquons les théories physiques à l'origine du mouvement brownien. Nous donnons un aperçu des différents modèles utilisés en biophysique et en physique pour décrire la sous-diffusion et la super-diffusion. Nous expliquons les scénarios biologiques asssociés aux différents types de diffusion.

Part I

Chapter 4 Nous introduisons la nouvelle statistique T_n. Nous développons un test à trois décisions pour classer une trajectoire suivant le modèle de diffusion le plus approprié. Notre hypothèse nulle suppose que la trajectoire observée est brownienne. Nous étendons cette procédure pour tester une collection de trajectoires indépendantes dans le cadre des tests multiples. Notre procédure contrôle le critère du FDR introduit par Benjamini and Hochberg [1995] au niveau α.

Chapter 5 Nous évaluons les différentes procédures de test sur des simulations en dimension deux. En particulier, nous estimons la puissance de nos tests sous différents modèles paramétriques de diffusion utilisés en biophysique. Nous comparons nos résultats à la méthode de Feder et al. [1996] basée sur le MSD. Nous analysons des données réelles acquises par microscopie TIRF décrivant le processus d'exocytose en dimension deux et trois.

Part II

Chapter 6 Nous présentons une nouvelle méthode pour détecter les instants de transition, c'est-à-dire lorsque la particule passe d'un mode de diffusion à un autre. Il s'agit d'un algorithme séquentiel basé sur la statistique T_n calculée sur des fenêtres glissantes le long de la trajectoire. La taille de la fenêtre est le seul paramètre à optimiser. Si la trajectoire est entièrement brownienne, nous contrôlons la probabilité de détecter un (faux) point de rupture au niveau α.

Chapter 7 Nous évaluons notre algorithme séquentiel sur des simulations et des données réelles. Nous comparons ses performances à celles de deux autres méthodes proposées par Türkcan and Masson [2013] et Monnier et al. [2015]. Finalement, nous montrons que notre méthode présente de meilleurs résultats sur les simulations considérées. Son temps de calcul est aussi négligeable par rapport aux autres méthodes.

Part III

Chapter 8 Nous présentons le modèle de simulation du logiciel Fluosim (IINS, université de Bordeaux 2). Ce logiciel simule les mouvements de molécules dans une cellule présentant des zones de confinement (microdomaines). Nous exposons de manière rigoureuse le modèle mathématique sous-jacent. Nous modélisons la proportion de particules piégées dans les microdomaines par des équations différentielles. Nous proposons un schéma de simulation pour évaluer la méthode proposée dans le Chapitre 9.

Chapter 9 Nous proposons une méthode pour détecter automatiquement les zones de piégeage dans lesquelles les particules sont confinées. Nous utilisons l'algorithme de clustering DBSCAN [Ester et al., 1996] combiné avec notre procédure de test pour détecter les zones avec une grande concentration de particules sous-diffusives. Nous évaluons notre méthode sur des données issues du simulateur FLUOSIM.

Chapter 10 Nous rappelons les enjeux de cette thèse et les contributions apportées. Des perspectives méthodologiques sont évoquées et d'autres champs d'application de nos méthodes sont mentionnés.

1 Preamble

In this thesis, we are interested in quantifying the dynamics of intracellular particles (e.g., biomolecules) inside living cells. A cell is a complex environment composed of lots of structures in interaction with each other. They continuously exchange biological material directly via the cytosol or via networks of polymerised filaments namely the microtubules, actin filaments and intermediate filaments. The dynamics of these proteins determine the organization and function of the cell [Bressloff, 2014, Chapter 9]. Then, inference on the modes of mobility of molecules is central in cell biology since it reflects the interaction between the structures of the cell.

In quantitative microscopy, there are multiple techniques that allow particle motion analysis. The most popular are fluorescence recovery after photobleaching (FRAP), correlation spectroscopy-based techniques (FCS) and single-particle tracking (SPT). While FRAP and FCS average the mobility of a very large number of molecules, SPT is attractive since it can reveal individual dynamics, close to molecular resolution. In this thesis, we analyse trajectories of individual proteins or molecules. These trajectories are computed from conventional or super-resolution microscopy image sequences. We use the Lagrangian setting, that is, we analyse the motion of individual particles (*e.g.*, proteins or molecules) along their trajectories. This concept is opposed to the Eulerian paradigm in which the motion is described as vector field based on the average motion of particles computed over local regions.

As the interior of a living cell is a fluctuating environment, we model the trajectories of particles with stochastic processes with continuous paths. Diffusions belong to this class of processes and can model a large range of intracellular movements. They are widely used in the biophysical literature [Qian et al., 1991, Saxton and Jacobson, 1997]. Biophysicians distinguish four main types of diffusions, namely Brownian motion (also referred to as free diffusion), superdiffusion, confined diffusion and anomalous diffusion. Trajectories illustrating these four type of diffusion are represented in Figure 1.1. These different diffusions correspond to specific biological scenarios. A particle evolving freely inside the cytosol or along the plasma membrane is modelled by free diffusion. Its motion is due to the constant collisions with smaller particles animated by thermal fluctuations. Then, the particle does not travel along any particular direction and can take a very long time to go to a precise area in the cell. Active intracellular transport can overcome this difficulty so that motion is faster and direct specific. The particles (called in this context cargo) are carried by molecular motors along microtubular filament networks. Superdiffusions model the motion of molecular motors and their cargo.

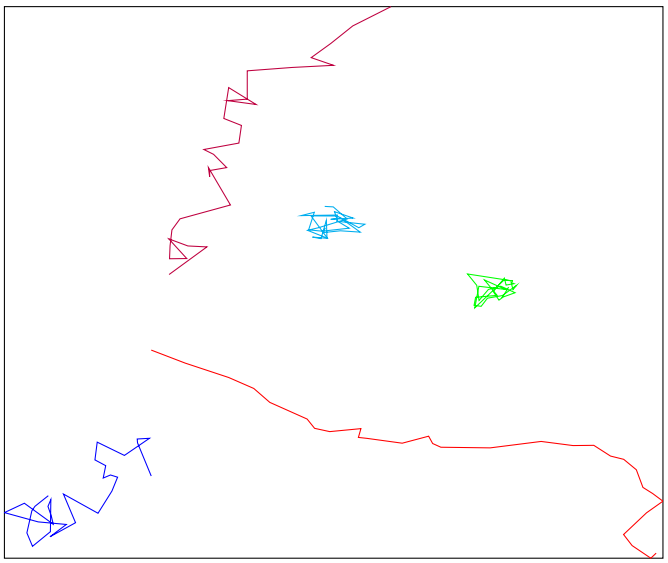


Figure 1.1: Representative trajectories from simulated data. The blue trajectory is Brownian; the purple trajectory is from a Brownian motion with drift (4.2.3) and illustrates superdiffusion; the red trajectory is from a fractional Brownian motion (2.5.7) (parameter $\mathfrak{h} > 1/2$) and illustrates superdiffusion; the cyan trajectory is from an Ornstein-Uhlenbeck process (3.3.14) and illustrates confined diffusion; the green trajectory is from a fractional Brownian motion (2.5.7) ($\mathfrak{h} < 1/2$) and illustrates anomalous diffusion. The parameters controlling the processes are given in Table (5.1).

Confined or restricted diffusion [Metzler and Klafter, 2000, Hoze et al., 2012] is characteristic of trapped particles: the particle encounters a binding site, then it pauses for a while before dissociating and moving away. Anomalous diffusion includes particles which encounters dynamic or fixed obstacles [Saxton, 1994, Berry and Chaté, 2014], or particles slowed by the contrary current due to the viscoelastic properties of the cytoplasm. In the sequel, we will not distinguish confined and anomalous diffusion and consider that both are subdiffusion. In fact, as we will use a non-parametric setting, discriminating confined diffusion from anomalous diffusion is not feasible, especially on short trajectories. A classification of protein trajectories into the three types of diffusion is shown in Figure 1.2. This classification is obtained with our three-decision test procedure exposed in Chapter 4.

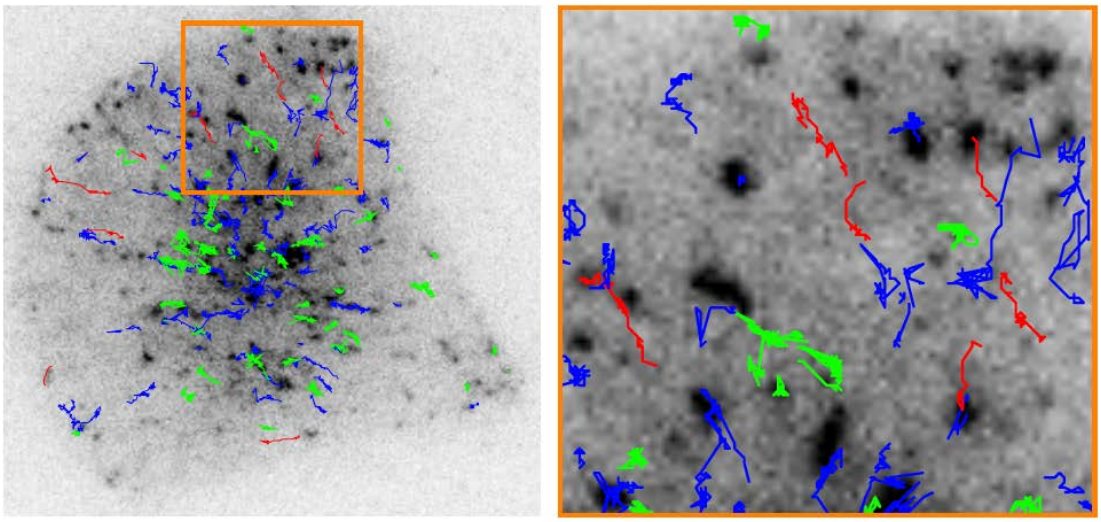


Figure 1.2: Classification of two-dimensional trajectories from the Rab11 protein sequence in a single cell observed in TIRF microscopy (Courtesy of UMR 144 CNRS Institut Curie PICT IBiSA). We use the three-decision test procedure developed in Chapter 4 at level $\alpha = 5\%$. The Brownian trajectories are in blue, the subdiffusive trajectories in green and the superdiffusive trajectories in red.

1.1 Problematic

Mean Square Displacement

In biophysics, the different types of diffusions are characterised by the mean square displacement (MSD) [Qian et al., 1991]. Given a particle trajectory $(X_t)_{t>0}$, the MSD is defined as the function,

$$\mathrm{MSD}(t) = \mathbb{E}\left(\|X_{t+t_0} - X_{t_0}\|^2\right), \tag{1.1.1}$$

where $\|\cdot\|$ is the euclidean norm and $\mathbb{E}$ is the expectation of the probability space. The MSD of Brownian motion is linear ($\mathrm{MSD}(t) \propto t$). This property makes the MSD a popular criterion to analyse intracellular motion as Brownian motion is the process of reference. The typical MSD curves of the different diffusion models are represented in Figure 1.3. In practical imaging, we observe the successive positions of a single particle $X_{t_0}, X_{t_1}, \ldots, X_{t_n}$ in the two or three dimensions at equispaced times, that is $t_{i+1} - t_i = \Delta$. The MSD is estimated at lag j by:

$$\widehat{\mathrm{MSD}}(j\Delta) = \frac{1}{n-j+1} \sum_{k=0}^{n-j} \|X_{t_{k+j}} - X_{t_k}\|^2. \tag{1.1.2}$$

Computing the estimator (1.1.2) at different lag j gives an estimation of the MSD function (1.1.1). Then the simplest rule to classify a trajectory is based on a fit of the MSD function (1.1.1) to $t \to t^{\beta}$. Feder et al. [1996] states that the trajectory is subdiffusive if $\beta < 0.9$, superdiffusive if $\beta > 1.1$ and Brownian if $0.9 < \beta < 1.1$. If $\beta < 0.1$ it states that the particle does not move, see Figure 1.4.

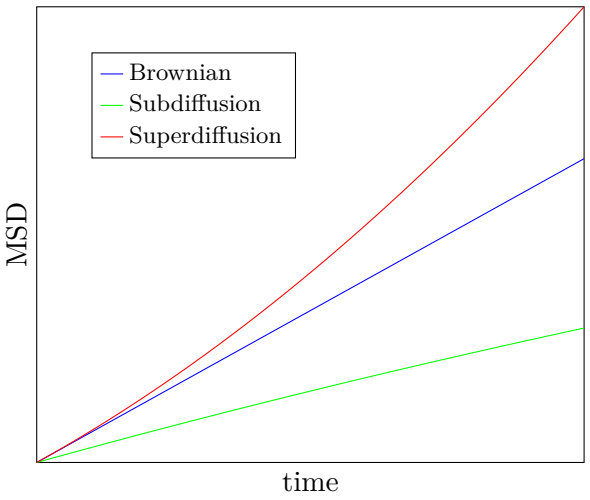


Figure 1.3: Typical MSD curves of the different diffusion types.

Limitations

The criterion of the MSD has some limitations. First the MSD statistic is a summary statistic, and is not sufficient to characterize the dynamics of the trajectory. Accordingly, Gal et al. [2013] present several other statistics which can be associated to MSD for trajectory analysis. Lund et al. [2014] propose a decision tree for selection motion model combining MSD, Bayesian information criterion and the radius of gyration. Lysy et al. [2016] present a likelihood-based inference as an alternative to MSD for the comparison between two models of subdiffusions: fractional Brownian motion and a generalized Langevin equation. They consider a Bayesian model to estimate the parameter of the diffusion and they use the Bayes factor to compare the models.

Second, the variance increases with the time lag. Figure 1.4 illustrates this problem in the case of Brownian trajectories. It suggests that the classification of Feder et al. [1996] based on parameter β overdetects subdiffusion and superdiffusion while it is Brownian motion. Moreover the MSD variance is also severely affected at short time lags by dynamic localization error and motion blur. Michalet [2010] details an iterative method, known as the Optimal Least Square Fit (OLSF) for determining the optimal number of points to obtain the best fit to MSD in the presence of localization uncertainty.

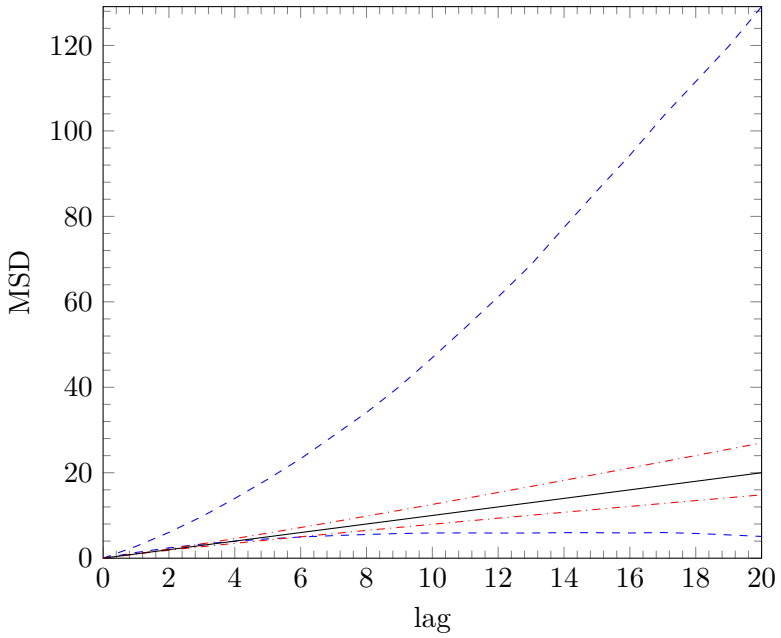


Figure 1.4: A classification rule for motion modes from MSD. The dashdotted lines are the bound defined by Feder et al. [1996], $t \to t^\beta$, $\beta = 0.9$ and 1.1. The dashed lines are the pointwise high probability interval of 95% associated to the empirical MSD curve for a standard Brownian motion trajectory of length $n = 30$. The bounds of the interval are the 2.5% and 97.5% empirical quantile of (1.1.2) and are computed by Monte Carlo simulation from 10 001 Brownian trajectories of size $n = 30$.

In order to take account of the variance of the MSD estimate, several authors use a set of independent trajectories rather than single trajectories. These trajectories may have different lengths but are assumed to have the same kind of motion. For instance, Pisarev et al. [2015] consider weighted-least-square estimate for β by estimating the variance of pathwise MSD. Their motion model selection is then based on the modified Akaike's information criterion. Monnier et al. [2012] propose a Bayesian approach to compute relative probabilities of an arbitrary set of motion models (free, confined, anomalous or directed diffusion). In general, this averaging process can lead to oversimplication and misleading conclusions about the biological process [Gal et al., 2013].

1.2 Contributions

In this thesis, we introduce a new statistic T_n that circumvents some of the aforementioned limitations of the MSD. The statistic T_n is defined as the standardized largest distance covered by the particle from its starting point. We interpret this measure as

follows:

1. if the value of T_n is low, it means that the process stayed close to its initial position and the particle may be trapped in a small area or hindered by obstacles (subdiffusion);

2. if the value of T_n is high, the particle went far to its initial position and the particle may be driven by a motor in a certain direction (superdiffusion).

First we use this statistic in order to classify individual trajectories into the three types of diffusion of interest. Secondly, we develop an algorithm based on T_n to detect change points along the trajectory. Finally, we apply our classification method in the context of spatial analysis.

Classification of the Trajectories

We want to classify the particle trajectories observed in living cells into the three types of diffusion namely Brownian motion, superdiffusion and subdiffusion. To this end, we develop a three-decision test procedure [Shaffer, 1980] based on the statistic T_n. The null-hypothesis is that the observed trajectory is generated from a Brownian motion and the two distinct alternatives are subdiffusion and superdiffusion. Then, we study the asymptotic behaviour of our procedure under the null hypothesis and four parametric models illustrating superdiffusion and subdiffusion and which are commonly considered in the biophysics literature. We also derive a multiple test procedure in order to apply simultaneously the test on a collection of independent trajectories which are tracked inside the same living cell. This procedure is an extension of the procedure of Benjamini and Hochberg [2000] to three decision tests. It allows to control the false discovery rate (FDR).

Detection of Change of Dynamic over Time

When we observe a long trajectory (more than 100 points), it is possible that the particle switches mode of motion over time. Then, fitting a single model to the trajectory can be misleading. We propose a method to detect the change points: the times at which a change of dynamics occurs.

A large range of change point detection method exist, developed in different contexts. Page [1954] introduced the well-known CUSUM test to detect changes in a parameter θ (as the mean) which is assumed to be piece-wise constant over time. Spokoiny [2009] proposes a method based on local parametric assumptions to detect changes in non-stationary time series. Change point in diffusion process has also been studied. For example, Pollak and Siegmund [1985] estimates the change in the drift of Brownian motion. Another approach is to use hidden Markov model [Rabiner and Juang, 1986].

Monnier et al. [2015] assume that the observed trajectories can be modelled by a mixture of K Brownian motion with different drift and diffusion coefficient values. The hidden states are the values of the parameters of these Brownian with drift.

In our framework, we consider that the particle switches between the three types of diffusion aforementioned; we want to detect at which times the changes happen in this non-parametric framework. To our knowledge, there exists no change point detection method addressing this issue. Then, we developed a sequential algorithm based on the statistic T_n which is computed on local windows along the trajectory. We use a particular sequential scheme adapted from the algorithm of Cao and Wu [2015]. In case the trajectory is fully Brownian, the probability to detect falsely a change point is controlled at level α.

Spatial Analysis

Regulation of cellular physiological processes such as synaptic transmission, relies on molecular interactions (binding and unbinding) at specific places and involves trafficking in confined local microdomains. Hoze et al. [2012] model these microdomains as potential wells which attract intra-cellular particles. The authors use an Eulerian method based on the non-parametric estimation of the drift parameter of the underlying diffusion process. This method needs a high concentration of particles over the spatial domain of interest to be meaningful. This situation is not always available experimentally. Moreover, due to its Eulerian approach, this method can not capture a mixture of different dynamics occurring at the same location; it will average the different motions, potentially leading to false conclusions. As an alternative, we define a new procedure to detect microdomains. We use a clustering algorithm coupled with our test procedure to detect the zones with a high concentration of subdiffusive particles. More specifically, we choose the DBSCAN algorithm, designed by Ester et al. [1996], as it can distinguishes true clusters from noise. Other clustering approaches can be used in this context as the *a-contrario* method [Cao et al., 2007]. We will only consider DBSCAN for simplicity. We assess the proposed method using the software FLUOSIM developed by M. Lagardere and O.Thoumine (Institut Interdisciplinaire de Neuroscience (IINS), Université de Bordeaux 2). This software is designed to simulate the molecular dynamics in the context of fluorescence microscopy.

1.3 Organisation of the Thesis

The thesis comprised three parts. Part I presents the statistic T_n and the three-decision test procedures to classify the particles trajectories into three groups. In Part II, we develop the sequential algorithm to detect motion switching. In part III, we estimate the microdomains where particles are confined. The thesis organization is synthetically presented below.

Chapter 2 We introduce the probabilistic concept of diffusion presented in Karlin [1981], Klebaner et al. [2012]. First, we define the notion of stochastic processes. Then, we put an emphasis on Brownian motion and connect this process to diffusions. Finally we present an extension of Brownian motion, namely fractional Brownian motion.

Chapter 3 We give the physical foundations which leads to the concept of Brownian motion. We give an overview of the models used in biophysics and physics for depicting subdiffusion and superdiffusion. We also described the underlying biological scenarios associated to the different modes of diffusion.

Part I

Chapter 4 We introduce the statistic T_n. We develop a three-decision test to classify a trajectory into one of the three types of diffusion aforementioned. Our null hypothesis is that the trajectory is Brownian. We also derive a multiple test procedure to classify a set of independent trajectories while controlling the criterion of false discovery rate (FDR) introduced by Benjamini and Hochberg [1995].

Chapter 5 We assess our test procedures on simulations in the two-dimensional case. In particular, we estimate the power of the procedures on different parametric diffusion processes used in biophysics. We compare our results to an approach based on the MSD. We also analyse real data depicting the exocytosis process in two and three dimensions.

Part II

Chapter 6 We provide a new method to detect the times at which the particle changes of diffusion mode. It is a sequential algorithm based on the statistic T_n computed on local windows along the trajectory. The size of the window k is the only parameter of the procedure. In the case of a fully Brownian trajectory, the probability to detect falsely a change point is controlled at level α.

Chapter 7 We assess our sequential algorithm on simulations and real data. We compare its performances to two different competitive procedures respectively designed by Türkcan and Masson [2013] and Monnier et al. [2015]. Our method outperforms the

others on simulations. Moreover, the computational cost is very small compared to the other methods.

Part III

Chapter 8 We present the software FLUOSIM (IINS, Université de Bordeaux 2). This software simulates the dynamics of molecules in an environment with local microdomains where the particles can be trapped. We give the underlying mathematical framework associated to the simulator and derive differential equations giving the proportion of trapped and free particles. We design a simulation scheme helpful to assess the method described in Chapter 9.

Chapter 9 We propose a method for detecting the trapping areas or microdomains where the particles are confined. We use the clustering algorithm DBSCAN [Ester et al., 1996] coupled with our test procedure to detect the zones with a high concentration of subdiffusive particles. We evaluate the method on data simulated with FLUOSIM.

Chapter 10 We summarize our contributions and give few methodological perspectives. We emphasize other possible applications of the methods developed in this thesis.

2 Introduction to Stochastic Processes and Diffusions

In this chapter, we present the probabilistic tools in order to define diffusion processes. As explained in Chapter 1, such processes are of great importance for modelling intracellular dynamics. To this end, we focus on d-dimensional processes with $d = 2$ or $d = 3$. We note that the biophysic literature uses the word diffusion in a very broad sense [Meroz and Sokolov, 2015]. Here we introduce the probabilistic concept of diffusion presented in Karlin [1981] and Klebaner et al. [2012]. First, we define the notion of stochastic processes. Then, we put an emphasis on Brownian motion, the cornerstone process which allows to build all the diffusion processes. We describe diffusion processes driven by Brownian motion. Finally, we deal with an extension of Brownian motion, namely fractional Brownian motion [Mandelbrot and Van Ness, 1968]; we present quickly diffusion processes driven by fractional Brownian motion.

2.1 Stochastic Process

Let $(\Omega, \mathcal{F}, P)$ a probability space where Ω is the sample space, $\mathcal{F}$ a field and P a probability measure. A d-dimensional *stochastic process* is a function:

$$\begin{aligned} I \times \Omega &\to \mathbb{R}^d \\ (t, \omega) &\mapsto X(t, \omega) \end{aligned} \tag{2.1.1}$$

where I is a time interval. We note this application $(X_t)_{t \in I}$ or simply (X_t). We present briefly stochastic processes from two angles.

Let $t \in I$, the application,

$$\begin{aligned} \Omega &\to \mathbb{R}^d \\ \omega &\mapsto X(t, \omega) \end{aligned} \tag{2.1.2}$$

is the random state of the process at time t. It is a random variable defined on $(\Omega, \mathcal{F}, P)$. Then, a stochastic process can be seen as the collection of random variables $\{\omega \mapsto X(t, \omega),\, t \in I\}$.

Let $\omega \in \Omega$, the application

$$\begin{aligned} I &\to \mathbb{R}^d \\ t &\mapsto X(t, \omega) \end{aligned} \tag{2.1.3}$$

is called a *trajectory* or a *path* of the stochastic process $(X_t)_{t\in I}$.

A stochastic process may be seen as an application from Ω to the set of functions from $I = [0,T]$ to $\mathbb{R}^d$. As previously mentioned, we consider only the stochastic processes whose trajectories are continuous, that is for almost $\omega \in \Omega$ $t \to X_t(\omega)$ is continuous.

Finite-Dimensional Distribution

A stochastic process may be seen as a random variable from $(\Omega, \mathcal{F}, P)$ to the measurable space,

$$\left(\mathfrak{F}([0,T],\mathbb{R}^d), \otimes_{t\in[0,T]}\mathcal{B}_d\right),$$

where $\mathfrak{F}([0,T],\mathbb{R}^d)$ is the set of functions from $[0,T]$ to $\mathbb{R}^d$, $\mathcal{B}_d$ is the Borelian sigma-algebra and $\otimes_{t\in[0,T]}\mathcal{B}_d$ is the sigma-algebra generated by all the finite dimensional cylindrical sets of $\mathfrak{F}([0,T],\mathbb{R}^d)$. Then the stochastic process X induces a probability measure on $\left(\mathfrak{F}([0,T],\mathbb{R}^d), \otimes_{t\in[0,T]}\mathcal{B}_d\right)$ which is defined through the finite-dimensional distribution.

Now we define the concept of finite-dimensional distribution. Let $J = \{t_0, t_1, \ldots, t_n\}$ such that $t_i \in I$ and $t_0 < t_1 <, \ldots, < t_n$. We note,

$$X_J = (X_{t_0}, \ldots, X_{t_n}), \tag{2.1.4}$$

the random vector whose components $X_{t_i} \in \mathbb{R}^d$. The distribution μ_J of X_J is the joint distribution:

$$\mu_J(A) = P(X_{t_0} \in A_0, \ldots, X_{t_n} \in A_n), \tag{2.1.5}$$

where $A_i \in \mathbb{R}^d$ and $A = A_0 \times \cdots \times A_n$.

The *finite-dimensional distributions* of X is the family of distributions $\{\mu_J | J \text{a finite set of } I\}$. If the finite-dimensional distributions μ_J satisfy a technical criterion called consistency then the Kolmogorov extension theorem guarantees the existence of a stochastic process X with finite-dimensional distributions μ_J on $(\Omega, \mathcal{F}, P)$ [Gallardo, 2008, Chapter 1, Section 1.1].

Filtered Probability Space

We state previously that a stochastic process can be seen as a collection of random variables defined on $(\Omega, \mathcal{F}, P)$. More precisely the random variable (2.1.2) is defined on $(\Omega, \mathcal{F}_t, P)$ where $\mathcal{F}_t \subset \mathcal{F}$. This reflects that the outcome of the random variable (2.1.2) depends on what happened before t, that is on the historic of the process until time t. The fact that $\omega \mapsto X(t,\omega)$ is $\mathcal{F}_t$-measurable and not $\mathcal{F}$-measurable can be compared to the fact that a stochastic process is determined by all the finite-dimensional distributions and not only the set of marginal distribution $P(X_t \in A_i)$.

Then we define the concept of filtration. A *filtration* $\mathbb{F}$ is a family $(\mathcal{F}_t)$ of increasing fields on $(\Omega, \mathcal{F})$ that is $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$ for $s < t$. $\mathbb{F}$ specifies how the information is

revealed over time. The property that a filtration is increasing corresponds to the fact the information is not forgotten. Finally, a stochastic process X is called adapted to a filtration $\mathbb{F}$ if, for all t, the random variable $\omega \mapsto X(t, \omega)$ is $\mathcal{F}_t$-measurable.

2.2 Brownian Motion

The observation of the erratic motion of a pollen particle suspended in a fluid by the botanist R. Brown in 1828 marks the first step in the development of the Brownian motion theory. In 1905, Einstein argued that the movement of the particle is due to its bombardment by the particles of the fluid; he obtained the equations of Brownian motion. The underlying probability theory was derived by N. Wiener in 1923 that is why Brownian motion is also known as the Wiener process. In this section, we define the one-dimensional Brownian motion and characterize it as a Gaussian process. Then, we define the $d-$dimensional Brownian motion.

Definition

The *one-dimensional Brownian motion* (B_t) is a stochastic process with the following properties:

- (B_t) is a process with *independents increments.* For all $t > s$, $B_t - B_s$ is independent of the field $\mathcal{F}_s$ generated by the historic of the process $(B_u)_{u\in[0,s]}$ until the time s.
- For all $t > s$, $B_t - B_s$ has normal distribution with mean 0 and variance $t - s$.
- The paths of (B_t) are almost surely continuous.

Gaussian Process

A *Gaussian process* is a process for which all the finite-dimensional distributions are multivariate normal. We have the following theorem:

Theorem 2.2.1. *A Brownian motion started at zeros is a Gaussian process with zero mean and covariance function* $\min(t, s)$*. Conversely, a Gaussian process with zero mean and covariance* $\min(t, s)$ *is a Brownian motion.*

Multivariate Brownian Motion

As we already stated, we are interested in modelling the trajectories of particle in dimension 2 and 3. We define the *d-dimensional Brownian motion* $(d \geq 1)$ as the random vector $B_t = (B_t^1, \ldots, B_t^d)$ where all coordinates B_t^i are independent one-dimensional Brownian motions.

2.3 Diffusion Process

We present briefly the family of stochastic processes of interest in this thesis, namely the diffusion processes. First, we recall the Markov property which is a central notion for defining the diffusion processes. Then, we give the definition of diffusions and some characterizations of theses processes.

Markov Property

The Markov property states that if we know the present state of the process, the future behaviour of the process is independent of its past. For instance, a simple model of weather forecast assumes that the probability to have rain at day j given the information of the weather on the previous days is the same as the probability to have rain at day j given the restricted information of the weather at day $j-1$. Let note (X_i) the process giving the weather at each day i and note k the modality corresponding to rain. In this discrete set up, the Markov property can be written as:

$$P(X_j = k|X_{j-1}, \dots, X_0) = P(X_j = k|X_{j-1}). \tag{2.3.1}$$

As we work with stochastic processes defined continuously in time, the historic of the process given by $X_{j-1}, \dots, X_0$ in the discrete case is replaced by the field $\mathcal{F}_t$ at time t. Then, a d-dimensional continuous stochastic process (X_t) is Markovian if:

$$P(X_{t+s} \in A|\mathcal{F}_t) = P(X_{t+s} \in A|X_t), \tag{2.3.2}$$

where $A \in \mathbb{R}^d$. Then we have the following theorem:

Theorem 2.3.1. *The Brownian motion (B_t) has the Markov property.*

Remark 2.3.1. *Another difference (apart from the conditioning) between Equations* (2.3.1) *and* (2.3.2) *is the different nature of the events $\{X_j = k\}$ and $\{X_{t+s} \in A\}$. It is due to the fact that in Equation* (2.3.1) *the state space of the stochastic process (modality of weather) is countable while the state space of the stochastic process is the whole space $\mathbb{R}^d$ (not countable) in* (2.3.2) .

Diffusions

A *diffusion process* (X_t) is a continuous time process which possesses the Markov property and for which the sample paths are continuous. Moreover, every diffusion process satisfies three key conditions see [Karlin, 1981, Chapter 15, Section 1]. The first condition states that large displacements of magintude exceeding $\epsilon > 0$ are very unlikely over sufficiently small intervals,

$$\lim_{\Delta \to 0} \frac{1}{\Delta} P(\|X_{t+\Delta} - X_t\| > \epsilon | X_t = x) = 0, \quad \forall \epsilon > 0, \quad \forall x \in \mathbb{R}^d, \tag{2.3.3}$$

where $\|\cdot\|$ denotes the Eucidean norm. In other words, condition (2.3.3) prevents the diffusion process from having discontinuous jumps. The two last conditions characterize the mean and the variance of the infinitesimal displacements and affirm the existence of the limits:

$$\lim_{\Delta\to 0} \mathbb{E}(X_{t+\Delta} - X_t | X_t = x) = \mu(x,t), \quad \forall x \in \mathbb{R}^d, \tag{2.3.4}$$

$$\lim_{\Delta\to 0} \mathbb{E}((X_{t+\Delta} - X_t)(X_{t+\Delta} - X_t)^\top | X_t = x) = \sigma^2(x,t), \quad \forall x \in \mathbb{R}^d, \tag{2.3.5}$$

where $\top$ denotes the transpose operator; $\mu(x,t) : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^d$ is the drift parameter; $\sigma^2(x,t) : \mathbb{R}^d \times \mathbb{R}^+ \to S_+^d$ is the diffusion coefficient where S_+^d is the set of positive semi-definite matrix of size d.

In particular, Brownian motion is a diffusion process : its drift is the null function, and its diffusion coefficient is constant.

2.4 Stochastic Differential Equation (SDE)

The most common approach for defining diffusion processes is to see them as the solution of stochastic differential equations.

Physical Model

Initially diffusion models were developed to describe the motion of a particle in a fluid submitted to a deterministic force due to the fluid and a random force due to random collisions with others particles. That is why we model efficiently the motion of intracellular particles with diffusion. Let $X_t \in \mathbb{R}^d$ be the position of the particle at time t and (B_t) a $d-$dimensional Brownian motion; assume that $X_t = x$. Then the displacement of the particle between t and $t + \Delta$ is approximately given by:

$$X_{t+\Delta} - x \approx \mu(x,t)\Delta + \sigma(x,t)(B_{t+\Delta} - B_t). \tag{2.4.1}$$

The component $\mu(x,t)\Delta$ is the displacement due to the fluid where the velocity of the fluid is given by the drift $\mu(x,t)$. The term $\sigma(x,t)(B_{t+\Delta} - B_t)$ expresses the random component of the motion due to random collisions. More specifically the collisions increased with the temperature of the fluid; the influence of temperature is modelled by the diffusion coefficient $\sigma(x,t)$. We note that the model (2.4.1) implies that, due to the normality of the Brownian increment, the displacement of the particle $X_{t+\Delta} - x$ is approximated by a Gaussian random variable of mean $\mu(x,t)\Delta$ depending on the drift and of variance $\sigma(x,t)\sqrt{\Delta}$ depending on the diffusion coefficient.

Heuristically, a *stochastic differential equation* is obtained from Equation (2.4.1) by replacing Δ by dt, $(B_{t+\Delta} - B_t)$ by dB_t and $X_t + \Delta - X_t$ by dX_t. Then we have the following definition:

Definition 1. *Let (B_t) be a $d-$dimensional Brownian motion. Let $\mu : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma(x,t) : \mathbb{R}^+ \times \mathbb{R}^d \to \mathcal{M}^d$ be given functions ($\mathcal{M}^d$ denoting the set of square matrix of size d). A stochastic differential equation (SDE) is defined as:*

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t, \tag{2.4.2}$$

where (X_t) is the unknown process. The function μ is referred to as the drift while the function σ is called the diffusion coefficient.

Solution of SDE

There are two types of solutions respectively called *strong* and *weak solutions*. A strong solution is a weak solution but the reverse is false.

Definition 2. *Let $\mathcal{F}_t$ the field induced by the initial condition X_0 and the Brownian motion (B_t) which drives the stochastic differential* (4.1)*. We say that Equation* (4.1) *has a strong solution (X_t) on the probability space $(\Omega, \mathcal{F}, P)$ with respect to (B_t) and initial condition X_0 if the stochastic process X_t satisfies* (4.1)*, has continuous paths and that X_t is $\mathcal{F}_t$-measurable for all t.*

The fact that X_t is $\mathcal{F}_t$-measurable is crucial. It means that X_t depends only on the historic of the Brownian motion which drives the stochastic differential equation and the initial condition. Then we can interpret X_t as an output of the system parametrized by $\mu(x,t)$ and $\sigma(x,t)$ whose input is the Brownian motion (B_t). It reflects the principle of causality of the system. If X_t could depend on the future, that is on B_s with $s > t$, causality would fail.

The concept of strong solution relies on the fact that the Brownian motion is given. A weak solution of a SDE consists in building at the same time a couple of processes (X_t, B_t) where (X_t) is a solution of the SDE driven by the Brownian (B_t). We will not give the exact definition of weak solution as it has technical points not of interest for the understanding of the concept.

Then the solution of the stochastic differential equation is written as:

$$X_t = X_0 + \int_0^t \mu(X_s, s)ds + \int_0^t \sigma(X_s, s)dB_s. \tag{2.4.3}$$

We note that the fact that the two integrals are defined is equivalent to the fact that X_t is (strong or weak) solution. In particular the integral with integrand dB_t is a random variable $\mathcal{F}_t$-measurable. Details of the construction of such integrals is given in [Klebaner et al., 2012, Chapter 4].

2.5 Fractional Brownian Motion

Fractional Brownian motion (fBm) was introduced to model scale-invariant phenomenons processes showing long-range dependence. Kolmogorov [1941] developed a turbulence theory based on two hypothesises of scale invariance. In his study of long-term storage capacity and design of reservoirs, Hurst [1951] observed hydrological events invariant to changes in scale. Mandelbrot and Van Ness [1968] defined (and named) fractional Brownian motion. They presented it as: *"fBm of exponent $\mathfrak{h}$ is a moving average of $dB(t)$, in which past increments of $B(t)$ are weighted by the kernel $(t-s)^{2\mathfrak{h}-1}$."* This kernel is at the origin of the long range dependence property (for a certain choice of parameter $\mathfrak{h}$). The parameter $\mathfrak{h}$ is known as the Hurst index or Hurst parameter. In this section, we define fractional Brownian motion and give its main properties. Fractional Brownian motion is then defined in dimension d.

Self-Similarity and Fractional Brownian Motion

A real-valued stochastic process (X_t) is *self-similar* with index $\mathfrak{h} > 0$ ($\mathfrak{h} - ss$) if, for any $a > 0$ the processes (X_{at}) and $(a^{\mathfrak{h}} X_t)$ have the same finite dimensional distributions. Then, a Gaussian $\mathfrak{h} - ss$ process $(B_t^{\mathfrak{h}})$ with stationary increments and Hurst index $0 < \mathfrak{h} < 1$ is a *fractional Brownian motion*.

Now we give some properties of the fBm. First, the fBm has continuous paths. We have $\mathbb{E}(B_t^{\mathfrak{h}}) = 0$ for all t. It is said to be standard if the variance of $B_1^{\mathfrak{h}}$ is equal to one. For the standard fBm we have:

$$\mathrm{Cov}(B_t^{\mathfrak{h}}, B_s^{\mathfrak{h}}) = \frac{1}{2}(|t|^{2\mathfrak{h}} + |s|^{2\mathfrak{h}} - |t-s|^{2\mathfrak{h}}) \tag{2.5.1}$$

Then we can show that a fBm with $\mathfrak{h} = 1/2$ is simply a (one-dimensional) Brownian motion.

Long Range Dependence

A stationary time series $(X_n)_{n\in\mathbb{N}}$ exhibits *long-range dependence* if $\mathrm{Cov}(X_n, X_0) \to 0$ as $n \to \infty$ but,

$$\sum_{n=0}^{\infty} |\mathrm{Cov}(X_n, X_0)| = \infty. \tag{2.5.2}$$

In other words the covariance between X_0 and X_n tends to 0 but so slowly that their sum diverges. Then, we define the stationary process known as fractional Gaussian noise:

$$X_k = B_{k+1}^{\mathfrak{h}} - B_k^{\mathfrak{h}}, \quad k \in \mathbb{N}, \tag{2.5.3}$$

where $(B_t^{\mathfrak{h}})$ is a standard fBm of Hurst index $\mathfrak{h}$. Due to the properties of fBm the fractional Gaussian noise (X_n) is a stationary centered Gaussian process with auto-covariance function:

$$\gamma(k) = \mathbb{E}(X_{i+k}X_i) = \frac{1}{2}(|k+1|^{2\mathfrak{h}} + |k-1|^{2\mathfrak{h}} - 2|k|^{2\mathfrak{h}}). \tag{2.5.4}$$

Then for $k \neq 0$ we can show that $\gamma(k) = 0$ if $\mathfrak{h} = 1/2$, $\gamma(k) < 0$ if $0 < \mathfrak{h} < 1/2$ and $\gamma(k) > 0$ if $1/2 < \mathfrak{h} < 1$. Now, for $\mathfrak{h} = 1/2$ we have:

$$\gamma(k) = \mathfrak{h}(2\mathfrak{h}-1)|k|^{2\mathfrak{h}-1} + o(1), \tag{2.5.5}$$

where $o(1) \to 0$ as $k \to \infty$. Consequently $\gamma(k) \to 0$ as $k \to \infty$ for $0 < \mathfrak{h} < 1$. From Equation (2.5.5) we deduce:

$$\sum_{k=0}^{\infty} \gamma(k) = \infty, \ 1/2 < \mathfrak{h} < 1,$$
$$\sum_{k=0}^{\infty} \gamma(k) < \infty, \ 0 < \mathfrak{h} < 1/2.$$

Consequently, if $1/2 < \mathfrak{h} < 1$, fractional Gaussian noise (hence fBm) (X_n) exhibits long range dependence.

Stochastic Integration and Fractional Brownian Motion

As stated in the introduction, Mandelbrot and Van Ness [1968] define the fBm as a moving average of dB_t. Decreusefond et al. [1999] shows that fBm can be written as the following stochastic integral driven by Brownian motion:

$$B_t^{\mathfrak{h}} = \int_0^t K_{\mathfrak{h}}(t,s)dB_s, \tag{2.5.6}$$

where the properties and analytical form of function $K_{\mathfrak{h}}(t,s)$ (called kernel) are given in [Decreusefond et al., 1999].

Multivariate Fractional Brownian Motion

Coutin and Qian [2002] give the following definition of a $d-$dimensional fractional Brownian motion:

Definition 3. *A fractional Brownian motion in dimension $d > 1$ is the random vector $B_t^{\mathfrak{h}} = (B_t^{\mathfrak{h},1}, \ldots, B_t^{\mathfrak{h},d})$ where all coordinates $B_t^{\mathfrak{h},i}$ are independent one-dimensional fractional Brownian motions of Hurst parameter $0 < \mathfrak{h} < 1$.*

Again a $d-$dimensional fBm reduces to a $d-$dimensional Brownian motion in the case $\mathfrak{h} = 1/2$.

SDE Driven by Fractional Brownian Motion

We can extent the stochastic differential equation (4.1) to define a ($d-$ dimensional) stochastic differential driven by a ($d-$dimensional) fBm of Hurst index $0 < \mathfrak{h} < 1$:

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t^{\mathfrak{h}}. \tag{2.5.7}$$

The same concepts of strong and weak solutions hold for the SDE (2.5.7). The SDE driven by Brownian motion (4.1) is of the form of the SDE (2.5.7) with $\mathfrak{h} = 1/2$.

In the rest of this thesis, we will call diffusion any processes solution of (2.5.7). We note that it does not match with the definition of [Karlin, 1981, Chapter 15, Section 1] given in Section 2.3, as the Markov property no longer holds due to the correlations between the fBm increments.

2.6 Summary

In this chapter, we presented Brownian motion from a probabilistic perspective. This process is of paramount importance in mathematics, physics and biophysics. It will be the process of reference in this thesis. We also presented the concept of diffusions. These processes can be seen as the solution of stochastic differential equations (SDE). Throughout this thesis, we will define the diffusion of interest through SDE. Fractional Brownian motion (fBm) is also introduced. In this manuscript, we will consider fBm as well as diffusion driven by fBm for modelling particle trajectories.

In the next chapter, we give the physical derivation of Brownian motion. We will also describe the motion models used in biophysics for describing intracellular dynamics, with a particular emphasis on the diffusion models defined in this chapter.

3 Diffusion for Modelling Intracellular Trajectories

In this chapter, we present the three main types of diffusion studied in biophysics to model intracellular motion, namely Brownian motion, subdiffusion and superdiffusion. We also described the different biological scenarios associated to each mode of diffusion. First, we present the physical models underlying Brownian motion. More specifically, we introduce the theory of Einstein [1905] and the Langevin approach. Then, we present subdiffusion processes which is often split in two parts: anomalous and confined diffusion. Finally, we deal with superdiffusion.

3.1 Einstein's Approach

In this section, we present the approach of Einstein [1905] introduced for modelling the motion of *"small suspended particles"* in a liquid. We develop the concept of Brownian motion in the exact same way as Einstein [1905]. First we depict the related physical experiment. Secondly, we show that the concentration of suspended particles is governed by a diffusion in the sense of Fick. Finally, the motion of individual suspended particles is modelled by a process corresponding to Brownian motion.

Physical Context

Einstein considers a particular physical situation. In first place, he assumes that z moles of a chemical specie is dissolved in a liquid of volume V. He also supposes that the solute is confined in a volume $V^\star$ separated from the pure solvent by a wall that is permeable to the solvent but not to the solute. In this situation, the solute produces a pressure on the wall called the osmotic pressure. Provided $z/V^\star$ is small enough, that is the solute concentration is low, we have:

$$pV^\star = RTz, \tag{3.1.1}$$

where p is the osmotic pressure, R is the gas constant and T is the temperature. Secondly, instead of the solute, Einstein considers suspended particles. Now the wall is permeable to the solvent but not to the particles. In this case, the theory of thermodynamics do not expect that the suspended particles will produce an osmotic pressure on the wall. However, according to the molecular-kinetic of heat, the only difference between

a dissolved molecule and a suspended body is their size. Then, Einstein points out that both the dissolved molecules and the suspended particles should produce the same osmotic pressure as long as their number is equal. Then he assumes that *"the suspended bodies perform an irregular, albeit very slow, motion in the liquid due to the liquid's molecular motion"*. This motion –we will see later that it corresponds to Brownian motion– is at the origin of the osmotic pressure. In fact, when the moving particles bounce on the wall, they exert a pressure as in the case of the solute. Then, we can derive a similar equation as (3.1.1):

$$pV^\star = RT\frac{n}{N}, \tag{3.1.2}$$

where n is the number of suspended particles and N the Avogadro number. Then n/N is the number of moles of the suspended particles.

In the sequel, for sake of simplicity, Einstein [1905] derives his theory in one dimension. In other words, the motion of the particles is along the x-axis and consequently we are only interested in the x-component of the forces applied on the particles.

Fick's Diffusion

In this paragraph, we are interested in the evolution of the concentration in space and time $\nu(x,t) = n(x,t)/dx$ where $n(x,t)$ in the number of suspended particles at time t in the small volume dx. Einstein [1905] assumes that a force K, depending on the position but not on the time, acts on each particle.

First, at the equilibrium we have:

$$K\nu - \frac{\partial p}{\partial x} = 0, \tag{3.1.3}$$

that is the force K and the force induced by the pressure p compensate each other. Using the definition of ν and Equation (3.1.2), we can rewrite Equation (3.1.3) as:

$$K\nu - \frac{RT}{N}\frac{\partial \nu}{\partial x} = 0. \tag{3.1.4}$$

On the other hand, the concentration ν is governed by a diffusion in the sense of Fick [1855]. In this case, diffusion refers to the evolution of a macroscopic quantity as the heat in a metal or the concentration of a chemical specie in a liquid. It is characterised by the two laws of Fick [1855]. Once combined, they give the diffusion equation which is written in our case as:

$$\frac{\partial \nu}{\partial t} = D\frac{\partial^2 \nu}{\partial^2 x}, \tag{3.1.5}$$

where D is the diffusion coefficient characterising the diffusion.

Now, to fully determined the diffusion of ν we need to derive D as a function of the parameters of the problem. To this end, we use the first law of Fick [1855] stating

that *"the diffusion flux between two points of different concentrations in the fluid is proportional to the concentration gradient between these points"*. In our case it can be written as:

$$J = -D\frac{\partial \nu}{\partial x}, \tag{3.1.6}$$

where J is the diffusion flux and D is the diffusion coefficient characterising the diffusion. Now we must derive the diffusion flux J that is the number of particles going through an area of unit one per unit of time. Einstein [1905] assumes that the suspended particles are spheric of radius a. Additionally, if the liquid has coefficient of viscosity k, then the force K gives to each particle the velocity,

$$\frac{\nu K}{6\pi k a}. \tag{3.1.7}$$

Consequently the diffusion flux is:

$$J = \frac{\nu K}{6\pi k a}. \tag{3.1.8}$$

In fact, a dimension analysis reveals that the inverse of a volume ($\nu = n/V^\star$) multiplied by a velocity (Equation (3.1.7)) defines a flux.

Finally, the first law of Fick [1855] gives:

$$\frac{\nu K}{6\pi k a} = -D\frac{\partial \nu}{\partial x}. \tag{3.1.9}$$

From Equations (3.1.4) and (3.1.9), the Fick's diffusion governing ν has for diffusion coefficient:

$$D = \frac{RT}{N 6\pi k a}. \tag{3.1.10}$$

Brownian Motion

Finally, Einstein [1905] models the *"disordered motions"* due to thermal molecular agitation of the n suspended particles. More importantly, Einstein links these individual motions to the Fick's diffusion examined in the previous paragraph. He assumes that the motions of individual particles are independent from each other. Moreover, he assumes that the displacements of a same particle on consecutive time intervals are independent as long as these time intervals are not too small. Then, in the following, we denote Δ the length of the time interval which is small compared to the observable time intervals but still satisfy the independence property of displacements. We recall that the displacements occur along the x-axis only. We denote Δ_x the displacement occurring during the period Δ. Einstein [1905] assumes that Δ_x is a random variable whose distribution function ϕ is symmetric. Then, the probability that a particle experiences a displacement

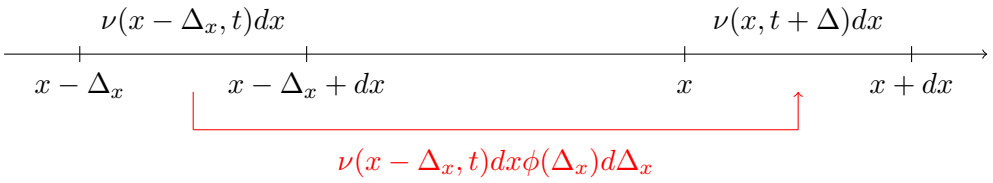


Figure 3.1: Scheme illustrating the transfer of particles from $x-\Delta_x$ to x between the times t and $t+\Delta$. There are $\nu(x-\Delta_x,t)dx$ particles in $[x-\Delta_x, x-\Delta_x+dx]$ at time t. Among them a proportion of $\phi(\Delta_x)d\Delta_x$ jump to $[x, x+dx]$ between t and $t+\Delta$. Integrating over all the displacements Δ_x, we obtain $\nu(x,t+\Delta)dx$ particles at time t in $[x, x+dx]$.

lying between u and $u+du$ is $\phi(u)du$. The average number of particles experiencing such a displacement during a period Δ is:

$$dn = n\phi(u)du. \tag{3.1.11}$$

Now, we can deduce the number of particles $\nu(x,t+\Delta)dx$ from the the numbers of particles at time t and ϕ. In Figure 3.1, we show how the particles go from $x-\Delta_x$ at time t to x at time $t+\Delta$ using Equation (3.1.11). Integrating over all the possible displacements we get:

$$\nu(x,t+\Delta)dx = dx.\int_{\mathbb{R}} \nu(x-\Delta_x,t)\phi(\Delta_x)d\Delta_x. \tag{3.1.12}$$

As Δ is small we can expand $\nu(x,t+\Delta)$ as:

$$\nu(x,t+\Delta) = \nu(x,t) + \Delta\frac{\partial\nu(x,t)}{\partial t}.$$

We also expand the left side of Equation (3.1.12) in Taylor series:

$$\int_{\mathbb{R}} \nu(x-\Delta_x,t)\phi(\Delta_x)d\Delta_x = \nu(x,t)\times 1 + \frac{\partial\nu(x,t)}{\partial x}\times 0 + \frac{\partial^2\nu(x,t)}{\partial^2 x}\int_{\mathbb{R}}\frac{\Delta_x^2}{2}\phi(\Delta_x)d\Delta_x,$$

where we use that $\int\phi(u)du = 1$ as ϕ is a distribution function and $\int u\phi(u)du = 0$ as ϕ is symmetric. We can equalize the right side of the two previous equations according to the equality given in Equation (3.1.12). Then, we deduce that ν respects the diffusion equation (3.1.5) predicted by the theory of Fick [1855] with diffusion coefficient given by:

$$D = \frac{1}{\Delta}\int_{\mathbb{R}}\frac{\Delta_x^2}{2}\phi(\Delta_x)d\Delta_x. \tag{3.1.13}$$

Therefore with a specific definition of the individual motion of n independent particles, Einstein [1905] shows that the concentration of such particles follows the Fick's equation.

At this step Einstein [1905] only assumed that the displacement of each particle over consecutive time intervals –for intervals not too small– are independent random variables from a symmetric distribution ϕ. Consequently, the particle motion fulfils the independence property of the Brownian increment, see Section 2.2. For the moment, we can not see why the displacement of the particles should be Gaussian as for Brownian particle. This link can be made by solving the diffusion equation (3.1.5).

We need additional conditions to solve Equation (3.1.5). Until this point, we have used the same coordinate system for all the particles. As there are independent of each other, we can define one coordinate system for each particle. Einstein [1905] states that the center of gravity of each particle at time $t = 0$ is the origin of their coordinate system. Then $\nu(x,t)dx$ now denotes the number of particles whose displacements between the times 0 and t is comprised between x and $x+dx$. In other words, x denotes the displacement and not the absolute position in a common coordinate system any more. Function ν still verify Equation (3.1.5) under this new scheme. Now we have the straightforward conditions:

$$\begin{aligned} \nu(x,0) &= 0, \quad \forall x \neq 0 \\ \int_{\mathbb{R}} \nu(x,0)dx &= n, \end{aligned} \tag{3.1.14}$$

Finally the solution of the diffusion equation (3.1.5) with conditions (3.1.14) is:

$$\nu(x,t) = n\frac{e^{\frac{-x^2}{4Dt}}}{\sqrt{4\pi Dt}}, \tag{3.1.15}$$

with x interpreted as a displacement as we have just said. With this meaning of x, $e^{-x^2/(4Dt)}/\sqrt{4\pi Dt}dx$ is the probability that the displacement of a single particle lies in $[x, x+dx]$. Therefore, the particle displacement is Gaussian. We also know that the displacements over consecutive time intervals are independent. Then the motion of the suspended particles defined by Einstein [1905] correspond to the Brownian motion defined in Section 2.2. Therefore the physical derivation of Brownian motion by Einstein [1905] is equivalent to the so-called Wiener process in mathematics. Due to the physical constraints, the diffusion coefficient D has a particular value given by Equation (3.1.10).

We can extend this theory to the d-dimensional case ($d = 2, 3$). In this context, each component follows a one-dimensional Brownian motion and the components are independent from each other. Not surprisingly, it corresponds to the Definition 3 of multi-dimensional Brownian motion.

Remark 3.1.1. *We note that, in this thesis, in case of the one-dimensional Brownian motion (B_t) the diffusion coefficient σ is defined as $\sigma = Var(B_1)$. Then we have the relationship $\sigma = 2D$.*

Remark 3.1.2. *From Equation* (3.1.11) *and* (3.1.15) *and the definition of ϕ we deduce that $\phi(x) = e^{-x^2/(4D\Delta)}/\sqrt{4\pi D\Delta}$. It is coherent with the equality* (3.1.13).

3.2 Langevin's Approach

Physicists define the motion of suspended particles in another way using the approach of Langevin [1908] (see [Kou, 2008] and [Schuss, 2009, Chapter 1]). This motion is sometimes refer to as Brownian motion which can be confusing. In this section, we present this alternative approach. First, we introduce the underlying physical model and the corresponding hypothesises about the particle motion. Secondly, we show that, in this case, the particle movement is governed by a well known stochastic differential equation. Thirdly, we explain why the particle motion defined by Einstein [1905] and by Langevin [1908] are mixed up. Finally, we explain which concept of Brownian motion we will use in the thesis. In this section, we derive the model directly in dimension d.

Langevin Equation

Langevin [1908] characterizes the particle motion through the $d-$dimensional (Langevin) equation:

$$m\frac{dv(t)}{dt} = -\zeta v(t) + L(t), \tag{3.2.1}$$

where $v : \mathbb{R}^+ \mapsto \mathbb{R}^d$ is the velocity of the particle, m its mass, $\zeta > 0$ the friction coefficient and $L : \mathbb{R}^+ \mapsto \mathbb{R}^d$ a random force resulting from the collisions with the surrounding particles. In case of spherical particles of radius a immersed in a liquid of viscosity coefficient k, the friction coefficient is $\zeta = 6\pi ka$ where k is the viscosity coefficient of the surrounding liquid.

Uhlenbeck and Ornstein [1930] constrained $L(t)$ with two additional assumptions. First, the mean of $L(t)$ over a large number of independent colliding particles is 0, that is $\mathbb{E}(L(t)) = \mathbf{0}_d$, where $\mathbf{0}_d$ is the null vector of $\mathbb{R}_d$. In their physical model, Uhlenbeck and Ornstein [1930] also assume that the colliding particles are similar to the particle of interest and have same initial speed v_0. Secondly, the autocorrelation function is given by:

$$\mathbb{E}(L(t)L(s)^T) = \sigma\delta(t-s)\mathbf{I}_d, \tag{3.2.2}$$

where $\sigma > 0$ is a constant, δ is the Kronecker function and $\mathbf{I}_d$ the identity matrix of size d. The idea is that each collision is practically instantaneous and that successive collisions are uncorrelated. Actually, Uhlenbeck and Ornstein [1930] originally model the autocorrelation function as a function of $t - s$ with a sharp peak of width equal to the duration of a single collision. The autocorrelation (3.2.2) is preferred nowadays [Van Kampen, 1992, chapter 9]. Such a force $L(t)$ is called a Langevin force.

Ornstein-Uhlenbeck Process

We did not fully define the stochastic process $L(t)$ as we provide only information on its first and second moment. Such a process is known as white noise in statistics. If

we further assume that $L(t)$ is Gaussian, we entirely define this process as a Gaussian process is determined by its first two moments. Then, $L(t)$ is called a Gaussian white noise. As explained in [Karlin, 1981, Chapter 15, Section 14], the Gaussian white noise $L(t)$ can be informally defined as the derivative of the Wiener process –equivalently the mathematical Brownian motion defined in Section 2.2 – $L(t) = \sigma dB_t/dt$. We use the word informally as in fact the Wiener process is nowhere differentiable. Finally, we can rewrite the Langevin equation (3.2.1) as the $d-$dimensional stochastic differential equation:

$$mdv(t) = -\zeta v(t)dt + \sigma dB_t. \tag{3.2.3}$$

The solution of the stochastic equation (3.2.3) is known as the Ornstein-Uhlenbeck process. It is a Gaussian process with:

$$\mathbb{E}(v(t)) = \mathbf{0}_d, \tag{3.2.4}$$

$$\mathbb{E}(v(t)v(s)^T) = \frac{\sigma^2}{2\zeta m} e^{-(\zeta/m)|t-s|}\mathbf{I}_d. \tag{3.2.5}$$

Waterston and Rayleigh [1892] states that, at the equilibrium (that is as $t \to \infty$), the mean square velocity verifies:

$$\lim_{t\to\infty} \mathbb{E}(\|v(t)\|_2^2) = d\frac{k_B T}{m}, \tag{3.2.6}$$

where k_B is the Boltzmann constant and T is the temperature. Each component of the velocity vector has the same variance, so that:

$$\lim_{t\to\infty} \mathbb{E}(v_i(t)^2) = \frac{k_B T}{m}, \quad i = 1, \ldots, d. \tag{3.2.7}$$

Then, equalizing the variances of $v_i(t)$ obtained with Equation (3.2.5) with $t = s$ and obtained with Equation (3.2.7), we have the relationship:

$$\sigma = \sqrt{2\zeta k_B T}. \tag{3.2.8}$$

Finally, the Brownian motion of Langevin [1908] is defined as:

$$X_t = \int_0^t v(s)ds \tag{3.2.9}$$

where $v(t)$ is the Ornstein-Uhlenbeck process solution of the SDE (3.2.3). Due to the Gaussian nature of $v(t)$, (X_t) is also a Gaussian process.

Mean Square Displacement

One reason explaining the confusion between the particle motion respectively defined by Einstein [1905] and Langevin [1908] is that they both exhibit a linear mean square displacement asymptotically. In the case of the d-dimensional Brownian motion of Einstein [1905], we can easily show that the mean square displacement is:

$$\begin{aligned}\mathbb{E}(\|X_t - X_0\|^2) &= d2Dt \\ &= d\frac{2RT}{N6\pi ka}t \\ &= d\frac{2k_BT}{\zeta}t,\end{aligned} \tag{3.2.10}$$

where $k_B = R/N$ is the Boltzmann constant and $\zeta = 6\pi ka$ is the friction coefficient.

In the case of the motion defined by Langevin [1908] (assuming $X_0 = 0$ for simplicity) we have:

$$\begin{aligned}\mathbb{E}(\|X_t - X_0\|^2) &= \sum_{i=1}^{d} \mathbb{E}\left(\int_0^t \int_0^t v^i(s)v^i(u)dsdu\right) \\ &= d\int_0^t \int_0^t \mathbb{E}(v^1(s)v^1(u))dsdu \\ &= d\frac{2k_BT}{\zeta}\left(t - \frac{m}{\zeta}(1 - e^{-(\zeta/m)t})\right) \\ &= d\frac{2k_BT}{\zeta}t + o(t)\end{aligned} \tag{3.2.11}$$

where $o(t) \to 0$ as $t \to \infty$.

Choice of the Definition of Brownian Motion

Each approach relies on different physical models. We emphasize that the Brownian motion of Einstein [1905] (corresponding to the Wiener process) is nowhere differentiable and then has a rough (but still continuous) path. On the other hand, the particle motion defined by Langevin [1908] is differentiable due to its definition as the integration of the Ornstein-Uhlenbeck process (Equation (3.2.9)). Then its path is smooth. Bressloff [2014] argues that both processes can be used to model intracellular dynamics in the case where the particle evolves freely inside the cytosol or along the plasma membrane.

> In this thesis, Brownian motion will refer to the motion defined by Einstein [1905]. It corresponds to the mathematical Brownian motion defined in Section 2.2 called also Wiener process in the mathematical literature.

3.3 Subdiffusion

Subdiffusion, which includes confined diffusion and anomalous diffusion, are the translations of several biological scenarios. In this section, we present models associated to these two types of diffusion. We note that certain models are called diffusion while there are not solutions of SDE. In this thesis, we will not distinguish confined and anomalous diffusion and consider that both are subdiffusion.

Anomalous Diffusion

In biophysics, [Saxton and Jacobson, 1997, Meroz and Sokolov, 2015], an *anomalous diffusion* (X_t) is characterized by a MSD which is proportional to the monome t^β,

$$\mathbb{E}(\|X_t - X_0\|^2) \propto t^\beta, \tag{3.3.1}$$

with $\beta < 1$. The first two presented models are solutions of a SDE driven by fBm (2.5.7) (the first being simply fBm). Then we present other type of processes used in biophysics.

Fractional Brownian motion As a particle moves through the cytoplasm, the latter pushes it back, due to macromolecular crowding and the presence of elastic elements generating correlations in the particle's trajectory [Jeon et al., 2011]. A fBm with Hurst index $0 < \mathfrak{h} < 1/2$ is a good candidate to model this situation. First, it is straightforward to show that its MSD is given by (3.3.1) with $\beta = 2\mathfrak{h} < 1$ (see Equation (2.5.1)). Secondly, we saw in Section 2.5 that fBm has its increments negatively correlated when $0 < \mathfrak{h} < 1/2$. As an example, Weber et al. [2010] study the mechanisms underlying subdiffusive motion in live Escherichia coli cells thanks to fluorescently labeled chromosomal loci and RNA-protein particles. They conclude that the observed motion was well modelled by fBm.

Generalized Langevin equation (GLE) As we have just explained, particles can be slowed by the contrary current due to the viscoelastic properties of the cytoplasm. This time we are interested in long-time correlations (and not just correlations) in diffusive motion. Then, Kou [2008] models such phenomenon with a stochastic differential equations driven by the fBm with Hurst index $1/2 < \mathfrak{h} < 1$; in fact we saw in Section 2.5 that in this case fBm exhibits long range dependence. Then, Zwanzig [2001] and Chandler [1987] proposed the generalized Langevin equation (GLE):

$$m\frac{dv(t)}{dt} = -\zeta \int_{-\infty}^{t} v(u)K(t-u)du + G(t), \tag{3.3.2}$$

where, in comparison with the Langevin equation (3.2.1), $G(t)$ is a noise having memory replacing the memoryless white noise $L(t)$; the velocity is convolved with a kernel K.

These two features make the solution of the Equation (3.3.2) a non-Markovian process. We note that both K and G must appear in the equation in order to fulfil a physical constraint comparable to Equation (3.2.6) (also called fluctuation-dissipation principle in [Chandler, 1987]):

$$\mathbb{E}(G(t)G(s)^T) = 2\zeta k_B T K(t-s)\mathbf{I}_d. \tag{3.3.3}$$

Not surprisingly, we observe that if we choose $K = \delta$ –the Dirac function– we find that the GLE (3.3.2) is equivalent to the Langevin equation (3.2.1) and the condition on the second moment (3.3.3) is equivalent to the condition (3.2.2). Kou [2008] chooses to define $G(t)$ as fractional Gaussian noise (2.5.3) with Hurst index $1/2 < \mathfrak{h} < 1$ for exhibiting long range dependence. From condition (3.3.3), they deduce the kernel K (noted now $K_\mathfrak{h}$):

$$K_\mathfrak{h}(t) = 2\mathfrak{h}(2\mathfrak{h}-1)|t|^{2\mathfrak{h}-2}. \tag{3.3.4}$$

Then the related stochastic differential equation is:

$$mdv(t) = -\zeta\left(\int_{-\infty}^{t} v(u)K(t-u)du\right)dt + \sigma dB_t^\mathfrak{h}, \tag{3.3.5}$$

where $\sigma = 2\zeta k_B T$ and $(B_t^\mathfrak{h})$ is a fBm with $1/2 < \mathfrak{h} < 1$. Finally, Kou [2008] shows that the integrated process $X_t = \int v(u)du$ verifies as $t \to \infty$:

$$\mathbb{E}(\|X_t - X_0\|^2) \propto t^{2-2\mathfrak{h}}, \tag{3.3.6}$$

It fulfils the MSD condition (3.3.1) asymptotically with $\beta = 2 - 2\mathfrak{h} < 1$ for $1/2 < \mathfrak{h} < 1$.

Remark 3.3.1. *Kou [2008] studies only one-dimensional process. Here, we explain how we can extend the models of Kou [2008] in higher dimensions. It is quite natural to define physical Brownian motion in higher dimensions as a stack of one-dimensional physical Brownian motion. It is what we implied writing Equation* (3.2.5) *with* $\mathbf{I}_d$*. In fact, in this case, the Langevin force* $L(t)$ *is modelled as a white noise and the component of* $d-$*dimensional white noise are independent. However, when we use the GLE* (3.3.2)*, we can wonder if the components of the noise* G *are necessarily independent. For instance, we could create some correlations through the kernel* K*. Here, for simplicity, we considered that all the components were independent and shared the same (one-dimensional) kernel.*

Continuous time random walk (CTRW) Intracellular particles can also bind to molecular complexes. Then, the particle motion is a permanent switch between binding events and movement toward another spot where it can bind again. Scher and Montroll [1975] introduce the continuous time random walk (CTRW) to model anomalous transport properties of charge carriers in amorphous materials. In their framework, the electron dynamics are successively trapped in different energy wells; the

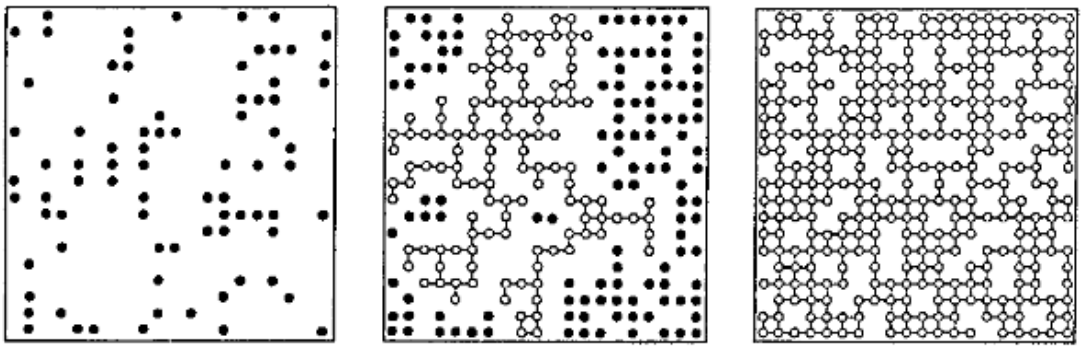


Figure 3.2: Percolation clusters on a square lattice for different values of p. We use a 20×20 lattice. From left to right percolation clusters obtained with $p = 0.20, 0.59, 0.80$ (the middle image correspond to the case $p = p_c$). Sites belonging to finite clusters are marked by full circles, and sites on the 'infinite' clusters are marked by open circles. (extracted from Havlin and Ben-Avraham [1987]).

total time spent in the trapped states is much larger than the time spent in free motion. In this model, a particle performs random jumps whose step length is generated by a probability density with finite second moments. The waiting times between jumps are assumed to be distributed according to a probability distribution $\psi(t)$. If $\psi(t)$ has a finite first moment that is $\int t\psi(t)dt < \infty$ then the mean square displacement of the CTRW is linear in time. For instance, we can use the exponential distribution:

$$\psi(t) = (1/\tau)e^{-t/\tau}, \ t > 0, \tag{3.3.7}$$

where $\tau > 0$ is called the characteristic time. We note that, in this case, the random walk has the Markov property (due to the memoryless property of the exponential distribution). On the contrary, if $\int t\psi(t)dt = \infty$ the mean square displacement of the CTRW is given by (3.3.1). A typical choice is a power law distribution:

$$\psi(t) = 1/(1 + t/\tau)^{1+\beta}, \ t > 0, \tag{3.3.8}$$

with $\tau > 0$ the characteristic time and $0 < \beta < 1$.

In neurobiology, Zhizhina et al. [2015] propose to investigate CTRW to model the axon growth. The growth of an axon to its target is guided by chemical signals from the cellular environment. The authors describe this interaction by a random waiting time thereby defining a CTRW. They observe that "normal" axons and "mutant" axons are driven by CTRW with different waiting time distribution.

Random walk on fractal The inner environment of a cell is crowded with small solutes and macromolecules which occupy 10-50% of the volume [Dix and Verkman,

2008]. If the concentration of obstacles is sufficiently high, the mean square displacement of the particle is given by Equation (3.3.1) [Havlin and Ben-Avraham, 1987, Saxton, 1994]. In this case, the domain where they evolve develops a fractal-like structure. Then, a popular model is the random walk on percolation clusters [Havlin and Ben-Avraham, 1987].

For simplicity, we present the model on a 2−dimension square lattice. Each vertex of the lattice has probability $1 - p$ to be an obstacle that is the particle can not go on this kind of vertex. The other vertices can be occupied by particles. They form connected clusters on which particles are assumed to undergo a random walk. In this very case, there exists a critical probability $p_c = 0.592745$ below which there exists only finite clusters and above which there exists one infinite cluster (see Figure 3.2) [Havlin and Ben-Avraham, 1987]. When $p = p_c$, the random walk on the infinite cluster have its MSD given by Equation (3.3.1) [Havlin and Ben-Avraham, 1987]. In the literature of diffusion on fractals, they parametrize the MSD (3.3.1) by $\beta = d/d_w$ where d is the dimension and d_w a parameter called the fractional dimension of the random walk. In the two-dimensional case ($d = 2$), the fractional dimension of the random walk on a square lattice with $p = p_c$ is $d_w = 2.8784$ [Grassberger, 1999] leading to $\beta = 0.6948$. Havlin and Ben-Avraham [1987] also consider other choices of p and random walks on both the finite and infinite percolation clusters. However, in these cases, the MSD is not a power function. As another two-dimensional example, the fractional dimension of a random walk on the Sierpinski gasket fractal gives $d_w = 2.32$ (then $\beta = 0.8621$)[Havlin and Ben-Avraham, 1987]. Berry and Chaté [2014] argues that the exponents β observed from real experiments span a wide range of values and that random walks on fractal can not model all these possibilities. Then some authors [Berry and Chaté, 2014, Saxton, 1994] prefer relying on Monte-Carlo simulations with different designs of obstacles (mobiles or not) to propose a model explaining the observed power function form of the MSD.

Confined Diffusion

In biophysics [Saxton and Jacobson, 1997, Monnier et al., 2012], a *confined diffusion* (X_t) is characterized by a MSD of the form:

$$\mathbb{E}(\|X_t - X_0\|^2) = \frac{r_c^2}{a}(1 - be^{-c\sigma^2/(2r_c^2)}), \tag{3.3.9}$$

where parameters r_c is the characteristic size of the region of confinement, a is a scale parameter and b and c depends on the shape of the region. Parameter $\sigma > 0$ is the constant diffusion coefficient. We present two models of confined diffusion and give their mean square displacements. For the first model, the MSD (3.3.9) is a simplification of the true MSD. We find the MSD (3.3.9) for a particular case of the second model. We note that parameter a does not appear in Saxton and Jacobson [1997], Monnier et al. [2012]. We use this extra scale parameter a to have the common expression (3.3.9) for

the MSD of the two presented models.

Diffusion within confined geometries The plasma membrane is parceled up into compartments where proteins undergo short-term confined diffusion. More specifically these compartments are separated by the actin-based membrane skeleton [Kusumi et al., 2005]. Then, the motion can be modelled by the SDE (4.1) adding boundary conditions. Equation (3.3.9) is based on the first term of the exact series solution of the MSD of a Brownian particle trapped in a square or circular corral (in dimension 2) or in a sphere (in dimension 3) [see Kusumi et al., 1993, Saxton, 1993]. As an example, Bickel [2007] shows that, for a certain type of boundary condition, the MSD of a Brownian motion confined in a circular domain of radius r_c is given by:

$$\mathbb{E}(\|X_t - X_0\|^2) = r_c^2 \left(1 - 8\sum_{i=1}^{\infty} \exp\left[-\iota_{1i}^2 \frac{t}{\tau}\right] \frac{1}{\iota_{1i}^2(\iota_{1i}^2 - 1)}\right), \tag{3.3.10}$$

where $0 < \iota_{1,1} < \iota_{1,2} < \dots$ are the positive zeros of J_1', the first derivative of the Bessel function of order one J_1 and $\tau = 2r_c^2/\sigma^2$ is the characteristic time. We note that, as expected, the MSD saturates to r_c^2 in the long-time limit $t \gg \tau$. Then, Equation (3.3.9) is the first term of the sum (3.3.10) with $a = 1$, $b = 8/(\iota_{11}^2(1 - \iota_{11}^2))$ and $c = \iota_{11}^2$. Parameters σ and r_c are unchanged in the two equations (3.3.10) and (3.3.9).

Diffusion in a potential well We can state that a particle is attracted by an external force modelled by a potential well U. Originally, Kramers [1940] introduced such a model for describing chemical reactions. His model can be seen as the ($d-$dimensional) Langevin equation (3.2.3) (written here as a SDE) with an extra term depending on U:

$$mdv(t) = -\zeta v(t)dt - \nabla U(X_t) + \sqrt{2\zeta k_B T} dB_t, \tag{3.3.11}$$

where ∇ denotes the gradient operator. Now we make other assumptions on Equation (3.3.11) to obtain a process with the MSD (3.3.9). First, we suppose that the viscosity is very large, that is the friction coefficient ζ tends to infinity. Then, the acceleration term $mdv(t)$ is negligible. This corresponds to the so-called overdamped condition in physics [Van Kampen, 1992]. The model reduces to:

$$\zeta dX_t = -\nabla U(X_t) + \sqrt{2\zeta k_B T} dB_t, \tag{3.3.12}$$

where $dX_t = v(t)dt$. Now, we assume that the potential U is uni-modal; in other words the particle is trapped in a single domain. In this case, U can be approximated by a polynomial of order 2. For simplicity, suppose that the potential is given by the following polynomial:

$$U(x_1, \dots, x_d) = (1/2)\sum_{i=1}^{d} k_i(x_i - \theta_i)^2, \tag{3.3.13}$$

where $k_i > 0$, $\theta_i \in \mathbb{R}$ and d is the dimension of the process. Then the SDE (3.3.12) turns into:

$$dX_t^i = -\lambda_i(X_t^i - \theta_i)dt + \sigma dB_t^i, \qquad i = 1, \dots, d, \tag{3.3.14}$$

where $\sigma = \sqrt{2k_BT\zeta}$ and $\lambda_i = k_i/\zeta > 0$. As in the case of Equation (3.2.3), the solution of the SDE (3.3.14) is the Ornstein-Uhlenbeck process (different parametrization compared to the SDE (3.2.3) with the extra parameters θ_i though). The parameter k_i measures the strength of attraction of the potential (related to the potential depth) while $\theta = (\theta_1, \dots, \theta_d)$ is the equilibrium position of the particle. As we already mentioned, the Ornstein-Uhlenbeck is a Gaussian process with normal stationary distribution. In the case of the Ornstein-Uhlenbeck (4.2.1), the mean and covariance of the stationary distribution are:

$$\mathbb{E}(X_t) = \theta, \tag{3.3.15}$$

$$\mathrm{Cov}(X_t, X_s) = \frac{\sigma^2}{2}\begin{pmatrix} (1 - e^{-\lambda_1|t-s|})/\lambda_1 & & 0 \\ & \ddots & \\ 0 & & (1 - e^{-\lambda_d|t-s|})/\lambda_d \end{pmatrix}. \tag{3.3.16}$$

The MSD of the Ornstein-Uhlenbeck process (4.2.1) is given by:

$$\mathbb{E}(\|X_t - X_0\|^2) = \sigma^2(1 - e^{-\lambda t})\sum_{i=1}^{d}(1/\lambda_i), \tag{3.3.17}$$

when X_0 is drawn with the stationary distribution. When $\lambda_i = \lambda$ for $i = 1, \dots, d$ Equation (3.3.17) reduces to:

$$\mathbb{E}(\|X_t - X_0\|^2) = \frac{d\sigma^2(1 - e^{-\lambda t})}{\lambda}. \tag{3.3.18}$$

Then, we obtain the MSD (3.3.9) with $r_c^2 = \sigma^2/(2\lambda)$, $a = 2/d$ and $b = c = 1$.

As an example, Hozé [2013][Chapter 2, Section 2.9] studies the postsynaptic AMPA-type glutamate receptor (AMPAR), a protein involved in the fast excitatory synaptic transmission. AMPAR plays a crucial part in many aspects of brain functions including learning, memory and cognition. Aberrant AMPAR trafficking is implicated in neurodegenerative process [Henley et al., 2011]. Hozé [2013][Chapter 2, Section 2.9] uses the overdamped Equation (3.3.12) with a polynomial of order 2 for the potential U to model potential wells attracting AMPAR in the synapses.

3.4 Superdiffusion

We note that less attention has been paid to superdiffusion in biophysics. We present here the most popular models.

Brownian with Drift

At the macroscopic level, the main type of active intracellular transport involves molecular motors which carry particles (cargo) along microtubular filament tracks. The molecular motors and their cargo undergo superdiffusion on a network of microtubules in order to reach a specific area quickly. The molecular motor moves step by step along the microtubules thanks to a mechanicochemical energy transduction process. A single step of the molecular motor is modelled by the so-called Brownian ratchet [Reimann, 2002]. When we observe the motion of the molecular motor along a filament on longer time-scales (several steps), its dynamic can be approximated by a Brownian motion with constant drift (also called directed Brownian) [see Peskin and Oster, 1995, Elston, 2000].

The Brownian motion with drift is solution of the SDE :

$$dX_t^i = v_i dt + \sigma dB_t^{1/2,i}, \qquad i = 1, \dots, d, \tag{3.4.1}$$

where $v = (v_1, \dots, v_d) \in \mathbb{R}^d$ is the constant drift parameter modelling the velocity of the molecular motor. Then the MSD of the directed Brownian motion is given by:

$$\mathbb{E}(\|X_t - X_0\|^2) = \|v\|_2^2 \, t^2 + d\sigma^2 t, \tag{3.4.2}$$

the linear component coming from the Brownian part while the quadratic part is due to the constant drift. In absence of the Brownian component the MSD is quadratic, the motion is described as ballistic that is the particle goes straight.

Anomalous Superdiffusion

Anomalous superdiffusions are the analogue to anomalous subdiffusion. Then the MSD of an *anomalous superdiffusion* (X_t) is characterized by a MSD which is proportional to the monome t^β,

$$\mathbb{E}(\|X_t - X_0\|^2) \propto t^\beta, \tag{3.4.3}$$

with $1 < \beta < 2$.

Fractional Brownian motion Superdiffusion can also be modelled by the fractional Brownian motion with Hurst parameter $1/2 < \mathfrak{h} < 1$. In fact, we know that the MSD of the fBm is given by Equation (3.4.3). However, we note that in biophysics the use of the fractional Brownian motion is mainly related to subdiffusion.

3.5 Summary

In this chapter, we presented the three types of diffusion of interest in this thesis, namely Brownian motion, subdiffusion and superdiffusion. For each diffusion type, we gave examples of models used in biophysics. There exists a wide variety of models for subdiffusion and superdiffusion. We emphasized that, in biophysics, some processes are

considered as subdiffusive or superdiffusive even if there are not diffusion according to the probabilistic definition, see Section 2.3. As an example, continuous time random walks (CTRW) are not diffusions since their paths are not continuous.

In the next part, we define a test procedure to classify the observed trajectories into the three diffusion types. Then, we define subdiffusion and superdiffusion as solution of stochastic differential equations. However, we note that, in principle we can adapt our test to deal with a larger range of stochastic processes, including random walks. Such refinement is out of the scope of this thesis. Throughout this thesis, we will evaluate the proposed methods on diffusions presented in this chapter. We will use the Ornstein-Uhlenbeck process and the fBm (with $0 < \mathfrak{h} < 1/2$) for modelling subdiffusion. We will use the Brownian with drift (4.2.3) and the fBm (with $1/2 < \mathfrak{h} < 1$) for modelling superdiffusion.

Part I

A Statistical Test for the Classification of Trajectories

4 Test Procedures

We suppose that the trajectory $\mathbb{X}_n = (X_{t_0}, \ldots, X_{t_n})$ is generated from some unknown d-dimensional ($d = 2$ or $d = 3$) diffusion process (X_t) solution of the SDE (4.1). We note that we will emphasize the dependence on d on the notations only if necessary. Our procedure allows to test from which type of diffusion the observed trajectory is generated. We derive two tests:

1. H_0 *"(X_t) is a Brownian motion" versus* H_1 *"(X_t) is a subdiffusion",*
2. H_0 *"(X_t) is a Brownian motion" versus* H_2 *"(X_t) is a superdiffusion".*

Then, we aggregate the two procedures to build a three-decision procedure. Finally we propose a multiple test procedure to test a set of independent trajectories. This procedure allows to control the number of false detections that is the number of trajectories detected as non-Brownian while they are.

4.1 Model

We observe the successive positions of a single particle in a d-dimensional space ($d = 2$ or 3) at times $t_0, t_1, \ldots t_n$. We suppose the lag times between two consecutive observations is a constant Δ. The observed trajectory of the particle is:

$$\mathbb{X}_n = (X_{t_0}, X_{t_1}, \ldots, X_{t_n}), \tag{4.1.1}$$

where $X_{t_i} \in \mathbb{R}^d$ is the position of the particle at time $t_i = t_0 + i\Delta$, $i = 0, \ldots, n$.

The discrete trajectory is generated by a stochastic process $(X_t)_{t_0 \le t \le t_n}$ with continuous path and which is a solution of the stochastic differential equation (SDE),

$$dX_t = \mu(X_t)dt + \sigma dB_t^{\mathfrak{h}}, \qquad t \in [t_0, t_n], \tag{4.1.2}$$

where $B_t^{\mathfrak{h}} = (B_t^{\mathfrak{h},i})_{i=1\ldots d}$ are unobserved fractional Brownian motion of unknown Hurst parameter $\mathfrak{h}$, that is the processes $B_t^{\mathfrak{h},i}$ are independent and are standard fractional Brownian motion. The unknown paremeters of the model are the Hurst parameter $\mathfrak{h} \in (0, 1)$, the diffusion coefficient $\sigma > 0$ and the drift term $\mu : \mathbb{R}^d \to \mathbb{R}^d$. The SDE admits a unique strong solution according to a given initial condition X_{t_0}, if the parameters of the model fulfil Assumption 1 or Assumption 2.

Assumption 1. *If $h \in [1/2, 1)$, μ fulfils the linear growth hypothesis,*

$$\exists K > 0, \quad \forall x \in \mathbb{R}^d \qquad \|\mu(x)\| \leq K(1 + \|x\|), \tag{4.1.3}$$

and the Lipschitz condition,

$$\exists M > 0, \quad \forall x \in \mathbb{R}^d \qquad \|\mu(x) - \mu(y)\| \leq M\|x - y\|, \tag{4.1.4}$$

where $\|\cdot\|$ is the euclidean norm of $\mathbb{R}^d$.

Assumption 2. *If $h \in (0, 1/2)$, for all $x = (x^1, x^2, \dots x^d) \in \mathbb{R}^d$, the drift term μ is rewritten as*

$$\mu(x) = \left(\mu_1(x^1), \dots, \mu_d(x^d)\right), \tag{4.1.5}$$

where, for all $i = 1 \dots d$, $\mu_i : \mathbb{R} \to \mathbb{R}$ fulfils the linear growth hypothesis,

$$\exists K > 0, \quad \forall x^i \in \mathbb{R} \qquad |\mu_i(x^i)| \leq K(1 + |x^i|), \tag{4.1.6}$$

and the Lipschitz condition,

$$\exists M > 0, \quad \forall x^i \in \mathbb{R} \qquad |\mu_i(x^i) - \mu_i(y^i)| \leq M|x^i - y^i|. \tag{4.1.7}$$

We denote by $\mathcal{L}$ the set of functions μ verifying Assumption 1 or 2. Assumption 1 is sufficient to ensure that the SDE (4.1) admits a strong solution when $1/2 < \mathfrak{h} < 1$ [Mishura, 2008, Chapter 3]; Nualart and Ouknine [2002] show that, under Assumption 2, the SDE (4.1) admits a strong solution when $0 < \mathfrak{h} \leq 1/2$. In the following, $P_{\mathfrak{h},\mu,\sigma}$ denotes the measure induced by the stochastic process (X_t) solution of (4.1). This measure comprises all the finite-dimensional distributions of the process. We also note $\mathcal{P} = \{P_{\mathfrak{h},\mu,\sigma} : 0 < \mathfrak{h} < 1, \mu \in \mathcal{L}, \sigma > 0\}$ the set of solutions of the SDE (4.1).

Remark 4.1.1. *We adopt the large-sample scheme to derive asymptotic properties of the test procedure presented in 4, that is the inter-observation time Δ remains fixed and the number of observations n tends to infinity. Other schemes exist (see [Fuchs, 2013, Section 6.1.3]) as the high-frequency scheme for which Δ tends to zero while the duration of observation is fixed. In the experimental context of microscopic sequences, Δ is the resolution of the microscopy device while n is the number of frames during which we track the particle. The resolution of the microscopy device is fixed during the experiment. Moreover, in an ideal situation, we track the particle during an infinite time of observation therefore the number of frames n tends to infinity. Then, the large-sample scheme is the most realistic scheme in our context. [Fuchs, 2013, Section 6.1.3] also emphasizes that the large-sample scheme is the most realistic in real applications while the high-frequency scheme is convenient from a theoretical point of view.*

4.2 Parametric Diffusion of Interest

In this section, we mention parametric diffusions of interest which can be used to model subdiffusion or superdiffusion, the alternative hypothesises of our test. We recall that the null hypothesis of the test is the mathematical Brownian motion or Wiener process. In this chapter, we will derive asymptotic properties of our test under these parametric alternatives. In Chapter 5, we will assess our test based on simulations of these parametric diffusion processes. As these diffusions have already been presented in Chapter 3, we just give their name and related SDE.

Subdiffusion

For modelling subdiffusion, we use the Ornstein-Uhlenbeck process:

$$dX_t^i = -\lambda(X_t^i - \theta_i)dt + \sigma dB_t^{1/2,i}, \qquad i = 1, \dots, d \tag{4.2.1}$$

where $\lambda > 0$ and $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$.

We also consider the fractional Brownian motion:

$$dX_t^i = \sigma dB_t^{\mathfrak{h},i}, \qquad i = 1, \dots, d, \tag{4.2.2}$$

where $0 < \mathfrak{h} < 1/2$.

Superdiffusion

For modelling superdiffusion, we use the Brownian motion with drift solution of the SDE:

$$dX_t^i = v_i dt + \sigma dB_t^{1/2,i}, \qquad i = 1, \dots, d, \tag{4.2.3}$$

where $v = (v_1, \dots, v_d) \in \mathbb{R}^d$.

Superdiffusion will also be modelled by the fractional Brownian motion with Hurst parameter $1/2 < \mathfrak{h} < 1$.

4.3 The Test Statistic

Let us consider the standardized maximal distance T_n of the process from its starting point:

$$T_n = \frac{D_n}{\sqrt{(t_n - t_0)\hat{\sigma}_n^2}}, \tag{4.3.1}$$

where D_n is the maximal distance of the process from its starting point,

$$D_n = \max_{i=1,\dots,n} \|X_{t_i} - X_{t_0}\|, \tag{4.3.2}$$

and $\hat{\sigma}_n$ is a consistent estimator of σ. The choice of $\hat{\sigma}$ is discussed in Section 4.6. If T_n is low, it means the process stays close to its initial position during the period $[t_0, t_n]$: it is likely that it is a subdiffusion. On contrary, if T_n is large, it means the process goes away from its starting point as a superdiffusion does with high probability. It is worth noting that T_n can be related to the mean maximum excursion second moment proposed by Tejedor et al. [2010] as an alternative to MSD. Now, this new measure T_n introduces an order in the diffusion processes solution of the SDE (4.1). Then, it allows to classify them into the different classes of diffusion *i-e* free diffusion, superdiffusion and subdiffusion. We want to build a test whose null hypothesis is that the trajectory comes from a Brownian motion, the gold standard process in biophysics. As a consequence, T_n must be a pivotal statistic under the hypothesis H_0 that is the trajectory is Brownian.

Lemma 4.3.1. *Let $\hat{\sigma}_n$ be a consistent estimator of σ such that the distribution of $\hat{\sigma}_n/\sigma$ does not depend on σ. If (X_t) is a Brownian Motion, the distribution of T_n does not depend on σ.*

Let $q_n(\alpha)$ the quantile of T_n of order $\alpha \in (0,1)$ when (X_t) is a Brownian motion. From Lemma 4.3.1, $q_n(\alpha)$ does not depend on σ.

4.4 Two Hypothesis Test Procedures Derived from the Test Statistic

First, we define $\phi_{1,\alpha}$ the hypotheses test associated to H_0 versus H_1 at level $\alpha \in (0,1)$. The procedure $\phi_{1,\alpha}$ is defined through its critical region,

$$\mathcal{R}_{1,\alpha} = \{T_n < q_n(\alpha)\}, \tag{4.4.1}$$

as the following,

$$\phi_{1,\alpha}(\mathbb{X}_n) = \begin{cases} 1 & \text{if } \mathbb{X}_n \in \mathcal{R}_{1,\alpha}, \\ 0 & \text{otherwise.} \end{cases}$$

Then T_n has probability α to lie in the critical region (4.4.1). According to Lemma 4.3.1, the level of the test $\phi_{1,\alpha}$ is α,

$$\sup_{\sigma>0} P_{1/2,0,\sigma}\left(T_n < q_n(\alpha)\right) = \alpha. \tag{4.4.2}$$

In a similar way, we can perform the test $\phi_{2,\alpha}$ by replacing subdiffusion by superdiffusion in the alternative hypothesis. The associated critical region is :

$$\mathcal{R}_{2,\alpha} = \{T_n > q_n(1-\alpha)\}. \tag{4.4.3}$$

Table 4.1: The three kinds of error in a three-decision test procedure.

	Decision		
Truth	Do not Reject H_0	Decide H_1	Decide H_2
H_0 True	No error	Type I	Type I
H_1 True	Type II	No error	Type III
H_2 True	Type II	Type III	No error

4.5 A Three-Decision Test Procedure

From the two tests $\phi_{1,\alpha/2}$ and $\phi_{2,\alpha/2}$, we define a new procedure ϕ as follows,

$$\begin{cases} \text{we decide } H_1 \text{ if } \mathbb{X}_n \in \mathcal{R}_{1,\alpha/2}, \\ \text{we decide } H_2 \text{ if } \mathbb{X}_n \in \mathcal{R}_{2,\alpha/2}, \\ \text{we do not reject } H_0 \text{ otherwise.} \end{cases} \tag{4.5.1}$$

This procedure is well defined since the intersection of the critical region $\mathcal{R}_{1,\alpha}$ and $\mathcal{R}_{2,\alpha}$ is empty. This procedure is a three-decision test procedure and admits three kinds of errors, see Table 4.1.

The first kind of errors is to reject the null hypothesis H_0 while H_0 is actually true. The probability that this error occurs is the level of the test which is defined as,

$$\sup_{\sigma>0} \mathbb{E}_{1/2,0,\sigma} \left(\phi_{1,\alpha} + \phi_{2,\alpha}\right) = \alpha. \tag{4.5.2}$$

We only control the occurrence of this first kind of error. Then we draw attention that acceptance of H_0 "(X_t) is a free diffusion" does not necessarily demonstrate that H_0 is true. It only means that data do not show any evidence against the null hypothesis. At the end, we reject this assumption in direction to one of the alternatives at level $\alpha/2$. The second type of errors occurs when we do not reject the null hypothesis while one of the alternatives is true. The last type of errors is to reject the null hypothesis in favour to a wrong alternative. In the literature of three-decision test such an error is called a Type III error, see for example Rasch [2012] and references therein.

4.6 Choosing the Estimator of σ

Ideally, we would like to find an estimator of σ which is consistent according to the *large-sample scheme* under the hypotheses H_0, H_1 and H_2, and satisfies the assumption

that the distribution of $\hat{\sigma}_n/\sigma$ is free of σ under H_0. However, the *large-sample scheme* is not favourable to get an estimator with such properties. For instance, Florens-Zmirou [1989] shows that the naive maximum likelihood estimator for the drift parameter has an asymptotic bias of the order of lag time Δ. Then, the *high-frequency scheme* and the *rapidly increasing design* turns out to be more convenient to provide consistent estimators. In fact, in the limit, these schemes correspond to the situation in which we have a continuous observation of the process on the time interval of observation. Jiang and Knight [1997] propose non parametric estimators of both the drift and the diffusion coefficient. The consistency of these estimators is proven under the high-frequency scheme only. Therefore, in this section, we discuss about the estimation of the diffusion coefficient under the *large-sample asymptotic*.

The first proposition to estimate σ may be :

$$\hat{\sigma}^2_{1,n} = \frac{1}{dn\Delta} \sum_{j=1}^{n} \|X_{t_j} - X_{t_{j-1}}\|^2, \tag{4.6.1}$$

where d is the dimension of the process. Even if the estimator (4.6.1) is strongly consistent under the *high-frequency scheme* for every process (X_t) solution of (4.1) [Basawa and Prakasa Rao, 1980, Lemma 4.2, p 212], Proposition 1 tells us that it is not the case under the *large-sample scheme*.

Proposition 1.

- *Under* H_0, $\hat{\sigma}_{1,n}$ *is strongly consistent and the distribution of* $\hat{\sigma}_{1,n}/\sigma$ *is free of* σ.
- *If* (X_t) *is an Ornstein-Uhlenbeck process* (4.2.1), $\hat{\sigma}^2_{1,n}/\sigma^2$ *converges in probability to* $(1 - e^{-\lambda\Delta})/(\lambda\Delta)$.
- *If* (X_t) *is a Brownian motion with drift* (4.2.3), $\hat{\sigma}^2_{1,n}/\sigma^2$ *converges almost surely to* $\Delta\|v\|^2/(d\sigma^2) + 1$.
- *If* (X_t) *is a fractional Brownian motion* (4.2.2), $\hat{\sigma}^2_{1,n}/\sigma^2$ *converges almost surely to* $\Delta^{2\mathfrak{h}-1}$.

A proof of Proposition 1 is given in Appendix A.2. Proposition 1 states that $\hat{\sigma}_{1,n}$ is adequate to our procedure under the null hypothesis. However $\hat{\sigma}_{1,n}$ is asymptotically biased under some alternatives. Notice that if (X_t) is an Ornstein-Uhlenbeck process (4.2.1), then $\hat{\sigma}^2_{1,n}$ underestimates σ^2 in average since $(1 - e^{-x})/x < 1$ for $x > 0$. Then T_n might be overvalued with this estimator, increasing Type II or type III error rate in our procedure. If (X_t) is a Brownian motion with drift (4.2.3), $\hat{\sigma}^2_1$ overestimates σ^2 in average. Then T_n might be overvalued with this estimator, increasing Type II or type III error rate. Similarly, if (X_t) is a fractional Brownian motion (4.2.2), $\hat{\sigma}^2_1$ underestimates σ^2 if $\mathfrak{h} < 1/2$, and overestimates σ^2 if $\mathfrak{h} < 1/2$.

The second suggestion to estimate σ may be based on the second order differences rather than the first order differences,

$$\hat{\sigma}_{2,n}^2 = \frac{1}{2dn\Delta} \sum_{j=1}^{n-1} \|(X_{t_{j+1}} - X_{t_j}) - (X_{t_j} - X_{t_{j-1}})\|^2. \tag{4.6.2}$$

As $\hat{\sigma}_{1,n}^2$, $\hat{\sigma}_{2,n}^2$ fulfils the assumption of Lemma 4.3.1 under H_0. This estimator has the advantage of decreasing the bias under some alternatives. For instance it removes the bias in the case of the Brownian motion with drift.

4.7 Approximation of the Distribution of the Statistic under the Null Hypothesis and Asymptotic Behaviour of our Procedure

Theorem 4.7.1 gives the asymptotic behaviour of our procedure under the null hypothesis.

Theorem 4.7.1. *Let (X_t) be a Brownian Motion on $\mathbb{R}^d$. Let $\hat{\sigma}_n$ be a consistent estimator of the diffusion parameter σ of (X_t). The test statistic T_n converges in distribution to $S_0^d = \sup_{0\leq s\leq 1} \left\|W_s^d\right\|$ as $n \to \infty$. Here (W_t^d) is a standard d-dimensional Brownian motion that is the Brownian motion of variance $\mathbf{I}_d$ and initialization $W_0 = \mathbf{0}_d$.*

A proof of Proposition 4.7.1 is given in Appendix A.1. We emphasize the dimension d in Theorem 4.7.1 only for distinguishing the two limit distributions S_0^d, $d = 2, 3$. The limit distribution of the test statistic under H_0 admits an analytical form in both cases $d = 2$ and $d = 3$. The cumulative distribution of S_0^2 (2-dimensional case) is given by [see Borodin and Salminen, 1996, Formulae.1.1.4, p. 280]:

$$x \in (0, +\infty) \to \sum_{k=1}^{\infty} \frac{2e^{-j_{0,k}^2/(2x^2)}}{j_{0,k} J_1(j_{0,k})},$$

where $x \geq 0$, J_ν the Bessel function of order ν and $0 < j_{\nu,1} < j_{\nu,2} < \ldots$ the positive zeros of J_ν.

The cumulative distribution of S_0^3 (3-dimensional case) is given by [see Borodin and Salminen, 1996, Formulae.1.1.4, p. 317]:

$$x \in (0, +\infty) \to \frac{2x}{\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\left(-\frac{(2k+1)^2 x^2}{2}\right).$$

Replacing the quantiles $q_n(\alpha)$ by the quantiles of S_0^d in our test procedure provides us a test of asymptotic level α.

Furthermore, Proposition 2 gives the asymptotic behaviour of the test statistic under parametric alternatives when the estimator $\hat{\sigma}_{1,n}$ is considered (see Appendix A.3 for a proof). More generally, as long as the estimator $\hat{\sigma}_n$ of the diffusion coefficient is such that $\hat{\sigma}_n/\sigma$ converges in probability to a positive constant whatever the dynamic of (X_t), then Proposition 2 holds.

Proposition 2. *Assume that we consider the estimator* (4.6.1) *in our procedure* (4.3.1).

- *If* (X_t) *is an Ornstein-Uhlenbeck process* (4.2.1)*,* T_n *converges in probability to* 0.
- *If* (X_t) *is a fractional Brownian motion* (4.2.2) *with* $0 < \mathfrak{h} < 1/2$*,* T_n *converges in probability to* 0.
- *If* (X_t) *is a fractional Brownian motion* (4.2.2) *with* $1/2 < \mathfrak{h} < 1$*,* T_n *converges in probability to* $+\infty$.
- *If* (X_t) *is a Brownian motion with drift* (4.2.3)*,* T_n *converges in probability to* $+\infty$.

Note that Theorem 4.7.1 and Proposition 2 allow us to control the error rates of type II and type III under parametric alternatives: the associated error rates converges to 0 with n. However, as in practice n may be small, the asymptotic approximation of the quantiles of T_n may not be accurate. Then the level of the test is no longer α. Since we are able to draw a sample from the distribution of T_n under H_0 (see Algorithm 1), we propose a Monte Carlo estimate of the quantile $q_n(x)$, $0 < x < 1$. This estimate is defined as the $[xN]^{\text{th}}$ order statistic, $q_n^{(N)}(x)$, of the sample $(T_n^{(1)}, \ldots, T_n^{(N)})$. Table 4.2 shows that there is a significant difference between asymptotic and non asymptotic quantiles. As expected, as $n \to \infty$, $q_n(\alpha)$ converges to $q(\alpha)$.

In dealing with a test, we can also be interested in computing the p-value. The p-value of the test H_0 *vs* H_1 (subdiffusion as the alternative) is defined as:

$$p_{1,n} = F_n(T_n), \tag{4.7.1}$$

where F_n denotes the cumulative distribution function (cdf) of T_n under H_0. The p-value of the test H_0 *vs* H_2 (superdiffusion as the alternative) is defined as:

$$p_{2,n} = 1 - F_n(T_n). \tag{4.7.2}$$

Testing the hypothesis H_0 *vs* the hypotheses H_1 or H_2 is more tricky as we use a two-sided test with a non-symmetric distribution. In this case we can define the p-value as :

$$p_n = 2\min\left\{p_{1,n}, p_{2,n}\right\}. \tag{4.7.3}$$

Doubling the lowest one-tailed p-value can be seen as a correction for carrying out two one-tailed tests.

Algorithm 1: Simulation of a N-sample $(T_n^{(1)}, \dots, T_n^{(N)})$ of the distribution of the statistic T_n under H_0.

Input: n, N
// the length n of the trajectory
// the number N of Monte Carlo experiments
Result: a N-sample $(T_n^{(1)}, \dots, T_n^{(N)})$
for *i=1* **to** N **do**
  // Simulation of a Brownian trajectory of size n, of variance $\sigma = 1$ and with resolution time $\Delta = 1$.
  initialization $Y_0^{(i)} = \mathbf{0}_d$;
  for *j=1* **to** n **do**
    Draw $\epsilon \sim \mathcal{N}(\mathbf{0}_d\,, \mathbf{I}_d)$;
    $Y_j^{(i)} = Y_{j-1}^{(i)} + \epsilon$;
  end
  // Computation of the test statistic
  Compute the ratio $T_n^{(i)} = D_n^{(i)}/\hat{\sigma}_n^{(i)}$ from $(Y_0^{(i)}, \dots, Y_n^{(i)})$;
end

We estimate F_n with the standard empirical distribution function estimated by Monte Carlo simulations using Algorithm 1.

$$\hat{F}_n(x) = N^{-1} \sum_{i=1}^{N} \mathbf{1}(T_n^{(i)} \leq x). \tag{4.7.4}$$

Then, we estimate the p-value (4.7.3) substituting $\hat{F}_n$ to F_n.

4.8 Multiple Test Procedure for a Collection of Trajectories

Trackers compute a collection of particle trajectories from a sequence of images. Then, it is desirable to decide the modes of mobility for a collection of particle trajectories. From now, we consider a collection $\mathcal{X}_m$ of m trajectories which are simultaneously observed. We denote by $\mathbb{X}_{n_k}^{(k)}$ the observations associated to the k^{th} particle:

$$\begin{aligned} \mathbb{X}_{n_k}^{(k)} &= \left(X_{t_0}^{(k)}, \dots, X_{t_{n_k}}^{(k)}\right), \qquad k = 1, \dots, m, \\ \mathcal{X}_m &= \left\{\mathbb{X}_{n_k}^{(k)},\ k = 1, \dots, m\right\}. \end{aligned} \tag{4.8.1}$$

In this section, we denote by $\mathbb{P}$ the probability distribution of the m-uplet stochastic processes $((X_t^{(k)}),\ k = 1 \dots m)$ and by $\mathbb{E}$ its associated expectation. We assume that

Table 4.2: Estimation of the quantiles of order $\alpha/2$ and $1-\alpha/2$ ($\alpha = 5\%$) for different trajectory lengths n in the two and three dimensions case. We use Algorithm 1 with $N = 1\,000\,001$ Monte-Carlo replications to estimate the quantiles.

		Trajectory size			
Dimension	Quantile order	10	30	100	asymp
2	2.5%	0.725	0.754	0.785	0.834
2	97.5%	2.626	2.794	2.873	2.940
3	2.5%	0.950	0.981	1.011	1.061
3	97.5%	2.969	3.127	3.197	3.268

the observed trajectories are independent, that means $\mathbb{P}$ belongs to the tensorial product of probabilities $\mathcal{P}$, (defined in Section 4.1) $\mathbb{P} \in \mathcal{P}^{\otimes m}$. For all trajectories $k = 1 \dots m$, we derive our trichotomy hypothesis test procedure: $H_0^{(k)}$ "$(X_t^{(k)})$ is a free diffusion" versus $H_1^{(k)}$ "$(X_t^{(k)})$ is a subdiffusion" or $H_2^{(k)}$ "$(X_t^{(k)})$ is a superdiffusion". We are faced with the problem of simultaneous tests when the rejections of null hypotheses $H_0^{(k)}$ are accompanied by claims of the direction of the alternative ($H_1^{(k)}$ or $H_2^{(k)}$). In this setup, multiple test procedures are preferable than single test procedures. Indeed, applying the procedure at level α for each trajectory produces in average a number of $m\alpha$ type I errors. A multiple testing procedure aims to control the number of false discoveries. We refer the reader to Shaffer [1995], Roquain [2011], Grandhi [2015] for a review.

A multiple testing procedure of m null hypotheses against two alternative hypotheses is a rule $\mathcal{R}_1(\mathcal{X}_m) \times \mathcal{R}_2(\mathcal{X}_m)$, where $\mathcal{R}_1(\mathcal{X}_m)$ and $\mathcal{R}_2(\mathcal{X}_m)$ are disjoint subsets of $\{H_0^{(1)}, \dots H_0^{(m)}\}$. For $i = 1, 2$, $\mathcal{R}_i(\mathcal{X}_m)$ is the set of the rejected hypotheses $H_0^{(k)}$ to the benefit of the alternative $H_i^{(k)}$. We may commit three kinds of errors in such a multiple testing procedure. Let us introduce the following notations before listing these errors. For a given $\mathbb{P} \in \mathcal{P}^{\otimes m}$, we denote by $\mathcal{I}(\mathbb{P})$ the subset of indexes $\{1, \dots m\}$ for which the hypothesis $(H_0^{(k)})$ is actually true and by $m_0(\mathbb{P})$ the unknown cardinal of the set $\mathcal{I}(\mathbb{P})$. We denote by $R = R_1 + R_2$ the observed number of null hypotheses which are rejected by the multiple testing procedure. Table 4.3 summaries the number of errors which may occur following a multiple testing procedure.

- We make a type I error on $H_0^{(k)}$ when we reject $H_0^{(k)}$ while it is a true null hypothesis. In this case, k belongs to the set $\mathcal{I}(\mathbb{P}) \cap (\mathcal{R}_1(\mathcal{X}_m) \cup \mathcal{R}_2(\mathcal{X}_m))$. The number of errors of first kind is $V = V_1 + V_2$.
- Type II error occurs when we do not reject a null hypothesis $H_{0,k}$ while $H_{0,k}$ is false ($k \notin \mathcal{I}(\mathbb{P})$). The number of errors of second kind is $T = T_1 + T_2$.
- The type III errors are directional errors: the index $k \notin \mathcal{I}(\mathbb{P})$ is correctly rejected

Table 4.3: Outcomes in testing m null hypotheses against two-alternatives. For $i = 1, 2$, R_i is the cardinal of $\mathcal{R}_i(\mathcal{X}_m)$. The variables $(S_i)_{i=1\ldots4}, (T_i)_{i=1,2}, U, (V_i)_{i=1,2}$ are not observed and depend on $\mathcal{X}_m$ and P.

Truth	**Decision**			
	Accept H_0	Accept H_1	Accept H_2	**Total**
H_0	U	V_1	V_2	$m_0(\mathbb{P})$
H_1	T_1	S_1	S_3	$m_1(\mathbb{P})$
H_2	T_2	S_4	S_2	$m_2(\mathbb{P})$
Total	$m - R_1 - R_2$	R_1	R_2	m

$(k \in \mathcal{R}_1(\mathcal{X}_m) \cup \mathcal{R}_2(\mathcal{X}_m))$, but for the wrong alternative. We mix up the alternatives deciding one while it is the other. The number of errors of third kind is $S = S_3 + S_4$.

To measure the type I error rate, it is common to consider the k-family-wise error rate (k-FWER) or the false discovery rate (FDR), see [Roquain, 2011] and references therein. In our settings, controlling the type I error rate is a first step, but it would be necessary to control type III errors as well. In the literature, the sum of the number of errors of first and third kind is controlled using the mixed-directional-family-wise error rate (mdFWER) or the mixed-directional-false discovery rate (mdFDR), see [Grandhi, 2015]. To our knowledge, the mdFWER and mdFDR are only controlled for the problem of testing null hypotheses against two-sided alternatives for finite-dimensional parameters, see for example [Guo and Romano, 2015] and references therein.

Biologists are interested in the proportions of each dynamic (subdiffusion, superdiffusion and Brownian motion) and their geographic location in the cell. In this context, controlling the FWER, that is the probability to make a single false discovery, is not relevant. That is why we focus on a procedure which enables to control the FDR. Guo and Romano [2015] also present several multiple test procedures associated to three-decision problems which aim to control the FDR. Their approach is different since the problem is rewritten as a problem which carries out $3m$ null hypotheses. Their proposed procedures control strongly the FDR only on $2m$ null hypotheses among the $3m$ under the dependence or independence of the test statistics. In this section, we propose to adapt the multiple testing procedures of Benjamini and Hochberg [1995] and Benjamini and Hochberg [2000] controlling the FDR that is the average proportion of false discoveries among the discoveries. We stress that our model is non-parametric. Then we will consider the control of the mdFDR or mdFWER for a next issue.

Let $p^{(k)}, p_1^{(k)}$, and $p_2^{(k)}$ be respectively the p-value (4.7.3), (4.7.1) and (4.7.2) associated to the k^{th} trajectory, $k = 1 \ldots m$. Let $p^{(1:m)} \leq p^{(2:m)} \leq \ldots \leq p^{(m:m)}$ be the ordered p-values, and $H_0^{(1:m)}, \ldots H_0^{(m:m)}$ the associated null hypotheses. The adaptation of the

Benjamini-Hochberg (BH) procedure is described in Procedure 1.

Procedure 1 (Adaptation of the Benjamini-Hochberg (BH) procedure)**.**

1. *Use the Benjamini-Hochberg procedure on the p-values* $(p^{(k)})_{k=1\dots m}$:
   Let $k^\star$ *be the largest* k *for which* $p^{(k:m)} \leq \frac{k}{m}\alpha$.
   $\mathcal{R}_\alpha(\mathcal{X}_m)$ *is the set of all hypotheses* $H^{(k:m)}$ *for* $k = 1, \dots, k^\star$.
2. *Let* $\mathcal{R}_{1,\alpha}(\mathcal{X}_m)$ *be the subset* $\mathcal{R}_\alpha(\mathcal{X}_m)$ *such that* $p_1^{(k)} < p_2^{(k)}$.
3. *Let* $\mathcal{R}_{2,\alpha}(\mathcal{X}_m)$ *be the subset* $\mathcal{R}_\alpha(\mathcal{X}_m)$ *such that* $p_1^{(k)} > p_2^{(k)}$.

The set $\mathcal{R}_\alpha(\mathcal{X}_m)$ is the set of all rejected null hypotheses for our trichotomy test. According to Finner and Roters [2001], we have,

$$\begin{aligned}\forall \mathbb{P} \in \mathcal{P}^{\otimes m}, \quad \mathrm{FDR}(\mathcal{R}_\alpha(\mathcal{X}_m), \mathbb{P}) &= \mathbb{E}\left(\frac{V}{\max(R,1)}\right) \\ &= \frac{m_0(\mathbb{P})}{m}\alpha.\end{aligned}$$

Then the FDR of Procedure 1 is controlled by α. Moreover the p-values $p_1^{(k)}$ and $p_2^{(k)}$ give the information to which side of the distribution F_{n_k} the associated test statistic $T_{n_k}^{(k)}$ is. The case of equality ($p_1^{(k)} = p_2^{(k)} = 1/2$) never occurs since such null hypothesis will not be rejected at the step 1 of the Procedure 1.

Actually, we may also use the adaptive BH procedure of Benjamini and Hochberg [2000] as the first step of Procedure 1. Then the Procedure 1 will be referred to as the adaptive (respectively standard) Procedure 1 when we use the adaptive (respectively standard) BH procedure as the first step. The adaptive BH procedure is more powerful than the standard BH procedure. It uses an estimation of the number of true null hypotheses $m_0(\mathbb{P})$ to increase the power of the BH procedure. Benjamini and Hochberg [2000] simply define the adaptive BH procedure by replacing m by an estimator $\hat{m}_0$ of m_0 in the BH procedure. The associated FDR is $(m_0/\hat{m}_0)\alpha$ and is less than α if $\hat{m}_0 \leq m_0$ almost surely. The procedure to estimate m_0 presented in [Benjamini and Hochberg, 2000] is made for $\hat{m}_0$ to be upward biased. This bias favours the control of the FDR at level α. Due to the fact that $\hat{m}_0$ does not fulfil the condition $\hat{m}_0 \leq m_0$ almost surely, we can not say that the adaptive BH procedure controls the FDR at level α theoretically. However simulations from Benjamini and Hochberg [2000] suggest that the adaptive BH procedure controls the FDR at level α.

4.9 Summary

In this chapter, we modelled the trajectories with diffusion processes. The two and three-dimensional cases were considered. We proposed a three-decision test to classify a single

trajectory into the three groups of diffusion, namely Brownian motion, subdiffusion and superdiffusion. The null hypothesis of the test supposes that the trajectory is Brownian, the alternative hypothesises corresponding to either superdiffusion or subdiffusion. We also provided a multiple test procedure to classify a collection of independent trajectories, controlling the false discovery rate [Benjamini and Hochberg, 1995] at level α.

In the next chapter, we evaluate our two test procedures on simulations, in the two-dimensional case. We consider parametric diffusion processes used in biophysics for modelling subdiffusion and superdiffusion. We also analyse real data depicting the exocytosis process in both the two and three-dimensional cases.

5 Simulation Study and Real Data Applications

We assess the power of our single test procedure (on a single trajectory) and our multiple test procedure (on a collection of trajectories) by Monte Carlo simulations. We consider parametric alternatives: the Ornstein-Uhlenbeck (4.2.1) and the fractional Brownian motion with Hurst index $0 < \mathfrak{h} < 1/2$ for subdiffusion processes (H_1); the Brownian motion with drift (4.2.3) and the fractional Brownian motion with Hurst index $1/2 < \mathfrak{h} < 1$ for superdiffusion processes (H_2). We restrict the simulations study to the two-dimensional case ($d = 2$). Then, we apply our procedure on two and three-dimensional real data. We compare our results with those obtained thanks to a method based on the mean square displacement (for the two-dimensional case only).

5.1 Power of the Test Procedure for a Single Trajectory

In Section 4, we studied the asymptotic distribution of the test statistic under the null hypothesis and parametric alternative hypotheses. More precisely, Proposition 2 states that the power of the test under parametric alternatives converges to 1 with n. Figure 5.1 shows the Monte Carlo estimates of the power under the parametric alternatives aforementioned in Proposition 2 in the two-dimensional case. For a fixed step of time Δ and a fixed diffusion coefficient σ, we vary the values of the other parameters and the length n of the trajectories. For each parametric alternatives of Proposition 2, we can use exact simulation schemes.

If (X_t) is an Ornstein-Uhlenbeck process (4.2.1) which is entered in its stationary regime, then the distribution of the test statistic does not depend on θ (see Appendix A.4). Figure 5.1(b) shows the plot of the power regarding the values of λ which models the strength of the restoring force toward the equilibrium position θ. Stronger is the force, more powerful is the test.

Furthermore if (X_t) is a Brownian motion with drift with parameters (v, σ) such that $\|v\| \sqrt{\Delta} > \sigma$, then the particle goes toward the direction of v while the Brownian random part of the SDE (4.2.3) does not affect much its trajectory (see Appendix A.4). The bigger is the norm of the drift parameter v, more powerful is the test, see Figure 5.1(a).

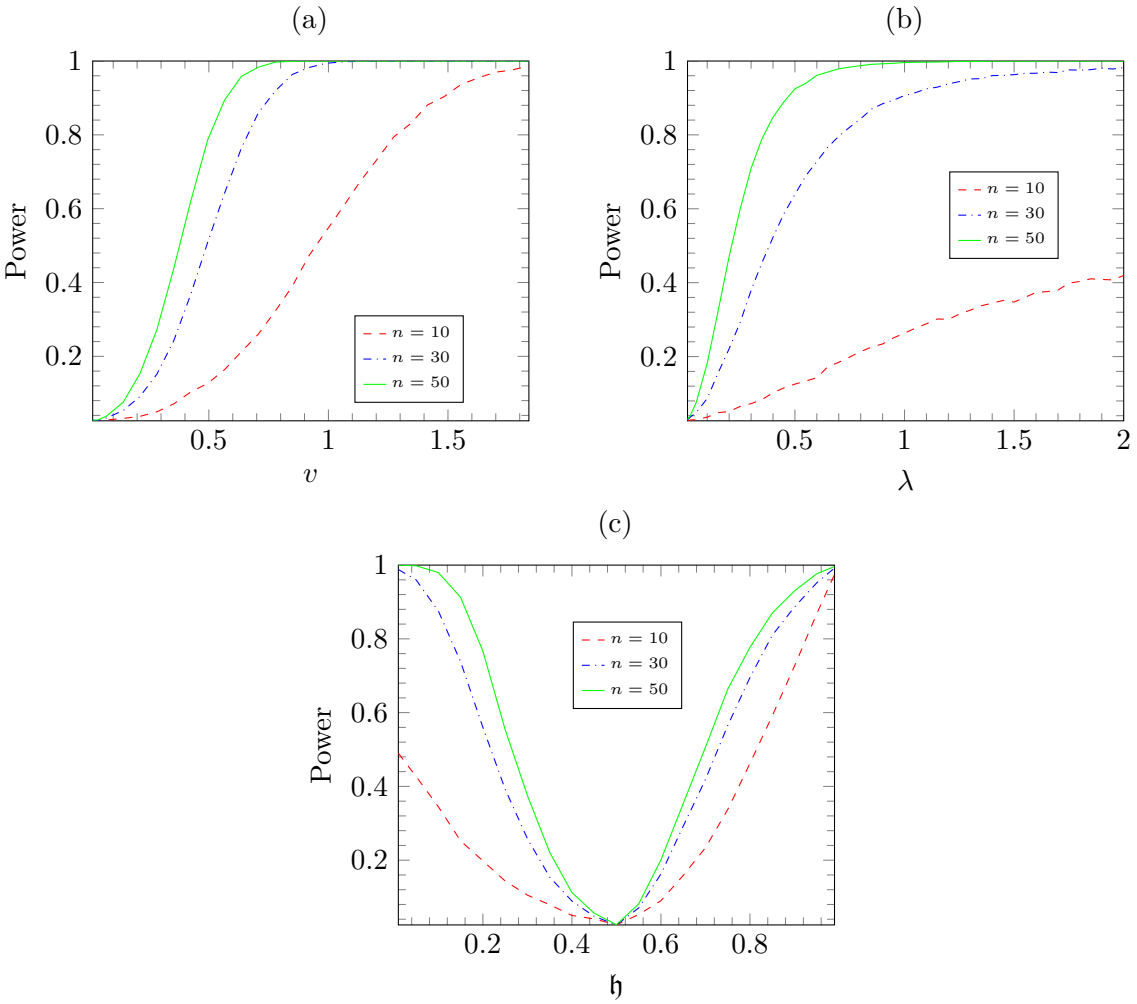


Figure 5.1: Monte Carlo estimate of the power of the test at level $\alpha = 0.05$ according to the trajectory length n and the parameter associated to the following two-dimensional parametric alternatives: (a) Brownian motion with drift (parameter $v = (v_1, v_2)$ such that $v_1 = v_2$); (b) the Ornstein-Uhlenbeck process (parameter λ) and (c) fractional Brownian motion (parameter $\mathfrak{h}$). We use 10 001 Monte Carlo replications for computing each point of the power curves.

Finally if (X_t) is a fractional Brownian motion, then the distribution of T_n depends only on the Hurst index $\mathfrak{h}$ (see Appendix A.4). Then the test procedure is equivalent to test the null hypothesis "$\mathfrak{h} = 1/2$" versus "$\mathfrak{h} \neq 1/2$", see Figure 5.1(c).

Table 5.1: Parameters used for simulating the alternative hypotheses. For simplicity we took $\sigma = 1$ for all processes (including Brownian motion). We choose $\Delta = 1$.

Hypothesis	Process	Parameter	Value
H_1	Ornstein-Uhlenbeck	λ	0.53
H_1	Fractional Brownian	$\mathfrak{h}$	0.13
H_2	Brownian motion with drift	$\|v\|$	0.66
H_2	Fractional Brownian	$\mathfrak{h}$	0.85

5.2 The Average Power and the mdFDR of the Multiple Test Procedure for a Collection of Trajectories

We recall that we restrict the simulation study to the two-dimensional case. The simulation settings are described as follows. According to experience, we choose the number of trajectories to be $m = 100$ or $m = 200$. All trajectories are assumed to have the same size $n = 30$, since this size is reasonable regarding real data. The diffusion coefficient σ and the lag-time Δ are set to 1. The collection of two-dimensional trajectories $\mathcal{X}_m$ is composed of :

- $m_0 < m$ Brownian trajectories (H_0),
- $(m-m_0)/2$ subdiffusive trajectories (H_1), half from an Ornstein-Uhlenbeck process with parameter $\lambda > 0$, half from a fractional Brownian motion with Hurst index $0 < \mathfrak{h} < 1/2$,
- $(m-m_0)/2$ superdiffusive trajectories (H_2), half from a Brownian motion with drift $v \in \mathbb{R}^2$, half from a fractional Brownian motion with Hurst index $1/2 < \mathfrak{h} < 1$.

The parameters to simulate these trajectories are given in Table 5.1. We take the parameters corresponding to a power of the single test procedure of 80%. Such parameters are used to produce Figure 1.1. This choice seems coherent in regards to trajectories from real data, see Figure 1.2. For a given m, the proportion of true null hypotheses H_0 varies: $m_0/m \in \{0, 0.2, 0.4, 0.6, 0.8\}$.

The mdFDR is a rate which controls the error of type I and type III. It is defined as $\mathbb{E}((V+S)/\max(R,1))$ (see Table 4.3). Table 5.2 shows that the Procedure 1 also controls the mdFDR. The mdFDR and FDR appear to be very close meaning that the number of type III errors is extremely low. Furthermore, the adaptive Procedure 1 (where m_0 is estimated) is less conservative than the standard Procedure 1. As expected, the FDR and mdFDR increase as the proportion of true null hypotheses increases.

Table 5.2: Monte Carlo estimate of the FDR and mdFDR for both standard and adaptive Procedure 1 at level $\alpha = 0.05$ in the two-dimensional case. The number of replications is 10 001. The error rate estimations are expressed in percentages.

		Standard		**Adaptive**	
m	m_0/m	FDR	mdFDR	FDR	mdFDR
100	0	0	0	0	0.2
	0.2	1	1	3.7	3.7
	0.4	2.1	2.1	4.2	4.2
	0.6	3.2	3.2	4.7	4.7
	0.8	4.1	4.1	4.8	4.8
200	0	0	0	0	0.4
	0.2	1	1	3.4	3.4
	0.4	2.1	2.1	4	4
	0.6	3.2	3.2	4.6	4.6
	0.8	4	4	4.7	4.7

To assess the performance of our multiple test procedure, we use the average power [Grandhi, 2015] :

$$\mathbb{E}\left(\frac{S_i}{m_i}\right), \quad i = 1, 2 \tag{5.2.1}$$

where m_i is the number of true alternatives H_i and S_i $(i = 1, 2)$ is defined in Table 4.3. In our simulation scheme, we set $m_i = (m - m_0)/2$. The average power is the expected proportion of hypotheses accepted as H_i among all true alternatives H_i. Average powers of the different simulations corresponding to different values of m_0/m and m are shown on Figure 5.3.

First, we can see that the powers of H_1 and H_2 are not very sensitive to the number of hypotheses m for both the standard Procedure 1 and the adaptive Procedure 1. Secondly, the adaptive Procedure 1 is more powerful than the standard Procedure 1 (red and blue dashed lines respectively above red and blue solid lines in Figure 5.3). The benefit of the adaptive Procedure 1 over the standard Procedure 1 decreases as the proportion of true null hypotheses m_0/m increases (solid and dashed line of same color getting closer as m_0/m increases in Figure 5.3). This is due to the fact that, as m_0/m tends to 1, m_0 and then $\hat{m}_0$ tend to m. As a result, the adaptive and standard Procedure 1 become similar.

Remark 5.2.1. *We observe that, given a certain procedure (standard or adaptive Procedure 1), the average power of H_1 is lower than the average power of H_2, see Figure 5.3. It is not due to the choice of parameters as both alternatives H_1 and H_2 are simulated to*

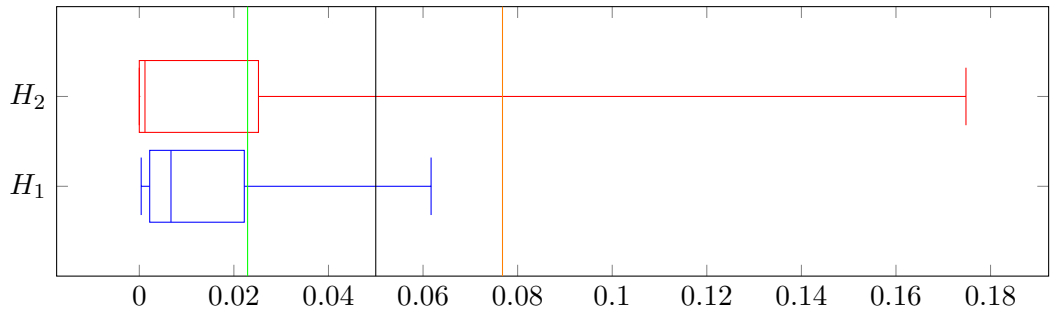


Figure 5.2: Boxplots of the p-value p_{30} (Equation (4.7.3)) under H_1 and H_2. We simulate a set of two-dimensional trajectories $\mathcal{X}_m$ with $m = 100$ and $m_0 = 20$ according to the simulation scheme described in Section 5. We plot the boxplot of the p-values $p_{30}^{(i:m)}$ corresponding to each true alternative hypothesises H_1 and H_2. The green (respectively orange) line is the threshold $h = p^{(k^*)}$ obtained by the first step of Procedure 1 (respectively Procedure 1) .The null hypothesis is rejected if the p-value is lower than h. The black line is the level $\alpha = 5\%$.

share the same power (80%) with the single test procedure. Actually, it comes from the fact that the p-values under H_2 are stochastically smaller than the p-values under H_1 (see Figure 5.2). Then, the true superdiffusive trajectories are more easily detected as non Brownian in the first step of the (adaptive) Procedure 1 than the true subdiffusive trajectories. We note that, if we use other parametric models for subdiffusion (H_1) and superdiffusion (H_2), we can have the opposite situation.

Finally, we compare the adaptive Procedure 1 to the MSD classification of Feder et al. [1996], based on a fit of the MSD curve to $t \to t^\beta$. We assess the two methods on a single collection of two-dimensional trajectories $\mathcal{X}_m$ with $m = 200$ and $m_0/m = 0.4$, composed of a mixture of Brownian motion, subdiffusion and superdiffusion as described at the beginning of this section. We get the confusion matrices Table 5.4 and 5.3 for respectively the adaptive Procedure 1 and the MSD method. The MSD method mixes up the Brownian trajectories with both subdiffusion and superdiffusion (see line 1 of Table 5.3). Another big issue is that 40% of the particles undergoing subdiffusion are considered as immobile by the MSD method. On the other hand, the adaptive procedure 1 detects well subdiffusion and superdiffusion in the setting of this simulation (line 2 and 3 of Table 5.4). More importantly, it controls the number of false discoveries through the FDR (line 1 of Table 5.4).

5.3 Real Data: the Rab11a Protein Sequence

Fluorescence imaging and microscopy has a prominent role in life science and medical research. It consists of detecting specific cellular and intracellular objects of interest at the diffraction limit (200 nm). These objects are first tagged with genetically engi-

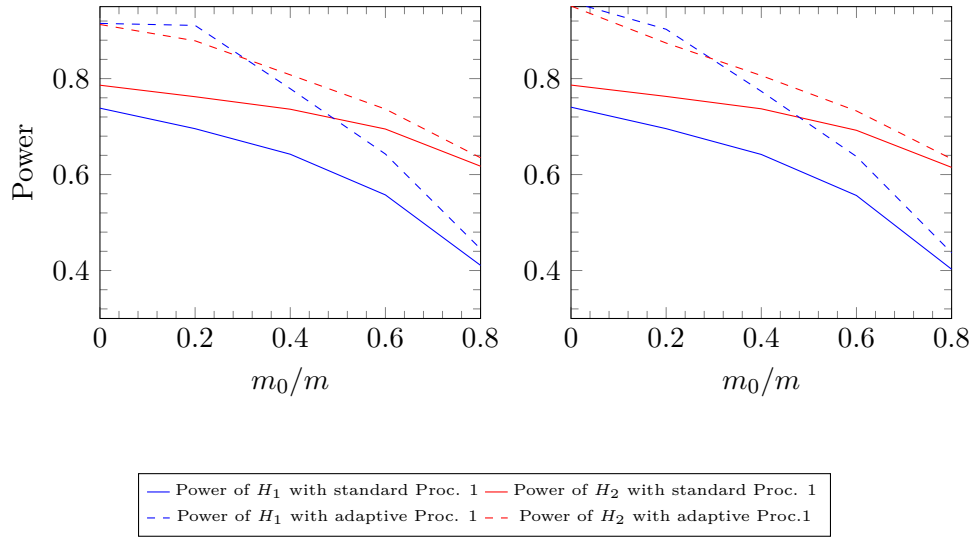


Figure 5.3: Monte Carlo estimate of the average power against the proportion of true null hypothesis m_0/m in the collection of hypotheses. On the left we test $m = 100$ hypotheses, on the right $m = 200$.

Table 5.3: Confusion matrix for the MSD method in the two-dimensional case

Ground truth/Test label	Brownian	Subdiffusion	Superdiffusion	Not moving
Brownian	19	45	36	0
Subdiffusion	0	60	0	40
Superdiffusion	3	0	97	0
Not moving	0	0	0	0

Table 5.4: Confusion matrix for the adaptive Proc.1 in the two-dimensional case

Ground truth/Test label	Brownian	Subdiffusion	Superdiffusion
Brownian	96	0	4
Subdiffusion	23	77	0
Superdiffusion	10	0	90

neered proteins that emit fluorescence. Then, they can be observed using wide field or confocal microscopy. Several image analysis methods have been developed to quantify intracellular trafficking, including object detection and tracking of fluorescent tags in cells [Chenouard et al., 2014, Kervrann et al., 2016]. In this section, we present the

biological process of interest, namely the exocytosis. Then, we apply our test procedure on both two and three dimensional real data.

Exocytosis Process

The exocytosis process is the mechanism of active transport of proteins out of the cell. Small structures, called the vesicles, travel from organelles to the cell membrane, propelled by motor activity. The vesicle fuses with the plasma membrane and delivers the transported protein in the extra-cellular medium, see Figure 5.4. Given computed trajectories, we investigate here the quantification of vesicles dynamics and trafficking. As explained earlier in the paper, the trajectories can be generally classified into three categories: Brownian motion, subdiffusion and superdiffusion.

As a model of exocytosis/recycling, we focus on the Rab11a protein. This protein is a member of the dynamic architecture of the complex molecular assembly which regulates recycling organelles trafficking. It plays an essential role in the regulation of late steps of vesicle recycling to the plasma membrane, namely the tethering-docking process [Schafer et al., 2014]. During exocytosis, Rab11a is attached to the vesicle membrane. Then, tracking Rab11a amounts to tracking the vesicle during the exocytosis phase. After the fusion of the vesicle to the cell membrane, Rab11a is recycled in the cytosol. During the recycling step, the tracking of Rab11a is not accurate as the proteins are detached from the vesicle and scatter around the cytosol, see Figure 5.4. It is currently under investigation. For that reason, we focus on the exocytosis process until the fusion time with the cell membrane.

The Two-Dimensional Rab11a Sequences

An illustration of a two-dimensional Rab11a sequence is shown in Figure 5.5 where the dark spots correspond to Rab11a vesicles in a "crossbow" micro-patterned shape cell. A typical image extracted from an image sequence is shown Figure 5.5. The image sequence is composed of 600 images of size 256×240 (1 pixel=160nm) acquired at 10 frames/s ($\Delta = 0.1s$). We tracked 1 561 trajectories with the multiple hypothesis tracking method with default parameters [Chenouard et al., 2013], available on the Icy software (`http:www.icy.org`). Now we explain how we select the trajectories of interest.

First, we discarded too small and too long trajectories corresponding to tracking errors in most cases.

Secondly, as we have just said in the previous subsection, we want to study Rab11a trajectories before the fusion of the vesicle to the membrane. We used a second molecular marker (Transferin Receptor (TfR)) to select trajectories related to the transport of vesicles until the fusion time. The transmembrane TfR protein is fluorescently labeled with a pH-sensitive probe, the pHluorin. Before the fusion time, the pH inside the vesicle is acidic, leading to a very low pHluorin photon emission. When the vesicle fuses to the

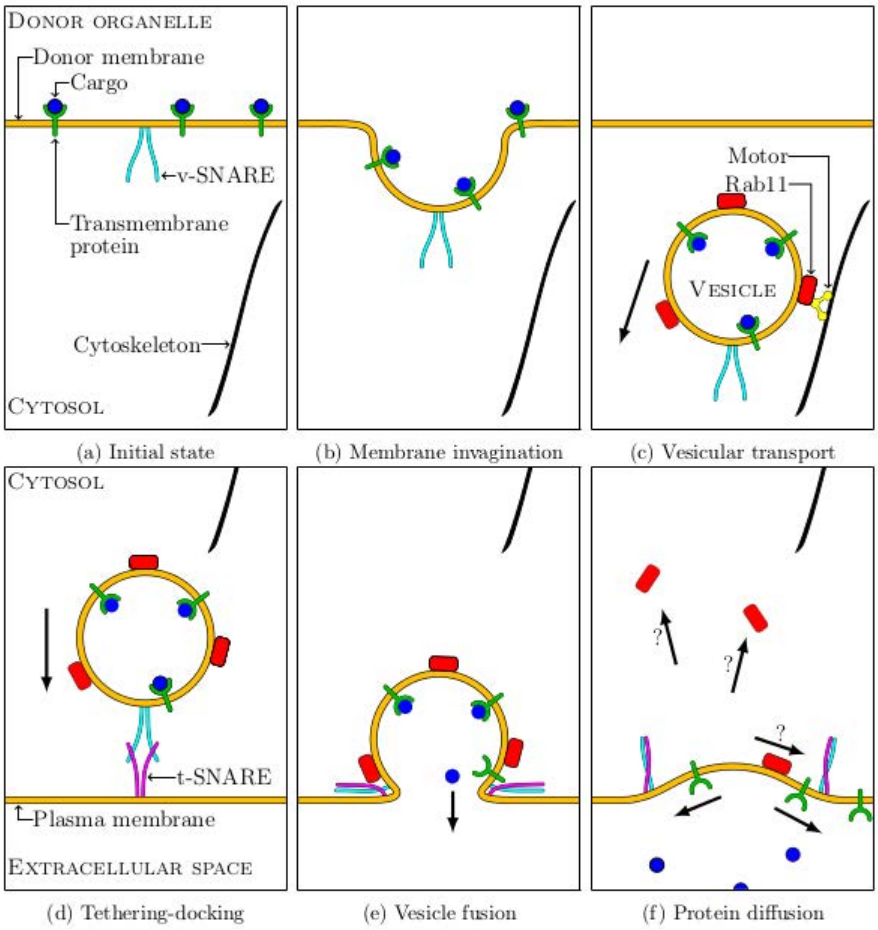


Figure 5.4: Main steps of the exocytosis process. Figure adapted from an original figure from Basset et al. [2015].

plasma membrane, the pHluorin gets exposed to the neutral extracellular medium and the fluorescence suddenly increases in the TIRF image plane. This feature allows us to detect the fusion time and the end of the exocytosis process [Basset et al., 2015]. Now the steps to select the trajectories of Rab11a undergoing exocytosis before the fusion time are described as follows.

1. We simultaneously observe two sets of trajectories: TfR and Rab11a trajectories.
2. We match each trajectory of Rab11a with the corresponding trajectory of TfR.
3. We cut the trajectory of Rab11a at the time when the matched trajectory of TfR starts (fusion time).

Table 5.5: Percentages of Brownian, superdiffusive and subdiffusive trajectories in the two-dimensional Rab11a sequence according to the different methods of classification.

Method	Brownian	Subdiffusion	Superdiffusion
Standard Proc. 1	80	16	4
adaptive Proc. 1	73	23	4
Single test	66	28	6
MSD	16	63	21

Finally, there is an additional step of selection of trajectories based on mathematical considerations. As we model particles motions with the diffusion processes 4.1, the particles are expected to move over time. Then, we have to get rid of the particles that do not move enough and consequently, can not be modelled by diffusion processes. In practice, we analyse only the trajectories with at least 20 distinct positions and the vesicles that stop at the same position less than $K = \lfloor n/10 \rfloor$ times (with n the length of the trajectory). In the case of the aforementioned image sequence, we end up with 166 trajectories whose median length is $n = 83$, once we went through the different steps of the selection process. The histogram of the trajectory sizes is given in Figure (a) 5.6. We also present the histogram of the test statistic T_n in Figure (b) 5.6.

In Figure 5.5, our results show that the four procedures – adaptive Procedure 1, standard Procedure 1, single test and MSD method – do not produce similar classification results visually. From the simulations, we found that the MSD method tends to wrongly over-detect subdiffusion and superdiffusion (see Tables 5.3 and 5.4). This is probably true also in the case of real Rab11a sequence. In Table 5.5, we give the proportion of each type of diffusion for the different methods aforementioned. The adaptive Procedure 1 tends to decrease the number of Brownian trajectories compared to the standard Procedure 1. It is not surprising as the adaptive Procedure 1 is defined to be more powerful than the standard Procedure 1: it rejects more easily the null hypothesis. This gain in power benefits to the alternative H_1 (subdiffusion). In fact we detect 23% of subdiffusion for the adaptive Procedure 1 against 16% for the standard Procedure 1 while both detect 4% of superdiffusion (see Table 5.5). The single test procedure detects even less Brownian motion but we know that it can not control the FDR. In Figure 5.5, the subdiffusion trajectories labelled with the test approach are more located in the center of the cell in a region corresponding to the Endosomal Recycling Compartment which is known to organize Rab11a carrier vesicles [Schafer et al., 2014]. It is also true for the subdiffusion trajectories labelled with the MSD analysis but we have just said that there is probably an over-detection of the subdiffusion with this method. We note that we carry the classification of trajectories with our different test procedures and the MSD method on multiple sequences of Rab11a protein, see Figure 5.7.

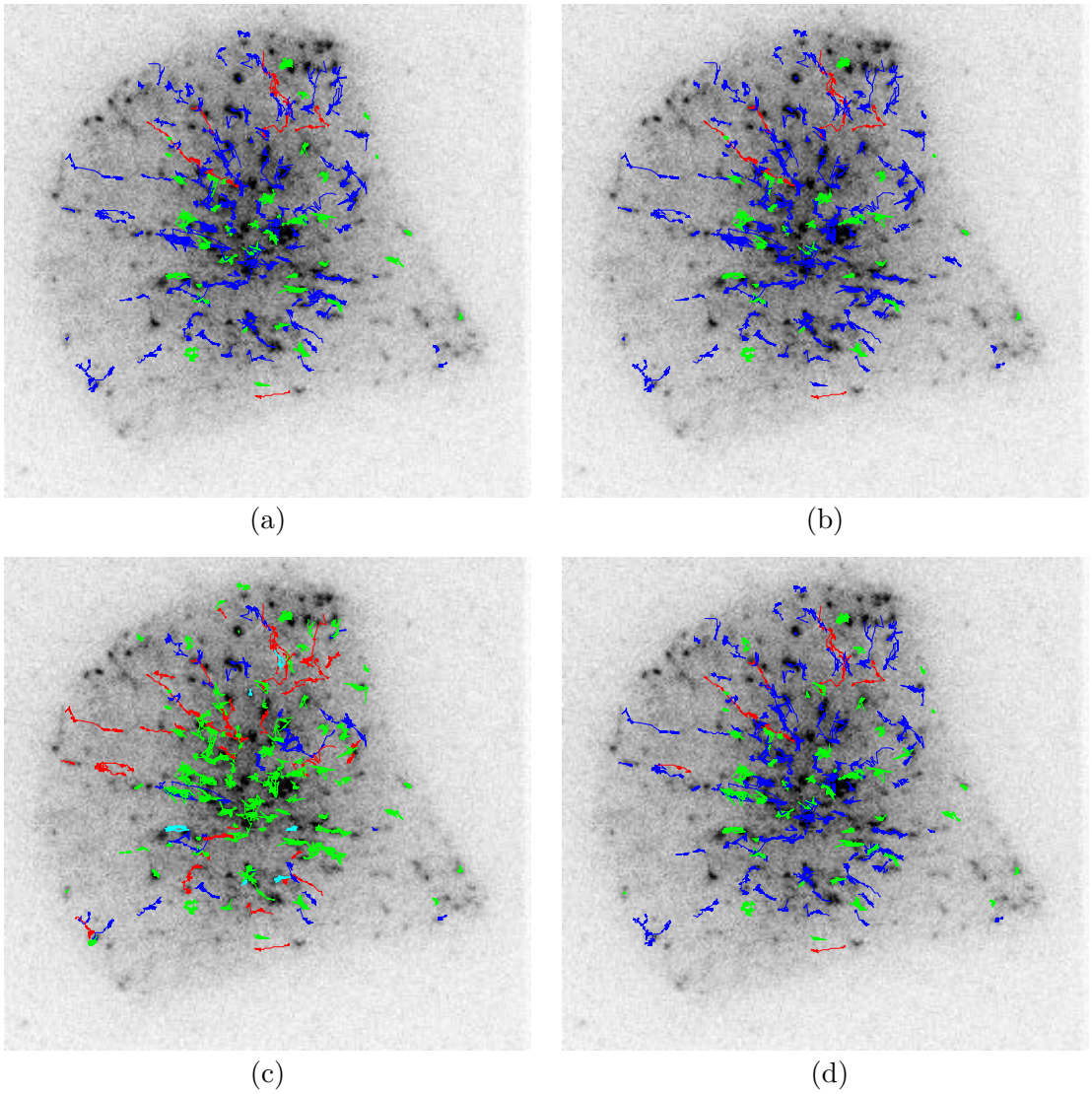


Figure 5.5: Map of the classification of the trajectories of the two-dimensional Rab11a sequence with (a) standard multiple test procedure 1, (b) its adaptative version, (c) MSD, (d) single test procedure. The colour code is: blue for Brownian motion, red for superdiffusion and green for subdiffusion, cyan for immobile particule (for the MSD method only).

The Three-Dimensional Langerin Sequence

We study the three-dimensional trajectories of vesicles containing the Langerin protein. Langerin is a C type lectin receptor almost exclusively expressed in Langerhans cells of the epidermis and is constitutively endocytosed and recycled [Gidon et al., 2012]. Gidon

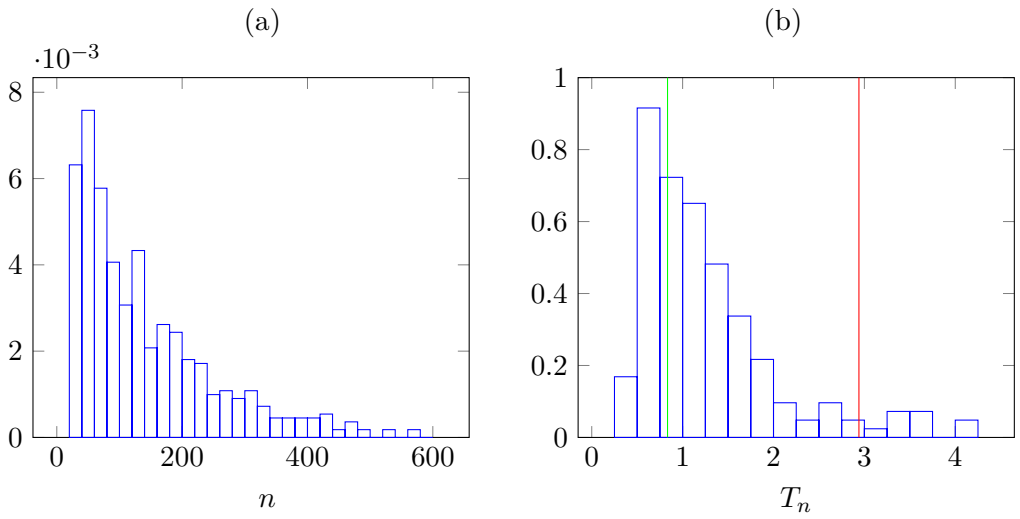


Figure 5.6: Histograms of the trajectory sizes n (a) and of the test statistics T_n (b) of the Rab11a 2D sequence. In Figure (b), the green (respectively red) vertical line represents the quantile of order 2.5% respectively 97.5% of the asymptotic distribution of T_n. We emphasize that the test statistics T_n whose histogram is given in (b) are computed from observed trajectories of different sizes n.

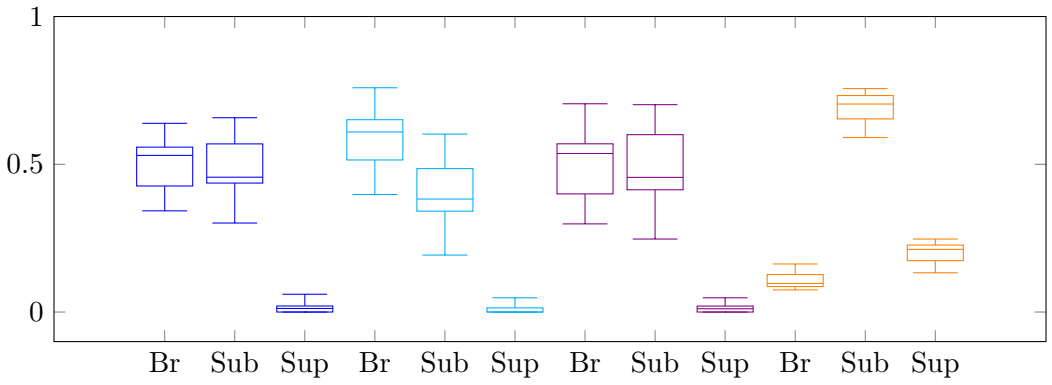


Figure 5.7: Boxplots of the proportions of Brownian, subdiffusion and superdiffusion computed from 12 two-dimensional Rab11a sequences. In blue proportions obtained with the single test procedure, in cyan with the Procedure 1, in violet with the adaptive Procedure 1 and in orange with the MSD method. Br stands for Brownian, Sub for subdiffusion and Sup for superdiffusion.

et al. [2012] show that a molecular complex containing Rab11a is necessary to sustain proper trafficking of Langerin and that Langerin delivery at the plasma membrane is always preceded by the docking and/or tethering of Rab11A/Rab11-FIP2 positive vesi-

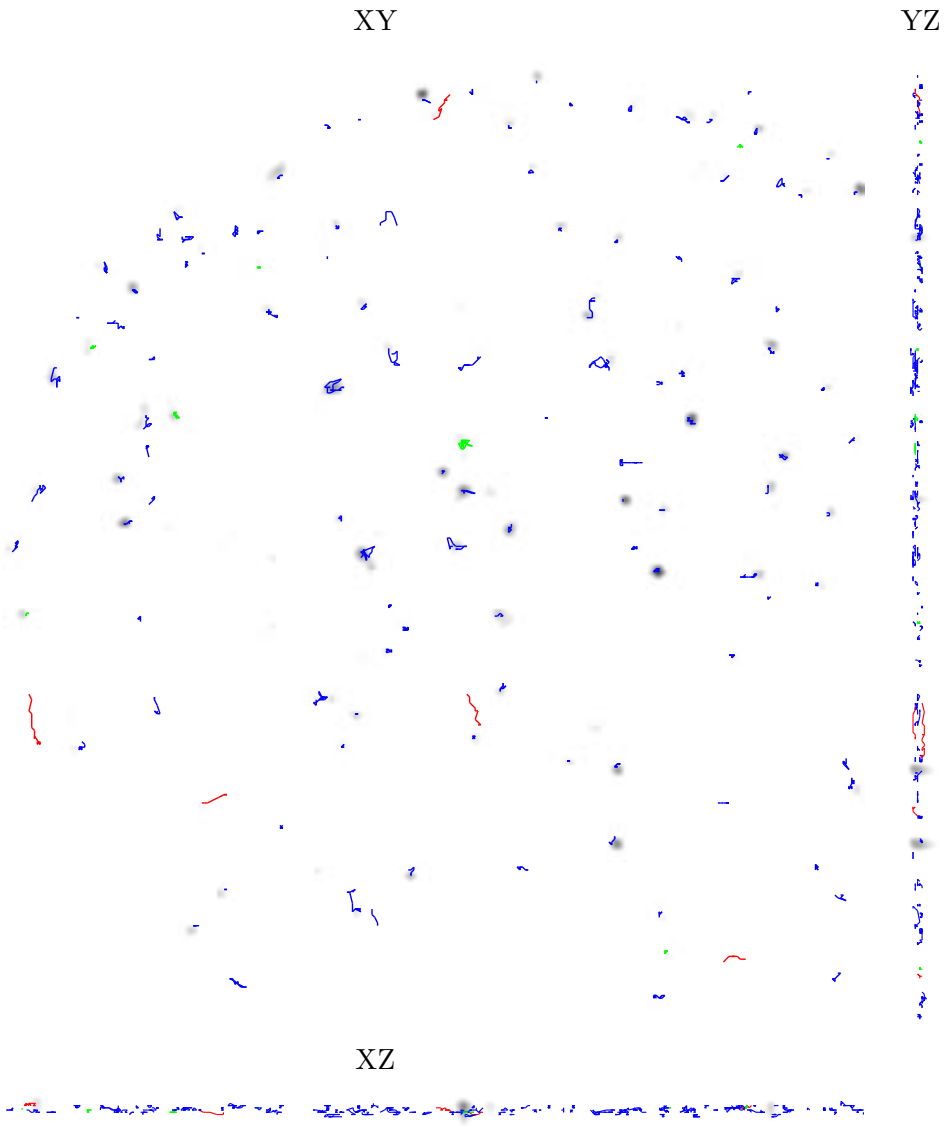


Figure 5.8: Orthogonal views of the three-dimensional Langerin trajectories classified with the adaptive Procedure 1. The colour code is: blue for Brownian motion, red for superdiffusion and green for subdiffusion Out of 123 trajectories, 111 are labelled as Brownian, 5 as superdiffusive and 7 as subdiffusive.

cles. Then, in this case, the tracking of trajectories of Rab11a is equivalent to track vesicles containing Langerin.

Here, the data consist in 31 three-dimensional volumes acquired at $\Delta = 0.1s$ with

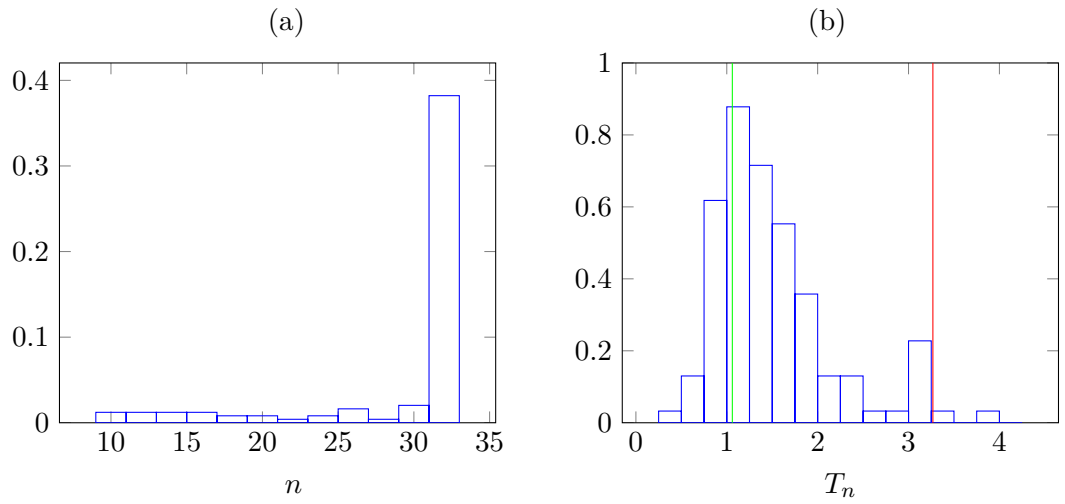


Figure 5.9: Histograms of the trajectory sizes n (a) and of the test statistics T_n (b) of the Langerin 3D sequence. In Figure (b), the green (respectively red) vertical line represents the quantile of order 2.5% (respectively 97.5%) of the asymptotic distribution of T_n. We emphasize that the test statistics T_n whose histogram is given in (b) are computed from observed trajectories of different sizes n.

3D TIRF[1] microscopy by incidence angle scanning and azimuthal averaging [Boulanger et al., 2014]. Each volume is composed of 20 images of size 402×402 pixels defining the XY plane. The Z resolution is much higher (about $50nm$) than the resolution in the X and Y direction (about $200nm$). We use the same pre-processing of the data as in Section 5. We classify 123 trajectories with the adaptive Procedure 1. The median length of the trajectories is $n = 31$ corresponding to the situation in which the trajectory is observed through the whole duration of the experiment. Results are shown on Figure 5.8. The histogram of the trajectory sizes is given in Figure (a) 5.9. We also give the histogram of the test statistic T_n in Figure (b) 5.9.

5.4 Summary

In this chapter, we evaluated both our single test procedure (4.5.1) and multiple test Procedure 1 on simulations in the two-dimensional case. To this end, we computed the power curves under the alternative of parametric diffusion models, namely the Ornstein-Uhlenbeck process, the fractional Brownian motion and the Brownian motion with drift. We showed on simulations that our test approach was more reliable than the method of

[1]Total Internal Reflection Fluorescence.

Feder et al. [1996] based on the MSD. We also studied real data depicting the exocytosis in both the two and three-dimensional cases.

In this part, we assumed that the particles were driven by the same diffusion process over time. In Part II, we relax this assumption and assume that a particle can switch between the three types of diffusion of interest (Brownian motion, subdiffusion and superdiffusion) over time.

Part II

Detection of Motion Switching along Particle Trajectories

6 A Sequential Algorithm to Detect Change Points

In this chapter, we use the test statistic proposed in Chapter 4 in a new setting, namely change point analysis. As intracellular transport presents a high heterogeneity of motions depending on the spatial location, the particle switches dynamics over time while crossing different areas of the cell. For instance, Lagache et al. [2009] model the dynamics of a virus invading a cell with successive switches between superdiffusion along the microtubules and Brownian motion in the cytosol. We develop here a sequential method for detecting the time at which an intracellular particle changes dynamic. More precisely, we are interested in changes from one type of diffusion (superdiffusion, subdiffusion or Brownian motion) to another type of diffusion.

6.1 Change Point Model

We observe a discrete trajectory $\mathbb{X}_n = (X_{t_0}, X_{t_1}, \dots, X_{t_n})$ with $t_i - t_{i-1} = \Delta$ as defined in Equation (4.1.1). We assume that the discrete trajectory is generated by a d-dimensional ($d = 2$ or $d = 3$) diffusion process (X_t) strong solution of the stochastic differential equation:

$$dX_t = \mu(X_t, t)dt + \sigma(t)dB_t^{\mathfrak{h}(t)}, \qquad t \in [t_0, t_n], \tag{6.1.1}$$

where $B^{\mathfrak{h}(t)}$ denotes a $d-$dimensional fractional Brownian motion of Hurst parameter $\mathfrak{h}(t)$; the unknown parameters of the model are the Hurst parameter function $\mathfrak{h} : \mathbb{R}^+ \to (0,1)$, the diffusion coefficient function $\sigma : \mathbb{R}^+ \to (0, \infty)$ and the drift term $\mu : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$.

Furthermore, we assume that there exists a sequence of N change points on $[t_0, t_n]$, namely $t_0 = \tau_0 < \tau_1 < \dots \tau_N < \tau_{N+1} = t_n$ such that,

$$\forall j \in \{0 \dots N\}, \forall x \in \mathbb{R}^d, \forall t \in [\tau_j, \tau_{j+1}), \quad \mu(x,t) = \mu_j(x) \tag{6.1.2}$$

$$\mathfrak{h}(t) = \mathfrak{h}_j \tag{6.1.3}$$

$$\sigma(t) = \sigma_j. \tag{6.1.4}$$

The number of change points N, the drift functions $(\mu_j)_{j=0\dots N}$ and the diffusion coefficient $(\sigma_j)_{j-0\dots N}$ are unknown. We note that as N is unknown the vector of change points $(\tau_j)_{j=1\dots N}$ is also unknown. We also assume that the drift terms μ_j

satisfy the Lipschitz and linear growth conditions of Assumption 1 or Assumption 2. Then, the stochastic differential equation (6.1.1) admits a strong solution on each interval $[\tau_j, \tau_{j+1})$. We extend by continuity the solution on each subinterval to get a solution on $[t_0, t_n]$. Moreover, we assume that $(\mathfrak{h}_j, \mu_j)$ and $(\mathfrak{h}_{j+1}, \mu_{j+1})$ are associated to different types of diffusion. We note that the parameter σ_j does not influence the type of diffusion. For example, $\mathfrak{h}_j = 1/2$ and $\mu_j(x) = 0$ define the Brownian motion on $[\tau_j, \tau_{j+1})$ then $(\mathfrak{h}_{j+1}, \mu_{j+1})$ must define a subdiffusion or superdiffusion on $[\tau_{j+1}, \tau_{j+2})$.

In the sequel, $P^{\boldsymbol{\tau}}_{\boldsymbol{\mathfrak{h}},\boldsymbol{\mu},\boldsymbol{\sigma}}$ denotes the measure induced by the stochastic process (X_t) solution of (6.1.1). We define the subscripts $\boldsymbol{\mathfrak{h}}, \boldsymbol{\mu}, \boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ as follows:

- $\boldsymbol{\tau} = (\tau_j)_{j=1\dots N} \in \mathbb{R}^{N+*}$ is the vector of change points with $\tau_1 < \tau_2 < \cdots < \tau_N$,
- $\boldsymbol{\mathfrak{h}} = (\mathfrak{h}_j)_{j=0\dots N} \in (0,1)^{N+1}$ is the vector of Hurst index,
- $\boldsymbol{\mu} = (\mu_j)_{j=0\dots N}$ is the set of $N+1$ drift functions from $\mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$,
- $\boldsymbol{\sigma} = (\sigma_j)_{j-0\dots N} \in \mathbb{R}^{(N+1)+*}$ is the vector of diffusion coefficients.

We set $(\boldsymbol{\mathfrak{h}}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\tau}) = (\mathfrak{h}, \mu, \sigma, \emptyset)$ if there is no change point.

Finally, we suppose that for each τ_j there exists $0 \leq j^\star \leq n$ such that $\tau_j = t_{j^*}$. It means that the change of motion occurs precisely at a sampling time. Then, we define the subtrajectory $\mathbb{X}^j_{n_j} = (X_{\tau_j}, \dots, X_{\tau_{j+1}})$ of size n_j generated with diffusion parameters $(\mathfrak{h}_j, \mu_j, \sigma_j)$.

We present a sequential procedure to estimate both the number of change points N and the vector of change points $(\tau_1, \dots, \tau_N)$. In the next section, we present the sequential procedure as a statistical test.

6.2 Null and Alternative Hypothesis of the Test

We adapt the sequential procedure proposed in [Cao and Wu, 2015] to our problem. In our setting, the sequence of p-values is replaced by the trajectory $\mathbb{X}_n$. Our global null hypothesis is that (X_t) is a Brownian motion on $[t_0, t_n]$:

$$H_0 : \mathbb{X}_n \text{ is generated from } (\sigma B_t)_{t_0 \leq t \leq t_n}. \tag{6.2.1}$$

Our alternative hypothesis is that there exist $\tau_0 = t_0 < \tau_1, \dots, \tau_N < \tau_{N+1} = t_n$ such that:

1. $\mathbb{X}_n = (\mathbb{X}^1_{n_1}, \dots, \mathbb{X}^N_{n_N})$ where the subtrajectory $\mathbb{X}^j_{n_j}$ is generated with diffusion parameters $(\mathfrak{h}_j, \mu_j, \sigma_j)$.
2. For all $j = 1, \dots, N$ diffusion parameters $(\mathfrak{h}_j, \mu_j, \sigma_j)$ and $(\mathfrak{h}_{j+1}, \mu_{j+1}, \sigma_{j+1})$ are associated to different types of diffusion (Brownian, subdiffusion or superdiffusion).

Remark 6.2.1. *The case where the whole trajectory is subdiffusive or superdiffusive belongs to the alternative hypothesis. In this case there is no change point ($\boldsymbol{\tau} = \emptyset$).*

In the next section, we present the sequential procedure. The parameters of this algorithm can be chosen such that we control the type I error of the aforementioned test at level α. In other words, with appropriate parameters, if the trajectory is fully Brownian, we will not detect any change point with probability $1 - \alpha$.

6.3 Procedure

Our procedure comprises three main steps:

1. detect the potential change points,
2. gather these potential change points in clusters; one cluster is assumed to contain a single change point,
3. estimate the change point in each cluster.

The critical parameter of our method is the size of the local window k (see Box Page 70). There are two parameters to detect the potential change point (γ_1, γ_2) and two parameters defining the clusters $(c, c^\star)$. We explain each step of our procedure in the next subsections.

Detecting the Potential Change Points

Let $1 \leq k \leq n/2$. We will discuss about the choice of k in Chapter 7. For all index i such that $t_k \leq t_i \leq t_{n-k}$, we consider two subtrajectories of size k starting at X_{t_i},

- the backward trajectory $\mathbb{X}_i^- = \{X_{t_i}, X_{t_{i-1}}, \dots X_{t_{i-k}}\}$,
- the forward trajectory $\mathbb{X}_i^+ = \{X_{t_i}, X_{t_{i+1}}, \dots X_{t_{i+k}}\}$.

We compute the test statistic (4.3.1) for the backward and forward trajectory as,

$$B_i = \frac{\max_{j=1,\dots,k} \left\| X_{t_{i-j}} - X_{t_i} \right\|}{\sqrt{(t_{i+k} - t_i)}\hat{\sigma}(t_{i-k} : t_i)}, \qquad A_i = \frac{\max_{j=1,\dots,k} \left\| X_{t_{i+j}} - X_{t_i} \right\|}{\sqrt{(t_{i+k} - t_i)}\hat{\sigma}(t_i : t_{i+k})}. \tag{6.3.1}$$

where $\hat{\sigma}(t_i : t_{i+k})$ (respectively $\hat{\sigma}(t_{i-k} : t_i)$) denotes the estimate of the diffusion coefficient from the forward trajectory $\mathbb{X}_i^+$ (respectively the backward trajectory $\mathbb{X}_i^-$). We note that if we use the standard estimate of the diffusion coefficient proposed in Section 4.6, we have $\hat{\sigma}(t_{i-k} : t_i) = \hat{\sigma}(t_i : t_{i-k})$. The denomination B_i (respectively A_i) is for *"Before time t_i"* (respectively *"After time t_i"*). We illustrate this sequential procedure on Figure 6.1.

Now, we want to compare the backward statistic B_i and the forward statistic A_i. The principle is that if the two values are in the same range of values, it is unlikely that time t_i is a change point. Then, we use two cut-off values $\gamma_1 < \gamma_2$ which depend on the parameters of the procedure, and we define the following step-function:

$$\phi(x;\gamma_1,\gamma_2) = \begin{cases} 1 & \text{if } x < \gamma_1 \\ 2 & \text{if } x > \gamma_2 \\ 0 & \text{otherwise.} \end{cases} \tag{6.3.2}$$

We have the following interpretation of the cut-off values: $\phi(A_i;\gamma_1,\gamma_2) = 0$ means that $\mathbb{X}_i^+$ is Brownian, $\phi(A_i;\gamma_1,\gamma_2) = 1$ means that $\mathbb{X}_i^+$ is subdiffusive and $\phi(A_i;\gamma_1,\gamma_2) = 2$ superdiffusive. Then we compute:

$$Q_i = \phi(A_i;\gamma_1,\gamma_2) - \phi(B_i;\gamma_1,\gamma_2), \quad i = k,\dots,n-k. \tag{6.3.3}$$

If $Q_i = 0$ it means that the statistics B_i and A_i belong to the same range of values defined by γ_1 and γ_2. Then, both $\mathbb{X}_i^+$ and $\mathbb{X}_i^-$ are from the same type of diffusion: it is unlikely that t_i is a change point. On the contrary, if $Q_i \neq 0$ the subtrajectories $\mathbb{X}_i^+$ and $\mathbb{X}_i^-$ are not from the same type of diffusion and t_i is a potential change point. The detection step is illustrated on a simulated trajectory in Figure 6.2.

Gathering the Potential Change Points into Clusters

A first option proposed by Cao and Wu [2015] is to consider that a cluster is composed of successive index i (in their context location in DNA sequence) such that $Q_i \neq 0$. Cao and Wu [2015] require the cluster to have a minimal size $r^\star$ set to $k/2$. However, in our case, this choice does not work well. Due to the high level of randomness of the stochastic processes modelling the trajectory, we observe rarely clusters of size $k/2$ of successive position i in the trajectory such that $Q_i \neq 0$. Then the procedure does not detect any change point (low power of the test). Also, optimizing the minimal size $r^\star$ is tricky and can lead to overdetection or underdetection depending on the situation. Therefore we choose an other way to build clusters. Even if it is hard to observe successive potential change points, we argue that a subset of indexes where the concentration of potential change point is high (even if there are not connected) is likely to contain a true change point. Then, we define a cluster of potential change points as a subset of index $\mathcal{M} = \{i,\dots,i+l\}$ such that:

$$\sum_{j=m}^{m+c-1} \mathbf{1}(Q_j \neq 0) \geq c^\star, \quad \forall m = i,\dots,i+l-c+1, \tag{6.3.4}$$

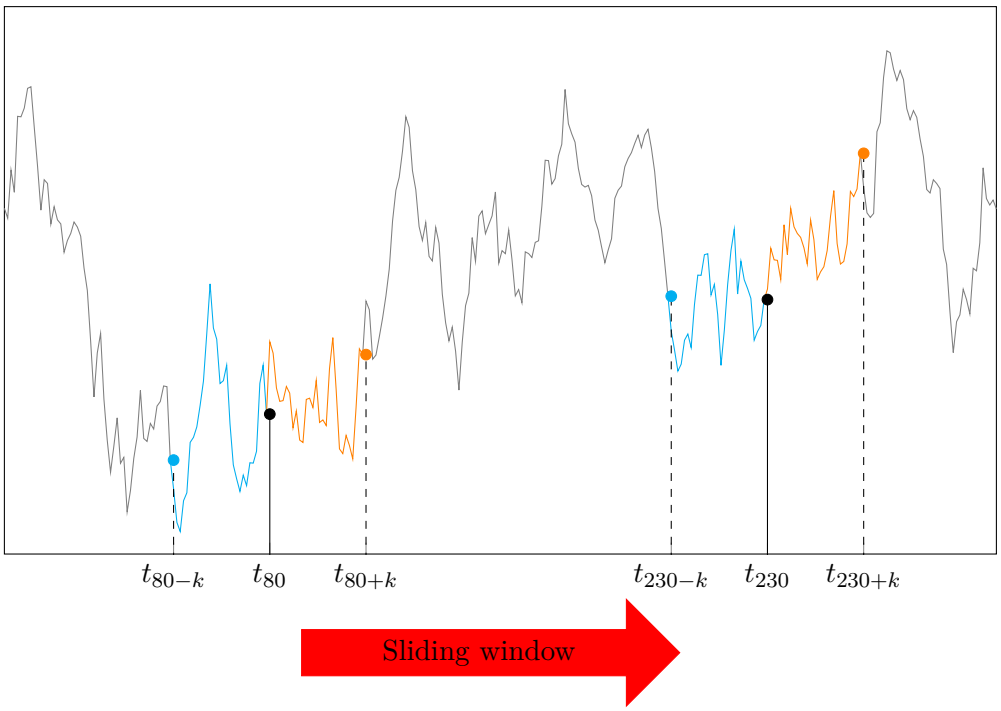


Figure 6.1: Illustration of the sequential procedure on a one dimension toy trajectory. Blue parts are the backward subtrajectories on which we compute B_i. Orange parts are the forward subtrajectories on which we compute A_i. The black points are the centres of the backward and forward subtrajectories. We shift the backward and forward subtrajectories all along the trajectory to compute the sequence of (B_i, A_i), as shown by the red arrow.

where c and $c^\star$ are tuning parameters. We set $c = k/2$, therefore the cluster has a minimal size of $k/2$ as in [Cao and Wu, 2015]. A cluster is created if there are at least $c^\star$ potential change points in a set of c successive points. The parameter $c^\star$ defines the minimum concentration of potential change point needed to build a cluster. Intuitively, we should have $c^\star \geq c/2$: the concentration of potential change points i $(Q_i \neq 0)$ is higher than the concentration of points i such that $Q_i = 0$. We set $c^\star = 0.75c$. We note that some points of the clusters are not potential change points $(Q_i = 0)$. We emphasize that the choice $c^\star = c$ is equivalent to build clusters as presented in [Cao and Wu, 2015].

To illustrate the construction of the clusters, we reproduce a portion of the sequence of $\mathbf{1}(Q_j \neq 0)$ computed on a trajectory simulated with the same parameters as the trajectory presented in Figure 7.1 (b):

0 0 0 1 0 0 1 0 0 1 0 1 0 1 1 0 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 0
1 1 1 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0.

Figure 6.2: Illustration of the detection step on a simulated trajectory. The trajectory is a zoom of the trajectory represented in Figure 7.1 (a) on the portion corresponding to the time interval $[50, 140]$. First the trajectory is Brownian, then after $\tau_1 = 100$ it undergoes Brownian motion with drift. The change point τ_1 is represented by the black dot. The blue (respectively orange) part corresponds to the subtrajectory $\mathbb{X}_{\tau_1}^-$ respectively $\mathbb{X}_{\tau_1}^+$. The size of the subtrajectories is $k = 30$. The gray part is the complementary part of the trajectory. The two circles are centred on X_{τ_1}. They represent the thresholds defining the different kinds of motion. The radius of the inner (respectively outer) circle is $\gamma_1\sigma\sqrt{k\Delta}$ (respectively $\gamma_2\sigma\sqrt{k\Delta}$). For the didactic purpose of the illustration, here we consider σ known (we estimate it in our method). The blue subtrajectory stays inside the inner circle: it is classified as subdiffusive. The orange subtrajectory goes outside the outer circle: it is classified as superdiffusive. If the maximum of one part had lied between the two circles it would have been classified as Brownian. The maximum of the blue and orange subtrajectories lie in different regions defined by the two limit circles, then $\tau_1 = 100$ is detected as a potential change point.

We use a window size of $k = 30$. Then the default parameters of the clustering step are $c = 15$ and $c^\star = 10$. The blue part defines a cluster. In fact, along this cluster, the condition (6.3.4) is respected. It turns out that the index of a true change point is contained in this cluster. We would not have detected this change point with the choice of Cao and Wu [2015] (corresponding to $c^\star = c$). To detect the same cluster with the choice of Cao and Wu [2015], we should have observed the following sequence of $\mathbf{1}(Q_j \neq 0)$:

0 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0.

Estimating the Change Point in each Cluster

Denote $\mathcal{M}_k$ the k^{th} cluster. We estimate the change point of cluster $\mathcal{M}_j$ by:

$$\hat{\tau}_j = t_{r_j}, \quad r_j = \max_{i \in \mathcal{M}_j} |B_i - A_i|. \tag{6.3.5}$$

We choose the point i of the cluster for which B_i and A_i are the most different. The rational of this idea is that, at the exact change point $\tau_j = t_{r_j}$, $\mathbb{X}_{r_j}^-$ and $\mathbb{X}_{r_j}^+$ are trajectories generated from different diffusion processes and thus B_{r_j} and A_{r_j} must be the most different. At a point t_i close to τ_j, the subtrajectories $\mathbb{X}_i^-$ and $\mathbb{X}_i^+$ are composed of a mixture of diffusion. Then B_i and A_i reflect this mixture and $|B_i - A_i| \leq |B_{r_j} - A_{r_j}|$.

Finally we can summarize the method as follows:

Procedure 2.

1. *For a chosen window size k compute B_i and A_i in* (6.3.1) *for $i = k, \ldots, n-k$.*
2. *For prespecified cut-off values $\gamma_1 < \gamma_2$ compute $Q_i = \phi(A_i; \gamma_1, \gamma_2) - \phi(B_i; \gamma_1, \gamma_2)$.*
3. *Decompose $\{k, \ldots, n-k\} = W_0 \cup W_1$ where $i \in W_0$ if $Q_i = 0$ and $i \in W_1$ if $Q_i \neq 0$..*
4. *Gather the potential change points, that is points t_i such that $Q_i \neq 0$, into clusters $\mathcal{M}_1, \ldots, \mathcal{M}_{\hat{N}}$ satisfying Equation* (6.3.4).
5. *For each $\mathcal{M}_j$ let $r_j = \max_{i \in \mathcal{M}_j} |B_i - A_i|$ then $\hat{\tau}_j = t_{r_j}$.*

The parameters of Procedure 2 are the size of the window k, the parameters defining the clusters c and $c^\star$ and the cut-off-values (γ_1, γ_2). We recommend to set $c = k/2$ and $c^\star = 0.75c$. A choice for the cut-off values (γ_1, γ_2) is given in Section 6.4. Then, the only free parameter to be set by the user is the window size k. The influence of parameter k is discussed in Chapter 7.

6.4 Cut-off Values

We choose γ_1 and γ_2 such that we control the type I error at level $0 < \alpha < 1$ that is:

$$P^{\emptyset}_{1/2,0,\sigma}(\exists i \in \{k, \ldots, n^\star\}, \quad \sum_{j=i}^{i+c-1} \mathbf{1}(Q_j \neq 0) \geq c^\star) \leq \alpha, \tag{6.4.1}$$

where $n^\star = n - k - c + 1$. We explain why controlling the probability in (6.4.1) at level α is equivalent to control the type I error at level α. The left hand side of Equation (6.4.1)

is the probability to build one cluster of minimal size c (in the sense of (6.3.4)) under H_0. With Procedure 2, we need to build a cluster of potential change points to detect a change point, otherwise no change point is detected. Then, controlling the probability in (6.4.1) at level α under H_0 is equivalent to control the probability to detect falsely a change point under H_0 at level α (definition of the type I error). Now we have the following proposition:

Proposition 3. *Let define* $d_i = \min(B_i, A_i)$ *and* $D_i = \max(B_i, A_i)$ *where* A_i *and* B_i *are the test statistics* (6.3.1)*, for* $i = k, \dots, n^\star$*. We define* $\gamma_1^\star$ *and* $\gamma_2^\star$ *as:*

$$\begin{aligned} P^{\emptyset}_{1/2,0,\sigma}\Big(\min_{i=k,\dots,n^\star} d_{i(c^\star/2)} < \gamma_1^\star\Big) &= \frac{\alpha}{2}, \\ P^{\emptyset}_{1/2,0,\sigma}\Big(\max_{i=k,\dots,n^\star} D_{i(c-c^\star/2)} > \gamma_2^\star\Big) &= \frac{\alpha}{2}, \end{aligned} \tag{6.4.2}$$

where $d_{i(c^\star/2)}$ *is the* $c^\star/2$ *smallest element of* $(d_i, \dots, d_{i+c-1})$ *and* $D_{i(c-c^\star/2)}$ *the* $c - c^\star/2$ *smallest element of* $(D_i, \dots, D_{i+c-1})$ *(equivalently the* $c^\star/2$ *greatest element). In other words,* $\gamma_1^\star$ *is the quantile of order* $\alpha/2$ *of the random variable* $\min_{i=k,\dots,n^\star} d_{i(c^\star/2)}$ *and* $\gamma_2^\star$ *is the quantile of order* $1-\alpha/2$ *of the random variable* $\max_{i=k,\dots,n^\star} D_{i(c-c^\star/2)}$*. With the choice of cut-off values* $\gamma_1^\star$ *and* $\gamma_2^\star$ *Procedure 2 with parameters* $(k, c, c^\star)$ *controls the type I error* (6.4.1) *at level* α*.*

A proof of Proposition 3 is postponed in Appendix B. We estimate $\gamma_1^\star$ and $\gamma_2^\star$ with Monte-Carlo simulation, see Algorithm 2.

When we analyse the proof of Proposition 3 (see Appendix B), we realize that the choice $(\gamma_1^\star, \gamma_2^\star)$ is not optimal. In particular, the bound in Equation (B.0.7) is loose. Then, we can see from simulations that the probability of type I error is controlled at a much lower level (about) than α (see Table 6.1). Consequently, we recommend to use the cut-off values verifying:

$$\begin{aligned} P^{\emptyset}_{1/2,0,\sigma}\Big(\min_{i=k,\dots,n^\star} d_{i(c^\star)} < \tilde{\gamma_1}\Big) &= \frac{\alpha}{2}, \\ P^{\emptyset}_{1/2,0,\sigma}\Big(\max_{i=k,\dots,n^\star} D_{i(c-c^\star)} > \tilde{\gamma_2}\Big) &= \frac{\alpha}{2}. \end{aligned} \tag{6.4.3}$$

We replace $c^\star/2$ in Equation (6.4.2) by $c^\star$. Then, it is straightforward to show that $\gamma_1^\star \leq \tilde{\gamma}_1$ and $\gamma_2^\star \geq \tilde{\gamma}_2$. We deduce from these inequalities that the power of Procedure 2, that is its ability to find a true change point, is higher with the choice $(\tilde{\gamma}_1, \tilde{\gamma}_2)$ than with $(\gamma_1^\star, \gamma_2^\star)$. In fact, a high value of γ_1 detects better subdiffusions than a low value of γ_1. The other way around, a low value of γ_2 detects better superdiffusions than a high value of γ_2. As stated before, the control the type I error constrains the choice of (γ_1, γ_2). We show from simulations that the choice $(\tilde{\gamma}_1, \tilde{\gamma}_2)$ still controls the type I error at level α (see Table 6.1). Estimations of $(\gamma_1^\star, \gamma_2^\star)$ and $(\tilde{\gamma}_1, \tilde{\gamma}_2)$ –obtained with the Monte Carlo Algorithm 2– are given in Table 6.2.

Table 6.1: Control of the type I error for different cut-off values (γ_1, γ_2). We estimate the probability of type I error with the proportion of trajectories with at least one change point detected among 100 001 Brownian trajectories. We use the default parameters for Procedure 2 that is $c = k/2$ and $c^\star = 0.75c$. The estimations are accurate at $\pm 0.14\%$.

		Probability of Type I error	
n	k	with $(\gamma_1^\star, \gamma_2^\star)$	with $(\tilde{\gamma}_1, \tilde{\gamma}_2)$
150	20	0.60	5.21
150	30	0.65	4.81
150	40	0.94	4.56
300	20	0.47	5.04
300	30	0.59	4.89
300	40	0.82	4.83

Table 6.2: Cut-off values of Procedure 2 for different sizes of trajectory n and sizes of window k. The cut-off values are estimated with Algorithm 2 using $N = 10\,001$ Monte-Carlo replications.

n	k	$\gamma_1^\star$	$\gamma_2^\star$	$\tilde{\gamma}_1$	$\tilde{\gamma}_2$
150	20	0.61	3.38	0.74	3.09
150	30	0.65	3.35	0.78	3.05
150	40	0.68	3.28	0.80	3.03
300	20	0.58	3.55	0.71	3.27
300	30	0.62	3.55	0.74	3.26
300	40	0.64	3.52	0.75	3.25

6.5 Summary

In this chapter, we assumed that a particle was able to switch between the three diffusion types (Brownian motion, subdiffusion and superdiffusion) over time. In this context, we developed an algorithm for detecting the change points corresponding to these switches. We designed a sequential algorithm in which we compute the test statistic T_n (4.3.1) on local windows along the trajectory. The sequential scheme is adapted from the algorithm of Cao and Wu [2015] which computes CUSUM-like statistics along genomic sequences. We perform an original clustering step for gathering the potential change points which are close from each other. Our algorithm is user-friendly as there is only one parameter to tune, namely the window size k.

In the next chapter, we assess our algorithm on Monte Carlo simulations. We study

Algorithm 2: Estimation of the cut-off values $\gamma_1^\star$ and $\gamma_2^\star$ by Monte Carlo simulations. For estimating $(\tilde{\gamma}_1, \tilde{\gamma}_2)$, one should turn $c^\star/2$ into $c^\star$ in this algorithm.

Input: n, k, α, N
// the length n of the trajectory
// the size k of the subtrajectories
// the level $\alpha \in (0,1)$
// the number N of Monte Carlo experiments
Result: $\hat{\gamma}_1(n,k,\alpha)$ $\hat{\gamma}_2(n,k,\alpha)$
for *i=1* **to** N **do**
  Generate $\mathbb{X}_n^i$ of size n from the null hypothesis (6.2.1) with $\sigma = 1$ and $\Delta = 1$;
  // Compute the statistics (6.3.1) along $\mathbb{X}_n^i$
  for *j=k* **to** *n-k* **do**
    Compute (B_j^i, A_j^i) from (6.3.1);
    Set $d_j^i = \min(B_j^i, A_j^i)$;
    Set $D_j^i = \max(B_j^i, A_j^i)$;
  end
  for *r=k* **to** *n-k-c+1* **do**
    Compute s_r^i the $c^\star/2$ smallest element of $(d_r^i, \ldots, d_{r+c-1}^i)$;
    Compute S_r^i the $c - c^\star/2$ smallest element of $(d_r^i, \ldots, d_{r+c-1}^i)$;
  end
  Compute $m_i = \min_r(S_r^i)$ and $M_i = \max_r(s_r^i)$;
end
Let $(\tilde{m}_1, \ldots, \tilde{m}_N)$ the sorted m_is and $(\tilde{M}_1, \ldots, \tilde{M}_N)$ the sorted M_is;
Set $\hat{\gamma}_1(n,k,\alpha) = \tilde{m}_{\lfloor(\alpha/2)N\rfloor}$ and $\hat{\gamma}_2(n,k,\alpha) = \tilde{M}_{\lfloor(1-\alpha/2)N\rfloor}$;

two different scenarios mimicking real biophysical processes. In particular, we give some insights of the impact of the window size k on the results.

7 Simulation Study

We assess the algorithm proposed in Chapter 6 on different simulation scenarios. We limit this study to the two-dimensional case. Subdiffusion (respectively superdiffusion) is modelled with the Ornstein-Uhlenbeck process (4.2.1) (respectively with the Brownian motion with drift (4.2.3)). We also compare our method to two others procedures, the first proposed by Türkcan and Masson [2013], the second by Monnier et al. [2015].

7.1 Performance of the Method

We simulate two different scenarios, see Table 7.1, where the particle motion switches at same change points. Subdiffusions are modelled by Ornstein- Uhlenbeck process:

$$dX_t^i = -\lambda(X_t^i - \theta_i)dt + \sigma dB_t^{1/2,i}, \qquad i = 1, 2, \tag{7.1.1}$$

where $\lambda > 0$ models the restoring force toward the equilibrium point $\theta = (\theta_1, \theta_2)$; $\sigma > 0$ is the diffusion coefficient. For modelling superdiffusion we use the Brownian motion with drift solution of the SDE:

$$dX_t^i = (v/\sqrt{2})dt + \sigma dB_t^{1/2,i}, \qquad i = 1, 2, \tag{7.1.2}$$

where $\sigma > 0$ is the diffusion coefficient and $v > 0$. Then the constant drift $\mathbf{v} = (v, v)/\sqrt{2}$ verifies $\|\mathbf{v}\| = v$.

For each scenario, we compute the performances of our procedure for different values of the parameters v (for the Brownian motion with drift) and λ (for the Ornstein-Uhlenbeck process). We assess the performances of our algorithm with respect to two criteria:

1. the number of change points detected,

2. the location of these change points.

Criterion 2 is assessed only on the trajectories for which we detect the right number of change points that is $N = 2$. We compute the average and standard deviation of the locations. We analyse the results of the simulation on the different scenarios in the next paragraphs.

Table 7.1: Simulation scenarios for the Monte Carlo study. The size of the simulated trajectories is $n = 300$. The change points occur at $\tau_1 = 100$ and $\tau_2 = 175$. We set $\sigma = 1$ for the diffusion coefficient and $\Delta = 1$ for the step of time. For the Ornstein-Uhlenbeck process (7.1.1), we define the equilibrium point as $\theta = X_{\tau_1}$ where X_{τ_1} is the position of the particle at τ_1.

Times	Scenario 1	Scenario 2
$[1, 100]$	Brownian	Brownian
$[101, 175]$	Brownian with drift	Ornstein-Uhlenbeck
$[176, 300]$	Brownian	Brownian

Scenario 1

First, we illustrate the scenario of simulation with Figure 7.1 (a) showing a trajectory simulated with Scenario 1. Table 7.2 gives us the results associated to Scenario 1 (see Table 7.1). We can see clearly that, as $\|v\|$ increases, the performance of the method increases with respect to both criteria. For a given window size k, we get:

1. the proportion of trajectories for which we detect the right number of change point ($\hat{N} - N = 0$) tends to 1 as v inceases.

2. given $\hat{N} - N = 0$, the bias and the variance of the estimated change point decrease to 0 as v inceases.

We also notice that for the window size $k = 20$, the performance of the algorithm is lower than for $k = 30$ and 40 except when $\|v\| = 2$. As the size of the window is too low, it is hard for the algorithm to detect a Brownian motion with drift with a low drift norm. In particular, when $\|v\| = 0.6$, it does not detect any change point in most cases; we note that $\hat{N} - N = -2$ for 42.2% of the trajectories. However, when the drift norm is high, a low window size performs as good as the larger ones (see the case $\|v\| = 2$). It performs even better if the change points τ_1, τ_2 are closer. In this case, a large window tends to mix up the two change points and consequently find only one. We can summarize this as follows: a large window size enables to detect well the change points associated with a small drift v if the change points are significantly separated while a small window is able to distinguish two close change points if the drift v is large enough.

Scenario 2

We illustrate the scenario of simulation with Figure 7.1 (b) showing a trajectory simulated with Scenario 2. Table 7.3 gives us the results associated to Scenario 2 (see Table 7.1). As in Scenario 1, for a window size $k = 20$ the performance of the algorithm

Table 7.2: Performance of the Procedure 2 for Scenario 1 (see 7.1) for different window sizes k and different values of the drift v. The computations are based on 1 001 simulated trajectories from Scenario 1. We compute the proportions of trajectories with $\hat{N} - N = -2$, $\hat{N} - N = \pm 1$, $\hat{N} - N = 0$ and $\hat{N} - N \geq 2$. The column τ_1 (respectively τ_2) gives the empirical average of the first (respectively second) detected change point on 300 trajectories among which we detect the right number of change points ($\hat{N} - N = 0$). The number in brackets is the empirical standard deviation of the estimate of τ_1 and τ_2 computed on these 300 trajectories.

		$\hat{N} - N$								
v	k	-2	-1	0	1	≥ 2	τ_1		τ_2	
0.6	20	42.2	14.1	34.7	5.7	3.3	126.3	(23.7)	153.7	(23.6)
0.6	30	20.9	16.9	55.9	5.5	0.8	115.0	(17.8)	162.8	(18.4)
0.6	40	11.8	18.6	67.6	1.8	0.2	109.4	(15.5)	168.2	(15.1)
0.8	20	6.5	12.9	54.4	17.3	8.9	117.4	(16.7)	157.5	(18.5)
0.8	30	1.6	6.5	84.1	6.3	1.5	107.3	(11.6)	170.1	(14.1)
0.8	40	0.3	4.1	93.2	2.3	0.1	104.7	(9.7)	172.4	(10.0)
1	20	0.2	3.7	63.2	21.6	11.3	108.3	(12.1)	168.2	(13.2)
1	30	0.0	1.9	93.8	3.5	0.8	102.9	(5.9)	173.9	(6.5)
1	40	0.0	0.1	97.7	1.9	0.3	103.2	(6.6)	174.5	(7.3)
2	20	0.0	0.0	96.6	2.4	1.0	101.4	(2.1)	176.0	(2.3)
2	30	0.0	0.3	96.8	2.5	0.4	101.2	(3.4)	175.9	(2.6)
2	40	0.0	0.1	99.2	0.5	0.2	101.5	(2.8)	175.8	(2.9)

increases as λ increases. However, it does not behave the same way if the window size is 30 or 40. For $k = 30$, the performance increases from $\lambda = 1$ to $\lambda = 2$ but remains the same for larger values of λ. For the window size $k = 40$, the proportion of trajectories with the correct number of detected change points dramatically drops from 83.6% with $\lambda = 1$ to 54.1% for $\lambda = 4$. At the same time, the proportion of trajectories with $\hat{N} - N = -1$ increases. It means that when λ becomes too high the algorithm mixes up the two change points and find only one. As λ is high (clear subdiffusion), we detect a potential change point very early in the trajectory: as soon as few points of the forward subtrajectory $\mathbb{X}_i^+$ enter in the subdiffusion regime ($t \geq \tau_1$) we classify it as subdiffusive. For example, if λ is big enough we can suppose that the subtrajectory of size k $\mathbb{X}_i^+ = (X_{t_i}, \dots X_{\tau_1}, X_{\tau_1+1}, X_{\tau_1+2})$ will be classified as subdiffusive with only three points in the subdiffusive regime. Then, we get a long sequence of potential change points. But as k is large, the forward subtrajectory has already reached the second change point τ_2.

Table 7.3: Performance of the Procedure 2 for Scenario 2 (see 7.1) for different window sizes k and different values of parameter λ. We use the same protocol as in Table 7.2.

		$\hat{N} - N$								
λ	k	-2	-1	0	1	≥ 2	τ_1		τ_2	
1	20	18.1	44.1	31.8	5.3	0.7	109.9	(20.7)	167.8	(17.9)
1	30	0.8	16.3	78.3	4.2	0.4	104.9	(8.7)	169.9	(9.3)
1	40	0.0	13.2	83.6	3.0	0.2	105.6	(10.8)	170.4	(11.6)
2	20	3.1	22.6	68.1	5.5	0.7	106.4	(8.5)	170.2	(8.1)
2	30	0.1	6.5	89.4	3.4	0.6	107.5	(8.7)	169.1	(8.2)
2	40	0.0	21.2	77.0	1.6	0.2	108.0	(12.7)	169.1	(12.8)
3	20	1.1	17.1	74.8	5.9	1.1	106.3	(5.6)	170.1	(7.9)
3	30	0.0	5.7	90.2	3.2	0.9	108.7	(8.8)	167.6	(8.7)
3	40	0.1	32.1	64.8	2.5	0.5	109.3	(12.9)	166.4	(13.5)
4	20	0.6	12.2	79.4	6.8	1.0	107.2	(6.5)	169.9	(8.4)
4	30	0.0	6.5	89.7	3.1	0.7	109.6	(9.6)	166.5	(9.2)
4	40	0.0	44.3	54.1	1.4	0.2	111.5	(13.3)	166.0	(13.3)

Consequently, it begins to detect potential change points corresponding to the second change point τ_2. As there is a single cluster of potential change points, the algorithm only detects one change point instead of the two expected. From our simulations, we observe that the change point detected is either close to τ_1 or τ_2: it estimated correctly one change point out of the two real change points.

The idea is that, in a way, a large λ (a very clear subdiffusion) makes the two change points get closer artificially. Then, a large window can not separate them. We note that, from our simulations, in the case of a change point between Brownian motion with drift and Brownian motion, we do not observe such a phenomenon (that is a fall of the proportions of trajectories with the right number of detected change points when v increases). However, when the change points are close and the window size k is large compared to the gap between the change points, the performance stops increasing (but does not fall) above a certain value of v.

Once the change points are estimated, we can label the type of diffusion on each subtrajectory $\mathbb{X}_{n_j}^j = (X_{\hat{\tau}_j}, \ldots, X_{\hat{\tau}_{j+1}})$. For a given scenario, value of parameter (λ or v) and size window k, we assessed the labelling of the subtrajectories. Results and details of the evaluation are given in Table 7.4.

Table 7.4: Proportions of trajectories (among the trajectories with $\hat{N} = N$) for which subtrajectories are correctly labelled, in scenario 1 and 2. The change points are detected and estimated with Procedure 2. The subtrajectories are labelled using the test 4.5.1 at level 5%. Columns 3 et 4 (respectively 4 and 5) correspond to scenario 1 (respectively scenario 2). For example, in scenario 1 with $v = 0.6$, when we use a window of size $k = 20$, 73.7% of the trajectories for which we detect $N = 2$ change points are labelled as Brownian on $[t_0, \hat{\tau}_1]$, superdiffusive on $[\hat{\tau}_1, \hat{\tau}_2]$ and again Brownian on $[\hat{\tau}_2, t_n]$.

	Scenario 1		Scenario 2	
k	v	% right label	λ	% right label
20	0.6	73.7	1	83.0
30	0.6	82.3	1	89.3
40	0.6	86.0	1	85.0
20	0.8	74.7	2	90.7
30	0.8	87.7	2	89.0
40	0.8	88.7	2	88.7
20	1.0	82.0	3	89.3
30	1.0	86.3	3	87.3
40	1.0	88.7	3	86.7
20	2.0	89.7	4	88.7
30	2.0	90.0	4	85.0
40	2.0	90.3	4	84.7

7.2 Comparisons with Competitive Methods

We compare our method to two other methods. The method of Türkcan and Masson [2013] detects change points between Brownian motion and confined motion in a potential well. The method of Monnier et al. [2015] detects change points between Brownian motion and Brownian motion with drift. We note that none of these methods deal with the three types of diffusion (Brownian motion, subdiffusion and superdiffusion) as we do. In this section, we present the two competitive methods and compare their performances to Procedure 2 on simulations. At the end of the section, we give a particular emphasis on the speed and stability of the different methods.

The Method of Türkcan and Masson [2013]

First the method of Türkcan and Masson [2013] is a parametric method. The two parametric models under concern are the Brownian motion and the Ornstein-Uhlenbeck pro-

Table 7.5: Comparison of Procedure 2 and the method of Türkcan and Masson [2013] on the simulation of Türkcan and Masson [2013]. We recall that the true change point is $\tau_1 = 250$.

Method	$\hat{N} - N$				τ_1	
	-1	0	1	≥ 2		
Procedure 2	19	77	3	1	240.5	(29.4)
Method of Türkcan and Masson [2013]	27	59	14	0	176.3	(53.7)

cess (called diffusion in a harmonic potential in [Türkcan and Masson, 2013]). Türkcan and Masson [2013] select the model that minimizes the BIC criterion. For detecting change points, the BIC criterion is computed on a sliding window along the trajectory. When the BIC indicates a switch of model and that the new model is confirmed in the next r steps of times, a change is assumed to occur.

We reproduce the simulation described in Türkcan and Masson [2013]. We simulate $N = 100$ trajectories of size $n = 500$. First the trajectory undergoes an Ornstein-Uhlenbeck process and at time $\tau_1 = 250$ it switches to a Brownian motion. The two processes share the same diffusion coefficient $\sigma = 0.4472$. The parameters of the Ornstein-Uhlenbeck process (4.2.1) is $\lambda = 7.3870$. The step of time is $\Delta = 0.05$. Results of the two methods are given in Table 7.5. We can see that our method show better results in both the number $\hat{N}$ of detected change points and in the location of the change points. We also emphasize that we do not set $r = 3$ as in [Türkcan and Masson, 2013] but we set $r = 51$ which corresponds to the size of the window. With $r = 3$, the method of Türkcan and Masson [2013] detects more than 4 change points in 91% of the trajectories. Actually, the method is able to detect the change point, if a collection of about $N = 50$ trajectories showing the same number of change points at the same location is available. Accordingly, it provides good results in average. However, such a situation is not realistic in practical imaging. In our scenarios, our non-parametric method outperforms the parametric method of Türkcan and Masson [2013].

The Method of Monnier et al. [2015]

Monnier et al. [2015] use two parametric models to fit the displacements of the particle: the Brownian motion and the Brownian motion with drift. We note that the Brownian motion can be seen as a Brownian motion with a null drift. The two models are actually a unique parametric model with parameters $v = (v_1, v_2)$ and σ (with $v = (0, 0)$ for the Brownian case). Then, Monnier et al. [2015] use hidden Markov models to fit the displacements of the particle over time. The hidden states are defined as a set of a drift parameter and diffusion coefficient $S_k = v_k, \sigma_k$. They estimate both the number of states K, the parameters (v_k, σ_k) and the successive (hidden) states along the trajectories.

They also add a constrained $v = 0$ for modelling Brownian motion. Model selection is used with a Bayesian criterion to select the best model. If we assume that $K \leq 2$ and also consider the constrained models with $v = 0$, the competing models are:

Model 1 a single state which is the Brownian with parameter σ_1,

Model 2 a single state which is the Brownian with drift with parameters (v_1,σ_1),

Model 3 two states which are two Brownian with parameter σ_1 and σ_2,

Model 4 two states which are one Brownian and one Brownian with drift with respective parameters σ_1 and (v_2, σ_2),

Model 5 two states which are two Brownian with drift with parameters (v_1, σ_1) and (v_2, σ_2).

In our experiment, we run the method of Monnier et al. [2015] on 100 simulated trajectories from Scenario 1. We assume $K \leq 2$ that is the competing models are the five models aforementioned. Results are given in Table 7.6. When $v = 0.6$ or 0.8, Monnier et al. [2015] detect no change point ($\hat{N} - N = -2$) for a large majority of the trajectories. In this case, the selected model can either have one state (models 1 and 2) or have two states but from the same type of diffusion (models 3 and 5). Actually, when $v = 0.6, 0.8$, the preferred model is Brownian only (model 1) for most of the trajectories (see Table 7.7). Then the drift is too low to select a model involving Brownian with drift. As expected, the performance of the method of Monnier et al. [2015] improves as v increases. The method detects the right number of change points for 96% of the trajectories when $v = 2$. When the method detects at least one change point, it means that the selected model is the model 5. Even when the right model is chosen, it can over-detect the number of change points (that is $\hat{N} - N \geq 1$). We have 9% of overdetection when $v = 1$. When the method detects the right number of change points ($\hat{N} - N = 0$), the location of the change points are very close to the true locations. For instance when $v = 2$, the average location of the first detected change point is 100 (which is exactly τ_1) and its standard deviation is 1.4. Finally, our non-parametric method detects better the change points when the drift is low ($v \leq 0.8$). The quality of detections are similar when the drift is high enough ($v = 2$). For $v = 1$, we have a larger proportion of trajectories detected with the right number of change points with our method except when we use a window size of $n = 20$ (63.2% with our method *versus* 68% with the method of Monnier et al. [2015]). The locations of the detected change points among the trajectories with $\hat{N} - N = 0$ are slightly more accurate with the method of Monnier et al. [2015].

Algorithmic Considerations

Finally, we compare the speed and stability of the different methods. The method of Monnier et al. [2015] is time consuming because of the estimation of the *a posteriori*

Table 7.6: Performance of the algorithm of Monnier et al. [2015] for Scenario 1 (see 7.1). The computations are based on 100 simulated trajectories from Scenario 1. We compute the proportions of trajectories with $\hat{N} - N = -2$, $\hat{N} - N = \pm 1$, $\hat{N} - N = 0$ and $\hat{N} - N \geq 2$. The column τ_1 (respectively τ_2) gives the empirical average of the first (respectively second) detected change point on the trajectories among which we detect the right number of change points ($\hat{N} - N = 0$). The number in brackets is the empirical standard deviation of the estimates of τ_1 and τ_2. We note that the empirical average and standard deviation estimate of τ_1 and τ_2 are not computed over the same number of trajectories for the different values of the drift v (causing the null standard deviation line 1).

	$\hat{N} - N$								
v	-2	-1	0	1	≥ 2	τ_1		τ_2	
0.6	99	0	1	0	0	93.0	(0.0)	177.0	(0.0)
0.8	82	0	15	1	2	96.0	(7.9)	173.4	(3.9)
1.0	23	0	68	7	2	99.9	(3.9)	174.9	(4.3)
2.0	0	0	96	1	3	100.0	(1.4)	175.0	(1.2)

Table 7.7: Selected models with the method of Monnier et al. [2015] on 100 simulated trajectories from Scenario 1. BR (respectively BRD) stands for Brownian (respectively Brownian with drift). For instance, when $v = 0.6$, the method of Monnier et al. [2015] states that the best fit is Brownian motion only for 97 trajectories; Brownian motion with drift only for 2 trajectories; a mix of Brownian and Brownian motion with drift for 1 trajectory.

v	BR	BRD	BR-BR	BR-BRD	BRD-BRD
0.6	97	2	0	1	0
0.8	74	8	0	18	0
1	18	5	0	77	0
2	0	0	0	100	0

distribution by the Metropolis-Hastings algorithm. Assuming $K \leq 2$, it took $115s$ in average to deal with one trajectory of the simulation presented in Table 7.6 (300 points) with four cores working in parallel on a Mac Book Pro version 10.10.1 equipped with 2.8 GHz Intel Core i7, 16 Gb of RAM. In comparison, our method takes less than $0.05s$ to process a trajectory without working in parallel. Both our procedure 2 and the method of Türkcan and Masson [2013] compute quantities on local windows (in our case the statistics 6.3.1, the BIC of different models for [Türkcan and Masson, 2013]). From this aspect, the complexity of these two algorithms is equivalent. However, Türkcan

and Masson [2013] needs to estimate the MAP (maximum *a posteriori*) to compute the BIC. They choose a complex likelihood to model the spatial heterogeneity of the motion. Therefore, they use quasi-Newtonian optimization to find the MAP which is the most time consuming step of their procedure. It took in average $11s$ to process a trajectory of the simulation presented in Table 7.6 (500 points) against less than $0.05s$ for Procedure 2. In term of stability, different runs of the method of Monnier et al. [2015] on the same trajectory can give different results (see Section 7.3). This is due to a bad convergence of the Metropolis-Hastings algorithm. In rare cases, the optimization step of Türkcan and Masson [2013] can fail. Procedure 2 does not suffer any of these problems as it does not involve any parameter inference.

7.3 Real Data

We use the same data as Monnier et al. [2015] depiciting long-range transport of mRNAs in complex with mRNA-binding proteins (mRNPs) (see Figure 7.2). In live neuronal cultures, endogenous β-actin mRNP particles alternate between Brownian motion and active transport. In case of active transport (superdiffusion), the particle is driven by molecular motors along microtubule tracks in the neuronal dendrites. The microscopic sequence was obtained using mRNA fluorescence labeling techniques. More specifically, in the experiment of Monnier et al. [2015], the MS2 bacteriophage capsid protein was tagged with a GFP (Green Fluorescence Protein). As the MS2 bacteriophage capsid protein binds to β-actin mRNP, it allows to track this latter.

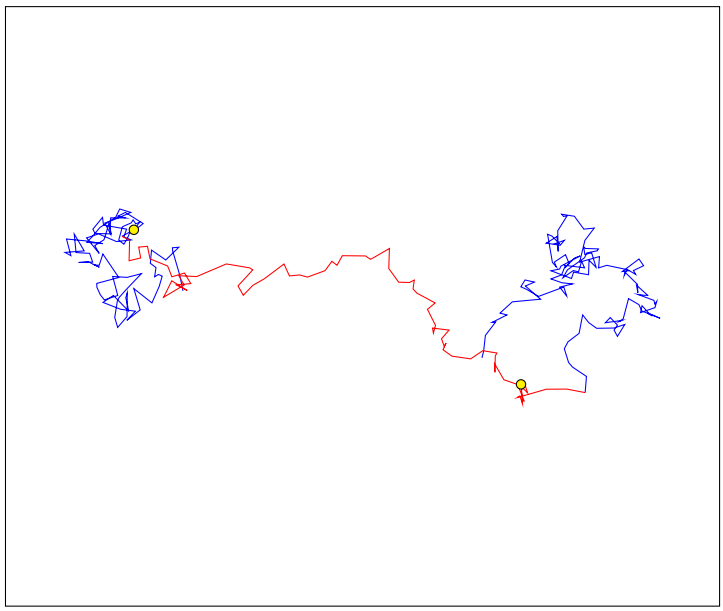
The time resolution of the sequence is $\Delta = 0.1s$. The space resolution is not given but when the Brownian motion with drift is chosen, Monnier et al. [2015] find a drift parameter with order of magnitude of $1\mu m.s^{-1}$. As before, we set the parameter $K = 2$ for the method of Monnier et al. [2015]. In this case, the model 3 (two Brownian motion with different diffusion coefficients) is selected by the method. Then, from our point of view, there are no change of dynamics. We note that we run 100 times the algorithm and did not get the same outcome each time. It is due to the fact that the inference is based on a Monte-Carlo Markov chains (MCMC) algorithm for computing the *a posteriori* estimates. Consequently, the selected model was not the same every times (92 times model 3, 7 times model 4, 1 time model 5). Then, the MCMC algorithm can show some problems of stability giving some contrary outcomes from one run to another.

In Figure 7.3, we show our results for two window sizes $k = 10$ and $k = 15$. We do not detect any change point for larger windows. With both window sizes, we detect approximately the same portion of the trajectory as superdiffusive. With the window size $k = 15$, we also detect a subdiffusive part in the trajectory.

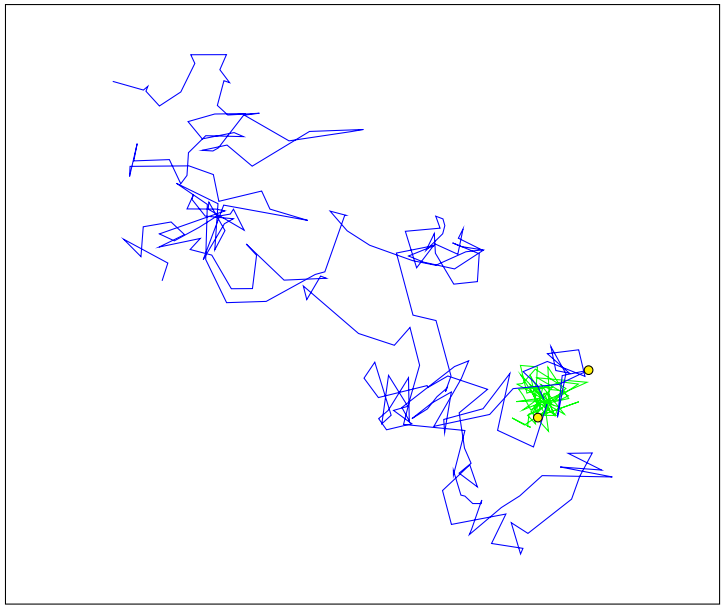
7.4 Summary

In this chapter, we evaluated our detection change point algorithm on two scenarios of simulation in the two-dimensional case. We used the Ornstein-Uhlenbeck process for modelling subdiffusion and the Brownian with drift for modelling superdiffusion. We also compared our method to i/ the method of Türkcan and Masson [2013] which detects switches between Brownian motion and subdiffusion, ii/ the method of Monnier et al. [2015] which detects switches between Brownian motion having different constant drifts. Our non parametric method outperformed the parametric methods of Türkcan and Masson [2013] and Monnier et al. [2015] on our simulation scenarios. We also considered real data depicting neuronal mRNPs (mRNAs in complex with mRNA-binding proteins). Other real data of interest can be considered. For instance, Dahan et al. [2003] reveal that the Glycine receptor –an inhibitory neurotransmitter receptor in the adult spinal cord– alternate between Brownian motion in the extra-synaptic domain and confined diffusion in the synaptic domain. This conclusion was made though the analysis of the MSD curve. Then, it will be of great interest to study these data with our method and compare the results of the two approaches.

So far we focused on the trajectories in themselves. In Part III, we consider the spatial distribution of particle dynamics inside a bound domain. More specifically, we are interested in detecting domains where the particles are attracted and thereby undergo subdiffusion. In biophysics, intracellular interactions, such as binding, take place in these particular domains.



(a)



(b)

Figure 7.1: Simulated trajectories. Figure (a), trajectory from Scenario 1 with $v = 0.8$. We detect $\hat{N} = 2$ change points $\hat{\tau}_1 = 99$ and $\hat{\tau}_2 = 169$ with a window of size 30. Figure (b) trajectory from Scenario 2 with $\lambda = 1$. We detect $\hat{N} = 2$ change points $\hat{\tau}_1 = 87$ and $\hat{\tau}_2 = 165$ with a window of size 30. The locations of the change points $X_{\hat{\tau}_1}$ and $X_{\hat{\tau}_2}$ are shown as yellow dots on the trajectories.

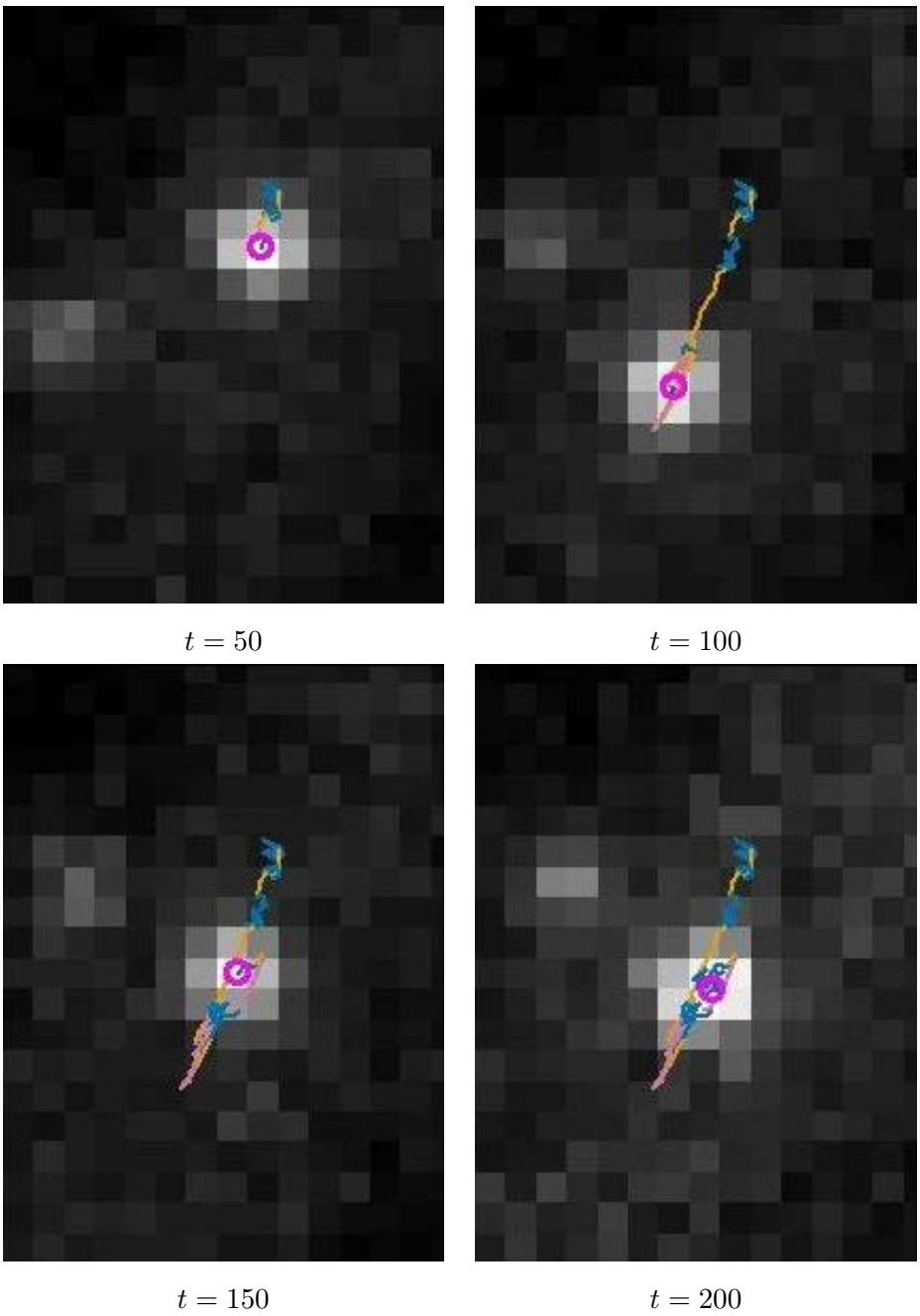
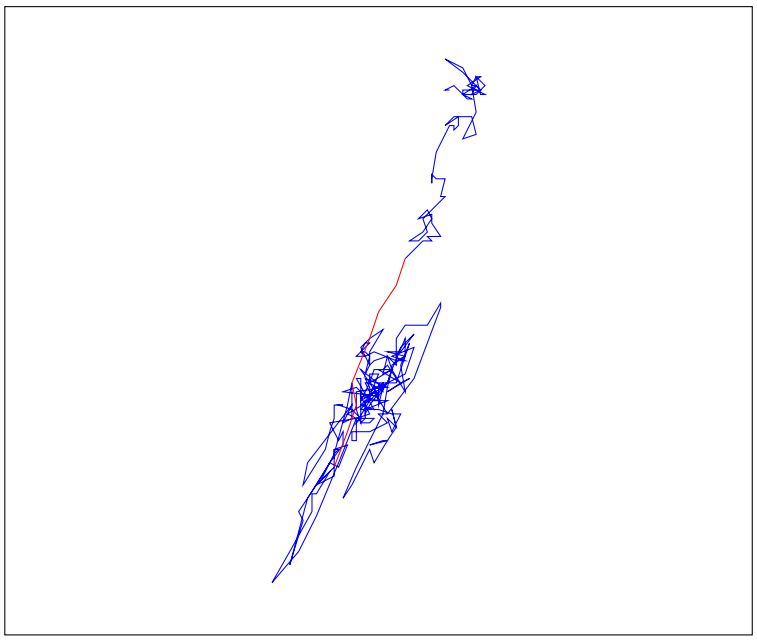
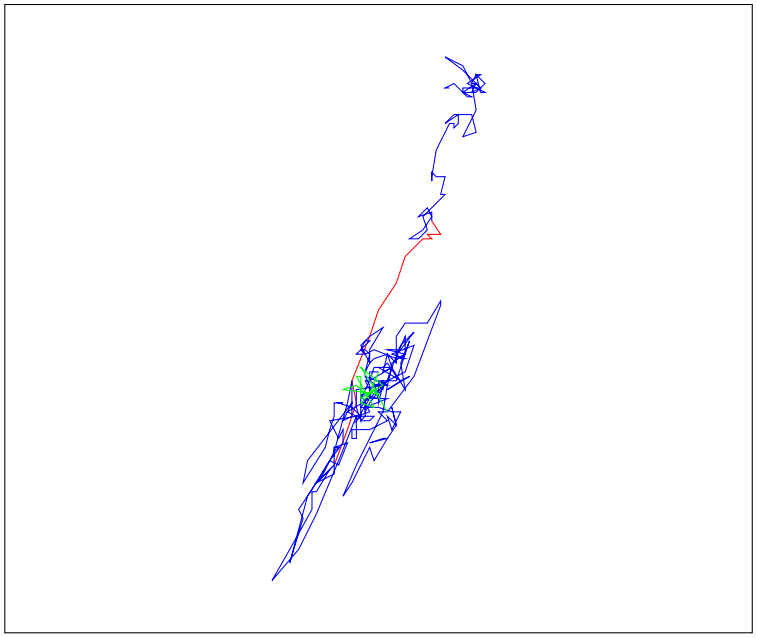


Figure 7.2: Observations at different times of the β-actin mRNP trajectories inferred by the hidden Markov model of Monnier et al. [2015]. Monnier et al. [2015] assume that there are $K \leq 3$ possible states (while we used $K \leq 2$ in our comparisons). They find three distinct motion states: one Brownian and two Brownian motions with two different drifts. The Brownian part is depicted in blue. The pink and orange part are associated with Brownian motion with drift (with different drift for each color). The purple circle marks the current position of the particle.



$k = 10$



$k = 15$

Figure 7.3: Change point detection on trajectories depicting neuronal mRNPs. The blue parts correspond to Brownian portions of the trajectory, red part to superdiffusive portions, green part to the subdiffusive portion. The detected change points are $\boldsymbol{\tau} = (67, 75)$ for $k = 10$ alternating between Brownian, superdiffusion and Brownian. The detected change points are $\boldsymbol{\tau} = (62, 75, 282)$ for $k = 15$ alternating between Brownian, superdiffusion, Brownian and subdiffusion.

Part III

Trajectory Clustering for Spatial Analysis of Dynamics

8 Simulation of Particles Trapped in Microdomains with FLUOSIM

In this chapter, we present the software FLUOSIM developed by M. Lagardere and O. Thoumine (IINS, University of Bordeaux 2). FLUOSIM was presented in the conference of MIFOBIO 2016 (see the conference website at http://gdr-miv.fr/mifobio2016/).This software simulates the dynamics of molecules in an environment with local microdomains where the particles can be trapped. We will use this model as a reference for studying the spatial distribution of motion in the cell. M. Lagardere provided us the software and explained us the simulation scheme. We give here the underlying mathematical framework associated to this simulation scheme (not studied by M. Lagardere formally). First, we explain how FLUOSIM describes the particle motion and the trapping process. Secondly, we derive differential equations that give the proportion of trapped and free particles. Finally, we design a simulation to assess our procedure for detecting the trapping areas (equivalently microdomains or confinement areas) presented in Chapter 9.

8.1 The FLUOSIM Model

FLUOSIM allows to study the population dynamics of intracellular particles. Therefore we observe a population of N independent trajectories, with N of order of magnitude of 10^3. Throughout this chapter, we denote $X_t^{(i)}$ the position of the i^{th} particle at time t. In some obvious cases, we will not specify the exponent i for the clarity of notations.

The particles undergo Brownian motion in confined geometries. Formally, we define the bounded region $\mathcal{D}$ where the particles evolve. Inside this region, one defines k subregions $\mathcal{S}_1, \ldots, \mathcal{S}_k$ in which particles can be trapped. In what follows, we denote $\mathcal{S} = \cup_j^k \mathcal{S}_j$. An example of configuration is shown in Figure 8.1 while a real example depicting AMPAR (postsynaptic AMPA-type glutamate receptor protein) is exhibited in Figure 8.2. The particles in $\bar{\mathcal{S}} = \mathcal{S}/\mathcal{D}$) undergo a Brownian motion with diffusion coefficient $\sigma_{\bar{s}}$. The particles are normally reflected at the boundaries of $\mathcal{D}$.

Inside a subregion $\mathcal{S}_j$, two types of motion can occur.

1. The first motion is confined Brownian motion normally reflected at the boundaries $\partial\mathcal{S}_j$ with diffusion coefficient $\sigma_t \leq \sigma_{\bar{s}}$. The particle is trapped in $\mathcal{S}_j$.

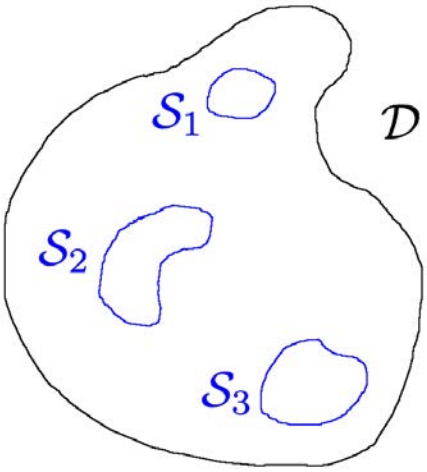


Figure 8.1: Example of configuration of trapping areas. Here, there are three trapping regions then $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$.

2. The second motion is Brownian motion with diffusion coefficient σ_s. The particle is not trapped in $\mathcal{S}_j$.

A particle in $\mathcal{S}_j$ can switch between two states: *trapped* or *non-trapped* (also denoted as *free*). We introduce the indicator variable:

$$\phi(X_t^{(i)}) = \begin{cases} 1 & \text{if particle } i \text{ is trapped at time } t, \\ 0 & \text{if particle } i \text{ is not trapped at time } t. \end{cases} \tag{8.1.1}$$

For a particle belonging to $\mathcal{S}$, we define the binding rate k^+ and the unbinding rate k^- as:

$$\begin{aligned} P(\phi(X_{t+h}) = 1 | \phi(X_t) = 0, X_{t+h} \in \mathcal{S}) = k^+ h + o(h), \\ P(\phi(X_{t+h}) = 0 | \phi(X_t) = 1, X_{t+h} \in \mathcal{S}) = k^- h + o(h), \end{aligned} \tag{8.1.2}$$

where $h > 0$ and $o(h)/h \to 0$ as $h \to 0$. We emphasize the fact that k^+ and k^- are two constants which do not depend on space. Then, a trapped particle close to the boundaries $\partial\mathcal{S}_j$ has the same probability to unbind to the trap region $\mathcal{S}_j$ than a trapped particle in the middle of the region $\mathcal{S}_j$.

In the settings of FLUOSIM, all the subregions $\mathcal{S}_j$ share the same parameters k^+, k^-, σ_t and σ_s. In our simulation scheme, we set $\sigma_{\bar{s}} = \sigma_s = \sigma_t$ to σ.

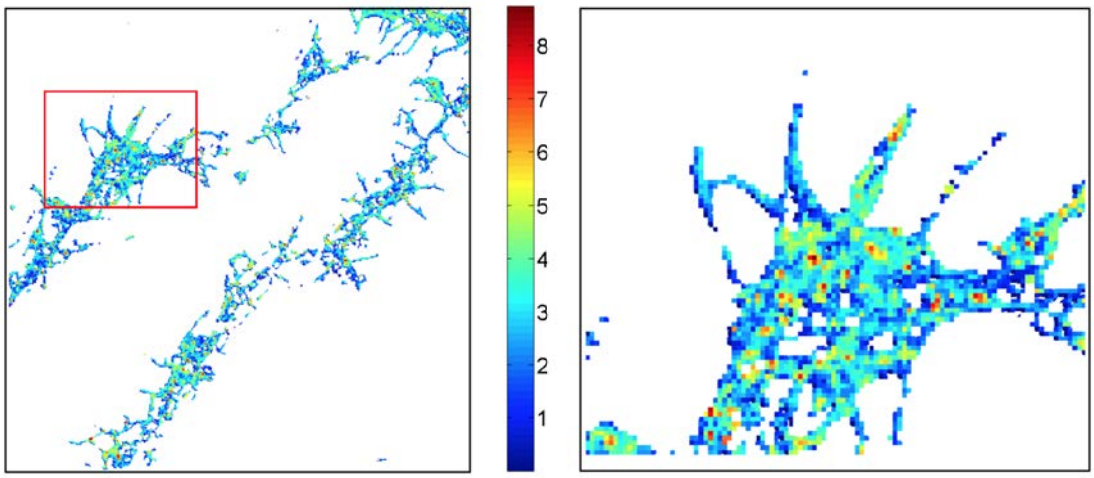


Figure 8.2: Density map of AMPAR (postsynaptic AMPA-type glutamate receptor) from [Hozé, 2013, Chapter 2]. AMPAR is a protein involved in the fast excitatory synaptic transmission. The hot spots correspond to trapping regions. [Hozé, 2013, Chapter 2] show that the trapping regions (called potential wells in their context) are located at the synapses.

8.2 Modelling the Proportions of Trapped Particles

In this situation, we model the evolution of the two populations of interest, the trapped (or bound) particles and the free particles in the domain $\mathcal{D}$. There are two causes that influence the population dynamics:

1. the dynamics of individual particles, here normally reflected Brownian motion in $\mathcal{D}$ and in $\mathcal{S}$ when trapped,

2. the trapping process.

We can see that the two processes are connected. The trapping process induces a different motion for the trapped particles; inversely the particles can be trapped only if their motion drive them inside $\mathcal{S}$. Now, we have the following proposition from Pinsky [2003]:

Proposition 4. *Let (X_t) be a normally reflected Brownian motion on a finite volume domain $\mathcal{D}$. The process (X_t) has a stationary distribution: the uniform distribution over $\mathcal{D}$ denoted $\mathcal{U}(\mathcal{D})$. Then, assuming the process (X_t) has reached its stationary distribution, we have for any $t > 0$ and $\mathcal{B} \subset \mathcal{D}$:*

$$P(X_t \in \mathcal{B}) = \frac{|\mathcal{B}|}{|\mathcal{D}|}, \tag{8.2.1}$$

where $|\mathcal{B}|$ denotes the area of domain $|\mathcal{B}|$.

Then, we make the assumption that initially all the particles follow the stationary distribution:

Assumption 3. *Initially, the particles are independently drawn from the uniform distribution over $\mathcal{D}$ that is:*

$$X_0^{(i)} \sim \mathcal{U}(\mathcal{D}), \quad i = 1, \ldots, N. \tag{8.2.2}$$

Instead of considering the exact dynamic of a particle, we simply model its probability to be in any trapping region $\mathcal{S}_j$ (equivalently to be in $\mathcal{S} = \cup_j^k \mathcal{S}_j$) by:

$$p_s \triangleq P(X_t^{(i)} \in \mathcal{S}) = \frac{|\mathcal{S}|}{|\mathcal{D}|} \quad i = 1, \ldots, N. \tag{8.2.3}$$

Equation (8.2.3) assumes that, at every time t, the spatial point process $(X_t^{(1)}, \ldots, X_t^{(N)})$ is a binomial point process over $\mathcal{D}$ of parameter N and which density function is the uniform density [Baddeley et al., 2007]. Then, when the number N of particles is large enough, the evolution of the two populations can be modelled by a system of differential equations depending on parameters p_s, k^+ and k^-. We denote $t \to b(t)$ the proportion of bound particles and $t \to f(t)$ the proportion of free particles. We have:

$$\begin{cases} \dfrac{db}{dt} & = \gamma_1 f(t) - \gamma_2 b(t), \\ b(t) + f(t) & = 1. \end{cases} \tag{8.2.4}$$

where γ_1 and γ_2 are respectively the global binding rate and global unbinding rate. By global, we mean that there are not defined given that the particle belongs to $\mathcal{S}$, as in the case of k^+ and k^-. There are defined as:

$$\begin{aligned} P(\phi(X_{t+h}) = 1 | \phi(X_t) = 0) &= \gamma_1 h + o(h), \\ P(\phi(X_{t+h}) = 0 | \phi(X_t) = 1) &= \gamma_2 h + o(h). \end{aligned} \tag{8.2.5}$$

Consequently, it defines $(\phi(X_t))$ as the continuous-time homogeneous Markov Chain with states $\{0, 1\}$ and infinitesimal generator parameters k^+ and k^- [Brémaud, 2013, Chapter 8, Section 2.2]. Finally, we can show that (see Appendix C):

$$\begin{aligned} \gamma_1 &= k^+ p_s, \\ \gamma_2 &= k^-. \end{aligned} \tag{8.2.6}$$

The solution of the system (8.2.4) in function of parameters p_s, k^+ and k^- is as follows:

$$\begin{cases} b(t) & = \left(b(0) - \dfrac{k^+ p_s}{k^+ p_s + k^-} \right) \exp(-(k^+ p_s + k^-)t) + \dfrac{k^+ p_s}{k^+ p_s + k^-}, \\ b(t) + f(t) & = 1. \end{cases} \tag{8.2.7}$$

From assumption 3, we have the constraint that $b(0) \leq p_s$, as only particles inside $\mathcal{S}$ can be trapped. The situation $b(0) = p_s$ matches with the situation where initially all the particles inside $\mathcal{S}$ are trapped. From Equations (8.2.7), the transitory regime is exponential and it converges to an equilibrium proportion:

$$\lim_{t\to\infty} b(t) = \frac{k^+ p_s}{k^+ p_s + k^-}. \tag{8.2.8}$$

If $b(0) < k^+p_s/(k^+p_s + k^-)$, function b is decreasing toward its equilibrium point $k^+p_s/(k^+p_s + k^-)$. If $b(0) > k^+p_s/(k^+p_s + k^-)$, function b is increasing toward its equilibrium point. If $b(0) = k^+p_s/(k^+p_s + k^-)$, function b is constant and equal to its equilibrium point. The characteristic time is defined as $\tau_c = 1/(k^+p_s + k^-)$. We can consider that the stationary regime is reached when $t > 5\tau_c$.

In the same way, we can model the proportions of trapped particles and free particles inside $\mathcal{S}$. We denote $t \to b_s(t)$ the proportion of trapped particles inside $\mathcal{S}$ and $t \to f_s(t)$ the proportion of free particle inside $\mathcal{S}$. We propose the following model:

$$\begin{cases} \dfrac{db_s}{dt} & = k^+ f_s(t) - k^- b_s(t) \\ b_s(t) + f_s(t) & = 1 \end{cases} \tag{8.2.9}$$

We can carry out the same study as previously. Again the solution is exponential. The equilibrium proportion of trapped particles inside any trapping regions $\mathcal{S}_j$ (same for all regions) is:

$$\lim_{t\to\infty} b_s(t) = \frac{k^+}{k^+ + k^-}. \tag{8.2.10}$$

8.3 Simulation with FLUOSIM

In this section, first we present a simulation scheme. Secondly, we evaluate the model (8.2.7) describing the evolution of the proportion of trapped and free particles in the whole domain $\mathcal{D}$ on the simulations. Finally, we assess the model (8.2.9) describing the evolution of the proportion of trapped and free particles in the set of trapping areas $\mathcal{S}$ on the simulations.

Simulation Scheme

We propose the following simulation settings. We design two trapping regions $\mathcal{S}_1$ and $\mathcal{S}_2$. We define the regions:

1. $\mathcal{D}$ is a square of radius 5 μm. We define the origin of the axis at the bottom left corner of $\mathcal{D}$.

Table 8.1: Parameters for the simulation with FLUOSIM.

Type	Parameters	Value
Biological	σ^2	$1\ \mu m^2.s^{-1}$
	k^+	$0.2\ s^{-1}$
	k^-	$0.05\ s^{-1}$
Microscopic	Δ	$0.1\ s$
	Δx	$0.025\ \mu m$

2. $\mathcal{S}_1$ is a circle of radius $r_1 = 0.65\ \mu m$ and center $\theta_1 = (2.5, 2.5)$.

3. $\mathcal{S}_2$ is a circle of radius $r_2 = 0.39\ \mu m$ and center $\theta_2 = (4, 4)$.

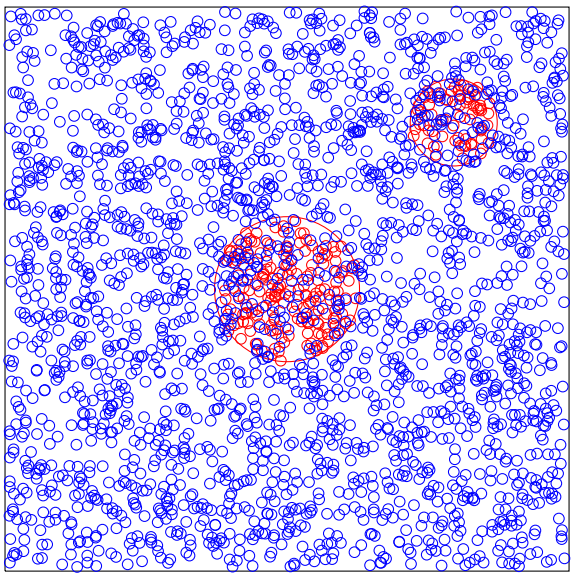
The parameters of simulation are given in Table 8.1. Initially, 2 500 particles are uniformly distributed over $\mathcal{D}$, fulfilling Assumption 3. FLUOSIM allows us to identify which particle is trapped at time t: we know $\phi(X_t^{(i)})$ that is if the particle i is trapped or not at time t. In Figure 8.3, we plot the positions of the particles labelled as free or trapped at time $t = 10\ s$ (transitory regime) and $t = 100\ s$ (stationary regime).

Evaluation of Model (8.2.7)

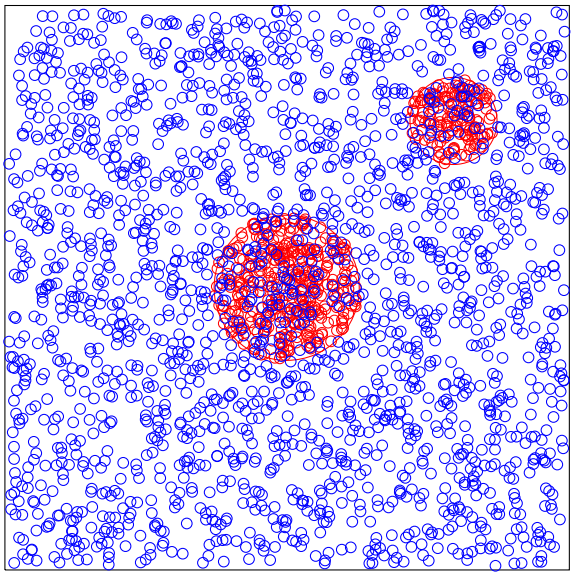
We compare the proportion of trapped particles $t \to b(t)$ given by the model (8.2.7) to the true proportion of trapped particles. The curves are given in Figure (8.4) (a). Visually, the fit of the transitory regime ($t < 5\tau_c = 77.5\ s$) is rather good . The stationary regime is not exactly the same as the one predicted by the model. The predicted equilibrium is 22.53% while the mean true proportion of trapped particles computed on the last 100 steps of time is 21.35%. The relative error of the model is 5.24%. This error is due to the fact that the model oversimplifies the underlying trapping process: the dynamic of the particles is only modelled by the parameter p_s. Moreover, by using the parameter p_s, we assume that particles are always in the stationary regime of a normally reflected Brownian motion on $\mathcal{D}$. This assumption holds at the beginning, as the particles are initially drawn from the uniform distribution on $\mathcal{D}$. As t increases, the trapping process makes this assumption fail. It explains the relative good fit of the transitory phase and the relative lack of fit of the asymptotic phase. However, we can consider that it is a rather good model considering its simplicity and parsimony in terms of the parameters.

Evaluation of Model (8.2.9)

We compare the proportion of trapped particles $t \to b_s(t)$ given by the model (8.2.9) to the true proportion of trapped particles inside $\mathcal{S}$. Interestingly, we observe the



$t = 10\ s$



$t = 100\ s$

Figure 8.3: Positions of the particles simulated with FLUOSIM at time $t = 10\ s$ and $t = 100\ s$. In blue free particles, in red trapped particles. The red circles defined the regions $\mathcal{S}_1$ (large disk) and $\mathcal{S}_2$ (small disk).

opposite situation compared to the case of $b(t)$ (proportion of trapped particles in the whole domain $\mathcal{D}$): there is a clear lack of fit during the transitory regime but a perfect fit during the stationary regime, see Figure (8.4) (b). We can see that the transitory regime has the same duration for $b(t)$ and $b_s(t)$. During this phase, the number of particles inside $\mathcal{S}$ increases due to the process of Brownian particles entering in $\mathcal{S}$ and getting trapped. However, in the model (8.2.9), it is assumed that the number of particles is constant (and large). Then, it explains the lack of fit during the transitory phase which matches with a period during which the number of particles increases in $\mathcal{S}$. On the contrary, once $b(t)$ has reached the stationary regime, the number of particles inside $\mathcal{S}$ is approximately constant (even if it is not the same particles that remain in $\mathcal{S}$ from one time to another). In the latter case, the model (8.2.7) is relevant and the stationary regime is well predicted by the model.

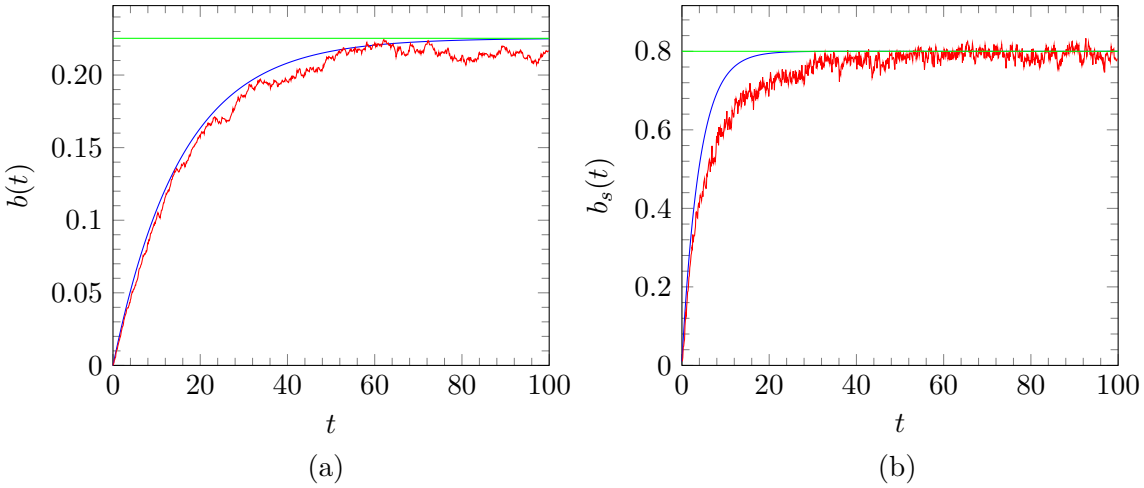


Figure 8.4: Evolution of the proportions of trapped particles over time. In Figure (a), proportions of trapped particles in the whole domain $\mathcal{D}$, in Figure (b) proportions of trapped particles in the trap region $\mathcal{S}$. In red the true proportion of trapped particles computed from the data, in blue the proportions computed with model (8.2.7), in green the asymptote of model (8.2.7). We computed the red curve over $N = 2\,500$ trajectories simulated with FLUOSIM with parameters given in Table 8.1.

8.4 Summary

In this chapter, we presented the underlying model of the simulator FLUOSIM. The formalization in a mathematical framework of the software allowed us to link the diffusion models used throughout this thesis to the trajectory dynamics generated by FLUOSIM.

We emphasize that M. Lagardere and O. Thoumine only gave us the software without any additional document. The particles generated in FLUOSIM can switch between the *trapped* and the *free* states. When there are trapped, they undergo Brownian motion in confined regions. When there are free, they are driven by Brownian motion only constrained by the boundary conditions on the limits of the domain. We modelled the proportions of trapped and free particles in the whole domain and in the trapping regions through two systems of differential equations. Then, we were able to predict the proportion of particles inside the trapping regions which where in the the *trapped* state.

The model of FLUOSIM defines local domains where particles are confined for a while. In the next chapter, we are interested in estimating such domains where particles undergo subdiffusion. We present a method to detect such areas. We evaluate the performances of the method based on clustering on the simulation described in Section 8.3 obtained with FLUOSIM. Our understanding of the generative process of FLUOSIM allows us to have better insights on the simulation results of our clustering approach.

9 A Method for Detecting Trapping Areas

In this chapter, we aim at detecting trapping areas (equivalently microdomains or confinement areas), that is regions where the particles are trapped and thereby undergo subdiffusion. In our context, it is expected that these areas contain a high concentration of particles detected as subdiffusive. Then, we use a clustering algorithm DBSCAN [Ester et al., 1996] coupled with our test procedure such areas. We evaluate the method on the simulation described in Section 8.3 obtained with Fluosim.

9.1 Model

We use a similar model as in Section 4.8. We observe a collection $\mathcal{X}_m$ of m d-dimensional trajectories which are simultaneously observed. We denote by $\mathbb{X}_{n_k}^{(k)}$ the observations associated to the k^{th} particle:

$$\begin{aligned} \mathbb{X}_{n_k}^{(k)} &= \left(X_{t_0}^{(k)}, \ldots, X_{t_{n_k}}^{(k)}\right), \qquad k = 1, \ldots, m, \\ \mathcal{X}_m &= \left\{\mathbb{X}_{n_k}^{(k)},\ k = 1, \ldots, m\right\}. \end{aligned} \tag{9.1.1}$$

We assume that each discrete trajectory is generated by a stochastic process $(X_t^{(k)})$ with continuous path defined on the spatial domain $\mathcal{D} \subset \mathbb{R}^d$ and which is a solution of the stochastic differential equation (SDE),

$$dX_t^{(k)} = \mu(X_t^{(k)})dt + \sigma^{(k)} dB_t^{(k),\mathfrak{h}^{(k)}}, \qquad t \in [t_0, t_{n_k}], \tag{9.1.2}$$

where $B_t^{(k),\mathfrak{h}}$ is a d-dimensional fractional Brownian motion of unknown Hurst parameter $\mathfrak{h}^{(k)}$. The unknown parameters of the model are the Hurst parameter $\mathfrak{h}^{(k)} \in (0,1)$, the diffusion coefficient $\sigma^{(k)} > 0$ and the drift term $\mu^{(k)} : \mathcal{D} \to \mathbb{R}^d$. We can also assume that the SDE (9.1.2) is constrained by some boundary conditions.

Even if our method can be used in two or three dimensions, in the rest of the chapter, we restrict our model to the two-dimensional case for sake of simplicity. Then, we use the simulation scheme of Section 8.3 to illustrate our method. More specifically, we assume that we observe 2 500 trajectories of size $n = 30$ simulated with the simulation scheme described in Section 8.3. The trajectories are observed once the equilibrium regime is reached (burning period of 100 s). We recall that in the simulation scheme 8.3 the trajectories can switch between two diffusion models:

1. confined Brownian in small microdomains (denoted $\mathcal{S}_j$) with normal reflection on the boundaries $\partial\mathcal{S}_j$,

2. Brownian motion in the whole domain $\mathcal{D}$ with normal reflection on the boundaries $\partial\mathcal{D}$.

These two models are included in Equation (9.1.2), adding the right boundary conditions. However, our model does not specify any switching between different diffusions. In fact, for trajectories of size $n = 30$, we can assume that no switching will occur due to the choice of the switching parameters k^+ and k^- given in Table 8.1. Finally, the simulation scheme of Section 8.3 will be refer to as the FLUOSIM simulation through this chapter.

9.2 Outline of the Procedure

We can use our test procedures to detect the confinement areas with the following four-step procedure.

1. We run a test procedure, our single test procedure 4.5.1 or our multiple test Procedure 1 on the collection of trajectories $\mathcal{X}_m$. As in Part 4, we denote $\mathcal{R}_1(\mathcal{X}_m)$ the set of trajectory indexes corresponding to the acceptance of hypothesis H_1 that is the trajectory is subdiffusive.

2. We choose a single point $\tilde{x}_i$ to represent each trajectory $\mathbb{X}_{n_i}^{(i)}$.

3. We partition the set $\mathcal{R} = \{\tilde{x}_i | i \in \mathcal{R}_1(\mathcal{X}_m)\}$ into clusters.

4. We use these clusters to define confinement areas.

The same scheme can be used to detect areas where superdiffusion or Brownian motion are the pilot dynamics. In step 3, we just have to replace $\mathcal{R}_1(\mathcal{X}_m)$ by respectively $\mathcal{R}_2(\mathcal{X}_m)$ (set of trajectory indexes corresponding to superdiffusion) or $\mathcal{R}_0(\mathcal{X}_m)$ (set of trajectory indexes corresponding to Brownian motion). In the case of superdiffusion, we could use the aforementioned method to detect potential actin filaments associated to active transport.

In the next sections, we will detail successively the steps 2, 3 and 4 of the method. Step 1 is straighforward as it is the procedure explained in Part 4. However, we can compare the outcomes of the method obtained with different test procedures. As already mentioned, we will use the simulation presented in Section 8.3 to illustrate the different steps of the methods.

9.3 Representative Points of Trajectories and Spatial Distribution of Detections

First, we propose different ways for representing trajectories. Then, we use these representations to study the spatial distribution of the particles detected as subdiffusive with our test procedures.

Representative Points

We need to define a single point to represent our trajectory. Then we define a function f as:

$$\begin{aligned} \mathbb{R}^{2\times n} &\to \mathbb{R}^2 \\ (x_1, \ldots, x_n) &\mapsto f(x_1, \ldots, x_n). \end{aligned}$$

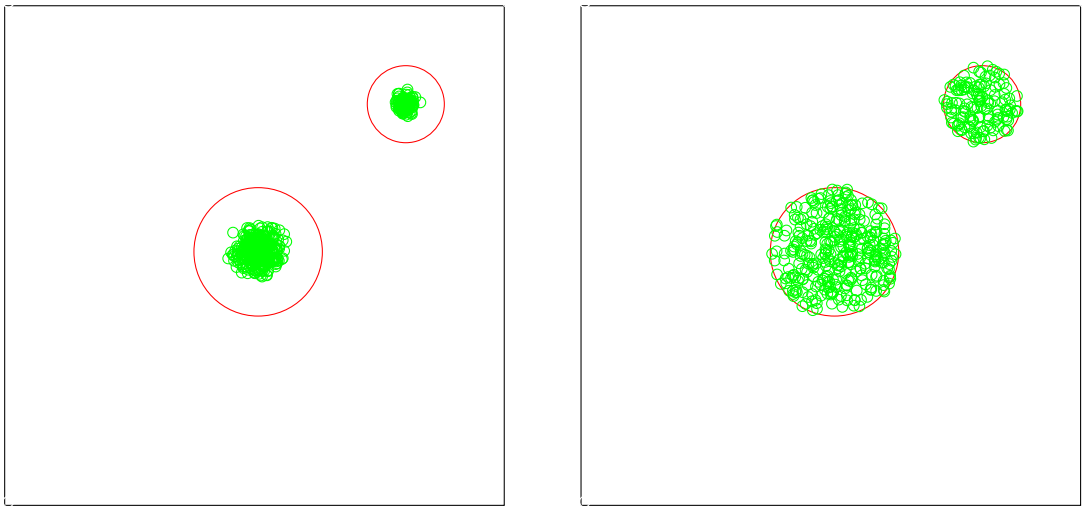
Here are some examples of f:

- $f(x_1, \ldots, x_n) = (1/n) \sum_{i=1}^{n} x_i$, the representative point is the mean point,
- $f(x_1, \ldots, x_n) = x(k)$, with $k \in \{1; \ldots, n\}$ the representative point is the k^{th} point of the trajectory.

In Figure 9.1, we show the trapped particles of the FLUOSIM simulation. As we already mentioned, the trapped particles are modelled by confined Brownian motion normally reflected at the boundaries of the trapping regions. In that case, we can see that different representative points of the trajectories have very different spatial distributions. From Figure 9.1, we choose to represent the trajectory $\mathbb{X}_n = (X_{t_0}, \ldots, X_{t_n})$ by the point $X_{\lfloor n/2 \rfloor}$ (related function $f(x_1, \ldots, x_n) = x(\lfloor n/2 \rfloor)$). In fact, we can see that the spatial distribution of the representative points $X^{(i)}_{\lfloor n/2 \rfloor}$ of the trapped trajectories $\mathbb{X}^{(i)}_n$ is uniform over the trapped region $\mathcal{S}_1$ and $\mathcal{S}_2$ (Figure 9.1 (b)). This is due to Proposition 4, as the trapped particles undergo a normally reflected confined Brownian motion in $\mathcal{S}_1$ and $\mathcal{S}_2$. On the other hand, the average position of the trapped trajectories are concentrated in the middle of the trapped regions (Figure 9.1 (a)). Therefore in the purpose of estimating $\mathcal{S}_1$ and $\mathcal{S}_2$, we should represent the trajectory i with $X^{(i)}_{\lfloor n/2 \rfloor}$ rather than by its average position.

Spatial Distribution of the Detections

We study the spatial distribution of the representative points of the particles detected as subdiffusive with our test procedures. Spatial statistics and point processes have been successfully investigated in image analysis for several decades (see [Mumford and Desolneux, 2010] and [Descombes, 2013] for a recent review and analysis). In what follows, we



(a) (b)

Figure 9.1: Spatial distribution of the trapped particles in the Fluosim simulation. Left the trajectories are represented by their mean point, right they are represented by $X_{n/2}$. A trajectory is considered trapped if it is trapped during the whole period of observation. Particle i is trapped if $\phi(X_{t_j}^{(i)}) = 1$, $j = 1, \ldots, n$ using definition of Equation (8.1.1). Red circles represent the boundaries of the trapping regions $\mathcal{S}_1$ (big circle) and $\mathcal{S}_2$ (small circle). The black square represents the boundaries of the whole domain $\mathcal{S}$.

do not consider generative models to represent the spatial distributions of trajectories represented by points. Instead, we propose to estimate clusters corresponding to aggregates of trajectories, with no prior information. In that sense, our approach is in the spirit of conventional clustering approaches, and is somehow related to the *a-contrario* modelling [Desolneux et al., 2003b, Cao et al., 2007, Desolneux et al., 2003a].

We define the set:

$$\mathcal{R} = \{X_{\lfloor n/2 \rfloor}^{(i)} | i \in \mathcal{R}_1(\mathcal{X}_m)\}, \tag{9.3.1}$$

where $\mathcal{R}_1(\mathcal{X}_m)$ is the set of indexes of the trajectories detected as subdiffusive by a test procedure. The scatter plot of $\mathcal{R}$ is presented in Figure 9.2. We test the trajectories of the FLUOSIM simulation with the single test procedure (Figure 9.2 (a)) and the adaptive Procedure 1 (Figure 9.2 (b)). Results of the two procedures in terms of numbers of true and false detections are also presented in Table 9.1. As expected, we detect more subdiffusive trajectories with the single test procedure (508 subdiffusive trajectories) than with the adaptive Procedure 1 (209 subdiffusive trajectories). We can see that the false detections of the single test procedure at level 5% are uniformly distributed over

the domain $\mathcal{D}/\mathcal{S}$ (points outside the red circles delimiting $\mathcal{S}_1$ and $\mathcal{S}_2$ in Figure 9.2 (a)). From Proposition 4 and Assumption 3 for a particle i not trapped in $\mathcal{S}_1$ or $\mathcal{S}_2$ we have:

$$X^{(i)}_{\lfloor n/2 \rfloor} \sim \mathcal{U}(\mathcal{D}/\mathcal{S}). \tag{9.3.2}$$

Secondly we can consider that the set:

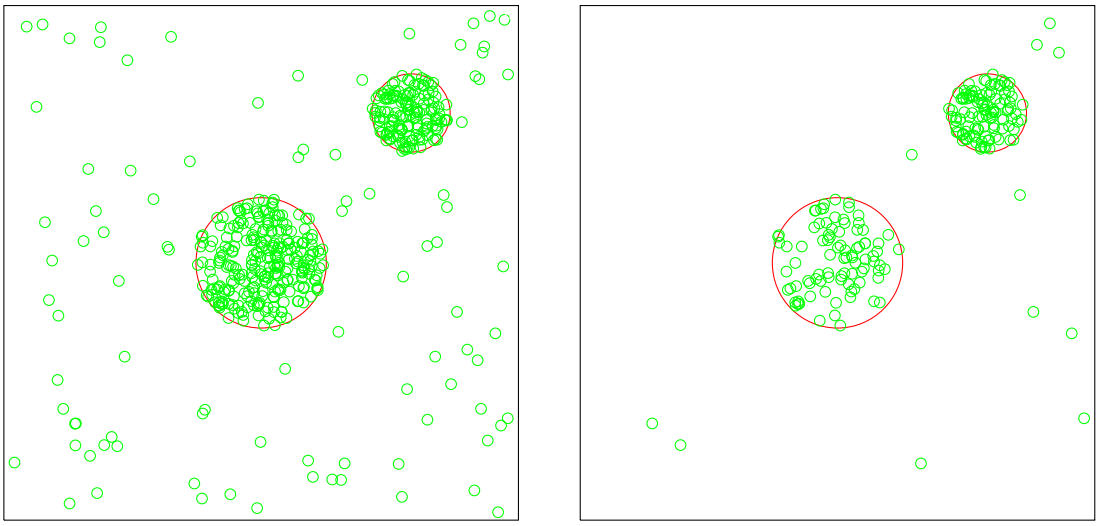
$$\{X^{(i)}_{\lfloor n/2 \rfloor} | i \in \mathcal{R}_1(\mathcal{X}_m) \text{ and trajectory } i \text{ not trapped}\}, \tag{9.3.3}$$

is a random subsample from the sample (9.3.2) of uniformly distributed points. Then the subsample (9.3.3) is also generated by a uniform distribution over $\mathcal{D}/\mathcal{S}$. Then, it explains why the false detections are uniformly distributed over $\mathcal{D}/\mathcal{S}$ in Figure 9.2 (a). We note that Equation (9.3.2) does not exactly hold because of the trapping regions. However, we can see from the simulation that it is a good approximation. When we use the adaptive Procedure 1, there are less false detections (few points outside the red circles in Figure 9.2 (b)). On the other hand, the adaptive Procedure 1 does not detect well the subdiffusive trajectories inside the biggest trapping region $\mathcal{S}_1$: there are significantly fewer points detected inside this region than with the single test procedure. This is due to the fact that the adaptive Procedure 1 is less powerful than the single test procedure. Now, it is very intuitive that a Brownian particle confined in a small area will be easier to detect as subdiffusive than if it is trapped in a large area. In the latter case, we will need to observe it over a much longer time to figure out that it is effectively confined in a domain. That is why the adaptive Procedure 1 detect better the particles trapped in $\mathcal{S}_2$, the smallest trapping region, than those trapped in $\mathcal{S}_1$. The single test procedure is powerful enough to detect well the trapped particles in the two regions.

From Figure 9.2, we can see that the trapping regions $\mathcal{S}_1$ and $\mathcal{S}_2$ correspond to high concentrations of particles detected as subdiffusive. However, because of the false detections, we detect subdiffusive particles outside the trapping regions. Then, we propose to use the clustering algorithm DBSCAN developed by Ester et al. [1996] to detect the trapping regions. It can detect the clusters corresponding to real trapping regions from the noisy points due to false detections. There exist alternative clustering algorithms able to detect clusters from noisy points [Desolneux et al., 2003b, Cao et al., 2007, Desolneux et al., 2003a]. Cao et al. [2007] propose an approach based on *a-contrario* models. The authors define noise as a background process characterised by its distribution π. This method allows a finer analysis of the clusters than DBSCAN through the concept of meaningful clusters [Desolneux et al., 2003b, Cao et al., 2007]. Then, the authors can build a hierarchy of clusters. We will focus only on DBSCAN for simplicity.

9.4 A Clustering Algorithm: DBSCAN

The algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) proposed by Ester et al. [1996] is a non parametric clustering algorithm. The clusters



(a) (b)

Figure 9.2: Spatial distribution of particles detected as subdiffusive with the single test procedure at 5% (left), with the adaptive Procedure 1 (right). Red circles represent the boundaries of the trapping regions $\mathcal{S}_1$ (big circle) and $\mathcal{S}_2$ (small circle). The black square represents the boundaries of the whole domain $\mathcal{D}$.

Table 9.1: Numbers of true and false detections in the FLUOSIM simulation.

Method	True H_1	False H_1	Total
Single test	390	118	508
adaptive Proc. 1	188	21	209

are defined through a notion of point concentration or point density. In this section, $\mathcal{R}$ denotes the set of detected points (9.3.1) as previously, but more generally it denotes any set of points x_i from which we want to detect clusters. In the same way, $\mathcal{D}$ denotes the whole domain of the FLUOSIM simulation but more generally denotes any space containing the set of points from which we want to detect clusters.

Algorithm

First, Ester et al. [1996] define the neighbourhood of $x \in \mathcal{R} \subset \mathcal{D}$ as:

$$N_\epsilon(x) = \{y | y \in \mathcal{R}/x, \quad d(x,y) \leq \epsilon\}, \tag{9.4.1}$$

where $\epsilon > 0$ is a parameter of DBSCAN to fix and d is a norm, for instance the Euclidean norm.

Secondly, the authors define a core point x as a point fulfilling the condition:

$$\#N_\epsilon(x) \geq n^\star, \tag{9.4.2}$$

where $n^\star$ is a parameter to fix and $\#A$ is the cardinal of the set A.

Then, DBSCAN scans all the points of $\mathcal{R}$ as shown in Algorithm 3:

Algorithm 3: Single scan approach of the algorithm DBSCAN.

```
Input: ϵ, n*, R
Result: a partition of R = ∪_{i=1}^{m} C_i
j=1;
for i=1 to #R do
    if x_i is unclassified And #N_ϵ(x) ≥ n* then
        Create a new cluster C_j;
        Expand the cluster from x_i with a rule expandCluster(x_i,R,ϵ,n*);
    end
    j=j+1;
end
```

A cluster can be created only from a core point. Now, we explain how DBSCAN expands a cluster from a core point. We need first to present the concepts of boarder points and noisy points introduced by Ester et al. [1996]. A boarder point x is a point such that:

$$\begin{cases} \exists y \in \mathcal{R}, \#\mathcal{N}_\epsilon(y) \geq n^\star \text{ and } x \in \mathcal{N}_\epsilon(y), \\ \#N_\epsilon(x) < n^\star. \end{cases} \tag{9.4.3}$$

In other words, a boarder point is a point belonging to the neighbourhood of a core point but which is not a core point itself. A noisy point is a point which is not in the neighbourhood of any core points and which is not a core point itself.

From a core point x, DBSCAN expands a cluster connecting all the core points and boarder points until the maximal cluster is built. More specifically Ester et al. [1996] build a cluster C from a core point x as the set of points $C = \{y | y \text{ is density reachable from } x\}$. A point y is said to be density reachable from x if there exists a sequence $x = x_1, x_2, \ldots, x_{p-1}, x_p = y \in \mathcal{R}$ such that:

$$\begin{cases} x_{i+1} \in \mathcal{N}_\epsilon(x_i), \\ \#N_\epsilon(x_i) \geq n^\star. \end{cases} \tag{9.4.4}$$

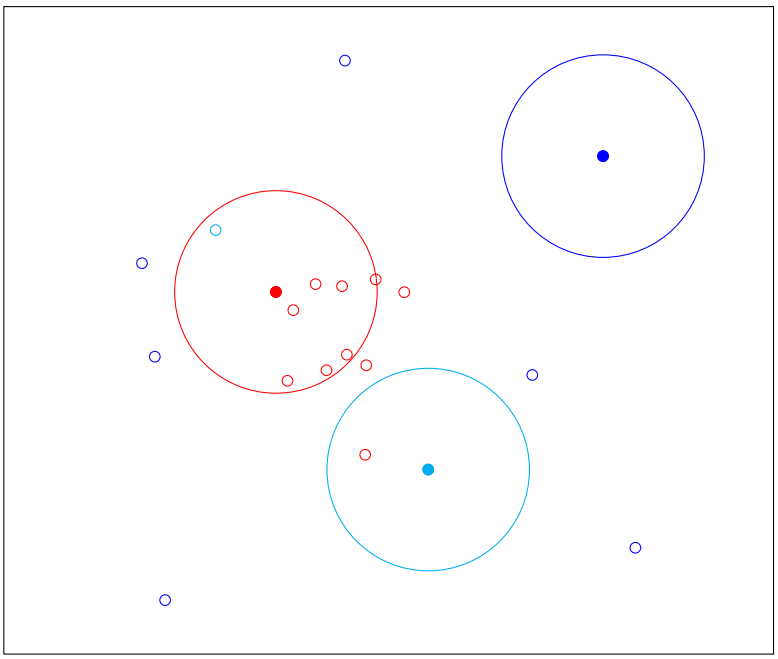


Figure 9.3: Classification of spatial points with DBSCAN with parameters $\epsilon = 0.15$ and $n^\star$. Red points are the core points, cyan points are the boarder points and blue points are the noise points with respect to $(\epsilon, n^\star)$. We plot a circle of radius ϵ centered on an element of each type of points. The centers are the points with filled circles of the color corresponding to the point type. We can see that the red circle contains more than $n^\star$ point (center excluded) then its center is a core point ; the cyan circle contains less than $n^\star$ point (center excluded) but contains a core point then its center is a boarder point; the blue circle contains less than $n^\star$ point (center excluded) and does not contain any core point then it is a noisy point. The points were simulated by a mixture of the uniform distribution on the square $[0, 1]^2$ and a bivariate normal distribution of mean $[0.5, 0.5]$ and covariance $0.05 * \mathbf{I}_2$. Out of the 20 points, 10 points were simulated by the uniform and 10 by the normal distribution.

We illustrate the DBSCAN algorithm on a very simple simulation in Figure 9.3. We highlight the definition of the different types of points (core, boarder and noisy points) of the method.

Selection of the DBSCAN Parameters

Now we must choose the parameters ϵ and $n^\star$ of the method. First Ester et al. [1996] note that, ideally, we should have a pair of parameters $(\epsilon, n^\star)$ adapted to each cluster, as each cluster does not have the same density (or concentration) of points. Obviously this information is available only once the clusters are found. Then, we have to use a single pair of parameters $(\epsilon, n^\star)$ for all the clusters. A couple $(\epsilon, n^\star)$ able to detect the

least dense cluster is a good choice as such parameters will also be able to detect denser clusters. Ester et al. [1996] argue that, in two-dimensional problems, we can set $n^\star = 4$ as a rule of thumb. Once the parameter $n^\star$ is fixed, we can choose ϵ using two different methods:

1. a data driven method based on the observed distribution of the $n^\star$ nearest neighbours,

2. a parametric method which models the distribution of the noisy points.

In the following, we present two data driven methods and two parametric method for selecting the parameter ϵ. Then, we compare the outcome of the DBSCAN algorithm according to the different choices of ϵ on the FLUOSIM simulation.

Data driven methods The data driven methods define ϵ from the sample $d_1(n^\star), \dots, d_m(n^\star)$ with $d_i(n^\star)$ denoting the distance of the $n^\star$ nearest neighbours of point $x_i \in \mathcal{R}$. Parameter ϵ is chosen from the sample $d_1(n^\star), \dots, d_m(n^\star)$. Denote $d_{(1)}(n^\star), \dots, d_{(m)}(n^\star)$ the increasing-ordered sample. Note that if we choose $\epsilon = d_{(i)}(n^\star)$, the i points corresponding to the values $d_{(1)}(n^\star), \dots, d_{(i)}(n^\star)$ will be core points while the other points will be either boarder or noisy points. The choice $\epsilon < d_1(n^\star)$ corresponds to the case where all the points are noisy points (no core points); the choice $\epsilon \geq d_m(n^\star)$ correspond to the case where all the points are core points. In the latter case, they all belong to one single cluster. Then, if we have an estimation of the number of noisy points $\hat{q}$, a natural choice proposed by Ester et al. [1996] is $\epsilon = d_{(m-\hat{q}+1)}(n^\star)$. Otherwise Ester et al. [1996] rely on a graphical approach to determine ϵ. The authors plot the sequence $u_1 = d_{(m)}(n^\star), \dots, u_m = d_{(1)}(n^\star)$ and ϵ is defined as the first u_i in the first valley of the sequence.

Alternatively, we propose to use the algorithm of Otsu [1979] on the sample $(d_1(n^\star), \dots, d_m(n^\star))$ to find the optimal ϵ. We briefly explain the method. Let $y_1, \dots, y_m$ a sequence of scalars and $y_{(1)}, \dots, y_{(m)}$ the sorted sequence. Otsu [1979] finds k such that the variance between the sets $A_1 = \{y_{(1)}, \dots, y_{(k)}\}$ and $A_2 = \{y_{(k+1)}, \dots, y_{(m)}\}$ is maximal. Note that the algorithm of Otsu [1979] is widely used for image segmentation. In the clusters, the distance to the $n^\star$ nearest neighbours are significantly lower than in the noisy points. Then, we expect the method of Otsu [1979] to distinguish between the two distributions and find a good candidate ϵ.

Parametric methods Daszykowski et al. [2001] assume implicitly that the noisy points are uniformly distributed and derive ϵ accordingly. They define ϵ as the α quantile of the distribution of the $n^\star$ nearest neighbours of m points drawn from a uniform distribution over $\mathcal{D}$. Daszykowski et al. [2001] choose $\alpha = 0.05$. If the observed noisy points are really drawn from a uniform distribution, none of them will be chosen as a core point with

approximately $1-\alpha$ probability. It is not an exact $1-\alpha$ probability for the following reasons.

1. There are $q \leq m$ noisy points and not m noisy points as in the method of Daszykowski et al. [2001].

2. The noisy points are uniformly distributed over $\mathcal{D}/\mathcal{S}(\cup C_i) \subset \mathcal{D}$ where $\mathcal{S}(\cup C_i)$ is the space delimited by the clusters C_i; then there are not uniformly distributed over the whole domain $\mathcal{D}$ as in the method of Daszykowski et al. [2001].

As q and the clusters C_i are unknown, we can not propose a method which exactly controls the probability to detect a noisy point as a core point. We note that Daszykowski et al. [2001] estimate ϵ with Monte-Carlo simulations. It can be time consuming if m is large.

Friedman et al. [1975] also study the distribution of the $n^\star$ nearest neighbours of m points uniformly distributed over a finite space $\mathcal{D}$. They assume that m is large enough to neglect boundary effects. Then, they state that the ratio of the volume of a d dimensional sphere centered at a point containing $n^\star$ neighbours and the volume of the whole space $\mathcal{D}$ is governed by a beta distribution f of parameters $(n^\star, m-n^\star)$:

$$f(x) = \frac{m!}{(n^\star-1)!(m-n^\star)!} x^{n^\star-1} x^{m-n^\star}, \quad 0 \leq x \leq 1. \tag{9.4.5}$$

Now, we define ϵ as Daszykowski et al. [2001]. Then, if we note $F^{-1}(\alpha)$ the quantile of order α of the distribution f, we get:

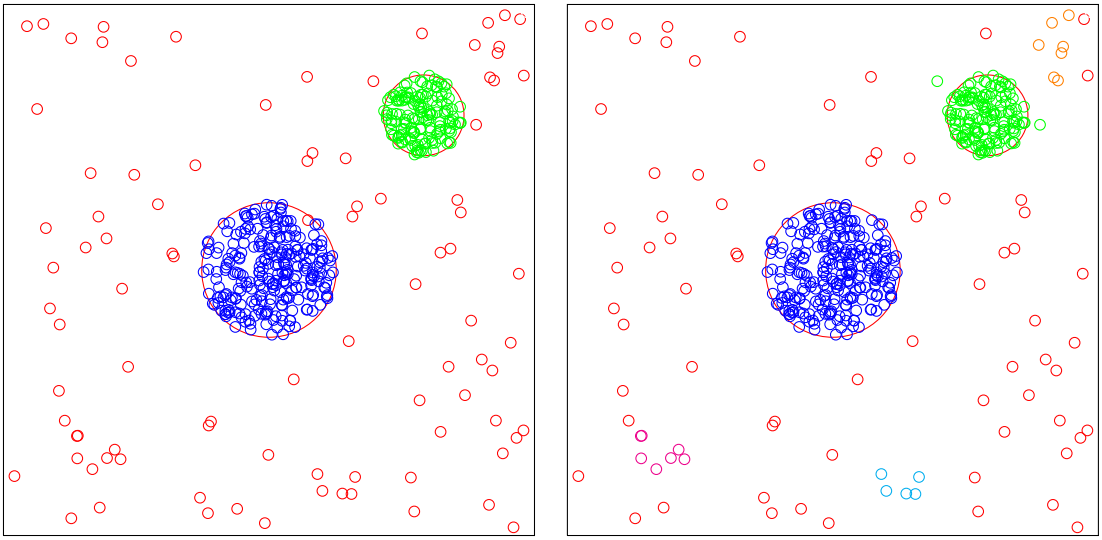
$$\epsilon = \left(\frac{|\mathcal{D}| F^{-1}(\alpha)}{\pi} \right)^{1/d}, \tag{9.4.6}$$

where $|\mathcal{D}|$ is the volume (or area in the two-dimensional case) of $\mathcal{D}$. In our case, we have $d = 2$ (3 if we work in three dimensions).

Comparison on simulations We apply the DBSCAN algorithm on the set $\mathcal{R}$ (9.3.1) obtained with the single test procedure (Figure 9.2 (a)). The parameter ϵ is estimated with the different methods aforementioned (see Table 9.2). We run DBSCAN with the different values of ϵ of Table 9.2; we always keep $n^\star = 4$. Results are shown on Figure 9.4. All the methods for estimating ϵ -except the one of Otsu [1979]- give similar estimations of ϵ (see Table 9.2). Consequently, these close values of ϵ give the exact same clusters when we run DBSCAN: we find two clusters corresponding to the two trapping region $\mathcal{S}_1$ and $\mathcal{S}_2$. When ϵ is given by the method of Otsu [1979], DBSCAN detects 5 clusters while there are only two trapping regions. It also includes two points in the cluster corresponding to $\mathcal{S}_2$ while there are out of the trapping region $\mathcal{S}_2$ (see the green cluster Figure 9.4 (b)). Then we prefer to use the others methods to estimate ϵ.

Table 9.2: Values of the parameter ϵ of DBSCAN obtained with different methods. The parameter ϵ is estimated assuming $n^\star = 4$ as advised by Ester et al. [1996]. The parameter ϵ is expressed in pixel units. We recall that the original image is 200×200 pixels.

Type	Method	ϵ
Data driven	Graphical method	5.64
	Otsu [1979]	12.39
Parametric	Daszykowski et al. [2001]	5.94
	Beta distribution	5.86



(a) (b)

Figure 9.4: Clusters detected by DBSCAN on the set of particles detected as subdiffusive with the single test procedure at 5%. To be accurate, we run the DBSCAN algorithm on the set $\mathcal{R}$ (9.3.1) obtained with the single test procedure. We use $n^\star = 4$. On the left, we choose $\epsilon = 5.86$ derived from the beta distribution. On the right, we pick $\epsilon = 12.39$ obtained with the method of Otsu [1979]. The noisy points are in red, points from the same clusters have the same color. Red circles represent the boundaries of the trapping regions $\mathcal{S}_1$ (big circle) and $\mathcal{S}_2$ (small circle). The black square represents the boundaries of the whole domain $\mathcal{D}$.

9.5 Estimation of the Shape of the Trapping Areas

In this section, we estimate the shapes $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_2$ of the confinement areas $\mathcal{S}_1$ and $\mathcal{S}_2$ from the clusters detected with DBSCAN. Xu et al. [1998] propose a grid- based approach to estimate the geometries of clusters. More specifically, they want to estimate the area occupied by the clusters. We use this approach to estimate the geometries of the clusters detected with DBSCAN. We note that, as we use a grid, the shapes of the clusters are approximated by polygons. The resolution r of the grid is the key parameter to optimize. If r is too large the shape of the cluster is poorly approximated while if it is too small the cluster may be split in disconnected polygons. First, suppose we detect only one cluster C of size n_c. As proposed by Xu et al. [1998], we set:

$$r = \max_{i=1,\dots,n_c} \tilde{d}_i(n^\star), \tag{9.5.1}$$

where $\tilde{d}_i(n^\star)$ is the distance of the $n^\star$ nearest neighbours of the i^{th} point of cluster C. The estimated shape of cluster C is given by the union of the cell grid containing at least one point of cluster C. The resolution parameter r given by (9.5.1) assures that the estimated shape is not split in several disconnected polygons. Now, suppose that we detect several clusters. Optimally, we use one resolution parameter per cluster. Then we get as many grids as clusters; we want to merge all the cluster shapes on a common grid. We use the finest gird. It is straightforward to translate the cluster shapes from a coarser grid to the finest grid. For the shape computed on the finest grid nothing has to be done. We note that, as we work with shapes (and not points anymore), we can not split the shapes into disconnected shapes going from a coarser grid to a finer grid. The estimation of the geometry of the cluster C_2 corresponding to the trapping region $\mathcal{S}_2$ is given in Figure 9.5.

We evaluate the quality of the estimation of the trapping regions $\mathcal{S}_1$ and $\mathcal{S}_2$ comparing the centroids and areas of the estimated regions to the true regions. We also compare the proportion of subdiffusive particles inside the estimated trapping regions to the true proportions. Results are given in Table 9.3. First, the estimated regions are close to the true regions in terms of areas and centroids (see Table 9.3). Secondly, the proportion of subdiffusive particles in the second estimated region $\hat{\mathcal{S}}_2$ is much closer to the ground truth proportion than the one of $\hat{\mathcal{S}}_1$. As the region $\mathcal{S}_1$ is larger than $\mathcal{S}_2$, it is harder to detect a subdiffusive particle trapped in $\mathcal{S}_1$ than in $\mathcal{S}_2$. In other words, when the alternative hypothesis is confined Brownian motion, the power of our test procedures (both single test procedure or adaptive Procedure 1) decreases as the size of the confinement domain increases.

Remark 9.5.1. *We note that the ground truth proportion of subdiffusive particle inside the true regions are around* 70% *in Table 9.3 while from Figure 8.4 (b) the proportion in the trapping regions at the equilibrium is around* 80%. *We simulate the trajectories at*

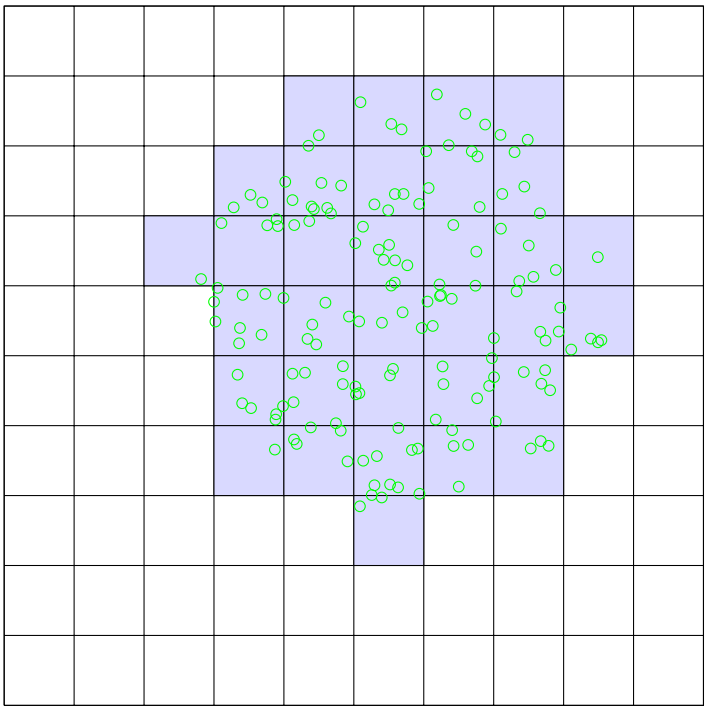


Figure 9.5: Estimation of the geometry of the cluster C_2 corresponding to the trapping region $\mathcal{S}_2$. The resolution of the grid is given by (9.5.1). In this case, it is $r = 5.2$ pixels. We recall that the original image is 200×200 pixels. The blue cells of the grid contain at least one point of the cluster C_2 (represented in green). The blue part is the polygon estimation of the geometry of C_2 corresponding to the trapping region $\mathcal{S}_2$.

the equilibrium (burning period of $t = 100s$) then the difference is due to something else. Actually, the ground truth proportion of Table 9.3 is computed considering that a particle is subdiffusive if it is trapped in the domain during the whole period of observation. On the other hand, in Figure 8.4, we compute this proportion at a single time t. We prefer using the former definition as it is clearly not possible to state if a particle is subdiffusive from the observation of a single point.

Table 9.3: Comparison of the estimated regions with the true regions using different features. For computing the ground truth proportion of subdiffusion inside the true regions we consider that a particle is subdiffusive if it is trapped in the domain during the whole period of observation.

	True regions		Estimated regions	
Property	$\mathcal{S}_1$	$\mathcal{S}_2$	$\hat{\mathcal{S}}_1$	$\hat{\mathcal{S}}_2$
Centroid	(2.5,2.5)	(4,4)	(2.5,2.6)	(4.0,3.9)
Area	1.34	0.48	1.37	0.58
Proportion of subdiffusion	70	69	54	72

9.6 Assessment of the Method on Another Example

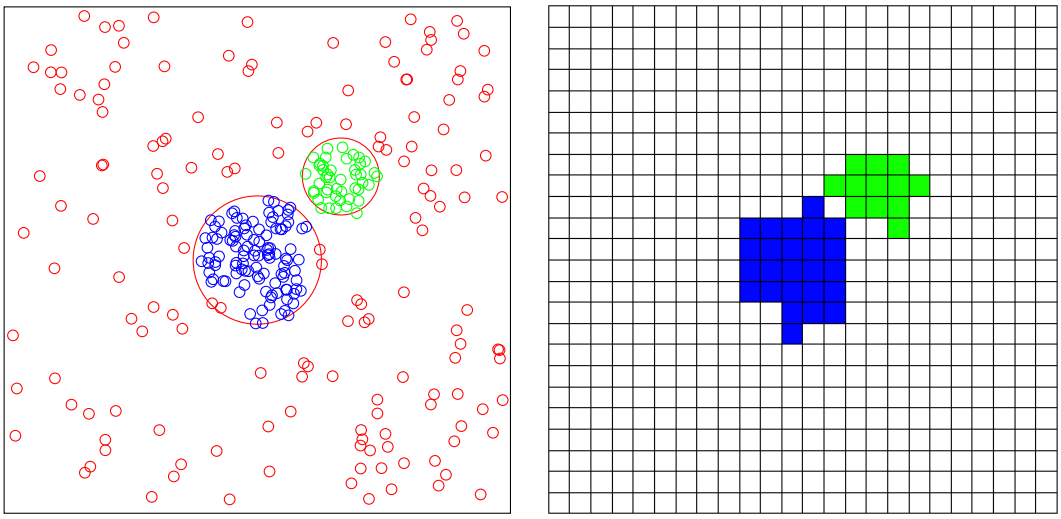
We evaluate the method for detecting microdomains on a more challenging example. We use a simulation scheme similar to the one described in Section 8.3. However, the region $\mathcal{S}_2$ is now the circle of radius $r_2 = 0.39\mu m$ and center $\theta_2 = (3.35, 3.35)$; we translate the original region $\mathcal{S}_2$ closer to $\mathcal{S}_1$. We also set $k^+ = 0.05\ s^{-1}$ while it was originally set to $0.2\ s^{-1}$ (see Table 8.1). Consequently, at the equilibrium, the proportion of trapped particles inside $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ is $k^+/(k^+ + k^-) = 0.5$ (see Equation (8.2.10)) against 0.8 in the previous simulation. We use a burning period of $t = 250\ s$ to be at the equilibrium. Therefore, with these new settings, the microdomains $\mathcal{S}_1$ and $\mathcal{S}_2$ are closer and the concentration of trapped particles in the microdomains is lower than in the first simulation scheme.

Despite the fact that the true trapping regions $\mathcal{S}_1$ and $\mathcal{S}_2$ are close, the DBSCAN algorithm –with $n^\star = 4$ and $\epsilon = 7.53$ derived from the beta distribution (9.4.5)– succeeds in discriminating two clusters corresponding to the trapping regions, see Figure 9.6 (a). We also present the regions estimated with the grid-based approach Figure 9.6 (b). Quantitative results about the estimation of the regions are given in Table 9.4.

Remark 9.6.1. *We can build artificial examples in which DBSCAN discriminates two clusters and at the same time the corresponding regions estimated with the grid-based approach overlap. Therefore, further development should be conducted to deal with this problem. For instance, another choice for the resolution* (9.5.1) *of the grid-based approach could be considered. We note that we do not have this problem in the example of this section, see Figure* (9.6) *(b).*

9.7 The Method of Hoze et al. [2012]

In this section, we present the method developed by Hoze et al. [2012] to detect microdomains. In their context, the microdomains, called potential wells, attract proteins



(a) (b)

Figure 9.6: Clusters detected by DBSCAN on the set of particles detected as subdiffusive with the single test procedure at 5% (a), estimation of the trapping regions with a grid-based approach (b) in a scenario where microdomains are closer to each other and less dense. For the clustering step (Figure (a)), we run the DBSCAN algorithm on the set $\mathcal{R}$ (9.3.1) obtained with the single test procedure. We use $n^\star = 4$. We choose $\epsilon = 7.53$ derived from the beta distribution (9.4.5). In Figure (a), the noisy points are in red, points from the same clusters have the same color. Red circles represent the boundaries of the trapping regions $\mathcal{S}_1$ (big circle) and $\mathcal{S}_2$ (small circle). The black square represents the boundaries of the whole domain $\mathcal{D}$. For the estimation of regions (Figure (b)), the resolution of the grid is $r = 8.6$ pixels of the original image. In Figure (b), the blue (respectively green) region correspond to the blue (respectively green) cluster of Figure (a). The blue (respectively green) region is the estimation of $\mathcal{S}_1$ (respectively $\mathcal{S}_2$).

Table 9.4: Comparison of the estimated regions with the true regions using different features in a scenario where microdomains are closer to each other and less dense.. For computing the ground truth proportion of subdiffusion inside the true regions we consider that a particle is subdiffusive if it is trapped in the domain during the whole period of observation.

	True regions		Estimated regions	
Property	$\mathcal{S}_1$	$\mathcal{S}_2$	$\hat{\mathcal{S}}_1$	$\hat{\mathcal{S}}_2$
Centroid	(2.5,2.5)	(3.35,3.35)	(2.6,2.8)	(3.5,3)
Area	1.34	0.48	1.15	0.55
Proportion of subdiffusion	46	41	52	45

such as the postsynaptic AMPA-type glutamate receptor (AMPAR). We note that Hoze et al. [2012] focus on the two dimensional case but the method can be extended to the three-dimensional case. For simplicity, we consider only the former case as Hoze et al. [2012]. First, we present the model of Hoze et al. [2012]. Secondly, we describe their two-step approach.

1. Hoze et al. [2012] estimate the drift vector field (equivalently the drift function) in a non parametric way from the observation of multiple trajectories,

2. The authors fit this vector field to a parametric drift function.

Overdamped Langevin Equation

Hoze et al. [2012] observe a collection of m independent trajectories $\mathcal{X}_m$ (see Equation (9.1.1)). They assume that all the trajectories $\mathbb{X}_{n_k}^{(k)} = \left(X_{t_0}^{(k)}, \ldots, X_{t_{n_k}}^{(k)}\right)$ are generated from the same diffusion process (X_t) solution of the overdamped SDE (already presented in Equation (3.3.12) with another parametrization):

$$dX_t = -\nabla U(X_t)dt + \sigma dB_t, \tag{9.7.1}$$

where ∇ is the gradient operator, $U : \mathbb{R}^2 \to \mathbb{R}$ is the potential function, and $\sigma > 0$ the diffusion coefficient. As usual, we note $\mu = -\nabla U$, defined in $\mathbb{R}^2$ with values in $\mathbb{R}^2$, the drift function (also called the drift vector field in [Hoze et al., 2012])

Non-Parametric Estimation of the Drift

Hoze et al. [2012] estimate the drift function μ in a non parametric way. We recall that the drift function is defined through the limit:

$$\mu(x) = \lim_{\Delta \to 0} \frac{1}{\Delta} E(X_{t+\Delta} - X_t | X_t = x). \tag{9.7.2}$$

Then, Hoze et al. [2012] define the estimator $\hat{\mu}$ of the drift μ as the empirical counterpart of (9.7.2). First, they define a square $S(x, r)$ or side r to reflect the local conditioning $X_t = x$ in Equation (9.7.2). The estimator $\hat{\mu}(x)$ at point $x \in \mathbb{R}^2$ is defined as the empirical average of all the displacements of particles starting from $S(x, r)$. An expression of $\hat{\mu}$ is derived in Appendix D. Figure 9.7 shows the estimated drift vector field $\hat{\mu}$ computed from the FLUOSIM simulation. We note that a similar approach can be used to infer the diffusion coefficient in case we assume that it is not constant over space but is defined as the function $\sigma : \mathbb{R}^2 \to \mathcal{M}_2$ where $\mathcal{M}_2$ is the set of squared matrix of size 2.

Parametric Fit of the Drift Vector Field

Hoze et al. [2012] assume that the potential U is a truncated polynomial of order 2:

$$U(x,y) = \begin{cases} A\left(\frac{(x-x_0)^2}{a} + \frac{(y-y_0)^2}{b} - 1\right) & \text{if} \quad \frac{x-x_0}{a} + \frac{y-y_0}{b} - 1 < 0, \\ 0 & \text{otherwise,} \end{cases} \tag{9.7.3}$$

where (x_0, y_0) is the attractor, $A > 0$ is the depth of the potential modelling the strength of the attractive force toward (x_0, y_0). Parameters (a, b) are the axis lengths of the ellipse of center (x_0, y_0) in which the particle is submitted to the attractive force. In fact the range of the attractive force is limited to the aforementioned ellipse. As already mentioned in Section 3.3 if potential U is a polynomial of order 2 not truncated the solution of the SDE (3.3.12) is the Ornstein-Uhlenbeck process. In this case, the attractive force has infinite range.

Finally, Hoze et al. [2012] assume that parameter (a, b) and (x_0, y_0) are known. In other words they suppose that the microdomain of attraction is known. They infer the depth A of potential U with the least- square estimator:

$$\hat{A} = \min_{A>0} \sum_{i=1}^{N_x} \left\| -\nabla U(x_i) - \hat{\mu}(x_i) \right\|^2, \tag{9.7.4}$$

where N_x is the number of points of the finite lattice $\mathcal{G} \subset \mathcal{D}$ on which we compute the drift vector field. In the case we can compute $\hat{\mu}$ on the whole space $\mathcal{D}$, we replace the finite sum in Equation (9.7.4) by an integral over $\mathcal{D}$. It is straightforward to compute the closed form of $\hat{A}$, see [Hoze et al., 2012].

9.8 Comparison of the Two Methods

In this section, we emphasize the assumptions, key parameters and properties of our method of detection of microdomains and the method of Hoze et al. [2012]. Our method is based on a Lagrangian approach while the method of Hoze et al. [2012] relies on an Eulerian strategy.

Our Procedure

Our method is based on the combination of the clustering algorithm DBSCAN [Ester et al., 1996] and our test procedures. The critical parameter in our method is the parameter ϵ of DBSCAN. Multiple choices have been studied in Section (9.4). Results on simulations (see Table 9.3) reveal that:

1. our method detects satisfyingly the shapes of the microdomains,

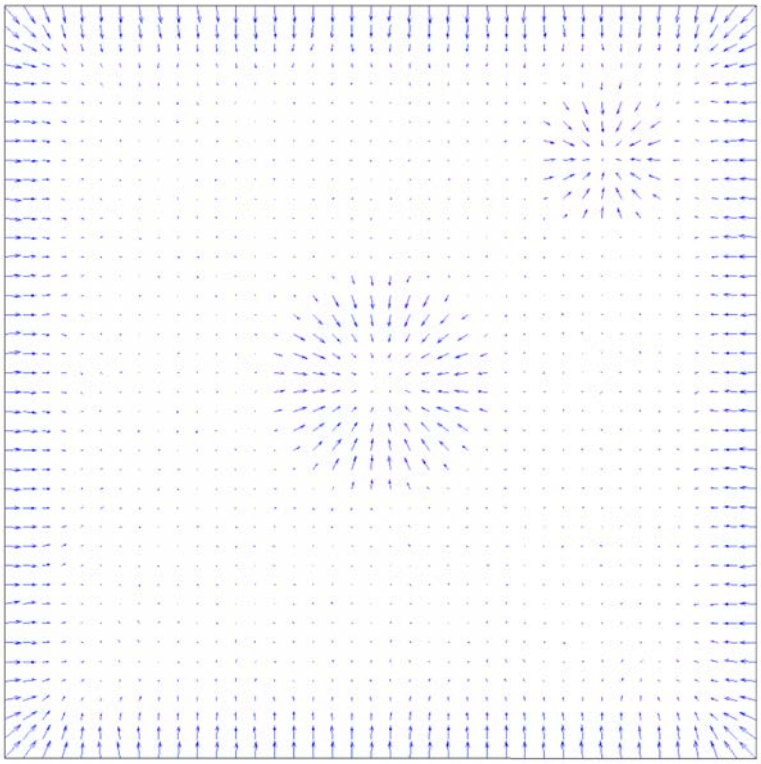


Figure 9.7: Drift field computed with the method of Hoze et al. [2012]. The FLUOSIM simulation comprises 30 frames of 200×200 pixels as the domain $\mathcal{D}$ is a square of $5\mu m$ and the spatial resolution is $\Delta x = 0.025\mu m$ (see Table 8.1). We compute $\hat{\mu}(x)$ on each point of this 200×200 lattice, see Equation (D.0.4). We set $r = 5$ pixels. We can see clearly the two microdomains $\mathcal{S}_1$ and $\mathcal{S}_2$: the vector field converges to the centres of these domains. We can see that the vector field is orthogonal to the boundaries of the square $\mathcal{D}$. It is due the normal boundary condition on $\partial\mathcal{D}$. Elsewhere, $\hat{\mu}(x) = \mathbf{0}_2$ reflecting Brownian motion.

2. our method is able to reliably estimate the proportion of subdiffusive and non-subdiffusive trajectories if the microdomain contains a mixture of trajectories of different diffusion types.

Finally our method can handle a general collection of trajectories $\mathcal{X}_m$. More specifically, the different trajectories do not need to have the same drift parameters and diffusion coefficients (see Equation (9.1.2)). However, their drift functions and diffusion coefficients must not depend on time. In such a case, a further development of the method is to use the detection of change point of Chapter 6, as a pre-processing step of our method. We split the trajectories into smaller trajectories showing the same diffusion type over time. Then, we come back to the same framework as the one studied in

this Chapter.

The Approach of Hoze et al. [2012]

The method of Hoze et al. [2012] is based on a first non-parametric estimation of the drift vector field followed by a parametric fit to a chosen drift function. The purpose of Hoze et al. [2012] is to estimate the parameter A (the depth of potential), assuming a parametric model for the trajectories, defined through the potential function (9.7.3). Then, the authors do not provide a method to detect confinement zone. We note that the estimation of the confinement zone is equivalent to estimate parameters (a, b, x_0, y_0) in their model (see Equation (9.7.3)). The authors assumes that these parameters are known.

Future development of the method of Hoze et al. [2012] can involve the estimation of parameters (a, b, x_0, y_0) in order to estimate the microdomains under the parametric model (9.7.3). However, in this case, the microdomains can only be approximated by ellipses of length axis (a, b) and center (x_0, y_0).

Another approach could be to detect microdomains directly from the non-parametric estimation of the drift, see Equation (D.0.4). In any cases, the non-parametric estimation of the drift is a key step. We must set the critical parameter r reflecting a balance between bias and variance (see Appendix D). Consequently the parameter r can be related to the bandwidth parameter in density estimation. [Hozé, 2013, Chapter 2, Section 2.9.1] states that their method is robust to the choice of r but do not provide a statistical method for choosing this parameter. We also note that the quality of the estimation of $\hat{\mu}(x)$ depends on the number of trajectory points close to x. Then, the locations where the density of trajectory points is low will give a poor estimate of μ. Consequently, the quality of estimation of the drift vector field varies across space which can be problem. However, the microdomains show a high concentration of trajectory points; the drift vector field should be well estimated there. Finally, the non-parametric estimation of the drift assumes that, locally in space, all the trajectories undergo the same diffusion process with the same drift function. Such an hypothesis can be a strong in practice. For instance, we can see in Figure 9.7 that the estimated drift field can not capture Brownian motion occuring in the microdomains, due to the estimation process based on averaging (see, Equation D.0.4).

Related Method

Masson et al. [2014] also propose an Eulerian approach to infer the drift vector field (and the diffusion coefficient vector field). The image is first decomposed into non overlapping blocks. All the subtrajectories inside a block of the partition (or mesh) are supposed to be driven by the same SDE. The drift and diffusion coefficient is supposed to be constant in each block. Then, the starting point is similar to the analysis proposed in

[Hoze et al., 2012]. However, instead of using non-parametric estimates of the drift and the diffusion coefficient, the authors use Bayesian inference to estimate the parameters. More specifically, they use an approximate Gaussian likelihood based on the Gaussian approximation of the SDE. This likelihood approach is flexible and can take into account localization errors. Jeffrey's prior is used as a default prior distribution for the drift and the diffusion coefficient. The motion in the different blocks are supposed independent. Accordingly, the *a-posteriori* distribution on the whole space is the product of the *a-posteriori* distributions of each block. An alternative option is to use smoothing priors to penalize strong gradients of the drift field (or the coefficient diffusion field) [El Beheiry et al., 2016]. In this case, the *a-posteriori* distribution is no longer the product of the *a-posteriori* distributions of each block and the estimation of the maximum *a-posteriori* (MAP) can be computationally costly. Once the local drifts (and diffusion coefficients) are estimated by the MAPs, Masson et al. [2014] propose to fit the estimated vector field to a parametric vector field as in [Hoze et al., 2012], (see Equation (9.7.4)). They use a least square estimator but add a penalization term to smooth the vector field compared to Hoze et al. [2012]. The main advantage of the Bayesian method of Masson et al. [2014] is that prior information – from a biologist expert for instance – about the drift and diffusion coefficient can be added. Finally Masson et al. [2014] emphasise that prior information –more specifically smoothing prior on the gradient field– can reduce the statistical error due to misconnections of particles during the tracking process.

9.9 Summary

In this chapter, we presented a method to detect the trapping regions where the particles are confined and thereby undergo subdiffusion. We combined the clustering algorithm DBSCAN [Ester et al., 1996] and our test procedure (4.5.1) to identify the areas with a high concentration of subdiffusive particles, corresponding to the trapping regions. We used a basic grid-based approach to estimate the trapping regions. We were also able to estimate the proportion of particles inside the trapping regions which were effectively confined in the domain. In fact, some particles can go through a trapping region without being trapped.

We compared our method to the Eulerian approach of Hoze et al. [2012]. Due to its averaging estimation process, Hoze et al. [2012] was not able to capture the motion of particles not confined in the trapping regions but still going through these regions. Each method have a critical parameter to set. In our case, the parameter ϵ defines the clusters. In [Hoze et al., 2012], r is a bandwidth parameter influencing the smoothness of the estimated drift field.

In future work, the non-parametric estimation of the drift field proposed in Hoze et al. [2012] can be combined to our own method to get better insights of the biological processes occurring in microdomains in practical imaging.

10 Conclusion

In this thesis, we developed several methods based on statistical testing to analyse the traffic of intracellular particles. Designing a statistical test implies to define the hypothesises and the test statistic. We proposed a new test statistic T_n (4.3.1). Our null hypothesis is that the particle is driven by Brownian motion, the motion of reference in biophysics for modelling intracellular motion [Qian et al., 1991], in mathematics for defining stochastic differential equations [Karlin, 1981] and in physics with the Langevin equation [Kou, 2008]. Then, we used this test in order to classify the observed trajectories according to their modes of diffusion. We also developed an algorithm for detecting the times at which a particle switches motion based on the test statistic T_n. Finally, we proposed a spatial interpretation of the outcome of our test on a collection of trajectories.

10.1 Contributions of the Thesis

Classification

We developed a three-decision test for classifying the particle trajectories observed in living cells into three types of diffusion: Brownian motion (null hypothesis), subdiffusion and superdiffusion (alternatives). On the one hand, we built a single test procedure for testing a single trajectory, on the other hand we proposed a multiple test procedure for testing a collection of trajectories. These procedures control respectively the type I error and the false discovery rate at level α. It is worth noting that the length of the trajectory n is taken into account in our classification rule. Our approach can be considered as an alternative to the MSD method. It gives more reliable results as confirmed by our Monte Carlo simulations and evaluations on real sequences of images depicting protein dynamics acquired with 2D and 3D TIRF microscopy. We implemented the test procedure in the Matlab package THOT (Testing HypOtheses for diffusion TricHotomy) available http://serpico.rennes.inria.fr.

Detection of Change of Dynamic over Time

We proposed a non parametric algorithm to detect the change points along a particle trajectory. These change points are defined as the times at which the particle switches between the three modes of diffusion mentioned in this thesis. When the trajectory is fully Brownian (our null hypothesis H_0), we control the probability to detect a false

change point at level α. Our procedure has a single parameter to choose: the size k of the local sliding window. We give guidelines on how to choose this parameter. We compared our method to the methods of Türkcan and Masson [2013] and Monnier et al. [2015]. None of the existing methods is able to distinguish the three types of motions, namely Brownian motion subdiffusion and superdiffusion. Secondly, we demonstrate that our procedure outperforms the two competitive procedures aforementioned. In addition, it is much faster than the two others which is a advantage when dealing with a large numbers of trajectories.

Clusters of Trajectories for Spatial Analysis

In the cell, there are microdomains where successively i/ particles are trapped ii/ they interact with other complex iii/they are released in the cytosol. We proposed a method for detecting these domains where particles are confined for a while. First, we represented each trajectory by a single point. Then, we tested the trajectories and obtained a map with the spatial distribution of the outcome of the test. We used a clustering algorithm to detect the clusters of subdiffusive trajectories. More specifically, we chose the DBSCAN clustering algorithm as it is able to discriminate noisy points (corresponding to trajectories outside microdomains) from true clusters (corresponding to subdiffusive trajectories inside microdomains). Finally, we estimated the contours of the clusters with a grid-based approach. This technique was validated on simulations produced by the FLUOSIM software. Our procedure gives good results as long as the microdomains is not too large.

10.2 Future Work and Extensions

Bayesian Tests

Further work involves adding prior information about the hypotheses to test. We can use a Bayesian framework to this end. Generally, the hypothesises are parametric and of the form: $H_0 : \theta \in \Theta_0$ *versus* $H_1 : \theta \in \Theta_1$. Prior information is introduced through [Kass and Raftery, 1995]:

1. prior probabilities affected to each hypothesis H_0 and H_1,
2. prior distributions over the set of parameters of each hypothesis $\pi_0(\theta),\ \theta \in \Theta_0$ and $\pi_1(\theta),\ \theta \in \Theta_1$.

Berger and Guglielmi [2001] developed a Bayesian test with a parametric null hypothesis against an alternative non-parametric hypothesis, which matches with our framework. Another approach consists in weighting the p-values obtained with a standard test procedure. It was proposed in the multiple test procedure of Holm [1979]. Giving a large

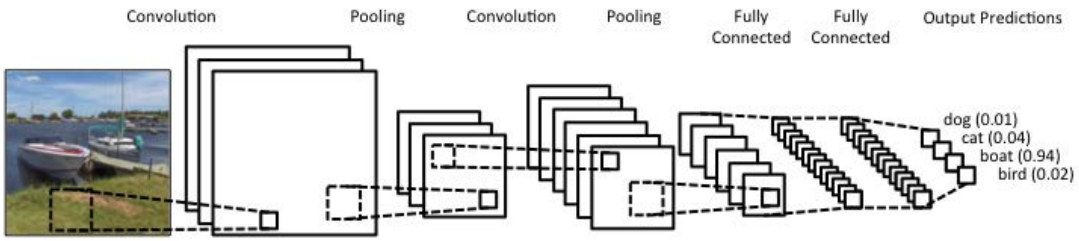


Figure 10.1: Example of architecture of a convolutional neural network (CNN). Source https://www.clarifai.com/technology.

weight to an hypothesis (or equivalently to the corresponding p-value) favours rejection while a small weight favours acceptance. Genovese et al. [2006] define a binary weighting scheme and use the weighted p-values as inputs of the algorithm of Benjamini and Hochberg [1995] to obtain a procedure that controls the false discovery rate. We emphasize this method with the example of the Rab6 protein trafficking studied in [Pécot et al., 2017]. The authors show that the Rab6 proteins move quite directly from the Golgi apparatus to the cell periphery until they enter a docking phase. Then, they mostly go back towards the cell center by following long and indirect trajectories. Therefore, we can give a weight favouring the superdiffusion hypothesis to the Rab6 proteins going towards cell periphery while we give a weight favouring the Brownian motion hypothesis to those going back to the Golgi apparatus. To conclude, we can improve our test procedures using prior information modelled in a Bayesian framework or by p-values weighting. A challenging task is to translate the Bayesian and p-values weighting methods to three-decision testing.

Convolutional Neural Networks

In this thesis, we proposed a model-based approach in the sense that Brownian motion was the null hypothesis or our tests. Machine learning and deep learning are another way to address the issue of trajectory classification. The intern V. Gleizes from INSA Rennes worked during two months on this topic. He used a convolutional neural network (CNN) [LeCun et al., 1995, Krizhevsky et al., 2012] and focused on the two-dimensional case. He considered 2D binary images depicting 2D paths corresponding to the observation of 2D trajectories over time (2D+Time), see Figure 10.3. These images are the inputs of the CNN.

A convolutional neural network is composed of successive layers. The layers are filters receiving an input from the previous layer and producing an output for the next one. In case of the CNN, there are three main types of layers. The convolutional layers apply convolution on the input image. The pooling or sub-sampling layers reduce the

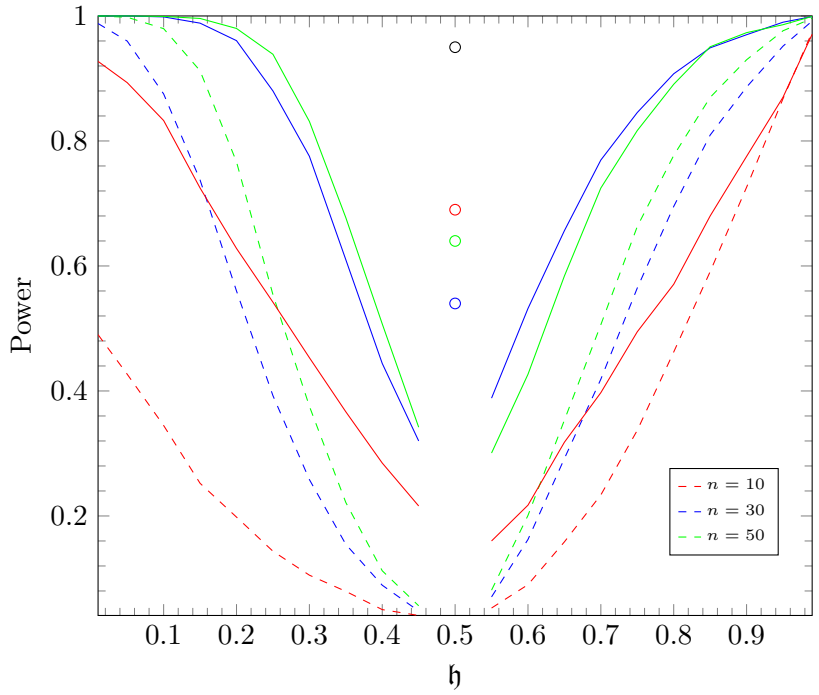
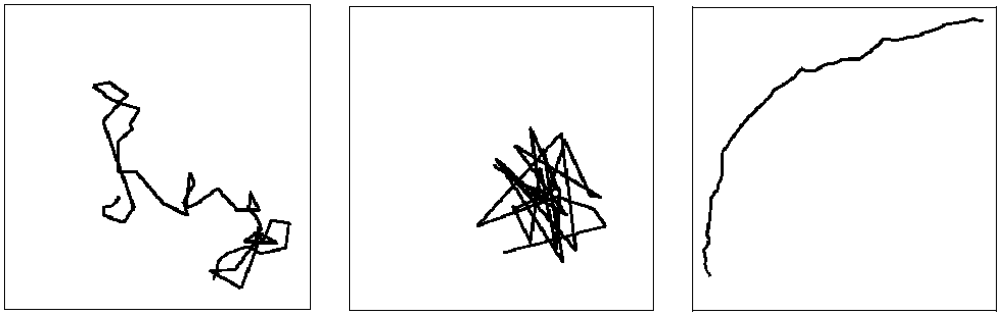


Figure 10.2: Comparison of the power curves of the fractional Brownian motion obtained with our test procedure and with CNN. The dashed curves are the power curves obtained with our three-decision test procedure at level $\alpha = 5\%$ (see Figure 5.1 (c)), the plain lines are the power curves obtained by V. Gleizes with CNN. If $0 < \mathfrak{h} < 1/2$ it is subdiffusion, if $1/2 < \mathfrak{h} < 1$ it is superdiffusion, if $\mathfrak{h} = 1/2$ it is Brownian motion. The CNN detection rate for Brownian motion is represented with dots at $\mathfrak{h} = 1/2$ (the color code still holds for the dots). Our test procedure detects well 95% of the Brownian trajectories at level $\alpha = 5\%$ for any size n (black dot).

dimension of the features representing the data. Then, it reduces the computational cost and prevent from overfitting. Finally, the fully connected layers apply non linear filters and allow to classify the original input image. An example of architecture is given in Figure 10.1. In practise, V. Gleizes designed its own architecture consisting in successively two convolutional layers, one pooling layer, two convolutional layers, one pooling layer, one fully connected layer of 128 neurones, one fully connected layer of 3 neurones.

CNN is a supervised classification method, then it must be trained with labelled images. In our case, CNN is trained with images depicting two-dimensional trajectories obtained by simulation. As a consequence we know their label. More specifically, CNN is learnt with images depicting Brownian trajectories, fractional Brownian trajectories with $\mathfrak{h} < 1/2$ for subdiffusion, fractional Brownian trajectories with $\mathfrak{h} < 1/2$ for superdiffusion. Then, he assessed the performances of the method on another set of Brownian and



(a) (b) (c)

Figure 10.3: CNN input images. The input image in Figure (a) represents a Brownian trajectory, the input image in Figure (b) a fBm with $\mathfrak{h} < 1/2$ modelling subdiffusion, and the input image in Figure (c) a fBm with $\mathfrak{h} > 1/2$ modelling superdiffusion. The dimensions of the images are 255×255. The trajectories have same size $n = 30$.

fractional Brownian trajectories. Results are presented in terms of power curves in Figure 10.2. It shows better perfomances than our test for detecting subdiffusion (fBm with $\mathfrak{h} < 1/2$) and superdiffusion (fBm with $\mathfrak{h} > 1/2$), especially on small trajectories. CNN detects well 69% (respectively 54% and 64%) of Brownian trajectories of size $n = 10$ (respectively $n = 30$ and $n = 50$). Surprisingly, the best detection rate is for the trajectory size $n = 10$. Our test is built such that we detect well 95% of Brownian trajectories at level $\alpha = 5\%$, irrespective to the trajectory size n.

In terms of computational cost, both the training and estimation phase are heavy for CNN. Moreover, it is required to estimate the CNN parameters for any trajectory size n. Also, V. Gleizes considered images of a fixed size. As the method is not scaled-invariant, future work could involve a multi-scale approach. On contrary, we derived asymptotic behaviour of our method (Theorem 4.7.1), so that we can use the asymptotic parameters for large n. Once the thresholds defining the rejection region of our test are computed, the classification is instantaneous in terms of computational cost. Finally, our test procedure is scale invariant in the sense that our test does not depend on the diffusion coefficient σ nor the step of time between two observations Δ. Then, the two approaches have pros and cons and could be combined to give better results.

Application to Spatial Ecology

In this thesis, the range of developed methods was applied to analyse intracellular trafficking. The same questions arise in ecology with the study of animal displacement. Each animal behaviour corresponds to a particular motion. For instance, when the rate of preys in an area decreases, predators move to the next prey patch in a relatively straight way; their motion can be modelled by a superdiffusion [Schick et al., 2008].

Another type of motion is referred to as area-restricted search. It includes foraging, breeding and resting behaviours [Bailey et al., 2009]. In this case, animals do not go in a specific direction but tend to increase their turning angle especially when they prey. Therefore, this type of motion can be modelled with subdiffusion or Brownian motion. Consequently, we are convinced that the statistical methods designed in this thesis can be helpful to analyse individual and collective animal movement.

Part IV

Appendices

A Convergence Results of the Single Test Procedure

A.1 Proof of Theorem 4.7.1

Proof of Theorem 4.7.1. Under the null hypothesis, $X_t/\sigma = B_t$ is a standard Brownian Motion. Let us introduce the following random variable,

$$\tilde{T}_n = \max_{k=1\ldots n} \left\| \frac{1}{\sqrt{n}} R_k \right\|, \tag{A.1.1}$$

where $R_k = \sum_{j=1}^{k} (B_{j\Delta} - B_{(j-1)\Delta})/\sqrt{\Delta}$. Since $\hat{\sigma}_n$ is a consistent estimator of σ and using the Slutsky Lemma, it remains to prove that $\tilde{T}_n$ converges in distribution to S_0. Using the fact that the increments of the Brownian process are independent and Gaussian, R_k is the sum of k independent identically $\mathcal{N}(0,1)$-distributed random variables. We define the following process,

$$W_t^{(n)} = \frac{1}{\sqrt{n}} R_{\lfloor nt \rfloor}, \qquad t \in [0,1],$$

where $\lfloor x \rfloor$ denotes the integer part of $x \in \mathbb{R}$. Then we get:

$$\tilde{T}_n = \sup_{t\in[0,1]} \|W_t^{(n)}\|_2. \tag{A.1.2}$$

Due to Donsker's Theorem [Billingsley, 2013, Theorem 8.2], $(W_t^{(n)})$ converges in distribution to the Wiener measure as $n \to \infty$ over the space of continuous function on $[0,1]$. Since $x \to \sup_{t\in[0,1]} \|x(t)\|$ is a continuous function on the space of continuous functions from $[0,1]$ to $\mathbb{R}$, $\tilde{T}_n$ converges in distribution to S_0. □

A.2 Proof of Proposition 1: the Convergence of the Estimator (4.6.1) of the Diffusion Coefficient

Notice that $\hat{\sigma}_n = \hat{\sigma}_{1,n}$ is strongly consistent under the null hypothesis due to the strong law of large numbers and the independence of the increments of the Brownian motion.

We focus now on the three alternatives. According to the alternative, we denote by $\mathbb{E}$ the expectation associated to the measure P of the solution of the related SDE ((4.2.2) or (4.2.1) or (4.2.3)).

Brownian with drift. We may rewrite the strong solution of the SDE (4.2.3) as,

$$X_{t_k} = X_{t_{k-1}} + v\Delta + \sigma\sqrt{\Delta}\epsilon_k, \qquad k = 1 \dots n,$$

where $\sqrt{\Delta}\epsilon_k = B_{t_k} - B_{t_{k-1}}$, and (B_t) is a standard Brownian motion. Then the random variables $Z_k = \|v\Delta + \sigma\sqrt{\Delta}\epsilon_k\|^2$, $k = 1 \dots n$, are positive independent identically distributed random variables, and admit a moment of order 1,

$$\mathbb{E}(Z_k) = \Delta^2\|v\|^2 + d\Delta\sigma^2.$$

Then according to the strong law of large numbers, $\hat{\sigma}_n$ converges almost surely to $\Delta\|v\|^2/d + \sigma^2$. □

Ornstein-Uhlenbeck process. Let (X_t) be an Ornstein-Uhlenbeck process (4.2.1). The SDE (4.2.1) admits a unique solution [Bressloff, 2014, Section 2.2.3]

$$X_t - X_s = (X_s - \theta)(e^{-\lambda(t-s)} - 1) + \sigma \int_s^t e^{-\lambda(t-u)} dB_u^{1/2}. \tag{A.2.1}$$

Then (X_t) is a stationary Gaussian process where transition density $p(s, x, t, y)$ is the density of

$$\mathcal{N}\left(x + (x-\theta)(e^{-\lambda(t-s)} - 1), \sigma^2(1 - e^{-2\lambda(t-s)})/(2\lambda)\mathbf{I}_d\right).$$

Then we get that,

$$\begin{aligned}\mathbb{E}(\|X_{t+\Delta} - X_t\|^2 \mid X_t = x) &= \int \|x - y\|^2 p(t, x, t+\Delta, y) dy, \\ &= \|x - \theta\|^2(e^{-\lambda\Delta} - 1)^2 + d\sigma^2(1 - e^{-2\lambda\Delta})/(2\lambda).\end{aligned}$$

Moreover the density μ of the stationary distribution of (X_t) is the Gaussian variable $\mathcal{N}(\theta, (\sigma^2\mathbf{I}_d)/(2\lambda))$. Then we obtain that,

$$\begin{aligned}\mathbb{E}(\|X_{t+\Delta} - X_t\|^2) &= \int \mathbb{E}(\|X_{t+\Delta} - X_t\|^2 \mid X_t = x)\mu(x)dx, \\ &= d\sigma^2(e^{-\lambda\Delta} - 1)^2/(2\lambda) + d\sigma^2(1 - e^{-2\lambda\Delta})/(2\lambda), \\ &= d\sigma^2(1 - e^{-\lambda\Delta})/\lambda.\end{aligned}$$

Now, according to [Bibby and Sørensen, 1995, Lemma 3.1], if (X_t) is a stationary diffusion, $\hat{\sigma}_n^2$ converges in probability to $\mathbb{E}(\|X_{t+\Delta} - X_t\|^2)/(d\Delta)$. We deduce the result. □

Fractional Brownian Motion. Let (X_t) be a fractional Brownian motion (4.2.2). Due to the self-similarity property and the stationary increments of the fractional Brownian motion, the following process,

$$W_t^{(n)} = \frac{X_{t_0 + n\Delta t} - X_{t_0}}{(n\Delta)^{\mathfrak{h}}\sigma}, \qquad t \in [0, 1],$$

is a standard fractional Brownian motion. The statistic associated to the quadratic variation of the process $(W_t^{(n)})$ may be defined as,

$$V_n = \frac{1}{n}\sum_{i=1}^{n} \frac{\|W_{i/n}^{(n)} - W_{(i-1)/n}^{(n)}\|^2}{\mathbb{E}\|W_{i/n}^{(n)} - W_{(i-1)/n}^{(n)}\|^2} - 1,$$
$$= \frac{\hat{\sigma}_n^2}{\sigma^2 \Delta^{2\mathfrak{h}-1}} - 1.$$

According to [Coeurjolly, 2001, Proposition 1], V_n converges almost surely to 0. Then we deduce that $\hat{\sigma}_n^2/\sigma^2$ tends to $\Delta^{2\mathfrak{h}-1}$ almost surely. □

A.3 Proof of Proposition 2: the Asymptotic Behaviour of the Test Statistic under Parametric Alternatives

Since the diffusion parameter σ is unknown, the test statistic (4.3.1) is normalized by an estimator of σ. Proposition 1 states that $\hat{\sigma}_n/\sigma$ converges in probability to a constant. Therefore, it is sufficient to study the asymptotic behaviour of the test statistic as if σ was known. Then, in this section, we consider the test statistic T_n as:

$$T_n = \frac{\max_{i=1,\dots,n} \|X_{t_i} - X_{t_0}\|}{\sigma\sqrt{t_n - t_0}}. \tag{A.3.1}$$

Brownian motion with drift (H_2). The process (X_t) is a Brownian motion with drift (4.2.3) and may be rewritten as,

$$X_{t_n} - X_{t_0} = v(t_n - t_0) + \sigma(B_{t_n} - B_{t_0}).$$

Using that (B_t) is a Brownian motion, the distribution of $B_{t_n} - B_{t_0}$ is $\mathcal{N}(\mathbf{0}_d, (t_n - t_0)\mathbf{I}_d)$. Then we have:

$$\mathbb{E}\left(\left\|\frac{X_{t_n} - X_{t_0}}{\sigma(t_n - t_0)} - \frac{v}{\sigma}\right\|^2\right) = \frac{d}{t_n - t_0}. \tag{A.3.2}$$

As $t_n - t_0 = n\Delta$, we deduce that $V_n = (X_{t_n} - X_{t_0})/(\sigma(t_n - t_0))$ converges in probability to v/σ. As the euclidean norm is a continuous function, the variable $\|V_n\|$ converges in probability to $\|v\|/\sigma > 0$. Then $\sqrt{n\Delta}V_n$ converges in probability to $+\infty$. Since T_n is lower bounded by $\sqrt{n\Delta}V_n = \|(X_{t_n} - X_{t_0})\|/(\sigma\sqrt{t_n - t_0})$, the proof is complete. □

The Ornstein-Uhlenbeck process (H_1). The process (X_t) is an Ornstein-Uhlenbeck process (4.2.1). We assume that the process is in its stationary regime, that means X_{t_0} is drawn from the stationary distribution that is $X_{t_0} \sim \mathcal{N}(\theta, \sigma^2/(2\lambda)\mathbf{I}_d)$. The SDE (4.2.1) admits an unique solution [Bressloff, 2014, Section 2.2.3]

$$X_t - \theta = (X_{t_0} - \theta)e^{-\lambda(t-t_0)} + \sigma \int_{t_0}^{t} e^{-\lambda(t-u)} dB_u^{1/2}. \tag{A.3.3}$$

Then we may bound the test statistic T_n by,

$$\frac{\|X_{t_0} - \theta\|}{\sigma\sqrt{n\Delta}} + \sum_{i=1}^{d} \max_{k=1\dots n} \frac{|X^i_{t_k} - \theta_i|}{\sigma\sqrt{n\Delta}}.$$

Since X_{t_0} is drawn from the stationary distribution, the term $\|X_{t_0} - \theta\|/\sqrt{n\Delta}$ converges in probability to zero.
Now we show that the second term in the previous equation tends to zero in probability as well. We introduce the variables (ξ^1_k, ξ^2_k) defined as,

$$\xi^i_k = (X^i_{t_k} - \theta_i)\sqrt{2\lambda}/\sigma, \qquad k = 1\dots n,\ i = 1, \dots, d.$$

Then for $i = 1, \dots, d$, the sequence $(\xi^i_k)_k$ is a standardized stationary normal sequence with covariance function,

$$r_k = \mathbb{E}\left(\xi^i_\ell \xi^i_{\ell+k}\right) = e^{-k\Delta}, \quad k \geq -\ell.$$

Let i be in $\{1, \dots, d\}$. Then $(a_n(\max_{k=1\dots n}(\xi^i_k) - b_n))_n$ converges in distribution according to [Leadbetter et al., 1983, Theorem 4.3.3], where $a_n = \sqrt{2\log(n)}$ and $b_n = a_n - (2a_n)^{-1}(\log\log(n) + \log(4\pi))$. We deduce that $\max_{k=1\dots n}(\xi^i_k)/\sqrt{n\Delta}$ converges in probability to 0. Moreover, since $(\xi^i_k)_k$ is a centred Gaussian process, then $\max_{k=1\dots n}(-\xi^i_k)/\sqrt{n\Delta}$ converges in probability to 0 by symmetry. Then we conclude that $\max_{k=1\dots n} |X^i_{t_k} - \theta_i|/\sqrt{n\Delta}$ converges in probability to 0. □

The fractional Brownian Motion (H_1). The process (X_t) is a fractional Brownian motion with $\mathfrak{h} \in (0, 1/2)$. From the property of self-similarity and stationarity of increments of the fractional Brownian motion, the following process,

$$Z^{(n)}_t = \frac{X_{tn\Delta+t_0} - X_{t_0}}{\sigma(n\Delta)^{\mathfrak{h}}}, \quad t \in [0, 1], \tag{A.3.4}$$

is a fractional Brownian motion. We rewrite the test statistic as,

$$T_n = \frac{1}{(n\Delta)^{1/2-h}} \max_{k=1\dots n} \|Z^{(n)}_{k/n}\|$$

Then T_n is bounded by,

$$\frac{1}{(n\Delta)^{1/2-h}} \sum_{i=1}^{d} \max_{k=1\dots n} |Z^{i,(n)}_{k/n}|,$$

where $Z^{(n)}_t = (Z^{1,(n)}_t, \dots, Z^{d,(n)}_t)$. The process $Z^{(n)}$ has a version with continuous path as a result of being γ-Holder continuous for any $\gamma < \mathfrak{h}$. Let $i \in \{1, \dots, d\}$ be fixed. Then the random variable $\max_{k=1\dots n} |Z^{i,(n)}_{k/n}|$ is bounded by,

$$M^{(n)}_i = \sup_{t\in[0,1]} |Z^{i,(n)}_t|,$$

which possesses an absolutely continuous density on $\mathbb{R}_+^*$ according to Zaïdi et al. [2003]. That means the sequence $\left(\max_{k=1\dots n} \|Z_{k/n}^{(n)}\|\right)_n$ is tight. Since $\mathfrak{h} < 1/2$, we deduce that T_n converges in probability to 0. □

The fractional Brownian Motion (H_2). The process (X_t) is a fractional Brownian motion with $\mathfrak{h} \in (1/2, 1)$. From the property of self-similarity we get that:

$$Y_n = \frac{\|X_{t_n} - X_{t_0}\|^2}{\sigma^2 (t - t_0)^{2\mathfrak{h}}} \sim \chi^2(d). \tag{A.3.5}$$

We observe that $T_n^2 \geq Y_n (n\Delta)^{2\mathfrak{h}-1}$. Let x be a positive constant. We have:

$$\begin{aligned} P\left(T_n < x\right) &\leq P\left(Y_n (n\Delta)^{2\mathfrak{h}-1} < x^2\right) \\ &\leq P\left(Y_n < x^2/(n\Delta)^{2\mathfrak{h}-1}\right). \end{aligned} \tag{A.3.6}$$

Since $\mathfrak{h} > 1/2$, $x^2/(n\Delta)^{2\mathfrak{h}-1}$ converges to 0 as $n \to \infty$. Then the right hand side of (A.3.6) converges to 0. That means $P(T_n < x)$ converges to 0 as $n \to \infty$: T_n converges to $+\infty$ in probability. □

A.4 Dependency of the Power on the Parameters of the Parametric Alternatives

Lemma A.4.1. *Let (X_t) be a Brownian motion with drift* (4.2.3)*. Let $\hat{\sigma}_n$ be the estimator of the diffusion coefficient defined in Equation* (4.6.1)*. The distribution of T_n* (4.3.1) *depends only on the parameter $v\sqrt{\Delta}/\sigma$ and the trajectory size n.*

Proof of Lemma A.4.1. We may rewrite the strong solution of the SDE (4.2.3) as,

$$X_{t_k} = X_{t_{k-1}} + v\Delta + \sigma\sqrt{\Delta}\epsilon_k, \qquad k = 1 \dots n,$$

where $\sqrt{\Delta}\epsilon_k = B_{t_k} - B_{t_{k-1}}$, and (B_t) is a standard Brownian motion. Then (ϵ_k) is a sequence of independent Gaussian variables $\mathcal{N}(0, 1)$. Furthermore, we have immediately:

$$X_{t_k} - X_{t_0} = vk\Delta + \sigma\sqrt{\Delta}\sum_{i=1}^{k} \epsilon_i, \qquad k = 1 \dots n.$$

Finally the test statistic T_n may be rewritten as,

$$T_n = \frac{\max_{k=1,\dots,n} \left\| k\frac{v\sqrt{\Delta}}{\sigma} + \sum_{i=1}^{k} \epsilon_i \right\|}{\sqrt{\frac{1}{2}\sum_{i=1}^{n} \left\| \frac{v\sqrt{\Delta}}{\sigma} + \epsilon_i \right\|^2}}.$$

As the distribution of (ϵ_k) is free of the parameters the distribution of T_n depends only on $v\sqrt{\Delta}/\sigma$. □

Lemma A.4.2. *Let (X_t) be a fractional Brownian motion* (4.2.2)*. Let $\hat{\sigma}_n$ be the estimator of the diffusion coefficient defined in Equation* (4.6.1)*. The distribution of T_n* (4.3.1) *depends only on the parameter $\mathfrak{h}$ and the trajectory size n.*

Proof of Lemma A.4.2. The fractional Brownian motion may be described by its incremental process [Taqqu, 2003]:

$$\epsilon_k = (X_{t_k} - X_{t_{k-1}})/(\sigma\Delta^{\mathfrak{h}}), \qquad k \geq 1, \tag{A.4.1}$$

where (ϵ_k) is a fractional Gaussian noise which is a stationary standardized Gaussian process with autocovariance function $\mathbb{E}(\epsilon_k\epsilon_{k+i}) = (1/2)(|i+1|^{2\mathfrak{h}} - 2|i|^{2\mathfrak{h}} + |i-1|^{2\mathfrak{h}})$. Finally the test statistic T_n may be rewritten as,

$$T_n = \frac{\max_{k=1,\dots,n} \left\| \sum_{i=1}^{k} \epsilon_i \right\|}{\sqrt{\frac{1}{2}\sum_{i=1}^{n} \|\epsilon_i\|^2}}.$$

Then the distribution of T_n depends only on the trajectory size n and on $\mathfrak{h}$ through the distribution of (ϵ_k). □

B Proof of Proposition 3

Proof. We suppose that the trajectory $\mathbb{X}_n$ is generated under the null hypothesis (6.2.1). For simplicity, we note P the probability under H_0 (noted $P^{\emptyset}_{1/2,0,\sigma}$ in Chapter 6). We want to show that under H_0, Procedure 2 with thresholds γ_1 and γ_2 defined in Proposition 3, controls the probability of the type I error at level α:

$$P\left(\exists i \in \{k, \dots, n^\star\}, \quad \sum_{j=i}^{i+c-1} \mathbf{1}(Q_j \neq 0\right) \geq c^\star) \leq \alpha \tag{B.0.1}$$

where $n^\star = n - k - c + 1$.
We express the event $\{Q_i \neq 0\}$ as:

$$\begin{aligned} \{Q_i \neq 0\} = &\{\phi(B_i) = 0, \phi(A_i) = 1\} \cup \{\phi(B_i) = 0, \phi(A_i) = 2\} \\ &\cup \{\phi(B_i) = 1, \phi(A_i) = 0\} \cup \{\phi(B_i) = 2, \phi(A_i) = 0\} \\ &\cup \{\phi(B_i) = 1, \phi(A_i) = 2\} \cup \{\phi(B_i) = 2, \phi(A_i) = 1\} \end{aligned} \tag{B.0.2}$$

Then we deduce the following inclusion:

$$\{Q_i \neq 0\} \subset \{\phi(B_i) = 1\} \cup \{\phi(B_i) = 2\} \cup \{\phi(A_i) = 1\} \cup \{\phi(A_i) = 2\} \tag{B.0.3}$$

Then from the definition of ϕ we can reexpress the right hand side of (B.0.3) and get the inclusion:

$$\begin{aligned} \{Q_i \neq 0\} &\subset \{B_i < \gamma_1\} \cup \{A_i < \gamma_1\} \cup \{B_i > \gamma_2\} \cup \{A_i > \gamma_2\} \\ &= \{\min(B_i, A_i) < \gamma_1\} \cup \{\max(B_i, A_i) > \gamma_2\} \end{aligned} \tag{B.0.4}$$

In the sequel we note $d_i = \min(B_i, A_i)$ and $D_i = \max(B_i, A_i)$. Then we have:

$$P(Q_i \neq 0) \leq P(\{d_i < \gamma_1\} \cup \{D_i > \gamma_2\}), \quad i = k, \dots, n^\star. \tag{B.0.5}$$

This implies the following:

$$P\Big(\sum_{j=i}^{i+c-1} \mathbf{1}(Q_j \neq 0) \geq c^\star\Big) \leq P\Big(\sum_{j=i}^{i+c-1} \mathbf{1}(\{d_j < \gamma_1\} \cup \{D_j > \gamma_2\}) \geq c^\star\Big) \tag{B.0.6}$$

Now we can bound the right-hand side of Equation B.0.6:

$$
\begin{aligned}
&P\Big(\sum_{j=i}^{i+c-1} \mathbf{1}(\{d_j < \gamma_1\} \cup \{D_j > \gamma_2\}) \geq c^\star\Big) \\
\leq &P\Big(\sum_{j=i}^{i+c-1} \mathbf{1}(\{d_j < \gamma_1\}) + \mathbf{1}(\{D_j > \gamma_2\}) \geq c^\star\Big) \\
\leq &P\Big(\sum_{j=i}^{i+c-1} \mathbf{1}(\{d_j < \gamma_1\}) \geq c^\star/2\Big) + P\Big(\sum_{j=i}^{i+c-1} \mathbf{1}(\{D_j > \gamma_2\}) \geq c^\star/2\Big)
\end{aligned}
\tag{B.0.7}
$$

Then we can express the right-hand side of Equation B.0.7 as:

$$
\begin{aligned}
&P\Big(\sum_{j=i}^{i+c-1} \mathbf{1}(\{d_j < \gamma_1\}) \geq c^\star/2\Big) + P\Big(\sum_{j=i}^{i+c-1} \mathbf{1}(\{D_j > \gamma_2\}) \geq c^\star/2\Big) \\
= &P(d_{i(c^\star/2)} < \gamma_1) + P(D_{i(c-c^\star/2)} > \gamma_2)
\end{aligned}
\tag{B.0.8}
$$

Finally we have:

$$
\begin{aligned}
&P(\exists i \in \{k, \dots, n^\star\}, \quad \sum_{j=i}^{i+c-1} \mathbf{1}(Q_j \neq 0) \geq c^\star) \\
= &P\Big(\bigcup_{i=k}^{n^\star} \Big\{ \sum_{j=i}^{i+c-1} \mathbf{1}(Q_j \neq 0) \geq c^\star \Big\}\Big) \\
\leq &P\Big(\bigcup_{i=k}^{n^\star} \Big\{ \sum_{j=i}^{i+c-1} \mathbf{1}(\{d_j < \gamma_1\}) \geq c^\star/2 \Big\}\Big) + P\Big(\bigcup_{i=k}^{n^\star} \Big\{ \sum_{j=i}^{i+c-1} \mathbf{1}(\{D_j > \gamma_2\}) \geq c^\star/2 \Big\}\Big) \\
= &P\Big(\bigcup_{i=k}^{n^\star} \{d_{i(c^\star/2)} < \gamma_1\}\Big) + P\Big(\bigcup_{i=k}^{n^\star} \{D_{i(c-c^\star/2)} > \gamma_2\}\Big) \\
= &P\Big(\min_{i=k,\dots,n^\star} d_{i(c^\star/2)} < \gamma_1\Big) + P\Big(\max_{i=k,\dots,n^\star} D_{i(c-c^\star/2)} > \gamma_2\Big) \\
= &\frac{\alpha}{2} + \frac{\alpha}{2} = \alpha
\end{aligned}
\tag{B.0.9}
$$

We go from line 2 to line 3 using Equations (B.0.6) and (B.0.7). We go from line 3 to line 47 using Equation (B.0.8). Finally, we go from line 5 to 6 using the thresholds γ_1 and γ_2 of Proposition 3. It finishes the proof. □

C Derivation of the Global Binding and Unbinding Rates

We show that under Assumption 3 the equalities (8.2.6) hold. Let begin with the equality involving γ_1.

$$\begin{aligned}
&P(\phi(X_{t+h}) = 1|\phi(X_t) = 0)\\
=&P(\phi(X_{t+h}) = 1, X_{t+h} \in \mathcal{S}|\phi(X_t) = 0) + P(\phi(X_{t+h}) = 1, X_{t+h} \in \bar{\mathcal{S}}|\phi(X_t) = 0)\\
=&P(\phi(X_{t+h}) = 1|\phi(X_t) = 0, X_{t+h} \in \mathcal{S})P(X_{t+h} \in \mathcal{S}|\phi(X_t) = 0) + 0\\
=&(k^+h + o(h))P(X_{t+h} \in \mathcal{S})
\end{aligned} \tag{C.0.1}$$

The particle can not be trapped outside $\mathcal{S}$. Then the second probability of the sum is zero. We go from line 3 to line 4 as the probability to be in region $\mathcal{S}$ at $t+h$ is independent from the event $\{\phi(X_t) = 0\}$ to be a free at t. Normally reflected Brownian motion in a domain $\mathcal{D}$ with finite volume has a stationnary distribution. This distribution is the uniform distribution over $\mathcal{D}$ [Pinsky, 2003]. Then with Assumption 3 we get:

$$\begin{aligned}
P(\phi(X_{t+h}) = 1|\phi(X_t) = 0) &= (k^+h + o(h))\frac{|\mathcal{S}|}{|\mathcal{D}|}\\
&= k^+\frac{|\mathcal{S}|}{|\mathcal{D}|}h + o(h)
\end{aligned} \tag{C.0.2}$$

We deduce that:

$$\gamma_1 = k^+\frac{|\mathcal{S}|}{|\mathcal{D}|} \tag{C.0.3}$$

Now we prove the second equality of (8.2.6) involving γ_2.

$$\begin{aligned}
&P(\phi(X_{t+h}) = 0|\phi(X_t) = 1)\\
=&P(\phi(X_{t+h}) = 0, X_{t+h} \in \mathcal{S}|\phi(X_t) = 1) + P(\phi(X_{t+h}) = 0, X_{t+h} \in \bar{\mathcal{S}}|\phi(X_t) = 1)\\
=&P(\phi(X_{t+h}) = 0|\phi(X_t) = 1, X_{t+h} \in \mathcal{S})P(X_{t+h} \in \mathcal{S}|\phi(X_t) = 1) + 0\\
=&(k^-h + o(h)) \times 1
\end{aligned} \tag{C.0.4}$$

We go from line 3 to line 4 as $\{\phi(X_t) = 1\} \subset \{X_{t+h} \in \mathcal{S}\}$. Then we get the result.

D Non Parametric Estimate of the Drift Function

We derive the estimate of the drift function μ defined as,

$$\mu(x) = \lim_{\Delta \to 0} \frac{1}{\Delta} E(X_{t+\Delta} - X_t | X_t = x), \tag{D.0.1}$$

under the model presented in Section 9.7. First we define the function

$$\tilde{\phi}(x, i, t) = \begin{cases} 1 & \text{if } X_t^{(i)} \in S(x, r) \\ 0 & \text{otherwise,} \end{cases} \tag{D.0.2}$$

where $X_t^{(i)}$ is the position of the i^{th} trajectory $\mathbb{X}_{n_i}^{(i)}$ at time t and $S(x, r)$ is a square centred in $x \in \mathbb{R}^2$ of side r. The function $\tilde{\phi}(x, i, t)$ indicates if the position $X_t^{(i)}$ of trajectory i is inside the square $S(x, r)$ at time t. Then we set:

$$N(x, r) = \sum_{i=1}^{m} \sum_{j=1}^{n_i - 1} \tilde{\phi}(x, i, t), \tag{D.0.3}$$

the number of points of the trajectory collection $\mathcal{X}_m$ falling in the square $S(x, r)$ during the whole period of observation. We note that a single trajectory can have multiple points inside $S(x, r)$ over time. Finally the estimate of the drift is given by:

$$\hat{\mu}(x) = \frac{1}{N(x, r)\Delta} \sum_{i=1}^{m} \sum_{j=1}^{n_i - 1} (X_{t_j + \Delta} - X_{t_j}) \tilde{\phi}(x, i, t_j) \tag{D.0.4}$$

which is the mean of all the displacements starting in the square $S(x, r)$.
From the law of large numbers we have:

$$\hat{\mu}(x) \to \frac{1}{\Delta} E(X_{t+\Delta} - X_t | X_t \in S(x, r)), \tag{D.0.5}$$

as $N(x, r) \to \infty$ noting that $N(x, r)$ is random as well. In other words, as the number of trajectory points falling inside $S(x, r)$ tend to infinity the convergence D.0.5 holds. We note that we also need $\Delta \to 0$ and $r \to 0$ for $\hat{\mu}(x)$ to converge towards $\mu(x)$. In a biological experiment, we can not make the temporal resolution Δ tend to 0 (see Remark 4.1.1). However, we can select parameter r. The choice of r is a compromise between bias and variance. A low r will reduce the bias of $\hat{\mu}(x)$ as $S(x, r)$ gets closer to x but at the same time few points will fall inside $S(x, r)$ and the variance of $\hat{\mu}(x)$ will increase.

List of Publications

International journals

- Briane V., C. Kervrann and M. Vimond. *A sequential algorithm to detect motion switching along intracellular particle trajectories.* Submitted.
- Briane V., C. Kervrann and M. Vimond. *A statistical analysis of particle trajectories in living cells.* In minor revisions in *Physical Review E.*

International conferences

- Briane V., C. Kervrann and M. Vimond. *Classification of particle trajectories in living cell.* Quantitative Bio Imaging (QBI), A&M University, College station, Texas, United-States, January 2017.
- Briane V., C. Kervrann and M. Vimond. *An adaptive statistical test to detect non Brownian diffusion from particle trajectories.* Spatial Statistic and Image Analysis for Biology (SSIAB), Rennes, France, May 2016.
- Briane V., C. Kervrann and M. Vimond. *An adaptive statistical test to detect non Brownian diffusion from particle trajectories.* In Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on (pp. 972-975). IEEE. International Symposium on Biomedical Imaging(ISBI), Prague, Czech republic, April 2016.

National conferences

- Briane V., C. Kervrann and M. Vimond. *Test statistique pour détecter les diffusions non browniennes.* In 48e Journées de Statistique (pp. 1-6). Journées de la Statistique (JDS), Montpellier, June 2016.
- Briane V., C. Kervrann and M. Vimond. *Classification of particle trajectories in living cell.* Molecule Trajectories in Cellular Spaces, France Bio Imaging (FBI), ENS, Lyon, November 2015.
- Briane V., C. Kervrann and M. Vimond. *A sequential algorithm to detect motion switching along intracellular particle trajectories.* Molecule dynamics in living cells, from models to experiments, GDR IMABIO, Genopolys, Montpellier, December 2017.

Bibliography

A. Baddeley, I. Bárány, and R. Schneider. Spatial point processes and their applications. *LECTURE NOTES IN MATHEMATICS-SPRINGER-VERLAG-*, 1892:1, 2007.

H. Bailey, B. R. Mate, D. M. Palacios, L. Irvine, S. J. Bograd, D. P. Costa, et al. Behavioural estimation of blue whale movements in the northeast pacific from state-space model analysis of satellite tracks. *Endangered Species Research*, 10(1):93–106, 2009.

I. Basawa and B. Prakasa Rao. *Statistical Inferences for Stochasic Processes*. Academic, 1980.

A. Basset, P. Bouthemy, J. Boulanger, F. Waharte, C. Kervrann, and J. Salamero. Detection and estimation of membrane diffusion during exocytosis in TIRFM image sequences. In *Proc. 12th IEEE International Symposium on Biomedical Imaging, ISBI 2015*, pages 695–698, Brooklyn, NY, USA, 2015.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83, 2000.

J. O. Berger and A. Guglielmi. Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, 96(453):174–184, 2001.

H. Berry and H. Chaté. Anomalous diffusion due to hindering by mobile obstacles undergoing brownian motion or ornstein-ulhenbeck processes. *Physical Review E*, 89 (2):022708, 2014.

B. M. Bibby and M. Sørensen. Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, pages 17–39, 1995.

T. Bickel. A note on confined diffusion. *Physica A: Statistical Mechanics and its Applications*, 377(1):24–32, 2007.

P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

A. Borodin and P. Salminen. *Handbook of Brownian Motion-Facts and Formulae*. Birkhäuser, 1996.

J. Boulanger, C. Gueudry, D. Münch, B. Cinquin, P. Paul-Gilloteaux, S. Bardin, C. Guérin, F. Senger, L. Blanchoin, and J. Salamero. Fast high-resolution 3d total internal reflection fluorescence microscopy by incidence angle scanning and azimuthal averaging. *Proceedings of the National Academy of Sciences*, 111(48):17164–17169, 2014.

P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.

P. C. Bressloff. *Stochastic Processes in Cell Biology*, volume 41. Springer, 2014.

F. Cao, J. Delon, A. Desolneux, P. Musé, and F. Sur. A unified framework for detecting groups and application to shape recognition. *Journal of Mathematical Imaging and Vision*, 27(2):91–119, 2007.

H. Cao and W. B. Wu. Changepoint estimation: another look at multiple testing problems. *Biometrika*, page asv031, 2015.

D. Chandler. Introduction to modern statistical mechanics. *Introduction to Modern Statistical Mechanics, by David Chandler, pp. 288. Foreword by David Chandler. Oxford University Press, Sep 1987. ISBN-10: 0195042778. ISBN-13: 9780195042771*, page 288, 1987.

N. Chenouard, I. Smal, F. De Chaumont, M. Maška, I. F. Sbalzarini, Y. Gong, J. Cardinale, C. Carthel, S. Coraluppi, M. Winter, et al. Objective comparison of particle tracking methods. *Nature Methods*, 11(3):281, 2014.

N. Chenouard et al. Multiple hypothesis tracking for cluttered biological image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 35(11): 2736–3750, 2013.

J.-F. Coeurjolly. Estimating the parameters of a fractional brownian motion by discrete variations of its sample paths. *Statistical Inference for stochastic processes*, 4(2):199–227, 2001.

L. Coutin and Z. Qian. Stochastic analysis, rough path analysis and fractional brownian motions. *Probability theory and related fields*, 122(1):108–140, 2002.

M. Dahan, S. Levi, C. Luccardini, P. Rostaing, B. Riveau, and A. Triller. Diffusion dynamics of glycine receptors revealed by single-quantum dot tracking. *Science*, 302 (5644):442–445, 2003.

M. Daszykowski, B. Walczak, and D. Massart. Looking for natural patterns in data: Part 1. density-based approach. *Chemometrics and Intelligent Laboratory Systems*, 56(2):83–92, 2001.

L. Decreusefond et al. Stochastic analysis of the fractional brownian motion. *Potential analysis*, 10(2):177–214, 1999.

X. Descombes. *Stochastic geometry for image analysis*. John Wiley & Sons, 2013.

A. Desolneux, L. Moisan, and J.-M. More. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):508–513, 2003a.

A. Desolneux, L. Moisan, and J.-M. Morel. Maximal meaningful events and applications to image analysis. *Annals of Statistics*, pages 1822–1851, 2003b.

J. A. Dix and A. Verkman. Crowding effects on diffusion in solutions and cells. *Annu. Rev. Biophys.*, 37:247–263, 2008.

A. Einstein. On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat. *Annalen der physik*, 17:549–560, 1905.

M. El Beheiry, S. Türkcan, M. U. Richly, A. Triller, A. Alexandrou, M. Dahan, and J.-B. Masson. A primer on the bayesian approach to high-density single-molecule trajectories analysis. *Biophysical journal*, 110(6):1209–1215, 2016.

T. C. Elston. A macroscopic description of biomolecular transport. *Journal of Mathematical Biology*, 41(3):189–206, 2000.

M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

T. J. Feder, I. Brust-Mascher, J. P. Slattery, B. Baird, and W. W. Webb. Constrained diffusion or immobile fraction on cell surfaces: a new interpretation. *Biophysical Journal*, 70(6):2767, 1996.

A. Fick. V. on liquid diffusion. *Philosophical Magazine Series 4*, 10(63):30–39, 1855.

H. Finner and M. Roters. On the false discovery rate and expected type i errors. *Biometrical Journal*, 43(8):985–1005, 2001.

D. Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics: A Journal of Theoretical and Applied Statistics*, 20(4):547–557, 1989.

J. H. Friedman, F. Baskett, and L. J. Shustek. An algorithm for finding nearest neighbors. *IEEE Transactions on computers*, 100(10):1000–1006, 1975.

C. Fuchs. *Inference for Diffusion Processes: With Applications in Life Sciences*. Springer Science & Business Media, 2013.

N. Gal, D. Lechtman-Goldstein, and D. Weihs. Particle tracking in living cells: a review of the mean square displacement method and beyond. *Rheologica Acta*, 52(5):425–443, 2013.

L. Gallardo. *Mouvement brownien et calcul d'Itô: cours et exercices corrigés*. Hermann, 2008.

C. R. Genovese, K. Roeder, and L. Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

A. Gidon, S. Bardin, B. Cinquin, J. Boulanger, F. Waharte, L. Heliot, H. Salle, D. Hanau, C. Kervrann, B. Goud, et al. A rab11a/myosin vb/rab11-fip2 complex frames two late recycling steps of langerin from the erc to the plasma membrane. *Traffic*, 13(6):815–833, 2012.

A. Grandhi. *Multiple Testing Procedures for Complex Structured Hypotheses and Directional Decisions*. New Jersey Institute of Technology, 2015.

P. Grassberger. Conductivity exponent and backbone dimension in 2-d percolation. *Physica A: Statistical Mechanics and its Applications*, 262(3):251–263, 1999.

W. Guo and J. P. Romano. On stepwise control of directional errors under independence and some dependence. *Journal of Statistical Planning and Inference*, 163:21–33, 2015.

S. Havlin and D. Ben-Avraham. Diffusion in disordered media. *Advances in Physics*, 36 (6):695–798, 1987.

J. M. Henley, E. A. Barker, and O. O. Glebov. Routes, destinations and delays: recent advances in ampa receptor trafficking. *Trends in Neurosciences*, 34(5):258–268, 2011.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

N. Hozé. *Modélisation et méthodes d'analyse de la diffusion et agrégation au niveau moléculaire pour l'organisation sous-cellulaire*. PhD thesis, Université Pierre et Marie Curie-Paris 6, 2013.

N. Hoze, D. Nair, E. Hosy, C. Sieben, S. Manley, A. Herrmann, J.-B. Sibarita, D. Choquet, and D. Holcman. Heterogeneity of ampa receptor trafficking and molecular

interactions revealed by superresolution analysis of live cell imaging. *Proceedings of the National Academy of Sciences*, 109(42):17052–17057, 2012.

H. E. Hurst. Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.*, 116:770–808, 1951.

J.-H. Jeon, V. Tejedor, S. Burov, E. Barkai, C. Selhuber-Unkel, K. Berg-Sørensen, L. Oddershede, and R. Metzler. In vivo anomalous diffusion and weak ergodicity breaking of lipid granules. *Physical Review Letters*, 106(4):048103, 2011.

G. J. Jiang and J. L. Knight. A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model. *Econometric Theory*, 13(05):615–645, 1997.

S. Karlin. *A Second Course in Stochasic Processes*. Academic, 1981.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

C. Kervrann, C. Sorzano, S. Acton, J.-C. Olivo-Marin, and M. Unser. A guided tour of selected image processing and analysis methods for fluorescence and electron microscopy. *IEEE Journal of Selected Topics in Signal Processing*, 10(1):6–30, 2016.

F. Klebaner et al. *Introduction to Stochastic Calculus with Applications*. Imperial College Press, 2012.

A. N. Kolmogorov. The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers. In *Dokl. Akad. Nauk SSSR*, volume 30, pages 299–303, 1941.

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

H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

A. Kusumi, Y. Sako, and M. Yamamoto. Confined lateral diffusion of membrane receptors as studied by single particle tracking (nanovid microscopy). effects of calcium-induced differentiation in cultured epithelial cells. *Biophysical Journal*, 65(5):2021, 1993.

A. Kusumi, C. Nakada, K. Ritchie, K. Murase, K. Suzuki, H. Murakoshi, R. S. Kasai, J. Kondo, and T. Fujiwara. Paradigm shift of the plasma membrane concept from the two-dimensional continuum fluid to the partitioned fluid: high-speed single-molecule tracking of membrane molecules. *Annu. Rev. Biophys. Biomol. Struct.*, 34:351–378, 2005.

T. Lagache, E. Dauty, and D. Holcman. Quantitative analysis of virus and plasmid trafficking in cells. *Physical Review E*, 79(1):011921, 2009.

P. Langevin. Sur la théorie du mouvement brownien. *CR Acad. Sci. Paris*, 146(530-533): 530, 1908.

M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes*. Springer, 1983.

Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

F. Lund et al. Spattrack: An imaging toolbox for analysis of vesicle motility and distribution in living cells. *Traffic*, 15(12):1406–1429, 2014.

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

B. B. Mandelbrot and J. W. Van Ness. Fractional brownian motions, fractional noises and applications. *SIAM review*, 10(4):422–437, 1968.

J.-B. Masson, P. Dionne, C. Salvatico, M. Renner, C. G. Specht, A. Triller, and M. Dahan. Mapping the energy and diffusion landscapes of membrane proteins at the cell surface using high-density single-molecule imaging and bayesian inference: application to the multiscale dynamics of glycine receptors in the neuronal membrane. *Biophysical journal*, 106(1):74–83, 2014.

Y. Meroz and I. M. Sokolov. A toolbox for determining subdiffusive mechanisms. *Physics Reports*, 573:1–29, 2015.

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

X. Michalet. Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. *Physical Review E*, 82(4): 041914, 2010.

X. Michalet and A. J. Berglund. Optimal diffusion coefficient estimation in single-particle tracking. *Physical Review E*, 85(6):061916, 2012.

Y. Mishura. *Stochastic Calculus for Fractional Brownian Motion and Related Processes*, volume 1929. Springer Science & Business Media, 2008.

N. Monnier, S.-M. Guo, M. Mori, J. He, P. Lénárt, and M. Bathe. Bayesian approach to msd-based analysis of particle motion in live cells. *Biophysical Journal*, 103(3): 616–626, 2012.

N. Monnier, Z. Barry, H. Y. Park, K.-C. Su, Z. Katz, B. P. English, A. Dey, K. Pan, I. M. Cheeseman, R. H. Singer, et al. Inferring transient particle transport dynamics in live cells. *Nature methods*, 12(9):838–840, 2015.

D. Mumford and A. Desolneux. *Pattern theory: the stochastic analysis of real-world signals*. CRC Press, 2010.

D. Nualart and Y. Ouknine. Regularization of differential equations by fractional noise. *Stochastic Processes and their Applications*, 102(1):103–116, 2002.

N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

T. Pécot, L. Zengzhen, J. Boulanger, J. Salamero, and C. Kervrann. Quantev: quantifying the spatial distribution of intracellular events. 2017. preprint on webpage at https://hal.inria.fr/hal-01575913/.

C. S. Peskin and G. Oster. Coordinated hydrolysis explains the mechanical behavior of kinesin. *Biophysical Journal*, 68(4 Suppl):202S, 1995.

R. G. Pinsky. Asymptotics of the principal eigenvalue and expected hitting time for positive recurrent elliptic operators in a domain with a small puncture. *Journal of Functional Analysis*, 200(1):177–197, 2003.

A. S. Pisarev, S. A. Rukolaine, A. M. Samsonov, and M. G. Samsonova. Numerical analysis of particle trajectories in living cells under uncertainty conditions. *Biophysics*, 60(5):810–817, 2015. ISSN 1555-6654. doi: 10.1134/S0006350915050176. URL http://dx.doi.org/10.1134/S0006350915050176.

M. Pollak and D. Siegmund. A diffusion process and its applications to detecting a change in the drift of brownian motion. *Biometrika*, 72(2):267–280, 1985.

H. Qian, M. P. Sheetz, and E. L. Elson. Single particle tracking. analysis of diffusion and flow in two-dimensional systems. *Biophysical Journal*, 60(4):910, 1991.

L. Rabiner and B. Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

D. Rasch. Hypothesis testing and the error of the third kind. *Psychological Test and Assessment Modeling*, 54(1):90–99, 2012.

P. Reimann. Brownian motors: noisy transport far from equilibrium. *Physics Reports*, 361(2):57–265, 2002.

E. Roquain. Type i error rate control in multiple testing: a survey with proofs. *Journal de la Société Française de Statistique*, 152:3–38, 2011.

M. J. Saxton. Lateral diffusion in an archipelago. single-particle diffusion. *Biophysical Journal*, 64(6):1766–1780, 1993.

M. J. Saxton. Anomalous diffusion due to obstacles: a monte carlo study. *Biophysical Journal*, 66(2 Pt 1):394, 1994.

M. J. Saxton and K. Jacobson. Single-particle tracking: applications to membrane dynamics. *Annual Review of Biophysics and Biomolecular Structure*, 26(1):373–399, 1997.

J. Schafer, N. Baetz, L. Lapierre, R. McRae, J. Roland, and J. Goldenring. Rab11-fip2 interaction with myo5b regulates movement of rab11a-containing recycling vesicles. *Traffic*, 15(3):292–308, 2014.

H. Scher and E. W. Montroll. Anomalous transit-time dispersion in amorphous solids. *Physical Review B*, 12(6):2455, 1975.

R. S. Schick, S. R. Loarie, F. Colchero, B. D. Best, A. Boustany, D. A. Conde, P. N. Halpin, L. N. Joppa, C. M. McClellan, and J. S. Clark. Understanding movement data and movement processes: current and emerging directions. *Ecology letters*, 11 (12):1338–1350, 2008.

Z. Schuss. *Theory and applications of stochastic processes: an analytical approach*, volume 170. Springer Science & Business Media, 2009.

J. P. Shaffer. Control of directional errors with stagewise multiple test procedures. *The Annals of Statistics*, pages 1342–1347, 1980.

J. P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995.

V. Spokoiny. Multiscale local change point detection with applications to value-at-risk. *The Annals of Statistics*, pages 1405–1436, 2009.

M. S. Taqqu. Fractional brownian motion and long-range dependence. *Theory and applications of long-range dependence*, pages 5–38, 2003.

V. Tejedor, O. Bénichou, R. Voituriez, R. Jungmann, F. Simmel, C. Selhuber-Unkel, L. B. Oddershede, and R. Metzler. Quantitative analysis of single particle trajectories: mean maximal excursion method. *Biophysical journal*, 98(7):1364–1372, 2010.

S. Türkcan and J.-B. Masson. Bayesian decision tree for the classification of the mode of motion in single-molecule trajectories. *PLoS One*, 8(12):e82799, 2013.

G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

N. G. Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

J. J. Waterston and L. Rayleigh. On the physics of media that are composed of free and perfectly elastic molecules in a state of motion. *Philosophical Transactions of the Royal Society of London. A*, 183:1–79, 1892.

S. C. Weber, A. J. Spakowitz, and J. A. Theriot. Bacterial chromosomal loci move subdiffusively through a viscoelastic cytoplasm. *Physical review letters*, 104(23):238102, 2010.

X. Xu, M. Ester, H.-P. Kriegel, and J. Sander. A distribution-based clustering algorithm for mining in large spatial databases. In *Data Engineering, 1998. Proceedings., 14th International Conference on*, pages 324–331. IEEE, 1998.

N. L. Zaïdi, D. Nualart, et al. Smoothness of the law of the supremum of the fractional brownian motion. *Elect. Comm. Probab*, 8:102–111, 2003.

E. Zhizhina, S. Komech, and X. Descombes. Modelling axon growing using ctrw. *arXiv preprint arXiv:1512.02603*, 2015.

R. Zwanzig. *Nonequilibrium statistical mechanics*. Oxford University Press, 2001.