# Inverse problems and data assimilation methods applied to protein polymerisation

Aurora Armiento


Université Paris Diderot

**École doctorale de sciences mathématiques de Paris centre**

# Thèse de doctorat

Spécialité : Mathématiques appliquées

---

# Problèmes inverses et méthodes d'assimilation de données appliquées à la polymérisation de protéines

---

présentée par

**Aurora Armiento**

Soutenue le 13 janvier 2017 devant le jury composé de :

| | |
|---|---|
| David Matthew Bortz | Rapporteur |
| Olivier Saut | Rapporteur |
| Béatrice Laroche | Examinatrice |
| Davy Martin | Examinateur |
| Luís Neves de Almeida | Examinateur |
| Marie Doumic | Directrice de thèse |
| Philippe Moireau | Directeur de thèse |

# Résumé

Cette thèse a pour objectif la mise en place d'une stratégie mathématique pour l'étude du processus physique de l'agrégation des protéines. L'étude de ce processus largement inconnu est particulièrement importante puisqu'il a été identifiée comme un élément clé d'une vaste gamme de maladies incurables, appelées maladies amyloïdes. Les maladies à prions appartiennent à cette classe et sont causées par l'agrégation d'une configuration mal pliée de la protéine prion. Notre travail contribue à la recherche sur les maladies à prions, en se concentrant sur deux types d'agrégats : les oligomères et les fibres. Les oligomères suspectés d'être les agrégats les plus toxiques sont étudiés dans la première partie de cette thèse. Nous fondons notre travail sur l'analyse de deux types de données expérimentales. D'une part, nous considérons les données de dispersion statique de la lumière (SLS), qui peuvent être interprétées biologiquement comme la mesure de la taille moyenne des oligomères et mathématiquement comme le deuxième moment de la concentration des agrégats. D'autre part, nous considérons les données de distribution de taille d'oligomère collectées à plusieurs instants en utilisant la Chromatographie d'Exclusion de Taille (SEC). Notre étude conduit à la conclusion importante selon laquelle au moins deux types différents d'oligomères sont présents. De plus, nous proposons une description de l'interaction entre ces oligomères en proposant pour la première fois un modèle à deux espèces. Notre modèle est composé d'un ensemble d'ODE avec les taux cinétiques comme paramètres. La description qualitative fournie par ce modèle a été couplée à l'information contenue dans les données expérimentales de SLS dans le cadre de l'assimilation de données. Au moyen de la méthode du filtre de Kalman étendue, nous résolvons un problème inverse non linéaire, estimant ainsi les coefficients cinétiques associés aux données expérimentales. Pour valider ce modèle, nous avons comparé notre estimation aux données expérimentales de SEC, en observant un très bon accord entre les deux. Notre caractérisation des espèces d'oligomères peut conduire à de nouvelles stratégies pour concevoir un premier traitement ciblé pour les maladies à prions. La méthodologie appliquée à l'étude des oligomères peut être considérée comme une première étape dans l'analyse des fibres. En raison des propriétés physiques de ces agrégats, des expériences moins nombreuses et moins précises peuvent être effectuées, et une approche mathématique peut donc apporter une contribution précieuse à leur étude. Notre contribution est de proposer une stratégie générale pour estimer l'état initial d'un système de fibres. Inspiré par la théorie de Lifshitz-

Slyozov, nous décrivons ce système par une équation de transport couplée à une équation intégrale. L'estimation est faite en utilisant quelques observations empiriques sur le système. Nous considérons le cas général d'observation d'un moment d'ordre $n$. Il est en effet possible de mesurer le moment d'ordre 1 par fluorescence de thioflavine T ou le moment d'ordre 2 par SLS. Nous proposons une solution théorique et numérique du problème d'estimation de la condition initiale dans le cas linéaire d'un système de dépolymérisation. En particulier, pour des taux de dépolymérisation constants, nous proposons une stratégie de régularisation par noyau, qui fournit une première caractérisation de l'estimation. Dans le cas de taux de dépolymérisation variables, nous proposons la méthode d'assimilation variationnelle 4d-Var et la méthode d'assimilation de données séquentielle du filtrage de Kalman. Ces deux méthodes sont plus générales et peuvent être facilement adaptée pour traiter différents problèmes. Ce problème inverse est particulièrement intéressant puisqu'il peut également être appliqué dans d'autres domaines tels que le cycle cellulaire ou la formation de poussière.

**Mots clefs :** les maladies à prions, amyloïde, oligomère, hétérogénéité, problème inverse, l'estimation d'état, l'identification des paramètres, l'assimilation des données, filtre de Kalman, 4d-Var, équation du transport

# Abstract

The aim of this PhD thesis is to set up a mathematical strategy to investigate the physical process of protein aggregation. The study of this largely unknown process is particularly important since it has been identified as a key feature of a wide class of incurable diseases, called amyloid diseases. Prion diseases belong to this class and are caused by the aggregation of a misfolded configuration of the prion protein. Our work contributes to the research on prion diseases, by focusing on two kinds of aggregates : oligomers and fibrils.

Oligomers, which are suspected of being the most toxic aggregates, are studied in the first part of this thesis. We base our work on the analysis of two types of experimental data. On the one hand, we consider Static Light Scattering (SLS) data, which can be interpreted biologically as the measurement of the average oligomer size and mathematically as the second moment of aggregate concentration. On the other hand, we consider oligomer size distribution data collected at several instants by using Size Exclusion Chromatography (SEC). Our study leads to the important conclusion that at least two different types of oligomers are present. Moreover, we provide a description of the interaction between these oligomers by proposing, for the first time, a two-species model. Our model is composed of a set of ODEs with the kinetic rates as parameters. The qualitative description provided by this model has been coupled to the information contained in the noisy experimental SLS data in a data assimilation framework. By means of the extended Kalman filter method, we solve a non-linear inverse problem, thereby estimating the kinetic coefficients associated to the experimental data. To validate this model we have compared our estimation to the experimental SEC data, observing a very good agreement between the two. Our oligomer species characterisation may lead to new strategies to design a first targeted treatment for prion diseases.

The methodology applied to the study of oligomers can be seen as a first step in the analysis of fibrils. Due to the physical properties of these aggregates, fewer and less precise experiments can be performed and so a mathematical approach can provide a valuable contribution to their study. Our contribution is to propose a general strategy to estimate the initial condition of a fibril system. Inspired by the Lifshitz-Slyozov theory, we describe this system by a transport equation coupled with an integral equation. The estimation is performed making use of some empirical observations on the system. We consider the general case of observing a moment of order $n$. It is indeed possible to measure the first moment by Thioflavine T fluorescence

or the second moment by SLS. We provide a theoretical and numerical solution of the initial condition estimation problem in the linear case of a depolymerising system. In particular, for constant depolymerisation rates, we propose a kernel regularisation strategy, that provides a first characterisation of the estimation. In the variable depolymerisation rates, we outline the variational data assimilation method 4d-Var. This method is more general and can be easily adapted to treat different problems. This inverse problem is particularly interesting since it can also be applied in other fields such as the cell cycle or dust formation.

**Keywords :** prion diseases, amyloid, oligomer, heterogeneity, inverse problem, state estimation, parameter identification, data assimilation, Kalman Filter, 4d-Var, transport equation

# Remerciements

En premier lieu, je tiens à remercier mes directeurs de thèse. Merci à Marie Doumic de m'avoir fait confiance d'abord en stage, puis en thèse et de m'avoir proposé un sujet si actif et fascinant. Elle m'a beaucoup aidée dans ce passage de la vie étudiante à celle de chercheur. Avec elle, j'ai pu apprendre comment « attaquer »un problème et comment s'approcher méthodiquement des résultats espérés. Merci Philippe Moireau de m'avoir guidé dans le monde de l'assimilation de données qui était pour moi un domaine quasiment inconnu. Merci pour toutes les discussions formatrices sur les maths comme sur la vie professionnelle. Merci de m'avoir toujours poussée à faire de mon mieux. Je tiens à vous remercier sincèrement de m'avoir encadrée si bien ; si je garde un si beau souvenir de ces années de thèse c'est en grande partie grâce à vous.

Mes remerciements vont également aux rapporteurs de thèse David M. Bortz et Olivier Saut pour leur disponibilité et pour avoir accepté de faire un travail de synthèse et de critique de grande qualité, conséquence évidente d'une lecture très attentive. Je remercie aussi les autres membres du jury Luis Neves de Almeida, Beatrice Laroche et Davy Martin pour le temps et l'intérêt qu'ils ont pu déployer aboutissant à cette soutenance de thèse.

Je tiens à remercier nos collaborateurs de l'INRA. Ce n'est que grâce à cette collaboration cruciale que ce travail peut se dire vraiment interdisciplinaire. Merci de m'avoir accueillie dans le labo et surtout pendant les manips. Ces journées m'ont ouvert une nouvelle perspective sur la polymérisation de protéines. Je veux, en particulier, remercier Stéphanie Prigent qui a toujours été très gentille et m'a beaucoup aidée au début de cette thèse en répondant patiemment à mon interminable liste de questions sur les aspects biologiques, mais surtout je dois remercier Human Rezaei pour son énergie et pour son enthousiasme à propos des résultats obtenus. Pour l'étendu de sa connaissance sur les prions il a été un point de référence tout le long de ce travail de thèse.

Merci à mes collègues moniteurs de Paris 7 et aux enseignants de filière Frédéric Hélein et Muriel Livernet dont l'enthousiasme et le temps consacrés à aider les étudiants dans leur parcours méritent d'être soulignés.

Pendant ces trois ans j'ai eu la chance de travailler au sein de deux équipes Inria (MAMBA et MΞDISIM) et dans quatre endroits différents (Inria Rocquencourt, Inria Saclay, Laboratoire Jacques Louis Lions et Paris 7), ce qui m'a donné l'occasion de connaitre des personnes

merveilleuses avec qui j'ai partagé des moments très agréables. Je veux remercier toute l'équipe MΞDISIM et en particulier Dominique Chapelle. L'atmosphère positive qui règne dans cette équipe est bien rare et le résultat du grand investissement de ses « chefs ». Je garderai précieusement le souvenir de l'équipe MAMBA qui m'a chaleureusement accueillie dans ses rangs. Une pensée particulière va à mes soeurs et mon frère de thèse. Même si la première année est désormais lointaine, je n'oublie pas les belles journées passées en compagnie des équipes ANGE et REO au bâtiment 16. Pour finir, je suis ravie d'avoir pu participer à la vie du LJLL, aux séminaires comme aux pauses café j'ai pu y apprendre beaucoup sur une grande variété de domaines et de méthodes, que j'en sorte surement enrichie.

Mes profonds remerciements vont à Richard James pour ses précieux conseils en anglais. Je suis vraiment fascinée par l'engagement et la passion qu'il met dans son travail.

Je n'aurais jamais pu mener ce travail de thèse à terme sans mes amis, qui ont toujours été à mon côté dans les moments de fête comme dans les moments de doute. *Acheter un poisson rouge, apprendre à préparer la pizza, l'agenda partagée, les fêtes de Noel, le diner de sofficini, le mariage au bord de mer, découvrir Bordeaux, les brunchs chez Marc, les confessions dans les tours de Jussieu, les vignobles de la Bourgogne, une soirée en piazza Vettovaglie comme au bon vieux temps, jouer aux cartes avec un verre de vin, s'étonner aux côtes de la Normandie, la première fois aux USA, le ramadan à l'appart, Copenhague, une pancarte pour Marseille, les appelles qui durent heures, l'opéra, s'acculturer sur la taille des bulles de champagne, « Peux-tu chanter joyeux anniversaire avec moi ? », les commentaires aux expos, « pour une fois qu'Andrea vient à Paris il faut sortir », l'irlandais en face, un cours de musique au bureau, le musée Horta, les pique-niques de l'été, les rigolades au bureau*...à vous, vous avec qui j'ai pu partager tout ça : MERCI !!

Pour finir je veux remercier ma famille. Je vous dois tout. Merci pour votre amour inconditionnel, merci pour vos conseils et pour votre support. Vous m'avez aidée énormément avec vos mots, vos silences et vos gestes. Vous êtes toujours prêts à faire et vous avez fait les plus grands sacrifices pour moi et je ne pourrai jamais vous remercier assez. J'espère seulement réussir à vous retourner tout l'amour que vous m'apportez. Cette thèse vous est dédiée.

# Table des matières

# Introduction

This PhD thesis is the result of the work I carried out under the surpervision of Marie Doumic, the Head of the Inria team MAMBA, and Philippe Moireau a senior researcher in the Inria team MΞDISIM. The main objective of this work has been to apply the mathematical framework of data assimilation to explore the largely unknown phenomenon of protein polymerisation.

## 0.1 Motivation

Protein polymerisation is a phenomenon of major importance, since it has been identified as one of the causes of a class of diseases called *amyloid diseases*. Protein aggregation occurs when some protein, which is naturally present in healthy organisms in a monomeric configuration, misfolds. In the misfolded (ill) configuration, proteins are able to bind to other proteins and, in this way, propagate the disease. In fact, when an ill protein aggregates with a healthy protein, it is able to make the healthy protein assume the misfolded configuration.

Our understanding of the mechanisms that propagates these diseases is far from complete and there is, as yet, no cure for these diseases. Many fundamental questions still need to be answered, and perhaps the most important question is which are the most toxic aggregates ? In other words, which aggregates can propagate the disease the fastest ? This question will be rephrased in our work as which aggregates are associated with the highest aggregation rate ?

Once these objects have been identified, a natural question arises. How can we attack or destroy them ? In order to design treatment, we need to know how these toxic aggregates interact with the other aggregates or with the monomeric proteins. How does the aggregation take place ? Do aggregates grow by sequential addition of monomers ? Alternatively, do they attach themselves to dimers, trimers or $i$-mers ? How many different kinds of aggregates are there ? How do these aggregates interact with each other ? Once large aggregates have formed, do they disintegrate, do they break up, or do they lose small pieces ? In the last case, what is the typical size of these small pieces ?

This is, of course, only a small selection of all the possible questions that are currently being investigated by biologists, physicists and mathematicians. Multidisciplinary collaboration is

essential if we wish to obtain a better understanding of this complex phenomenon that can be characterised by highly heterogeneous aggregates both in terms of structural configuration and in terms of size. In our work, we collaborate with a team of biophysicists from INRA. Thanks to this collaboration we have been able to study *in vitro* experiments on ovine prion (ovPrP) and propose an adequate mathematical model of their aggregation mechanisms.

## 0.2 A brief introduction to prion diseases

This section presents a brief overview of discoveries and hypotheses concerning prion diseases and, more generally, amyloid diseases. Our work aims at making a positive contribution to the decades of research that have been carried out on this important subject.

Prion diseases, also known as transmissible spongiform encephalopathies (TSEs), are fatal neurodegenerative disorders affecting both humans and animals. Prion diseases are caused by the aggregation of an abnormally folded cellular prion protein.

The prion protein, called $PrP^{C}$, is mostly expressed in the central nervous system but it can also be found in the lymphoreticular tissue, skeletal muscle, kidney, heart, skin, mammary gland, digestive tract and endothelia. The physiological function of this protein has not yet been precisely characterised.

The pathogenic form is denoted by $PrP^{Sc}$, with 'sc' standing for scrapie, the prototype prion disease occurring in sheep and goats. The toxicity of this $PrP^{Sc}$ is due to its ability to convert healthy $PrP^{C}$ cellular proteins into ill proteins and, consequently, to propagate the disease. $PrP^{Sc}$ has a strong tendency to aggregate and exhibits high resistance to heat and chemical denaturation [143, 29]. Its resistance to protease digestion and insolubility in non-ionic detergents makes it particularly difficult to attack.

The connection between prion diseases and the prion protein $PrP^{C}$ was well established in the work [144]. It was shown that mice developed resistance to experimental prion disease when the expression of the prnp gene, encoding for $PrP^{C}$, was reduced using a gene knockdown technique. However, the identification of the infectious agent and, in particular, understanding whether $PrP^{Sc}$ is able to cause the disease by itself or needs other cofactors remain controversial issues.

Initially, TSEs were associated with slow viruses because of the long incubation periods. Experiments designed to disrupt large molecules using electron beams were used in the study of TSEs. These experiments showed that the size of the infectious agent was very small and, tellingly, much smaller than a virus. Taking into account these results, G.W. Outram and A.G. Dickinson in 1979 [57] proposed the existence of *virinos*, small infectious particles with a vector-like nature. Focusing on scrapie, they conjectured that the scrapie agent bound to host proteins. In this way, it is seen as legitimate by the host and the immune response is not activated.

In 1967 J.S. Griffith and T. Alper, for the first time, presented the possibility that proteins can self-replicate without the presence of nucleic acids [79, 4]. This hypothesis was revolutionary and in fact went against the *central dogma of molecular biology*[1].

In 1982 S.B. Prusiner purified a mainly proteinaceous infectious agent. He coined the word *prion* from PRoteinaceous Infectious ONly [141]. For his work on prions, Prusiner won the

1. *"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid." [50].*

Nobel prize in Physiology or Medicine in 1997.

Today, Griffith's hypothesis, even if not completely proven, has been supported by several studies [106], and is generally accepted. Nevertheless, studies exploring the virus-like hypothesis are currently being carried out [115, 101].

Kuru, Creutzfeldt-Jakob disease (CJD), Gerstmann-Strussler-Scheinker syndrome (GSS), and familial fatal insomnia (FFI) have been proved to belong to prion diseases. Last year, in [145, 77] the authors demonstrated the existence of a new human prion ($\alpha$-*syn*) responsible for Multiple System Atrophy (MSA) in humans affected by Parkinson's disease. MSA has subsequently been identified as a prion disease. The wide spectrum of prion diseases has been explained by the variety of ways in which the $PrP^{C}$ can fold. Is has been proved that not all the misfolded forms are pathogenic, but each pathogenic strait corresponds to a precise disease phenotype [2, 143, 173].

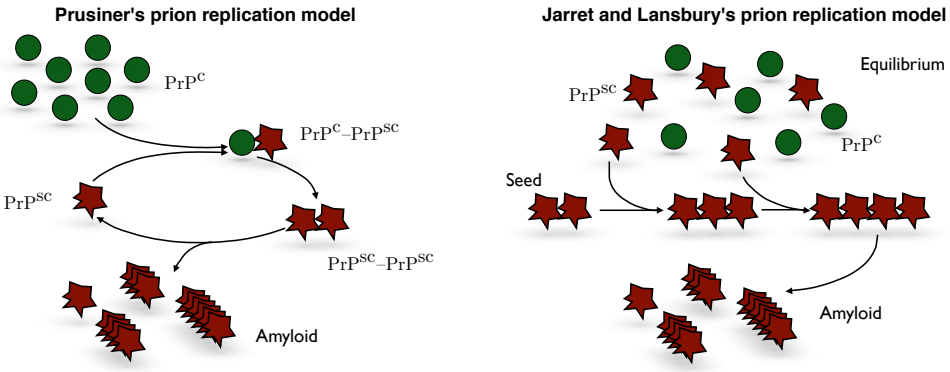
Like for sporadic CJD and MSA, other neurodegenerative diseases such as Alzheimer's disease (AD), fronto-temporal dementia (FTD), amyotrophic lateral sclerosis (ALS) and Parkinson's disease (PD) may be characterised by the aggregation of misfolded proteins into large ordered structures that accumulate in the brain. These aggregates can spread and progressively affect the whole neuronal network causing neuronal loss [49], the typical characteristic of all neurodegenerative diseases. The similarity to prion disease mechanisms explains the name 'prion-like' or 'prionoids' to denote the misfolded proteins [2]. More than 20 misfolded proteins have been identified in humans and animals, and include $\beta$-amyloid and $\tau$ responsible for AD, TDP-43 for ALS, and FTD and huntingtin for Huntington disease.

In contrast to prions, until recently we had no evidence for the transmissibility of prionoids. A preliminary study, published this year, suggests that Alzheimer's disease can spread from one person to another [101].

Prion diseases can be transmitted in two ways : *vertically* when the disorder is inherited and *horizontally* when infected by other animals from the same or different species. BSE, commonly known as "mad cow disease" was first reported in the UK in 1986 and soon spread worldwide, becoming endemic in many countries. In this case, the cause of the epidemic was attributed to the fact that the food given to cattle contained contaminated meat. BSE was then transmitted to humans (vCJD), cats and zoo animals.

The mechanisms leading to the conversion of $PrP^{C}$ into $PrP^{SC}$ and the capability of aggregating are still unclear. Research results indicate a common pattern for disease propagation in prion as well as in prion-like diseases. The proteins associated to the diseases, in their misfolded or partially unfolded configuration, can template to the normally folded proteins and interact with them to form cross-beta sheets, a common motif of regular secondary structure in proteins [2, 143, 57].

Two models have been suggested for this process. The first one was proposed by Jarrett and Lansbury (1993) for prion diseases [95]. Both $PrP^{C}$ and $PrP^{SC}$ are assumed to exist naturally in the human organism. The two forms are in a reversible thermodynamic equilibrium that is perturbed in the presence of $PrP^{SC}$ aggregates, which provides favourable conditions for the conversion from $PrP^{C}$ to $PrP^{SC}$. It should be pointed out that in this model the disease needs some $PrP^{SC}$ aggregates (*seeds*) to start. The second model was presented by Prusiner in 1991 [142]. In this second model, an initial slow reaction with high activation energy converts $PrP^{C}$ into $PrP^{SC}$. The protein $PrP^{SC}$ is able to template and it forms a dimer with $PrP^{C}$ ($PrP^{C}$–$PrP^{SC}$). In the dimeric configuration $PrP^{C}$ converts faster than the initial reaction. The newly created $PrP^{SC}$–$PrP^{SC}$ dimer then dissociates and allows the formation of new $PrP^{SC}$–$PrP^{C}$ dimers propagating the disease. In this case, $PrP^{SC}$ seeds are not necessary.




Once the $PrP^{SC}$ proteins start to aggregate, they form structures called *oligomers* or *polymers*, depending on the number of proteins composing the aggregates. Typically, $PrP^{SC}$ aggregates grow along a single axis creating organised filamentous structures called *fibrillary filaments* or *fibrils*. Fibrils can interact with each other and form higher order fibrillary aggregates called *amyloids* [140]. Fibrils can also break when they reach a critical length and in this way accelerate the propagation of the disease.

The toxicity associated to the deposits of protein aggregates in the tissues is still unknown [180]. However, the exponential production of $PrP^{SC}$ explains the rapid evolution of CJD which leads to death within a few months of the onset of the disease [152]. In scrapie, after an incubation period of 2-5 years, the affected animals die within 6 months of the onset.

Treatments preventing the formation of aggregates are likely to lead to positive results. In [76] the authors present the discovery of HSP104, HSP70, and HSP40 disaggregases, which are able to dissolve cytosolic aggregates such as yeast prions. It has been noticed that – overexpressing HSP104 – the yeast becomes immune to prions, as expected. However, the suppression of Hsp-104 also leads to the same result [174]. These studies suggest that a dynamical aggregation-fragmentation process is essential to explain prion disease.

In an attempt to cure prion and prion-like diseases, various approaches to design medical treatment are currently being studied all over the world [128, 100, 82, 28]. At present, the therapies in prion and prion-like diseases are only able to cure or at least relieve the symptoms. A better understanding of the mechanisms governing prion propagation is essential to design new treatments as this knowledge could then be used to interpret the prion-like diseases.

## 0.3 Objectives and contributions

The main objective of this thesis is to set up a mathematical strategy to investigate the physical process of protein aggregation. The aim is not to develop new methods but, rather, to apply existing methods to a new problem.

We focus on *prion diseases*, which belong to the class of amyloid diseases. We refer to experimental data on ovine prion proteins. In particular, we analyse the behaviour of prion oligomers. Increasing evidence has shown that the most infectious factor is the smaller subfibrillar oligomers formed by prion proteins [155].

We aim at designing a mathematical model able to capture the principal features of prion oligomer evolution and reproduce the *in vitro* experimental results.

With no existing model for this oligomer system, we have based our study on two essential elements. On the one hand, we consider the mathematical models proposed for larger prion aggregates called polymers or fibrils [129, 116, 139, 26, 179, 116, 78], on the other hand, the experimental data provided by the team led by Dr. Human Rezaei at INRA. The experiments carried out by this team were vital in order to gain a better understanding of this phenomenon. The value of these data comes from the fact that the experiments were designed in tandem with our collaborators to address the specific questions raised in the course of our research.

Our study reveals the need to model at least two oligomer species. We have called these two species *stable oligomers* and *unstable oligomers*. The former grow by gaining one monomer at a time and shrink by losing one monomer at a time. Unstable oligomers can gain or lose one monomer at a time, similarly to stable oligomers, and, most importantly, can completely disintegrate into monomers. The two species interact through the exchange of monomers. In fact, the oligomers of both species are made up of the same kind of monomers and in their evolution contribute to and draw from the same monomer reservoir.

All chemical reactions involving oligomers of size $i$ are associated to rates that are potentially dependent on the size. These rates govern the behaviour of the oligomer system and constitute the set of parameters of our model.

Many prion polymer models have been proposed in the literature. In particular, we find ordinary differential equation systems [129, 116]. The main drawback of such models is that they require a number of equations that is at least equal to the number of aggregate sizes. Therefore, they are unsuitable when studying large aggregates, due to their high computational cost. A big advantage in working on oligomers is that these aggregates are made up, at most, of several hundred monomers. Since the monomers are the composing unit of the oligomers, we define the oligomer size as the number of monomers forming the oligomers.

Inspired by the Becker-Döring theory [19], we propose an ODE model consisting of a differential equation for each oligomer size, for each oligomer species and a differential equation for the monomers. It reads as follows

$$\begin{cases} \dot{w}_i &= -k_{\mathrm{dis}i} w_i + k_{\mathrm{on}_{i-1}}^{w} w_{i-1} v - k_{\mathrm{on}_i}^{w} w_i v + k_{\mathrm{dep}_{i+1}}^{w} w_{i+1} - k_{\mathrm{dep}_i}^{w} w_i, & i \in [i_0, i_1] \\ \dot{y}_i &= k_{\mathrm{on}_{i-1}}^{y} y_{i-1} v - k_{\mathrm{on}_i}^{y} y_i v + k_{\mathrm{dep}_{i+1}}^{y} y_{i+1} - k_{\mathrm{dep}_i}^{y} y_i, & i \in [i_0, i_1] \\ \dot{v} &= \sum_{i=i_0}^{i_1} \big( - v(k_{\mathrm{on}_i}^{w} w_i + k_{\mathrm{on}_i}^{y} y_i) + k_{\mathrm{dep}_i}^{w} w_i + k_{\mathrm{dep}_i}^{y} y_i + i k_{\mathrm{dis}i} w_i \big). \end{cases}$$

Here $y_i$ and $w_i$ are the concentrations of the stable and unstable oligomers of size $i$, respectively, and $v$ is the concentration of isolated monomers. The range of sizes $[i_0, i_1]$ generally covers hundreds of sizes. The kinetic rates $k_{\mathrm{on}i}$, $k_{\mathrm{dep}_i}$ and $k_{\mathrm{dis}i}$ are the size-dependent rates associated to the polymerisation – *i.e.* sequential monomer addition – depolymerisation – *i.e.* sequential monomer loss – and disintegration, respectively.

When targeting real applications, we faced noisy and/or partial data. In order to produce a reliable tool, we need to treat these data accurately.

We aim at setting up a **data analysis** methodology.

Correctly achieving this objective plays a crucial role. It allows us to gain a deep understanding of the data and leads to a better comprehension of the physical process. The results of the data analysis are necessary to develop a model that matches biological reality. This objective includes analysing the reliability of the experimental data. In our work, we provide a mathematical approach to estimate the amount of error in the experimental data. This can also lead to establishing which data are descriptive of reality and which data should be neglected. Moreover, this methodology can be applied to the study of any kind of protein and, more generally, to any kind of molecule.

The achievement of the first two objectives naturally leads to the third objective

> Taking into account some measurements on the oligomer systems, access to a per-size and per-species description of the system evolution.

This objective can be equivalently presented in the form of an *inverse problem*, as follows

> Taking into account some measurements and a model describing the evolution of the oligomer system, estimating the initial condition and the parameters of the model.

To solve this estimation problem, we start by assuming some simplifying conditions on the system. We assume that at the beginning of each experiment, we have a ratio of stable oligomers over the totality of the oligomers that does not change with respect to size. Furthermore, we assume that the unstable oligomers are subject to disintegration much faster than any other process. Assuming, in addition, that the kinetic rates do not depend on size, we obtain a model that is completely determined by only four parameters. Specifically, we have three kinetic rates associated to the polymerisation, depolymerisation and disintegration processes and the initial ratio of stable oligomers. The final model reads

$$\begin{cases} \dot{y}_i &= k_{\text{on}} y_{i-1} v - k_{\text{on}} y_i v + k_{\text{dep}} y_{i+1} - k_{\text{dep}} y_i, & i \in [i_0, i_1], \\ \dot{w}_i &= -k_{\text{dis}} w_i, & i \in [i_0, i_1], \\ \dot{v} &= \sum_{i=i_0}^{i_1} \big( - v k_{\text{on}} y_i + k_{\text{dep}} y_i + i k_{\text{dis}} w_i \big), \\ w_i(0) &= (1-\alpha) u_i(0), \\ y_i(0) &= \alpha u_i(0), \\ v(0) &= 0, \end{cases}$$

where $\alpha$ is the initial ratio of stable oligomers and for all times $t$ and all sizes $i$, $u_i(t) = y_i(t) + w_i(t)$. In general, it is not possible to measure the quantities one wants to estimate directly. In our applications, we observe the variation in time of the average oligomer size. Such data can be collected by a static light scattering (SLS) device and the observations can be described mathematically as follows

$$z_{sls}(t) = \lambda_1 \left( v(t) + \sum_{i=i_0}^{i_1} i^2 u_i(t) \right) + \lambda_2 + \chi,$$

The parameters $\lambda_1$ and $\lambda_2$ depend on the experimental conditions and are unknowns. We assume that the data are affected by some additive noise $\chi$. The estimation of $\lambda_1$ and $\lambda_2$ and the analysis of the noise $\chi$ is performed using our data analysis methodology.

Coupling the data and the model in a *data assimilation* framework [54], we extract the useful information contained in both of these elements and thus realise the third objective. Simultaneously estimating the initial condition and the kinetic parameters, even with the simplified model, is no trivial problem. In fact, we are dealing with a non-linear problem and non-linearity represents a well-known difficulty when solving an inverse problem. We choose to apply the Extended Kalman Filter method to compute the estimations. As for all data assimilation methods in a non-linear setting, a critical point is the initialisation of the estimator. To define a good initial estimation, also called initial state *a priori*, we obtain a first candidate derived by a series of biological considerations and our modelling understanding.

We estimate the parameters associated to three sets of data, representative of the three main behaviours experimentally identified in oligomer observations. The estimation strategy is an important tool in the study of this phenomenon as it provides quantitative estimations of the kinetic parameters, which are impossible to measure otherwise.

A key aspect in data assimilation is to be able to validate the estimations, typically using extra data. We compare the estimated oligomer distributions to their empirical measurements collected by size exclusion chromatography (SEC) at certain instants. In order to be able to establish a degree of confidence in these comparisons, we performed a data analysis of the noise in the SEC data. The validation of the estimation suggests the validation of the model.

With this work we have provided – for the first time – a model for the evolution of ovPrP oligomers, based on the study of *in vitro* experimental data. This model is able to describe the most important processes occurring in ovPrP oligomer evolution in a simple way. This new tool may prove very useful for the study of other kinds of proteins or other kinds of aggregates.

In particular, our qualitative description of the oligomer system can be taken into account in the study of larger aggregates such as protein fibrils. Studying this kind of aggregates presents additional difficulties.

From an experimental point of view, there are very few techniques available to observe the evolution of fibrils. For instance, because of the large aggregate sizes, we cannot perform SEC to measure size distribution. One of the alternatives proposed is a microscopic analysis. Microscopic pictures of the fibril samples are taken (see, for instance, Figure 1). Then, the length of the fibrils is directly measured on the pictures. An example of the kind of data resulting from this type of analysis is given in Figure 2.

Because of the measurement process, we cannot access concentrations of polymers of sizes less than $145mer$. Moreover, two sources of noise must be considered : on one hand, instrumental error affecting the resolution of the pictures and, on the other hand, human error in rounding the measurements.

From a modelling point of view, due to the wide size range and the subsequent computational costs, ODE systems cannot be simulated. A continuous-size model is thus usually preferred. Taking into account only the so-called *primary pathways*, we consider only the polymerisation and depolymerisation processes. We refer to the Lifshitz-Slyozov theory. The system dynamics is thus given by a transport equation – describing the evolution of the aggregate concentrations – and an integral equation – describing the variation of the monomer

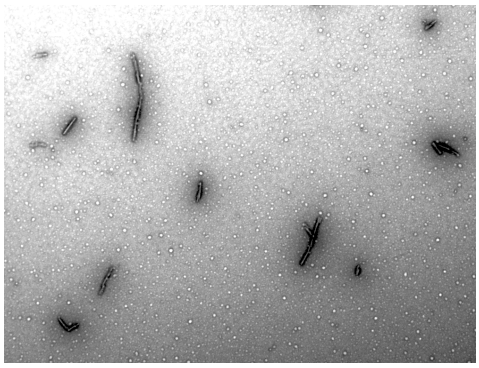


FIGURE 1 – Microscopic image of ovPrP fibrils. Source : Dr. Human Rezaei.

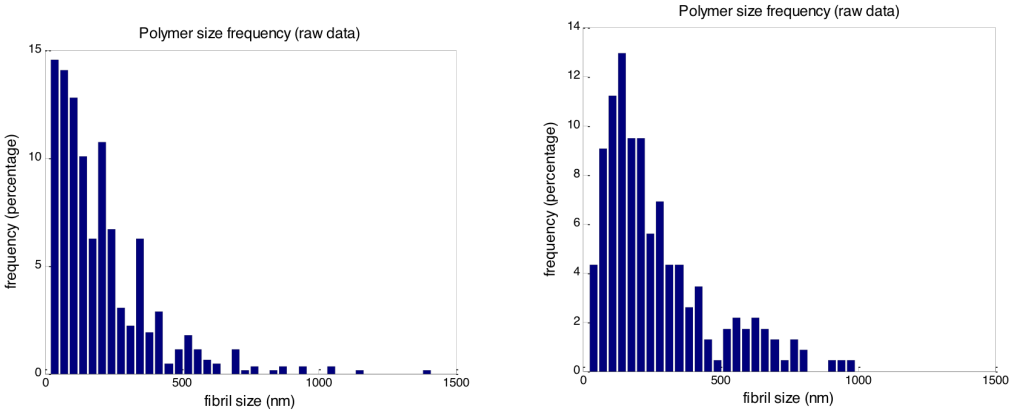



FIGURE 2 – Size Distribution obtained from microscopic pictures at time $t = 0$ on the left and time $t = 35min$ on the right. Source : Dr. Stéphanie Prigent.

concentration. Formally, it reads

$$\begin{cases} \partial_t u(x,t) + \partial_x(V(x,t)u(x,t)) &= 0, & \forall (x,t) \in \mathbb{R}^+ \times \mathbb{R}^+ \\ V(x,t) &= a(x)v(t) - b(x), & \\ v(t) + \int_0^\infty x u(t,x)\mathrm{d}x &= \rho > 0 & \forall t \in \mathbb{R}^+ \\ u_{|t=0} &= u_0, & \\ v_{|t=0} &= v_0, & \end{cases}$$

where $u(x,t)$ is the concentration of aggregates of size $x$ at time $t$, $v$ is the concentration of monomers, $a$ is the polymerisation rate and $b$ is the depolymerisation rate.

In the case of a depolymerising system, we can consider the system

$$\begin{cases} \partial_t u(x,t) - \partial_x(b(x)u(x,t)) &= 0, & \forall (x,t) \in \mathbb{R}^+ \times \mathbb{R}^+ \\ v(t) + \int_0^\infty x u(t,x)\mathrm{d}x &= \rho > 0 & \forall t \in \mathbb{R}^+ \\ u_{|t=0} &= u_0, & \\ v_{|t=0} &= v_0, & \end{cases}$$

Given that it is possible to measure the first moment of the function $u$ by Thioflavin T (ThT) fluorescence [18] and the second moment by Static Light Scattering [55, 165], we consider the general case of observing the variations in time of the moment of order $n$ of $u$. We model the observations as follows

$$z(t) = \lambda_1 \int_0^\ell x^n u(x,t)\mathrm{d}x + \lambda_2 + \chi(t),$$

where $\chi$ is some additive noise and $\lambda_1$, $\lambda_2$ are constants depending on the experimental settings.

We propose a methodology to estimate the initial condition of a physical system evolving with a backward transport equation, through the observation of one of its moments.

This inverse problem of initial condition estimations, has been analysed in the case of constant transport velocity. In this setting, we propose a kernel regularisation strategy. We describe how to set the optimal kernel with respect to the order of the observed moment and the amount of noise on the data. Furthermore, we provide an upper bound for the accuracy of the optimal estimation.

This first setting is useful as a reference case to study of the variable transport velocity case. To treat this second case, we propose a variational data assimilation strategy called 4d-Var. We show the equivalence of these two approaches in the case of constant velocity. This second method, however, is based on a formalism that can be more easily adapted to other problems. The 4d-Var method consists in minimising a least square criterion. The minimisation can be carried out using an adjoint method.

The two methods have been tested on synthetic data. The code implementing the 4d-Var has been included in the data assimilation library VerdandInMatlab, which is a Matlab module of the Verdandi C++ library [41].

A sequential data assimilation approach, the *Kalman Filter* method, has also been investigated as a supplementary strategy to tackle this estimation problem.

## 0.4 Outline of the present work

This thesis has been divided into two independent parts. The first part is organised in two chapters.

— **Chapter 1** In the first chapter of this thesis, we detail our study on ovPrP oligomers. We start by describing the experimental setting. We take into account two kinds of measurements : Static Light Scattering (SLS) [55, 165] and Size Exclusion Chromatography (SEC) [102]. The SLS test provides average measurements. Specifically, it can be related to the average cluster size. The SEC test is able to separate the aggregates of a given sample with respect to their size. For each size, the concentration of the objects of that size is given. After a brief description of the devices, we detail the data analysis. We focus on the estimation of unknown scaling parameters associated to the experimental conditions and the analysis of the noise.

A qualitative study of these data constitutes the starting point for the development of a mathematical model. We notice that the oligomers in three concentration regimes manifest different behaviours. The main challenge has been to create a model able to explain and reproduce all three behaviours.

Three main processes have been identified : polymerisation, depolymerisation and disintegration. We present biological or numerical evidence supporting this choice. Moreover, we illustrate the steps leading to the formulation of a two-species model in which one species is able to disintegrate while the other can polymerise and depolymerise.

We present the following inverse problem : Given the observations

$$z(t) = \breve{v}(t) + \sum_{i=i_0}^{i_1} i^2 \Big( \breve{y}_i(i) + \breve{w}_{0i} e^{-\breve{k}_{\mathrm{dis}} t} \Big) + \chi(t),$$

where $\chi$ is some additive noise, to find

$$\breve{k}_{\mathrm{on}}, \quad \breve{k}_{\mathrm{dep}}, \quad \breve{k}_{\mathrm{dis}}, \quad \breve{\alpha}$$

such that the solution of the dynamical system

$$\begin{cases} \dfrac{\mathrm{d}y_i}{\mathrm{d}t} &= -k_{\mathrm{on}} v y_i + k_{\mathrm{on}} v y_{i-1} - k_{\mathrm{dep}} y_i + k_{\mathrm{dep}} y_{i+1}, & i_0 \le i \le i_1, \\ \dfrac{\mathrm{d}w_{0i}}{\mathrm{d}t} &= 0, & i_0 \le i \le i_1, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= (i_0 - 1) k_{\mathrm{dep}} y_{i_0} + (-k_{\mathrm{on}} v + k_{\mathrm{dep}}) \sum y_i + k_{\mathrm{dis}} \sum i e^{-k_{\mathrm{dis}} t} w_{0i}, \\ \dfrac{\mathrm{d}k_{\mathrm{on}}}{\mathrm{d}t} &= 0, \\ \dfrac{\mathrm{d}k_{\mathrm{dep}}}{\mathrm{d}t} &= 0, \\ \dfrac{\mathrm{d}k_{\mathrm{dis}}}{\mathrm{d}t} &= 0, \end{cases}$$

with initial conditions

$$y_i(0) = \breve{\alpha}u(0), \quad w_{0i} = (1-\breve{\alpha})u(0), \quad k_{\text{on}}(0) = \breve{k}_{\text{on}}, \quad k_{\text{dep}}(0) = \breve{k}_{\text{dep}}, k_{\text{dis}}(0) = \breve{k}_{\text{dis}},$$

is the observed trajectory.

We present the Kalman filter and extended Kalman filter as strategies to estimate the state of a dynamical system from the knowledge of some noisy and/or partial measurements. We then describe the application of the Extended Kalman Filter (EKF) method to our problem. We discuss the choices leading to the definition of the initialisation of the extended Kalman estimator.

The results are presented and validated by comparing the estimated oligomer size distribution with the SEC data.

— **Chapter 2** The results of the work detailed in Chapter 1 are brought together in this second chapter in the form of the pre-printed article

*The mechanism of monomer transfer between two structurally distinct PrP oligomers*
*A. Armiento, P. Moireau, D. Martin, N. Lepejova, M. Doumic and H. Rezaei.*

In addition to what was presented in Chapter 1, this chapter points out the relevance of our results from a biological point of view.

The second part of this work is presented in the three following chapters.

— **Chapter 3** In the first chapter of this part, we present the state of the art of prion modelling. We aim at providing the context of the transport model that is taken into account in this part of the work. We focus on two theories : the Becker-Dö [19] theory and the Lifshitz-Slyozov theory [111]. The former is a discrete-size model. It consists of a set of ODEs modelling a polymerising-depolymerising system. This model is particularly interesting, being the one that inspired the oligomer model presented in the first part of this thesis. The second theory considers a continuous time-setting, and consists of two coupled equations. A transport equation describes the dynamics of the aggregate concentrations, while an integral equation – derived from the system mass conservation – links the monomer evolution to the aggregate concentrations. This model, as for the Becker-Dö model, represents a system in which aggregates attach or lose monomers. The rates at which these reactions happen govern the transport velocity. The continuous-size framework applies under the condition that the monomer size is asymptotically small compared to the average cluster size. The two methods are linked to each other. It has been proved that the solution of the Lifshitz-Slyozov model can be obtained as an asymptotic limit of the solutions of the Becker-Döring model, when the average size of the system tends to infinity [47]. Applying the Lifshitz-Slyozov theory to model depolymerisation experiments, we neglect the polymerisation term in the model. This assumption yields a linear backward transport equation.

We consider the observation of the moment of order $n$ of the concentration function, which is the solution of the transport equation. It is in fact possible to observe the first moment – corresponding to the total polymerised mass – by Thioflavin T (ThT) fluorescence [18] and the second moment – corresponding to the average cluster size – by Static Light Scattering [55, 165].

We set up the following inverse problem : Given the observations

$$z(t) = \lambda_1 \int_0^\ell x^n \breve{u}(x,t)\mathrm{d}x + \lambda_2 + \chi(t),$$

to find

$$\breve{\xi}$$

such that

$$\begin{cases} \dfrac{\partial \breve{u}}{\partial t}(x,t) - \dfrac{\partial}{\partial x}\big(b(x)\breve{u}(x,t)\big) = 0, \qquad x \in [0,\ell],\ t \geq 0, \\ \\ \breve{u}(\ell,t) = 0, \\ \\ \breve{u}(x,0) = u_\diamond + \breve{\xi}, \end{cases}$$

— **Chapter 4** The second chapter of the second part corresponds to the article published in the Journal of Theoretical Biology

*Estimation from moments measurements for amyloid depolymerisation*
*A. Armiento, M. Doumic, P. Moireau, and H. Rezaei.*

This work presents two general methodologies to solve the inverse problem presented in Chapter 3. The first belongs to the class of kernel methods [64]. This method requires an explicit theoretical relation linking the initial condition – which is the object of our estimation – to the noiseless data. In our case, when observing the moment of order $n$, we would need to derive the observations $n+1$ times. In the presence of noise, such an operation is not possible. Data are thus regularised by convolution with a kernel function, to have the desired regularity. We provide a characterisation of the kernel function, depending on the amount of noise in the data. This strategy has been investigated in the case of constant transport velocity.

The second method is a variational data assimilation method called 4d-Var [105]. It is designed for the general case of variable transport velocity. The initial condition estimation is obtained by minimising a least square criterion accounting for the error in the initial condition approximation and the discrepancy between the experimental observations and the observation simulated on the estimation. Moreover, we characterised the derivative of the criterion with respect to the initial condition by introducing the adjoint variable [23].

In the case of constant transport velocity, the equivalence between the two strategies is shown. Furthermore, the well-posedness of the inverse problem is investigated. Thanks to the linearity of the problem, it corresponds to verifying the so called *observability conditions*.

The methods have been numerically implemented and tested on synthetic data considering several amounts of noise in the data. We find that in both cases, the accuracy of the estimations deteriorates for increasing levels of noise and in increasing orders of the observed moment.

We also illustrate, using an example, how these methods can be applied to real data.

— **Chapter 5** The third and last chapter of the second part is dedicated to an overview of data assimilation methods. These methods are presented in a state-space formulation. Given the inverse problem introduced in Chapter 3, we define the operators associated to the model and the observations in this new formalism. The methods are presented in a continuous time setting with linear, time-independent infinite-dimensional operators. We focus on the 4d-Var method – belonging to the class of variational methods – and the Kalman filter method – belonging to the class of sequential methods.

## Première partie

# Data assimilation on an ODE model of polymerisation

# CHAPITRE 1

# Direct model and inverse problem solution for prion oligomer proteins

In this chapter we present the results of our research on the behaviour of ovine prion oligomers, namely ovPrP oligomers. As explained in the introductory chapter, the oligomers are protein aggregates made up of small number of monomers. These aggregates are particularly important as increasing experimental evidence is leading to the hypothesis that subfibrillar oligomers are the most infectious factor in prion diseases [93, 159]. Although much progress has been made in the study of prion oligomers, the biochemistry and the biophysics of prion oligomers remain unclear. In our research, we aim at understanding the main mechanisms occurring in oligomer evolution, as better knowledge of the oligomerisation pathways would help to clarify the pathological events at the molecular level. Our work can thus be seen as a first step on the path that will, hopefully, lead to the design of medical treatment targeting prion oligomers to cure prion diseases.

In this chapter we present the experiments and measurements performed by our collaborators in Dr. H. Rezaei's team at the INRA laboratories in Jouy-en Josas. Interdisciplinary collaboration between mathematicians, physicists and biologists has been crucial to obtain a better understanding of the subject and achieve significant progress.

When we started this project, the approach commonly adopted by biologists to study the dynamics and the aggregation-fragmentation mechanisms of prion oligomers was measuring the variations in the average molecular weight by a Static Light Scattering (SLS) device [55]. We based our considerations on this type of data and we represented the dynamics of the

oligomer system using a classical model which is known in the literature as the Becker-Döring system [19]. We assumed that oligomers can be involved in two kinds of processes : sequential polymerisation, *i.e.* gaining one monomer at a time, and depolymerisation, *i.e.* losing one monomer at a time.
By using the information from the experimental data and the Becker-Döring system, we have tried to answer some of the open biological/biochemical questions. In particular, we have focused on

1. estimating the oligomer size distribution at any time $t$,
2. identifying the kinetic parameters.

In fact, knowing the oligomer size distribution is helpful to distinguish between the formation pathways. Furthermore, as we prove in our work, it contains information on the structural heterogeneity of oligomers.
The kinetic parameters give information on the interactions between the oligomers and between the oligomers and the monomers. By analysing the kinetic rates, we can identify the objects that are more thermodynamically stable. The stability of assemblies is in fact important to predict the evolution of the oligomer population. It has been observed that aggregates with a low thermodynamical stability are prone to losing monomers. This implies that when high and low stable aggregates coexist, the low stable aggregates shorten, giving rise to a monomer reservoir that is used by the more stable aggregates [181].

Ideally, once we have a complete understanding of oligomer species and their behaviour, it will possible to conceive a treatment to attack the agent responsible for amyloid formation with a targeted strategy.

To achieve these objectives, we gather all the available sources of information and we connect them through a data assimilation framework. We adopt the Extended Kalman Filter method to perform the estimation. The accuracy of the estimation depends on the amount of error in the measurements and the definition of the initial oligomer size distribution. To gain some information regarding the initial distribution, a new set of experiments was performed and this time the size exclusion chromatography (SEC) measurements were collected. Coupling these data with the multi-wavelength static light scattering (MWLS) analysis [165, 55], we have access to the size distribution at different times.

The measurement of the initial oligomer distribution allows us to define a good estimation of the oligomer initial state. Having access to the oligomer distributions at successive times, we notice an unexpected behaviour of the oligomers that led us to question the model.

Thanks to our multidisciplinary collaboration, it was possible to formulate and evaluate a variety of mathematical models to represent our oligomer system. Some of the models were excluded because they did not match with biological evidence, while others were excluded because they were not able to reproduce the behaviours observed in the data. In this chapter we present the steps that guided us to the formulation of a two oligomer species model. In this model we consider two kinds of oligomers :

— ones that can polymerise and depolymerise, which we call *stable*
— and ones that can polymerise, depolymerise and, above all, disintegrate, which we call *unstable*.

We denote by $w_i$ and $y_i$ the concentrations for unstable and stable oligomers of size $i$, respectively. We assume that we have only one type of monomers and we call $v$ the monomer

concentration. The mathematical model results in the following set of ODEs

$$\begin{cases} \dot{w}_i &= -k_{\mathrm{dis}i} w_i + k_{\mathrm{on}_{i-1}}^{w} w_{i-1} v - k_{\mathrm{on}_i}^{w} w_i v + k_{\mathrm{dep}_{i+1}}^{w} w_{i+1} - k_{\mathrm{dep}_i}^{w} w_i, & i_0 \le i \le i_1, \\ \dot{y}_i &= k_{\mathrm{on}_{i-1}}^{y} y_{i-1} v - k_{\mathrm{on}_i}^{y} y_i v + k_{\mathrm{dep}_{i+1}}^{y} y_{i+1} - k_{\mathrm{dep}_i}^{y} y_i, & i_0 \le i \le i_1, \\ \dot{v} &= \sum_{i=i_0}^{i_1} \big( - v(k_{\mathrm{on}_i}^{w} w_i + k_{\mathrm{on}_i}^{y} y_i) + k_{\mathrm{dep}_i}^{w} w_i + k_{\mathrm{dep}_i}^{y} y_i + i k_{\mathrm{dis}i} w_i \big), \end{cases} \tag{1.1}$$

where the non-negative coefficients $k_{\mathrm{on}_i}^{a}, k_{\mathrm{dep}_i}^{a}, k_{\mathrm{dis}i}$, $a \in \{y, w\}$ are the kinetic rates of the polymerisation, depolymerisation and disintegration reactions, respectively. Furthermore, we denote by $i_0$ and $i_1$ the minimal and maximal oligomer size.

It is worth mentioning that it is not possible to empirically measure the kinetic rates. At the start of our project, the only vague information we had was that a classical range for the depolymerising rate is between $10^{-5}$ and $10^{-2}\mathrm{min}^{-1}$, commonly around $10^{-3}\mathrm{min}^{-1}$. With our approach we have been able to estimate all the kinetic rates precisely of the order of $10^{-1}\mathrm{min}^{-1}$.

This chapter contains four sections. In Section 1 we present and analyse the experimental data. In Section 2 we discuss and formulate our oligomer model. In Section 3 we introduce the Extended Kalman Filter and detail its application to estimating the kinetic rates and the initial oligomer size distribution. In Section 4, we discuss the results.

## 1.1 Experimental data analysis

In this section, we describe the formation of the oligomer system under study and how we can observe it experimentally. In particular, we focus on the experimental setting and we describe the methodology to analyse the experimental data.

### 1.1.1 In vitro oligomer formation

In the introductory chapter of this thesis, we presented the mechanisms of protein polymer formation *in vivo*. We can highlight three key pathogenic events in prion diseases : 1) the presence of misfolded (ill) proteins, 2) the misfolded proteins come into contact with healthy proteins and catalyse their conversion into the ill form 3) over time all the healthy proteins will be converted into ill proteins. To recreate these steps in *in vitro* experiments, scientists induce a partial unfolding of healthy full-length ovine $\mathrm{PrP}^{\mathrm{C}}$ proteins by thermal treatment. The partially unfolded proteins are then able to aggregate and form oligomers [150]. Depending on the way proteins aggregate with each other, we can distinguish between three structurally different oligomer species, as presented in [40]. Since the three oligomer size ranges do not overlap, it is possible to isolate one particular kind of oligomer by isolating clusters with sizes in a particular range. The selection of the clusters with respect to the size can be performed by a size exclusion chromatography test, which is detailed later in this section.

We report here some more technical information about the protocol and we refer to articles [62] and [40] for more details.
The *in vitro* oligomer formation starts with non-associated $\mathrm{PrP}^{\mathrm{C}}$ proteins (monomers). Full-length Ovine PrP 23-234 (Ala-136, Arg-154, Gln-171 variant) were produced in Escherichia coli and then purified. The proteins are incubated in a concentrated solution of the denaturant guanidinium chloride and then heated up to $37°C$. In this way, the proteins lose their

quaternary structure [1] (*denaturation*) and start to aggregate with each other forming oligomer structures. To have a detailed description of the purification process we refer to the publication [150]. The conversion of the misfolded proteins $PrP^{SC}$ into the oligomeric form is performed in $20mM$ sodium citrate buffer (pH 3.40). The $PrP^{SC}$ – at a final concentration of $50\mu M$ – is incubated in a Perkin Elmer GenAmp2400 thermocycler at $65°C$ for two hours. Homogeneous fractions of oligomers are then collected after separation by size exclusion chromatography (SEC), as first described in the work [62]. SEC is performed at $20°C$ using a TSK 4000SW ($7mm * 600mm$) gel-filtration column (Interchim, Montluon, France) with $20mM$ sodium citrate (pH 3.35). Protein elution is monitored by UV absorption at $280nm$. The oligomer size distribution is determined by SEC data coupled with static light scattering data. The light scattering test is performed with an in-lab device using $407nm$ laser beams in a $2mm$-path-length quartz cuvette. Kinetic experiments are performed according to a standardised methodology, as reported in the work [40] : $72°C$ in $20mM$ sodium citrate buffer (pH 3.40).

### 1.1.2 Biological experiments

FIGURE 1.1 – SLS device (left) and SEC device (right).

The experiment that we consider in our research is a *depolymerising experiment*. In this case, the term "depolymerising" does not imply that the only process involved is the depolymerisation, but rather the global size-reducing behaviour of the aggregates at the beginning of the experiments. The main reason for setting up this kind of experiment is that – since monomers are necessary for the growth of aggregates and we start with no monomers – we are able to decouple the growing-shrinking processes and have only shrinking processes, at least for a certain initial time lapse.

Once oligomers have been formed and isolated via size exclusion chromatography, we extract two samples that are put into two cuvettes of different volumes. We assume that the oligomers are homogeneously distributed and that any extracted sample is identically distributed. The two cuvettes are put in two static light scattering devices under the same

1. The *quaternary structure* is the arrangement of more than one polypeptide chain in a multi-subunit complex to form a fully functional protein.

temperature and pressure conditions. We perform a series of three experiments in which we put different total mass concentrations, namely $\rho$, in the smaller cuvette. We use the bigger cuvette as a pool from which we extract oligomer samples of concentration $\rho$ that are then used to perform the size exclusion chromatography test and obtain the size distribution. With this strategy, we have access to the size distribution at several times without perturbing the system in the smaller cuvette. Given the general consensus of the scientific community on this protocol, we assume that – since the oligomers in the two cuvettes are set in the same conditions – they undergo the same evolution process. We thus consider that the static light scattering measurements performed on the small cuvette and the size exclusion chromatography performed on the samples taken from the large cuvette are observations of the same system.

In the following sections, we describe the two kinds of data collected during the experiments on prion protein oligomers. An important feature of the aggregation-disaggregation processes is its dependence on the total monomer concentration. To take into account and better understand this dependence, three experiments were performed at three different total monomer concentrations. The total monomer concentration, denoted $\rho$, can be mathematically defined as follows

$$\rho = v + \sum_{i=i_0}^{i_1} i u_i, \tag{1.2}$$

where $v$ is the concentration of isolated monomers, while the second term indicates the concentration of the aggregated monomers. We call $u_i$ the concentration of oligomers of size $i$. Therefore, the quantity $iu_i$ corresponds to the concentration of monomers composing the oligomers of size $i$. We recall that the notation $i_0$ and $i_1$ stand for the minimal and maximal oligomer sizes, respectively.

The experimental concentrations were chosen to represent three different regimes

— $\rho = 1\mu M$ : low concentration regime

— $\rho = 3\mu M$ : medium concentration regime

— $\rho = 7\mu M$ : high concentration regime.

### 1.1.3 Size exclusion chromatography (SEC) data

*Chromatography* is a classical technique to separate macromolecules according to their specific properties. In our work we are interested in separating molecules on the basis of their size, leading us to consider the Size Exclusion Chromatography (SEC) test. This test is also called *gel filtration* because the aggregates pass through a gel grid, which is packed in a column. By controlling the degree of cross-linking of the grid, we can separate different sizes. For example, a loose grid can separate large molecules. An early application of the SEC test was carried out in 1955, see [102].

The excluded molecules are associated with an *elution volume*[2], which is the volume of buffer exiting the column before the molecules. Big molecules are not trapped in the gel structure and are thus selected first. The associated time elapsing before exiting the column is small and, equivalently, the elution volumes are small. The SEC device returns the weight concentration of the eluted molecules that is conventionally given with respect to the volume.

2. *Elution* is the process of extracting one material from another by washing with a solvent. The *elution volume* is the volume of solvent necessary to elute the material.

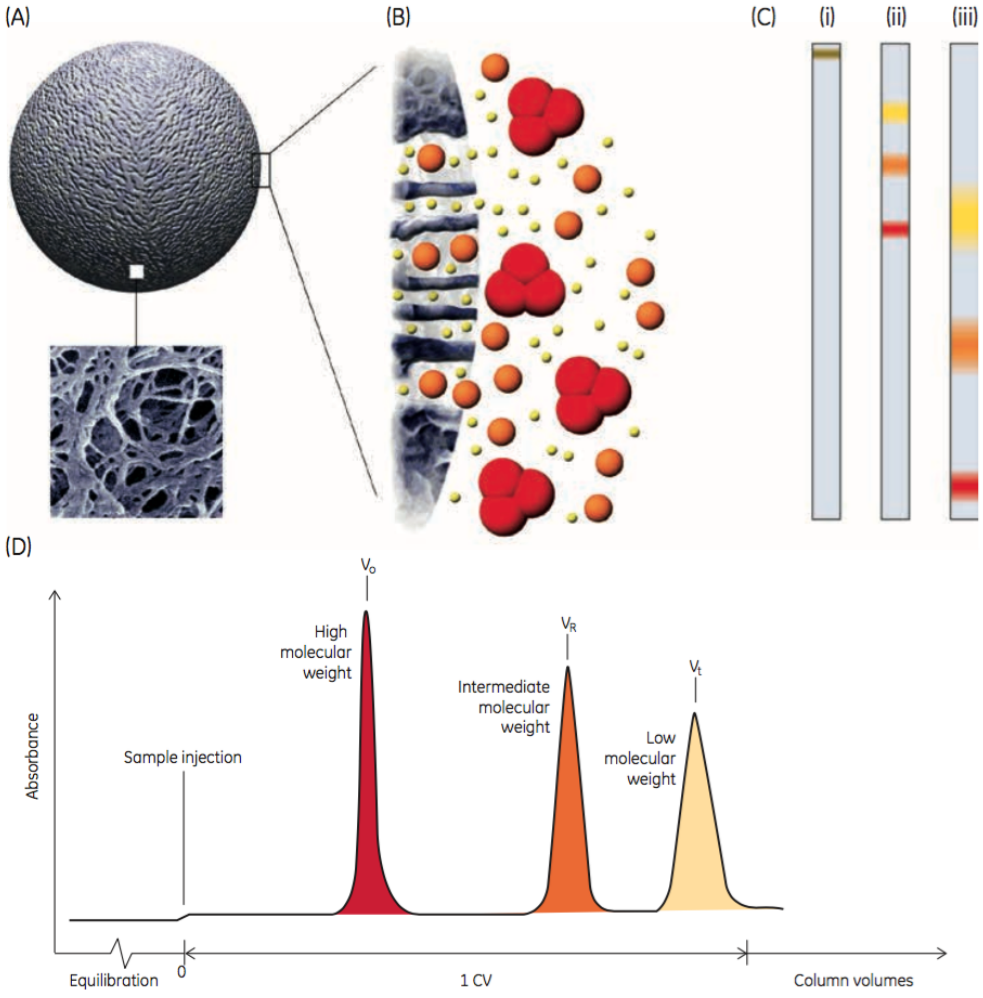



FIGURE 1.2 – (A) Schematic picture of a gel with an electron microscopic enlargement. (B) Schematic drawing of sample molecules diffusing in the gel pores. (C) Graphical description of molecules separating in the gel column. (D) Schematic chromatogram. Source : GE Healthcare Life Sciences, Size Exclusion Chromatography handbook.

We denote these data as $z_{\text{sec}}(V)$, where $V$ is the elution volume. To give an example, in Figure 1.2-(D) we show a typical result of SEC separation. When the data are represented with respect to the volume, the curves are independent of the speed at which the experiment has been carried out. Hence, it has conventionally been adopted in the literature to facilitate the analysis and comparison of the results.

In Figure 1.3, we present the experimental SEC data in three different concentration regimes : $\rho = 1,\ 3,\ 7\mu M$.

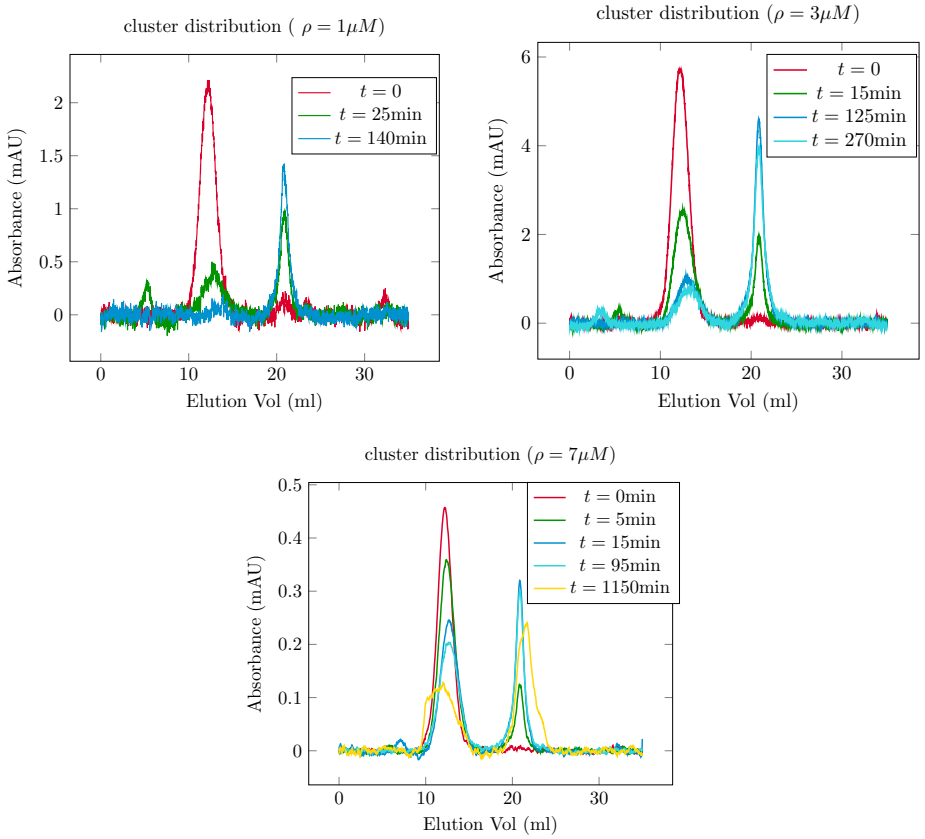



FIGURE 1.3 – Experimental SEC data relative to the total concentrations $\rho = 1,\ 3,\ 7\mu M$.

When observing these data, we can clearly distinguish two peaks. The first one, on the left, corresponds to the oligomer peak while the second one corresponds to the monomer peak. As mentioned before, oligomers – being bigger than monomers – correspond to smaller elution volumes. Moreover, we do not have objects with intermediate elution volumes because, if this had been the case, we would have observed other peaks as in Figure 1.2-(D). Therefore, in our mathematical model we should pay particular attention to ensuring that oligomers of intermediate sizes are not formed.

Thanks to multi-wavelength static light scattering (MWLS) analysis, [55], we can have an empirical Volume-Size transformation law. In Figure 1.4, we show the correspondence relation between oligomer sizes and elution volumes.

Coupling the SEC and MWLS data, we have access to the oligomer concentration with respect to the size of the oligomers. The oligomer size is defined as the number of monomers

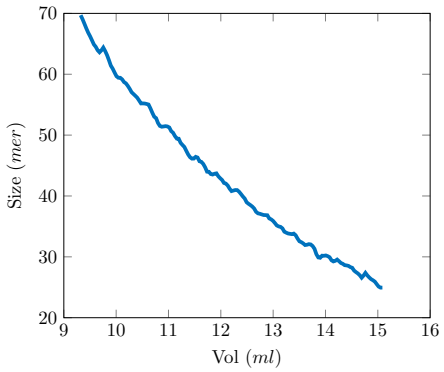



FIGURE 1.4 – Volume-Size Equivalence

composing the oligomer and it is expressed in $mer$, standing for "monomer". Denoting by $V$ the measured elution volume, we apply the change of variables $V = f(i)$, for the volumes $V \in [9.5, 15]$ml and the sizes $i \in [25, 70]mer$. In this way we obtain the oligomer weight concentration with respect to the size. By dividing each value by the corresponding size, we then obtain the oligomer concentration. Let $t$ be the time at which the chromatography test is performed, we introduce the notation

$$z_{\text{sec,o}}(i,t) = \frac{z_{\text{sec}}(f(i))}{i}, \quad \text{for } i \in [25,70]mer,$$
$$z_{\text{sec,m}}(V,t) = z_{\text{sec}}(V), \quad \text{for } x \in [17,25]\text{ml}.$$

Let us call $u_i(t)$ the concentration of oligomers of size $i$ at time $t$ and $v(t)$ the monomer concentration. Moreover, let $\lambda_o$, $\lambda_m$ be two unknown positive coefficients, we have :

$$z_{\text{sec,o}}(i,t) = \lambda_o u_i(t), \quad \text{for } i \in [25,70] \tag{1.3}$$

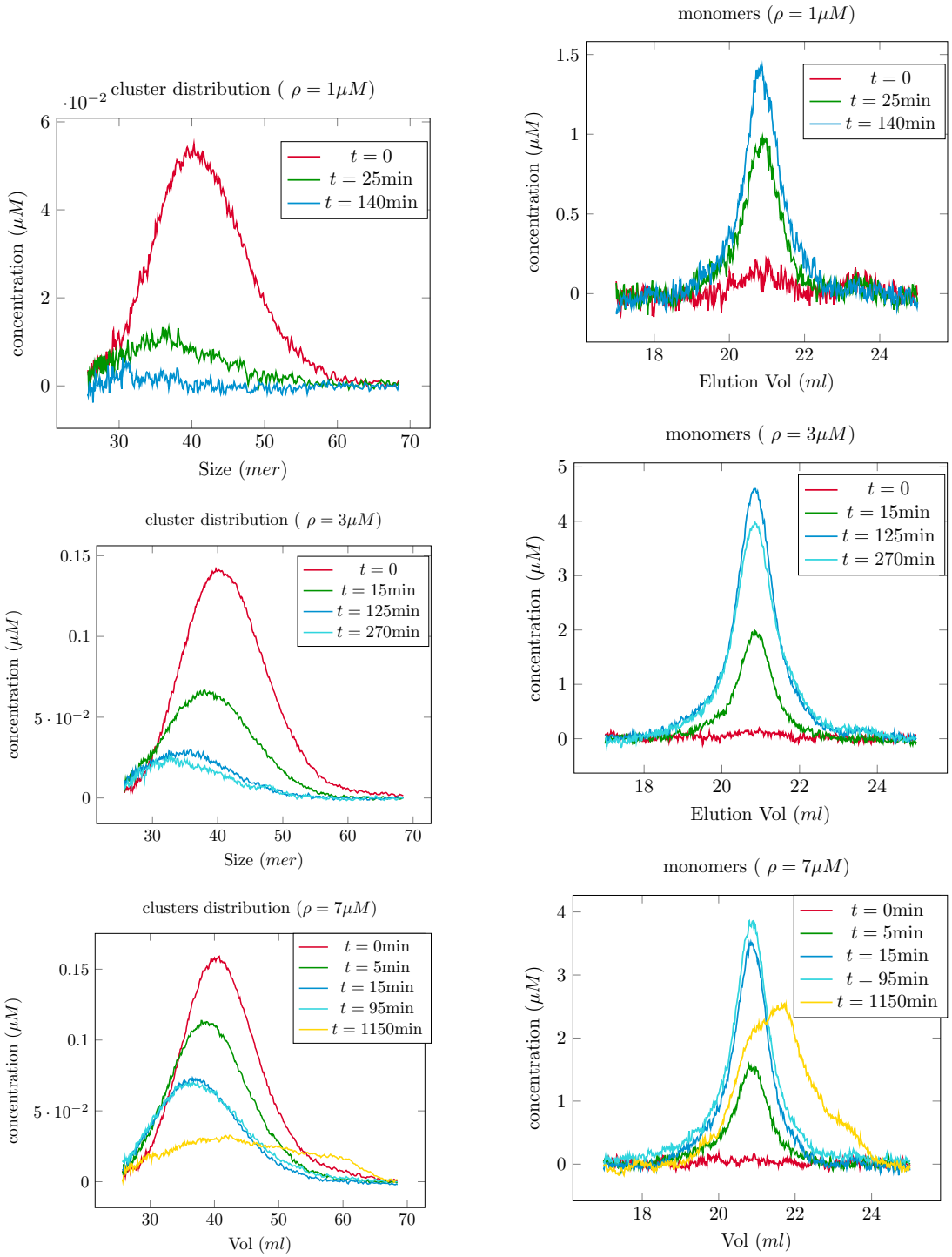
and

$$\int_{17}^{25} z_{\text{sec,m}}(V,t)\mathrm{d}V = \lambda_m v(t). \tag{1.4}$$

We assume that $i_0 = 25mer$ and $i_1 = 70mer$. Or, equivalently, we assume that the smallest and biggest detected sizes correspond to the smallest and biggest oligomer size. We notice that – given the physical properties of the SEC device – this assumption is true when we observe an oligomer system but it is not true if we observe larger aggregates. In fact, large aggregates are likely to be bigger than $100mer$ and the device is not able to separate polymers bigger than this size. These aggregates elute together at the end of the SEC test.

In Figure 1.5, we plot the functions $z_{\text{sec,o}}$ and $z_{\text{sec,m}}$ for different total concentrations.

**1.1.3.a SEC experimental constants** In this section we detail how to estimate the parameters $\lambda_o$ and $\lambda_m$ introduced in Equations (1.3)-(1.4).




(a) $z_{\text{sec,o}}$ (b) $z_{\text{sec,m}}$

FIGURE 1.5 – Left : Oligomer distribution computed from SEC data. Right : monomer peak in SEC data. From the top to the bottom we consider the total concentrations $\rho = 1, 3, 7\mu M$.

Since our experiments start with no monomers, $v(0) = 0$. Consequently, in this case Equation (1.2) reads

$$\sum_{i=i_0}^{i_1} i z_{\text{sec,o}}(i,t) = \lambda_o \sum_{i=i_0}^{i_1} i u_i(t) = \lambda_o \rho.$$

We conclude that

$$\lambda_o = \frac{\sum_{i=i_0}^{i_1} i z_{\text{sec,o}}(i,t)}{\rho}. \tag{1.5}$$

To estimate the coefficient $\lambda_m$, we consider an instant $t \neq 0$. We can write

$$\int_{17}^{25} z_{\text{sec,m}}(V,t)\mathrm{d}V = \lambda_m v(t) = \lambda_m \left( \rho - \lambda_o^{-1} \sum_{i=i_0}^{i_1} i z_{\text{sec,o}}(i,t) \right).$$

Therefore, we have

$$\lambda_{m|t} = \frac{\int_{17}^{25} z_{\text{sec,m}}(V,t)dV}{\rho - \lambda_o^{-1} \sum_{i=i_0}^{i_1} i z_{\text{sec,o}}(i,t)}, \tag{1.6}$$

where the subscript $\lambda_{m|t}$ highlights the dependency of the formula on time. Let $\{t_j\}_{j=1,\ldots,N_{\text{SEC obs}}}$ be the times of the SEC measurements, where $t_1 = 0$. We define

$$\lambda_m = \frac{1}{N_{\text{SEC obs}} - 1} \sum_{j=2}^{N_{\text{SEC obs}}} \lambda_{m|t_j}. \tag{1.7}$$

In the following table and in Table 1.1, we report the values obtained with formulas (1.5), (1.6) and (1.7).

| | | | | |
|---|---|---|---|---|
| $\rho = 1\mu M$ | $\lambda_{m\|25} = 1.5153$ | $\lambda_{m\|140} = 1.8049$ | | |
| $\rho = 3\mu M$ | $\lambda_{m\|15} = 1.4938$ | $\lambda_{m\|125} = 2.6642$ | $\lambda_{m\|270} = 2.3596$ | |
| $\rho = 7\mu M$ | $\lambda_{m\|5} = 0.9815$ | $\lambda_{m\|15} = 1.1031$ | $\lambda_{m\|95} = 1.5483$ | $\lambda_{m\|15} = 1.3046$ |

The coefficients $\lambda_o$ and $\lambda_m$ depend on the device. We do not expect them to vary greatly in the three experiments. We notice that, in the experiment at $\rho = 7\mu M$, the value of $\lambda_m$ is almost half as small as in the other two cases. This discrepancy has been explained (afterward) by considering that the samples used to perform the SEC test in this experiment had concentration of $3.5\mu M$ while we applied the formula (1.5) with the value $\rho = 7$. We can thus see how this approach is useful to reduce the errors arising from missing information about the experimental protocol or slight differences in the concentrations of the samples used in the SEC test.

**1.1.3.b SEC data noise** We move on now to analyse the effects of noise on the SEC data. SEC data are affected by two kinds of noise : an additive white noise and a variability in the position and the width of the peak.

The first kind of noise is related to the precision of the SEC device. We can easily identify the effects of this noise when looking at the base-lines of data in Figure 1.3 that correspond to

| $\rho = 1\mu M$ | $t = 0$min | $t = 25$min | $t = 140$min |
|---|---|---|---|
| $\sum_{i=i_0}^{i_1} i z_{\mathrm{sec,o}}(i,t)$ | 35.0279 | 6.9363 | 0.3611 |
| $\int_{17}^{25} z_{\mathrm{sec,m}}(V,t)\mathrm{d}V$ | 0.1879 | 1.2152 | 1.7863 |
| $\lambda_o = 35.0279, \lambda_m = 1.6601$ | | | |
| $\sum_{i=i_0}^{i_1} i z_{\mathrm{sec,o}}(i,t)/\lambda_o$ | 1.0000 | 0.1980 | 0.0103 |
| $\int_{17}^{25} z_{\mathrm{sec,m}}(V,t)\mathrm{d}V/\lambda_m$ | 0.1132 | 0.7320 | 1.0760 |
| $v(t) + \sum_{i=i_0}^{i_1} i u_i(t)$ | 1.1132 | 0.9300 | 1.0864 |

| $\rho = 3\mu M$ | $t = 0$min | $t = 15$min | $t = 125$min | $t = 270$min |
|---|---|---|---|---|
| $\sum_{i=i_0}^{i_1} i z_{\mathrm{sec,o}}(i,t)$ | 93.4175 | 42.0483 | 16.0868 | 12.6014 |
| $\int_{17}^{25} z_{\mathrm{sec,m}}(V,t)\mathrm{d}V$ | 0.283 | 2.464 | 6.616 | 6.124 |
| $\lambda_o = 31.1392, \lambda_m = 2.1725$ | | | | |
| $\sum_{i=i_0}^{i_1} i z_{\mathrm{sec,o}}(i,t)/\lambda_o$ | 3.0000 | 1.3503 | 0.5166 | 0.4047 |
| $\int_{17}^{25} z_{\mathrm{sec,m}}(V,t)\mathrm{d}V/\lambda_m$ | 0.1303 | 1.1343 | 3.0454 | 2.8188 |
| $v(t) + \sum_{i=i_0}^{i_1} i u_i(t)$ | 3.1303 | 2.4846 | 3.5620 | 3.2235 |

| $\rho = 7\mu M$ | $t = 0$min | $t = 5$min | $t = 15$min | $t = 95$min | $t = 1150$min |
|---|---|---|---|---|---|
| $\sum_{i=i_0}^{i_1} i z_{\mathrm{sec,o}}(i,t)$ | 97.9113 | 68.4282 | 42.2460 | 46.0280 | 37.2613 |
| $\int_{17}^{25} z_{\mathrm{sec,m}}(V,t)\mathrm{d}V$ | 0.2205 | 2.0689 | 4.3899 | 5.7430 | 5.6570 |
| $\lambda_o = 13.987, \lambda_m = 1.2344$ | | | | | |
| $\sum_{i=i_0}^{i_1} i z_{\mathrm{sec,o}}(i,t)/\lambda_o$ | 7.0000 | 4.8922 | 3.0203 | 3.2907 | 2.6639 |
| $\int_{17}^{25} z_{\mathrm{sec,m}}(V,t)\mathrm{d}V/\lambda_m$ | 0.1787 | 1.6761 | 3.5564 | 4.6525 | 4.5829 |
| $v(t) + \sum_{i=i_0}^{i_1} i u_i(t)$ | 7.1787 | 6.5682 | 6.5767 | 7.9432 | 7.2468 |

TABLE 1.1 – Numerical values computed from SEC data at concentrations $\rho = 1, 3, 7\mu M$, applying the definitions (1.5) and (1.7).

a zero concentration. Since SEC data are independent of the sample mass and concentration, we can assume that the resolution is the same for all the experiments.

To better understand the source of the second kind of noise, we recall that when oligomers pass through the SEC column, they are delayed causing the band broadening that can be observed in Figure 1.2-(C). Consequently, oligomers with the same size do not all leave the columns at the same moment and we observe broad peaks. This effect is unavoidable because it is linked to the diffusion of sample molecules inside and outside the gel medium.

A good SEC test has to guarantee sufficient selectivity and limit the peak broadening effects. Unfortunately, the resolution is influenced by too many factors – *e.g.* particle size, particle uniformity, column packing quality, volumes in system components, flow rate, sample volumes, viscosity, etc. – and it is therefore difficult to derive a mathematical description of the effect of this noise on SEC data.

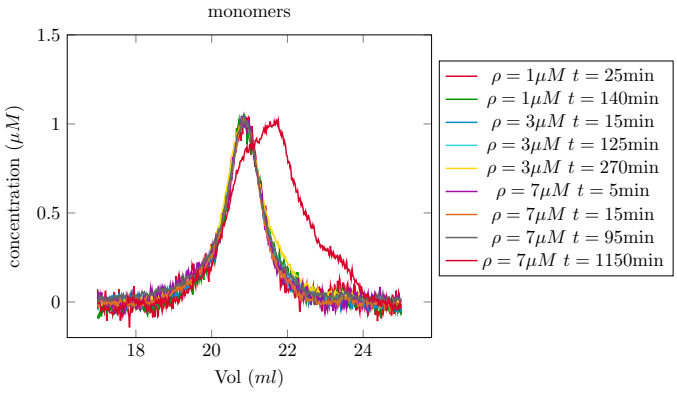



FIGURE 1.6 – Normalised monomer concentration data collected with the SEC device.

However, in the case of a single type of molecules, the broadening of the peak is a deterministic process that depends on the object's structure and the composition of the gel. In fact, if we compare the monomer peaks of our data – normalised to have the maximum one – we observe a general agreement, see Figure 1.6. We remark that the monomer peak observed at time $t = 1150$min and $\rho = 7\mu M$ has a different shape that may indicate a phenomenon of monomer degradation. The monomer peak shape can give some indication as to whether SEC data are trustworthy or not. Hence, we would assume the noise to be larger on the data corresponding to $t = 1150$min and $\rho = 7\mu M$.

This kind of argument cannot be applied to the oligomer peak since it corresponds to a group of objects and the peak shape changes over time. An empirical strategy to analyse the noise on these data would be to repeat the same experiment several times and compare the experimental oligomer peaks with an average peak. Without this additional information it is not possible to quantify the broadening effects on the oligomer peak during the experiments. However, for the following reason, we can have a high level of confidence in the initial oligomer distribution. Indeed, the initial samples are extracted from the same oligomer pool and then diluted to have the concentration $\rho = 1, 3, 7\mu M$. When we compare the normalised distribution of these three samples, we have a perfect match, see Figure 1.7. Hence we assume that there is no error in the initial oligomer peak shape. Furthermore, we can assume that the value of the total concentration $\rho$ is known exactly. We define the initial condition as the initial empirical oligomer peak rescaled to have its integral equal to $\rho$, and we can thus

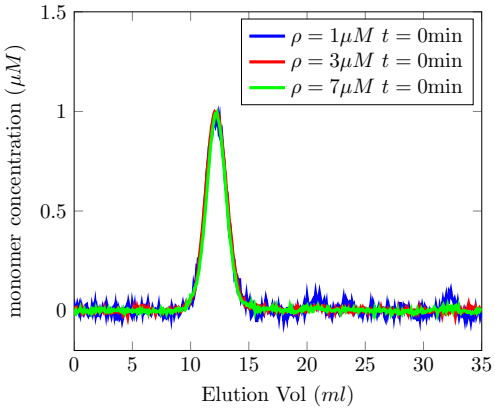
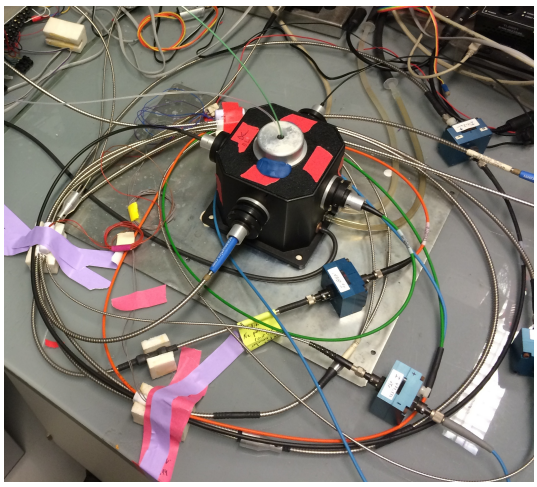


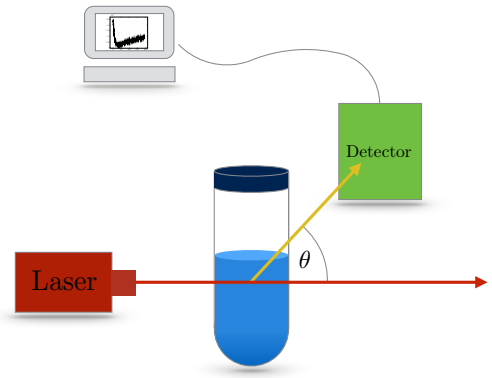

FIGURE 1.7 – Normalised initial oligomer concentration data collected with the SEC device.

assume that it is known precisely.

### 1.1.4 Static Light Scattering (SLS) data



(a) Light scattering instrument from Inra Laboratory




(b) Diagram of a set-up to measure the light scattering intensity

FIGURE 1.8 – SLS device

The second experiment involves Static Light Scattering (SLS) [55, 165]. Applying light scattering to chemical problems has become a popular technique to analyse macromolecular properties. To interpret experimental Static Light Scattering data, we begin with a brief overview of light scattering in order to convey a general understanding of the technique.

When light is sent through a material, several interactions are possible : fluorescence, transmission, absorption, and scattering. In particular *scattering* is the deflection of light from a straight trajectory, after encountering some physical object.

The intensity of the scattered light is a function of the molecular weight and concentration of the scattering object. To measure such intensity we use a Static Light Scattering device.

Typically, the solution of molecules is in a cylindric cuvette, integrated in the SLS device. A laser emits a beam of monochromatic light that hits the molecules. Both the power of the light source and the intensity of the light scattered from the sample are continuously detected and recorded. We call $\theta$ the angle formed between the axis which connects the cuvette and the detector and the direction of the laser beam. The intensity of the scattered light is usually expressed as a function of the scattering angle $\theta$. In Figure 1.8b, we present a simplified scheme of the experimental setup.

In our experiments, oligomer assemblies have been monitored by light scattering by incubating oligomer assemblies at different concentrations at $50°C$ in a quartz cuvette of $2mm$ path-length and illuminated by a Laser beam of $405nm$ and $50\mu m$ waist. The scattering angle was set to $\theta = 88°$. The scattered intensity was measured using a homemade device [39], see Figure 1.8a.

A mathematical description of the light scattering comes from the Rayleigh theory [169, 168, 164]. We report here some of the main steps that lead to the mathematical model in Equation (1.8). Let us start from the case of an incident unpolarised light scattered off a small particle in an ideal solution. The intensity of the scattered light is

$$I_{\text{scattered}} = I_{\text{laser}} \frac{8\pi^4 \alpha_p^2}{r^2 \lambda^2} (1 + \cos^2 \theta),$$

where $I_{\text{laser}}$ is the incident light intensity, $\alpha_p$ is a constant called polarizability which depends on the particle's characteristics. The scattered light is inversely proportional to the distance $r$ between the particle and the detector and to the light's wavelength $\lambda$.

We expect that the light intensity depends on the number of particles seen by the detector. Therefore, when we consider $nN_A$ particles ($N_A$ is Avogadro's number) in a volume $V$ the scattered light intensity becomes

$$I_{\text{scattered}} = \frac{nN_A}{V} I_{\text{laser}} \frac{8\pi^4 \alpha_p^2}{r^2 \lambda^2} (1 + \cos^2 \theta).$$

We assume here that the particles are randomly located and that we can consider them as independent sources of scattered light.

In the case of polymer particles in solution, the polarizability depends on molecular weight. Specifically, the polarizability of particles at concentration $c = \frac{nM}{V}$ is

$$\alpha_p = \frac{n_0 M}{2\pi N_A} \frac{dn_0}{dc},$$

where $n_0$ is the refractive index of the solvent and $\frac{dn_0}{dc}$ is the dependence of the refractive index with respect to the concentration. Substituting this value in the equation for $I_{\text{scattered}}$ we have

$$I_{\text{scattered}} = I_{\text{laser}} \frac{2\pi^4}{r^2 \lambda^2} \frac{n_0^2}{N_A} \left( \frac{dn_0}{dc} \right)^2 (1 + \cos^2 \theta) Mc = KMc,$$

where

$$K = I_{\text{laser}} \frac{2\pi^4}{r^2 \lambda^2} \frac{n_0^2}{N_A} \left( \frac{dn_0}{dc} \right)^2 (1 + \cos^2 \theta)$$

depends only on the experimental setting and, more importantly, it is independent of the concentration or molecular weight of the polymer system.

If the solution contains a mixture of different kinds of polymers, we have

$$I_{\text{scattered}} = K \sum_i c_i M_i,$$

where $c_i$ are the weight concentrations of the different kinds of objects and $M_i$ the relative molecular weights.

**Remark 1.1.4.1**
To apply this theory to our problem, we consider that we have only monomers and polymers in our solution. Our polymers differ only by their size, that is the number of monomers aggregated into a polymer. Hence, we consider that the $i$-th kind of object is the polymer of size $i$. We call $u_i$ the concentration of aggregates of size $i$. Since $c_i$ is the weight concentration we can write

$$c_i = i u_i M_{\text{monomer}}.$$

Moreover, the molecular weight of an oligomer of size $i$ is $i$ times the weight of a monomer, namely $M_i = i M_{\text{monomer}}$.

In conclusion, the experimental data recorded by the SLS device are a linear transformation of the second moment of the concentration distribution. Given that monomers correspond to the size $i = 1mer$, we have

$$z_{sls}(t) = \lambda_1 \left( v(t) + \sum_{i=i_0}^{i_1} i^2 u_i(t) \right) + \lambda_2 + \chi, \quad \chi \sim \mathcal{N}(0, \sigma^2), \tag{1.8}$$

where $i_0$ and $i_1$ are the sizes of the smallest and the biggest aggregates in the system, respectively. The parameters $\lambda_1$ and $\lambda_2$ depend on the experimental conditions and are unknowns. We assume that the data are affected by an additive Gaussian white noise $\chi$. We discuss the nature of this noise more fully in the following.

**Remark 1.1.4.2**
The SLS data are commonly read as the evolution of the average cluster size over time. In fact, the *average molecular weight* is defined as :

$$\langle Mw \rangle = \frac{\sum_i c_i M_i}{\sum_i c_i} = \frac{\sum_i c_i M_i}{c_{\text{tot}}}.$$

It is easy to link this quantity to the intensity of scattered light as follows

$$\langle Mw \rangle = \frac{I_{\text{scattered}}}{K c_{\text{tot}}}.$$

Taking into account Remark 1.1.4.2 and recalling the law of mass conservation expressed in Equation (1.2), we deduce that the *average cluster size* is defined as follows

$$\langle i \rangle = \frac{\langle Mw \rangle}{M_{\text{monomer}}} = \frac{I_{\text{scattered}}}{M_{\text{monomer}} K c_{\text{tot}}} = \frac{1}{\rho} \left( v + \sum_{i=i_0}^{i_1} i^2 u_i \right).$$

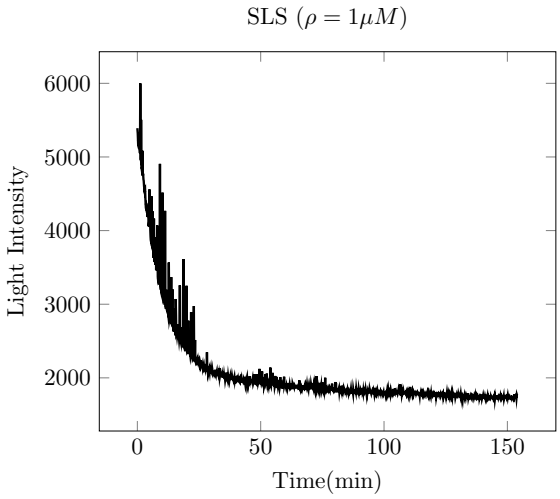



FIGURE 1.9 – SLS data at $\rho = 1\mu M$.

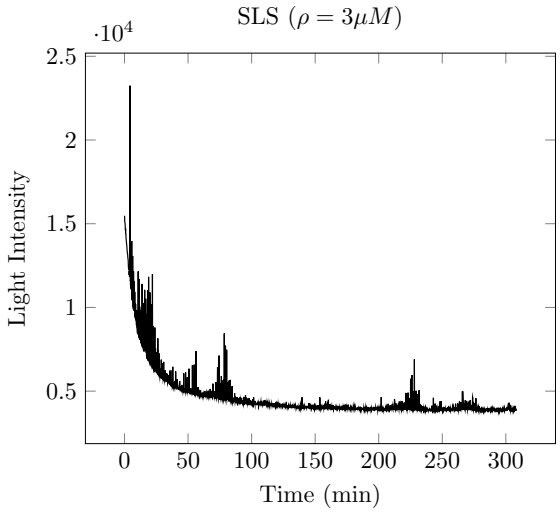
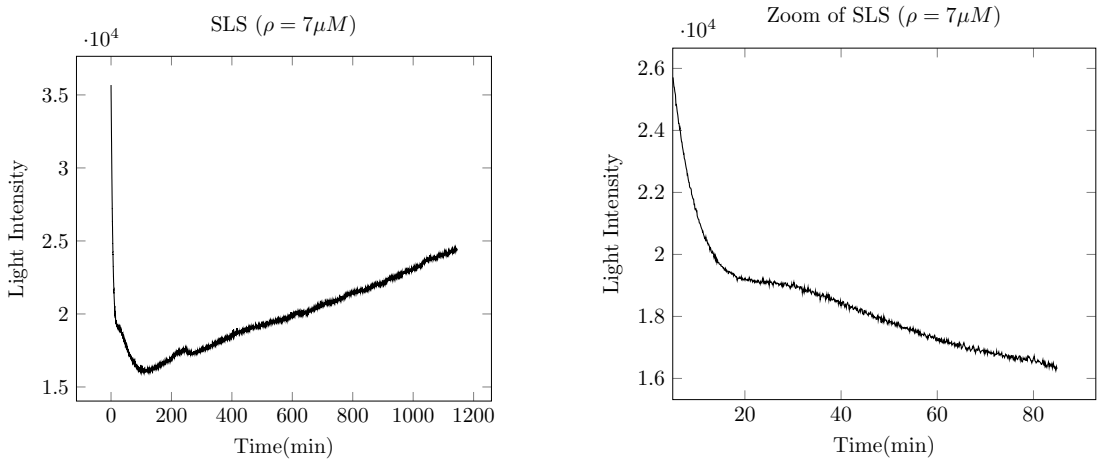



FIGURE 1.10 – SLS data at $\rho = 3\mu M$.




(a) (b)

FIGURE 1.11 – SLS data at $\rho = 7\mu M$. Left : data recorded at times in $[0, 1140]$min. Right : data recorded at times in $[5, 85]$min.

In Figures 1.9, 1.10 and 1.11 we report the experimental SLS data in the three concentration regimes $\rho = 1,\ 3,\ 7\mu M$.

In the $7\mu M$ case, we observe a depolymerisation phase followed by a polymerisation phase. We also notice a plateau phase between two depolymerisation phases, see Figure 1.11b.

To use these data, we need to estimate the unknown parameters $\lambda_1$, $\lambda_2$ and quantify the noise level. A precise estimation of these quantities is extremely important for the success of the data assimilation strategy. In the following, we detail the methodology applied to set the parameters $\lambda_1$, $\lambda_2$ and to analyse the noise.

**1.1.4.a SLS experimental constants** We start by establishing a criterion to define the coefficient $\lambda_2$. The description of SLS data given in (1.8) can be equivalently written as

$$z_{sls}(t) = \lambda_1 \left( v(t) + \sum_{i=i_0}^{i_1} i^2 u_i(t) \right) + \chi, \quad \chi \sim \mathcal{N}(\lambda_2, \sigma^2). \tag{1.9}$$

We thus consider $\lambda_2$ as the mean of the measurement noise. During the experimental phase, it was possible to perform the SLS test on a cuvette with no proteins in it, formally $\rho = 0\mu M$.

The data collected are presented in Figure 1.12. In this setting we have

$$z_{sls,\rho=0} = \chi.$$

Consequently, we can define

$$\lambda_2 = \text{mean}[z_{sls,\rho=0}]. \tag{1.10}$$

We compute the mean as follows : given $n$ data points $\{x_i\}_{i=1,\dots,n}$, the mean $\bar{x}$ is

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n}.$$

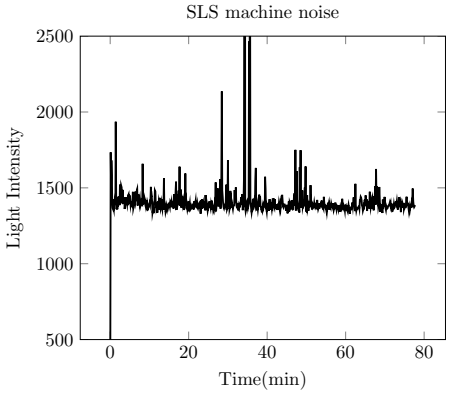



FIGURE 1.12 – SLS data on a cuvette containing no protein.

The estimated mean [3] of the collected data is

$$\lambda_2 = 1393.$$

We do not have experimental data from which we can derive an estimation on the multiplicative coefficient $\lambda_1$. At first, we decided to set this coefficient in order to guarantee coherence between SEC and SLS data. To do so, for each time $\{t_j\}$ at which we have SEC data, we compute the quantity

$$y_j = \frac{1}{\lambda_m}\int_{17}^{25} z_{\text{sec,m}}(V,t_j)\mathrm{d}V + \frac{1}{\lambda_o}\sum_{i=i_0}^{i_1} i^2 z_{\text{sec,o}}(i,t_j). \tag{1.11}$$

If experimental data were not affected by the noise we would have $y_j = \dfrac{z_{sls}(t_j)-\lambda_2}{\lambda_1}$. We take $\lambda_1$ as the factor that minimises the distances of the points $\{(t_j,y_j)\}_{j=1,\dots,N_{\text{SEC obs}}}$ from the curve $\dfrac{z_{sls}(t_j)-\lambda_2}{\lambda_1}$. Assuming that we may make a small error in recording the experimental time $t_j$, we want to compare the point $(t_j,y_j)$ to the values of the SLS taken at times in a neighbourhood of the time $t_j$. We thus define the distance of a point $(t,y)$ from a curve $f$ as follows

$$d[(t,y);f] = \min_{s\in[t-\varepsilon,t+\varepsilon]}\{(s-t)^2+(f(s)-y)^2\}. \tag{1.12}$$

Consequently, we set $\lambda_1$ to

$$\lambda_1 = \arg\min_c \left\{ \sum_{j=1}^{N_{\text{SEC obs}}} d\left[(t_j,y_j);\frac{z_{sls}-\lambda_2}{c}\right]\right\}.$$

In practice, we choose $\varepsilon$ such that we have 20 SLS data points in the range $[t-\varepsilon,t+\varepsilon]$. Applying this criterion for the three concentrations we obtain the coefficients

$$\lambda_1 = 100 \ \ (1\mu M), \qquad \lambda_1 = 111 \ \ (3\mu M), \qquad \lambda_1 = 125 \ \ (7\mu M).$$

3. The estimation is performed using the Matlab integrated function `normfit`.

This choice gives us a good agreement between the two classes of measurements. However, it lacks a biological explanation. This is why, even if we have found a good criterion, we question again the definition of this coefficient. We recall that all the oligomer distributions at time $t = 0$ correspond to the same distribution, rescaled by a factor depending on the total mass concentration, see Figure 1.7. We can assume that – once we have set the parameter $\lambda_o$ – we can have a high degree of confidence in the SEC data at time $t = 0$. Looking for the parameter $\lambda_1$ that makes the SLS data start from the second moment of the initial distribution given by the SEC data, we obtain

$$\lambda_1 = 93 \ \ (1\mu M), \qquad \lambda_1 = 109 \ \ (3\mu M), \qquad \lambda_1 = 114 \ \ (7\mu M).$$

We take the mean of these values and set

$$\lambda_1 = 105.$$

With this second approach, we obtain a value that is close to the value obtained with the first method. Now, however, it is supported by biological interpretation.

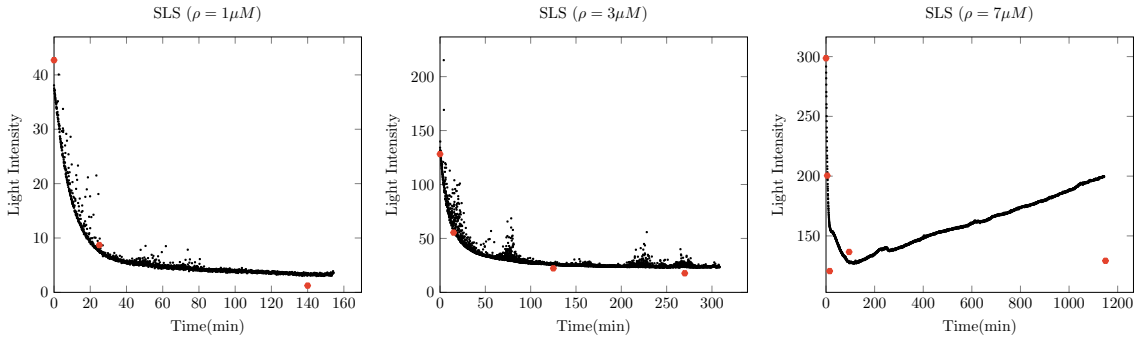
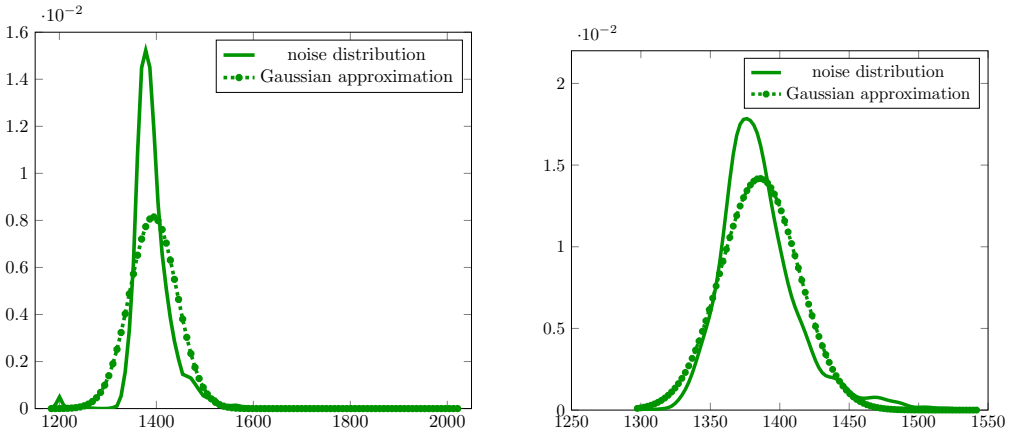



FIGURE 1.13 – Comparison between the data $\dfrac{z_{sls} - \lambda_2}{\lambda_1}$, with $\lambda_2 = 1393, \lambda_1 = 105$ (black curve), and the points $\{(t_j, y_j)\}_{j=1,\dots,N_{\text{SEC obs}}}$ defined in Equation (1.11) (red dots).

In Figure 1.13 we compare the transformed data $\dfrac{z_{sls} - \lambda_2}{\lambda_1}$ to the values of the second moment computed from the SEC data $\{(t_i, y_i)\}_i = 1, \dots, N_{\text{SEC obs}}$. We notice that we have a good agreement in the case of $\rho = 1$, $3\mu M$, while for $\rho = 7\mu M$ we have the point relative to $t = 1150$min far from the SLS curve. However – analysing the noise on the SEC data – we have already said that the noise on the SEC curve at $\rho = 7\mu M$ and $t = 1150$min is high. Therefore we place greater trust in the SLS data.

**1.1.4.b SLS data noise** As explained in [84], the intensity of the scattered light oscillates in time due to the Brownian diffusive motion of macromolecules. The movement causes intensity fluctuation. The SLS data in Figure 1.12 – collected by observing a cuvette without proteins – can be used as measurements of the noise.

A useful tool to display the distribution shape of a set of data is the *histogram* [30]. The histogram is built by dividing the data range of values into intervals (bins). At each interval a box is placed whose height depends on the number of data points falling into the interval




(a) Noise distribution estimated on the totality of the data

(b) Noise distribution estimated on the data observed between 50min and 75min.

FIGURE 1.14 – Comparison between the noise density (solid line) and the Gaussian function having the mean and variance of the data set. On the left we have the distribution on data observed in the period $[0, 75]$min and on the right in the period $[50, 75]$min

range. Given $y_i$ the observed data, $h$ the bin width and $y$ a point at which we want to estimate the density $f(y)$, the histogram estimate is

$$f_{\text{hist}}(y) = \sum_{i=1}^{n} I_h(y - y_i),$$

where $I_h$ is the characteristic function of the interval $[0, h]$, $I_h(x) = 1$ if $x \in [0, h]$ and 0 otherwise.

Moving from this basic idea, we see that it is possible to obtain a smooth estimation by replacing the characteristic function $I_h$ by a smooth *kernel function*. The kernel function is usually a symmetric probability density with zero mean [131]. In the following, we consider the normal density function with mean 0 and variance $h^2$. More precisely, we obtain the estimation

$$f_{\text{kernel fun}}(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h\sqrt{2\pi}} e^{\frac{(y-y_i)^2}{2h^2}}. \tag{1.13}$$

Furthermore, we estimate the mean and the variance of the data set as follows

$$\begin{aligned} \text{mean} \quad \mu &= \frac{1}{n} \textstyle\sum_{i=1}^{n} y_i, \\ \text{variance} \quad \sigma^2 &= \frac{1}{n-1} \textstyle\sum_{i=1}^{n} (y_i - \mu)^2. \end{aligned}$$

In Figure 1.14 we present a comparison between the noise distribution [4] and the Gaussian function relative to the estimated mean and variance.

We perform a chi-square statistical test on the default hypothesis of having a random sample normally distributed, with mean and variance estimated from the sample, against the

4. The noise distribution is computed according to Equation (1.13) by the Matlab integrated function `ksdensity`.

hypothesis of data not normally distributed. The result is that we can reject the normally distributed hypothesis at the 5% significance level.

When we perform our noise analysis in a zone without spikes, the error we make by approximating the noise distribution by a Gaussian function becomes smaller.

We analyse the noise on the SLS data relative to the concentrations $\rho = 1,\ 3,\ 7\mu M$. In Figures 1.15, 1.16, 1.17 we show that if we select a region in which there is no spike, we find that the noise has a normal distribution. We should point out that in these three cases – in contrast to the case of data in Figure 1.12 – we cannot perform the noise analysis directly. We consider a quadratic fit as an approximation of the noiseless data. Then, we estimate the distribution on the residuals. The noise analysis results are therefore strongly dependent upon the fitting strategy. In our work, we choose a polynomial fit. We compare the results when choosing polynomials of degrees 2, 3 and 4. We observe a global agreement in the three cases. In Figures 1.15, 1.16, 1.17 we present the results obtained using a quadratic fit.

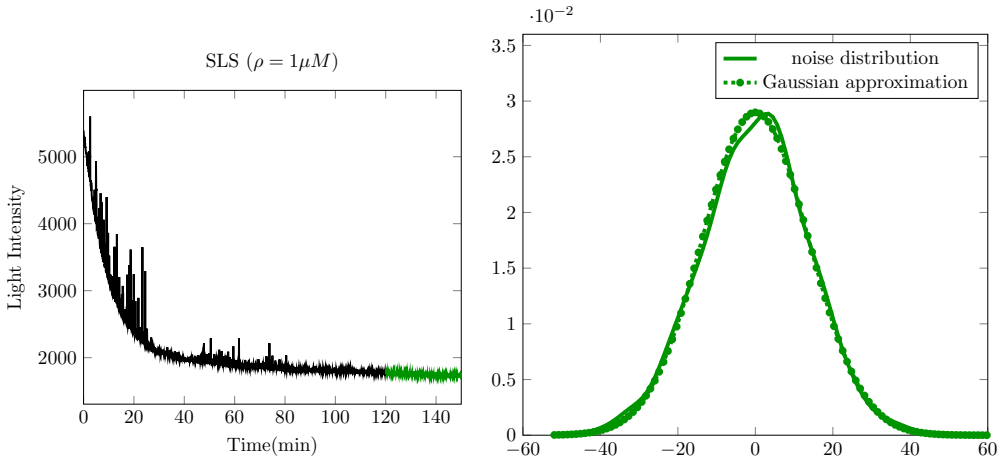



FIGURE 1.15 – Left : SLS data at $\rho = 1\mu M$. Right : Distribution of the noise on SLS data between 120 and 150 minutes (solid line) and Gaussian approximation of the noise distribution (dashed-dotted line).

The noise distribution in regions with spikes is of the kind shown in Figure 1.14. We notice that the spikes occur more at lower concentrations and that the presence of spikes or their intensity is inversely proportional to the concentration $\rho$. We deduce that there are at least two noise sources : an additive white Gaussian noise and a noise characterised by spikes. We decided to limit the influence of this second noise source by applying a low-pass filter on the data. In the rest of our study, we work on the filtered data and we assume them to be affected only by an additive white noise.

### 1.1.5 Normalised experimental data

In Figures 1.18, 1.19 and 1.20 we present an outline of the rescaled observation data to which we refer in the following. With a mild abuse of notation we refer to the SLS data as

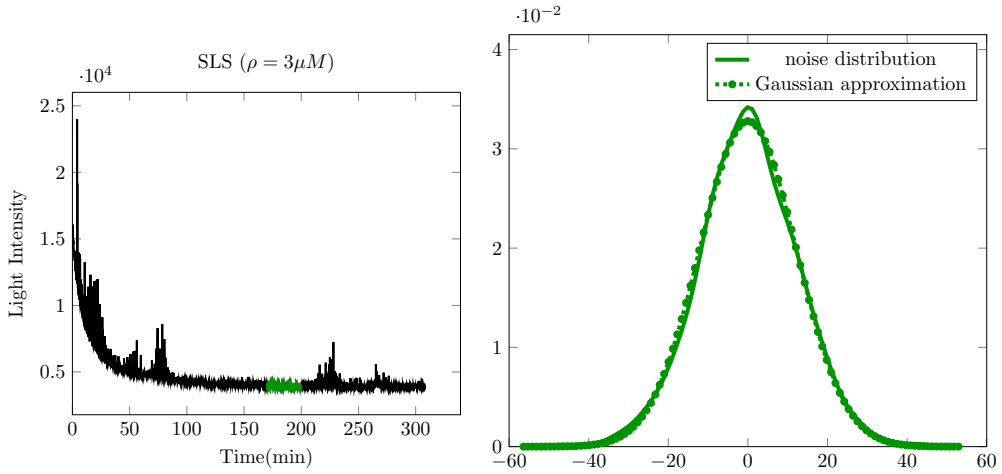



FIGURE 1.16 – Left : SLS data at $\rho = 3\mu M$. Right : Distribution of the noise on SLS data between 170 and 200 minutes (solid line) and Gaussian approximation of the noise distribution (dashed-dotted line).

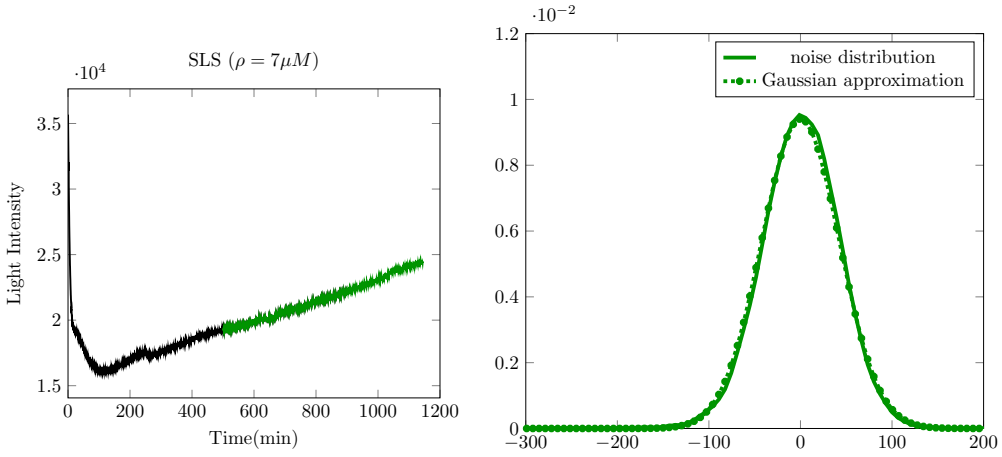
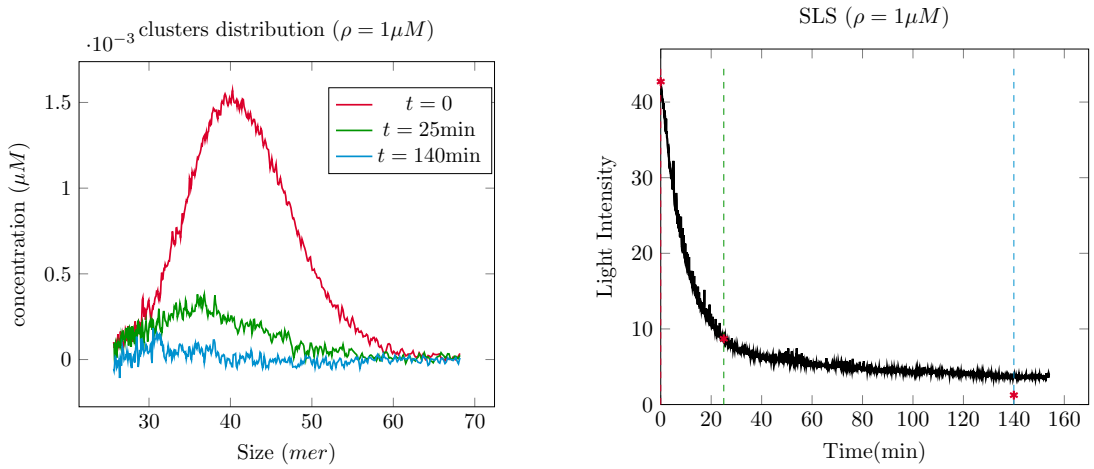



FIGURE 1.17 – Left : SLS data at $\rho = 7\mu M$. Right : Distribution of the noise on SLS data between 500 and 1190 minutes (solid line) and Gaussian approximation of the noise distribution (dashed-dotted line).

$$z_{sls} = \frac{z_{sls} - \lambda_2}{\lambda_1} = v(t) + \sum_{i=i_0}^{i_1} i^2 u_i(t). \tag{1.14}$$




(a) SEC data

(b) SLS data

FIGURE 1.18 – Experimental data on ovPrP oligomer size distribution, total concentration $\rho = 1\mu M$.

## 1.2 Design of a mathematical model

In this section we design a mathematical model step by step to describe the observed phenomena of *in vitro* prion oligomer polymerisation. Following a classical approach [163, 19, 81], we start by identifying the most important chemical reactions. We then transform these reactions into an ODE system. The reactions that we have selected are :

— the cluster's gain of monomers, namely **polymerisation**,

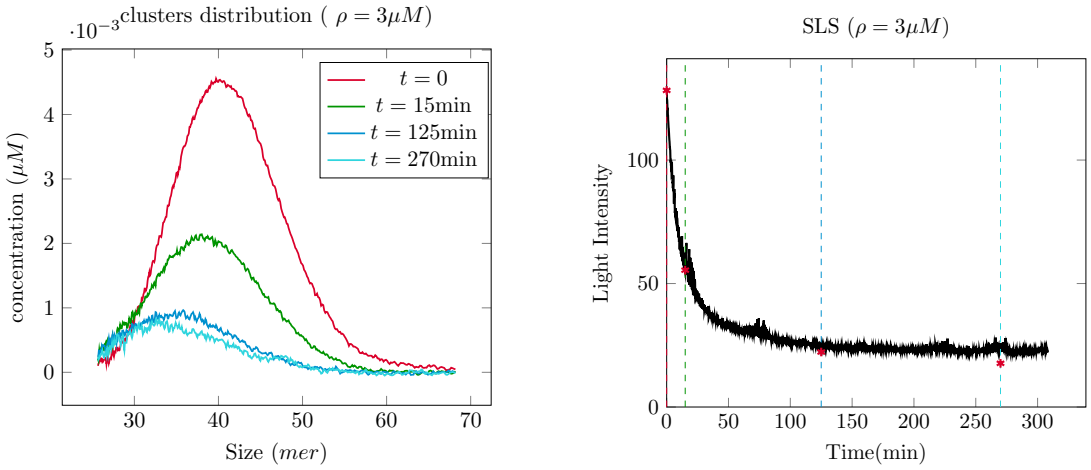
— the cluster's loss of monomers, namely **depolymerisation**,

— the cluster's decomposition into monomers, namely **disintegration**.

Polymerisation and depolymerisation reactions have been frequently considered in prion models [19]. The main novelty of our work is that we take the disintegration reaction into account.

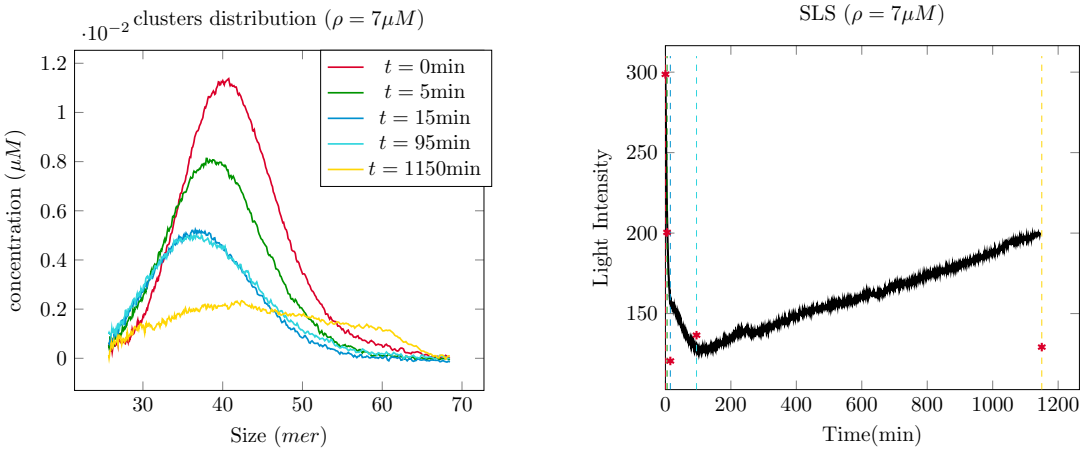
By observing the SLS data Figures 1.18b, 1.19b, 1.20b, we notice that the average oligomer size is both decreasing – as in the experiments at $\rho = 1\mu M$, $\rho = 3\mu M$ – and increasing – as in the experiment at $\rho = 7\mu M$. We thus need to take into account at least two classes of processes : one making the cluster grow in size and the other making the cluster reduce in size.

Moreover, we do not take into account the creation of new oligomers. The formation of new oligomers through monomer aggregation is in fact a very slow reaction that requires much longer times than the experimental time as presented in [40].




(a) SEC data (b) SLS data

FIGURE 1.19 – Experimental data on ovPrP oligomer size distribution, total concentration $\rho = 3\mu M$.




(a) SEC data (b) SLS data

FIGURE 1.20 – Experimental data on ovPrP oligomer size distribution, total concentration $\rho = 7\mu M$.

### 1.2.1 Oligomer size-increasing process

Several growth processes have been proposed in the literature [170, 62, 16]. We refer to the work [62] in which the authors propose a model of fibril elongation by attachment of small oligomers and the growth by polymerisation is proposed as a first simplification of a more complex model.

For the system considered, the hypothesis of growth by *polymerisation* is supported by biologists. We denote by $k_{\mathrm{on}i}$ the rate, or speed, at which an oligomer of size $i$, namely $o_i$, gains a monomer, namely $m$, becoming of size $i+1$. The reaction rates are always assumed to be positive or null. Formally we have

$$o_i + m \xrightarrow{k_{\mathrm{on}i}} o_{i+1}, \quad i_0 \leq i \leq i_1. \tag{1.15}$$

Thanks to the *law of mass action*, proposed by Guldberg and Waage in 1864 in [81], an elementary chemical reaction such as

$$A + B \underset{h}{\overset{k}{\rightleftharpoons}} C + D$$

can be modelled by the ordinary differential equation

$$\frac{\mathrm{d}[A]}{\mathrm{d}t} = -h[A][B] + k[C][D],$$

where $[A]$ is the concentration of $A$. Similar differential equations can be written for $B$, $C$ and $D$, see [118] for more details.

With $u_i$ and $v$ being the concentration of oligomers of size $i$ and isolated monomers, respectively, the system modelling the reactions (1.15) is

$$\begin{cases} \dfrac{\mathrm{d}u_i}{\mathrm{d}t} &= -k_{\mathrm{on}i} v u_i + k_{\mathrm{on}i-1} v u_{i-1}, \quad i_0 \leq i \leq i_i, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= -\sum_{i=i_0}^{i_1} k_{\mathrm{on}i} v u_i, \end{cases} \tag{1.16}$$

where $k_{\mathrm{on}i_0-1} = 0$. An easy calculation shows that the total monomer concentration $\rho$ – defined in equation (1.2) – is constant

$$\frac{\mathrm{d}(v + \sum_{i=i_0}^{i_1} i u_i)}{\mathrm{d}t} = \sum_{i=i_0}^{i_1-1} (-k_{\mathrm{on}i}) v u_i + i(-k_{\mathrm{on}i} v u_i + k_{\mathrm{on}i-1} v u_{i-1}) = 0.$$

To have an understanding of the system's behaviour over time, it might be useful to observe that the oligomer dynamics in System (1.16) can be formally seen as a first order approximation of the transport equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(k_{\mathrm{on}} v u) = 0, \tag{1.17}$$

where the function $u$ is such that $u(i,t) = u_i(t)$. More precisely, taking a Taylor expansion of $k_{\mathrm{on}} u$ around the size $i$, we have

$$k_{\mathrm{on}}(i+\delta i) u(i+\delta i, t) = k_{\mathrm{on}}(i) u(i,t) + \frac{\partial}{\partial x}\big(k_{\mathrm{on}}(i) u(i,t)\big)\delta i + \frac{1}{2}\frac{\partial^2}{\partial x^2}\big(k_{\mathrm{on}}(i) u(i,t)\big)\delta i^2 + o(\delta i^2).$$

Therefore, choosing $\delta i = -1$, we have

$$\frac{\mathrm{d}u_i}{\mathrm{d}t}(t)+k_{\mathrm{on}i}v(t)u_i(t)-k_{\mathrm{on}i-1}v(t)u_{i-1}(t) \approx \frac{\partial u}{\partial t}(i,t)+\frac{\partial}{\partial x}(k_{\mathrm{on}}(i)v(t)u(i,t))-\frac{1}{2}\frac{\partial^2}{\partial x^2}(v(t)k_{\mathrm{on}}(i)u(i,t)).$$

The evolution of the oligomer concentration can thus be approximated by the sum of two contributions :

— $\dfrac{\partial}{\partial x}(k_{\mathrm{on}}(i)v(t)u(i,t))$ describing a shift toward the right – that is toward the big sizes – at velocity $k_{\mathrm{on}}v$ ,

— $-\dfrac{\partial^2}{\partial x^2}(v(t)k_{\mathrm{on}}(i)u(i,t))$ describing the decrease of the peak value and the peak broadening.

When the polymerisation is the dominant process, we notice an increase in the oligomer average size. Therefore, we can capture this phenomenon by the analysis of the SLS data. We have an example in the case of the high concentration regime ($\rho = 7\mu M$) after 100 minutes. Moreover, when observing the SEC data in Figure 1.20a, we notice that the peak corresponding to $t = 1150$min has shifted toward the right.

These experimental data support the polymerisation hypothesis.

### 1.2.2 Oligomer size-reducing process

To select the most important size-reducing processes, we have considered, on the one hand, the models presented in literature and, on the other hand, the insights we get from scientists and experimental data.

We started from the work [62], in which the *disintegration* process was proposed as the main size-reducing process for ovPrP aggregates.

We call $k_{\mathrm{dis}}$ the disintegration rate. Thus, oligomers of size $i$ disassemble into $i$ isolated monomers at speed $k_{\mathrm{dis}i}$. As was done before, we can transform the chemical reactions

$$o_i \xrightarrow{k_{\mathrm{dis}i}} im, \qquad i_0 \leq i \leq i_1, \tag{1.18}$$

into the ordinary differential equations

$$\begin{cases} \dfrac{\mathrm{d}u_i}{\mathrm{d}t} &= -k_{\mathrm{dis}i}u_i, & i_0 \leq i \leq i_i, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= \sum_{i=i_0}^{i_1} ik_{\mathrm{dis}i}u_i. \end{cases} \tag{1.19}$$

Given the initial oligomer distribution $\{u_{0i}\}_{i_0 \leq i \leq i_i}$, the functions $u_i(t) = u_{0i}e^{-k_{\mathrm{dis}i}t}$ are solutions of System (1.19).

Plotting the oligomer concentration function $u(i,t) = u_i(t)$ at several times, we would see it flattening over time toward zero distribution. In fact, observing a system of disintegrating oligomers, after a certain time lapse, we would have only monomers. In Figure 1.21 we present a numerical example to illustrate this behaviour. In the example, we take $20 \leq i \leq 70$, $k_{\mathrm{dis}i} = k_{\mathrm{dis}} = 0.5$, and $u_{0i} = u_0(i)$ where $u_0(x) = e^{\frac{(x-45)^2}{2\cdot 5^2}}$.

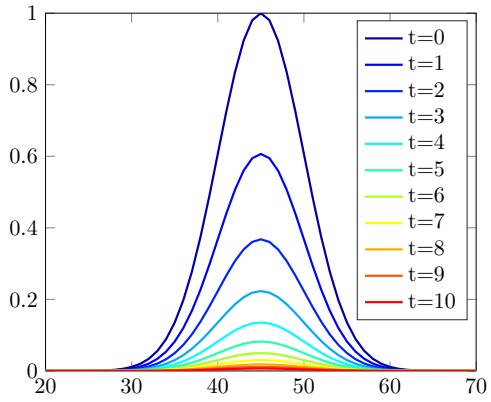



FIGURE 1.21 – Oligomer distribution evolution. Constant disintegration rate $k_{\text{dis}} = 0.5$, initial distribution $u_{0i} = e^{\frac{(i-45)^2}{2 \cdot 5^2}}$, for $20 \leq i \leq 70$.

If we assume that in our physical system oligomers can both polymerise and disintegrate, we get the model

$$\begin{cases} \dfrac{\mathrm{d}u_i}{\mathrm{d}t} &= -k_{\text{on}i} v u_i + k_{\text{on}i-1} v u_{i-1} - k_{\text{dis}i} u_i, \quad i_0 \leq i \leq i_i, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= \sum_{i=i_0}^{i_1} (-k_{\text{on}i} v u_i + i k_{\text{dis}i} u_i). \end{cases} \tag{1.20}$$

This model being a generalisation of models (1.16) and (1.19), we can reproduce both the decrease of concentration values due to disintegration and the shift toward the right of the oligomer concentration function. Furthermore, this model can also simulate the behaviour observed in Figure 1.18a. The left shift of the peak could in fact be obtained by considering higher disintegration rates for large sizes than for small sizes. In Figure 1.22 we present a numerical example in which we have $k_{\text{on}i} = 0$ and $k_{\text{dis}i} = 0.05 \cdot 2^{\frac{7i}{51}}$, for $20 \leq i \leq 70$.

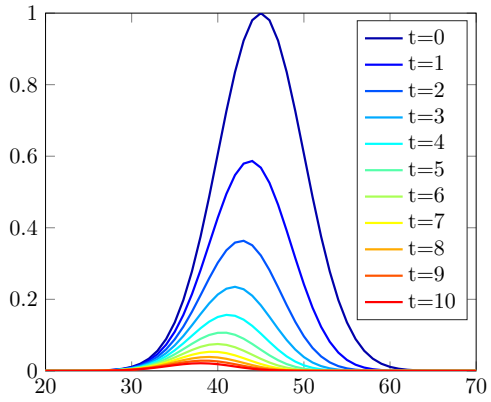



FIGURE 1.22 – Oligomer distribution evolution. Kinetic rates $k_{\text{on}i} = 0$ and $k_{\text{dis}i} = 0.05 \cdot 2^{\frac{7i}{51}}$ and initial distribution $u_{0i} = e^{\frac{(i-45)^2}{2 \cdot 5^2}}$, for $20 \leq i \leq 70$.

The analysis of the evolution at a medium concentration regime highlights a more interesting behaviour. We point out two features of the data in Figure 1.19a : 1) the concentration of small oligomers (between $20mer$ an $30mer$) increases over time ; 2) the distribution measured at 125min does not differ much from the distribution at time 270min.

From the second feature, we can conjecture that the distribution reaches a pseudo steady state. To model a steady state, we need to consider two processes that balance each other. In other words, we should consider reversible reactions.
We have assumed that monomers cannot polymerise. Therefore, it is not possible to balance the disintegration process.

However, we can balance the polymerisation process by introducing the *depolymerisation* process in our model. We call $k_{\mathrm{dep}_i}$ the rate at which an oligomer of size $i$ loses one monomer, becoming of size $i-1$. The depolymerisation reactions read as follows

$$o_i \overset{k_{\mathrm{dep}_i}}{\longrightarrow} o_{i-1} + m, \qquad i_0 \leq i \leq i_1. \tag{1.21}$$

The corresponding mathematical model is then

$$\begin{cases} \dfrac{\mathrm{d}u_i}{\mathrm{d}t} &= -k_{\mathrm{dep}_i} u_i + k_{\mathrm{dep}_{i+1}} u_{i+1}, \quad i_0 \leq i \leq i_i, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= \sum_{i=i_0}^{i_1} k_{\mathrm{dep}_i} u_i, \end{cases} \tag{1.22}$$

where $k_{\mathrm{dep}_{i_1+1}} = 0$.

As done before for the polymerisation process, we can consider System (1.22) as the first order discretisation of the backward transport equation

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial}{\partial x}(k_{\mathrm{dep}}(x) u(x,t)) = 0, \tag{1.23}$$

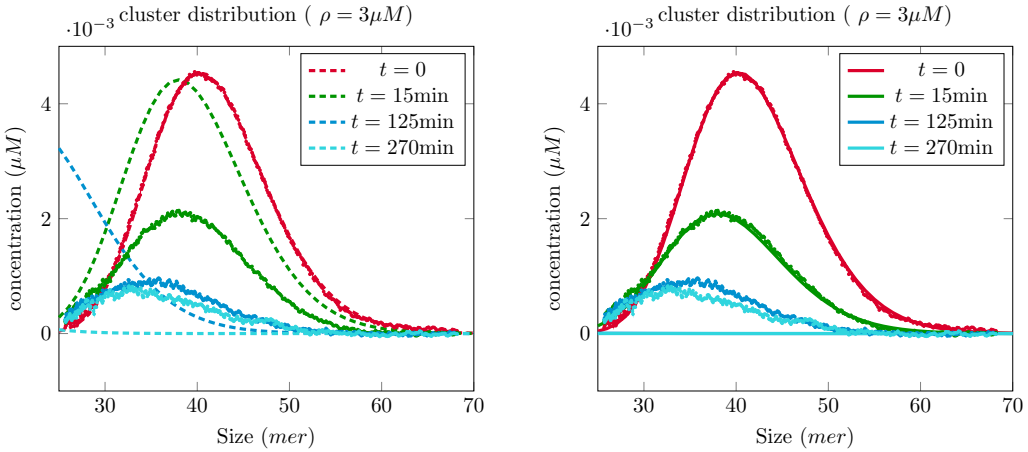
where $k_{\mathrm{dep}}(x)$ is the transport rate and we have $k_{\mathrm{dep}}(i) = k_{\mathrm{dep}_i}$ and $u(i,t) = u_i(t)$, for all $i_0 \leq i \leq i_1$.

### 1.2.3 One-species models

If we gather the three processes of polymerisation, disintegration and depolymerisation in a single model we obtain the ODE system

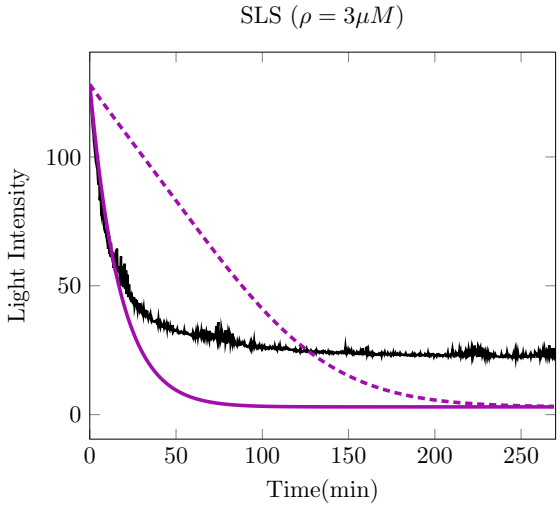
$$\begin{cases} \dfrac{\mathrm{d}u_i}{\mathrm{d}t} &= -k_{\mathrm{on}i} v u_i + k_{\mathrm{on}i-1} v u_{i-1} - k_{\mathrm{dep}_i} u_i + k_{\mathrm{dep}_{i+1}} u_{i+1} - k_{\mathrm{dis}} u_i, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= \sum_{i=i_0}^{i_1} (-k_{\mathrm{on}i} v u_i + k_{\mathrm{dep}_i} u_i + i k_{\mathrm{dis}i} u_i). \end{cases} \tag{1.24}$$

We have seen before that the polymerisation and disintegration processes alone cannot explain the empirical observations. The same holds for the polymerisation and depolymerisation processes alone. In fact, if we take $k_{\mathrm{dis}i} = 0$ for all $i$, we are not able to reproduce the rapid concentration decrease, noticed in SEC data. Let us consider the SEC measurements in the experiment at $\rho = 3\mu M$, Figure 1.19a. Starting from the observation at $t = 0$ as an initial




(a) Best fit of empirical size distribution data at $\rho = 3\mu M$ with $k_{\text{dis}i} = 0$ : parameters $\mathcal{P}_1 = \{k_{\text{on}i} = 0, k_{\text{dep}_i} = 0.16, k_{\text{dis}i} = 0\}$.

(b) Best fit of empirical size distribution data at $\rho = 3\mu M$ with $k_{\text{dis}} \neq 0$ : parameters $\mathcal{P}_2 = \{k_{\text{on}i} = 0, k_{\text{dep}_i} = 0.16, k_{\text{dis}i} = 0.05\}$.




(c) Comparison between SLS data and synthetic observations associated to parameters $\mathcal{P}_1$ (dashed line) and $\mathcal{P}_2$ (solid line).

FIGURE 1.23

condition, we want to fit the first oligomer distribution corresponding to the observation time $t = 15\text{min}$.

If we do not consider the disintegration process, the best result we can have is fitting the peak position but not the peak value, as shown in Figure 1.23a. When we include the effects of the disintegration, we obtain a good agreement both in peak position and in peak value, see Figure 1.23b. In conclusion, depolymerisation alone cannot explain the rapid decrease in oligomer concentration.

A further analysis shows that the disintegration, needed to fit the first curve, forces the oligomer system to vanish. We can observe this behaviour in Figures 1.23b, 1.23c. Therefore, we must modify the model to limit the effects of disintegration.

Several hypotheses have been formulated and tested. Substituting the term $k_{\text{dis}i}u_i$ in System (1.24) by a term of the form $k_{\text{dis}i}f_i(t)u_i(t)$, where $f_i$ are decreasing functions, we can slow down the disintegration. The system with this new term could thus represent ovPrP oligomer evolution better than System (1.24). This term can be obtained by considering a modified disintegration process. Let $\phi$ play the role of a catalyser, we have

$$o_i + \phi \quad \overset{k_{dis_i}}{\rightarrow} \quad \phi + i\,m, \quad i_0 \leq i \leq i_1.$$

Let $f$ be the concentration of $\phi$. These reactions correspond to the differential equations

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = -k_{\text{dis}i} f u_i \qquad i_0 \leq i \leq i_1.$$

We aim at defining a candidate or a set of candidates for $\phi$.

We start by considering the candidates $\{o_j\}_{i_0 \leq j \leq i_1}$ and hence the reactions

$$o_i + o_j \quad \overset{k_{dis_i}}{\rightarrow} \quad o_j + i\,m, \quad i_0 \leq i, j \leq i_1. \tag{1.25}$$

Taking into account the polymerisation and depolymerisation processes as well, we have

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = -k_{\text{on}i} v u_i + k_{\text{on}i-1} v u_{i-1} - k_{\text{dep}_i} u_i + k_{\text{dep}_{i+1}} u_{i+1} - k_{\text{dis}i} \left( \sum_{j=i_0}^{i_1} u_j \right) u_i.$$

At least in the cases of low and medium total monomer concentration regimes, we know that the quantity $\sum_{j=i_0}^{i_1} u_j$ decreases over time. With this hypothesis we are able to have a good fit of the SLS data for $\rho = 1\mu M$ and $3\mu M$. However, we are still not able to fit the distributions in the SEC data or to describe the evolution at high concentration regimes.

After discussing this model with the biologists who performed the experiments, we were able to discard it without any further analysis, since there are no chemical results supporting this assumption. In fact, reactions (1.25) imply that oligomers $o_i$ and $o_j$ bind together for a long enough time lapse (of the order of magnitude of $10^{-6}$ seconds) to be experimentally observed.

It is interesting to notice how the existence of interdisciplinary collaboration played a key role in the process of designing models. It gave a mutually better understanding of the physical phenomenon of ovPrP oligomer evolution.

For instance, one of the hypotheses proposed by the biologists consisted in having $\phi = nm$, where $n$ is a small integer ($n \leq 3$). When an oligomer gains $n$ monomers, it may either increase

its size, as in the polymerisation process, or it may become unstable and then disintegrate. The difference in behaviour is due to the point at which the monomer is attached. We notice that, under this assumption, we get the model

$$\begin{cases} \dfrac{\mathrm{d}u_i}{\mathrm{d}t} &= -k_{\mathrm{on}i}vu_i + k_{\mathrm{on}i-1}vu_{i-1} - k_{\mathrm{dep}_i}u_i + k_{\mathrm{dep}_{i+1}}u_{i+1} - k_{\mathrm{dis}i}u_iv^n, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= \sum_{i=i_0}^{i_1}(-k_{\mathrm{on}i}u_1u_i + k_{\mathrm{dep}_i}u_i + ik_{\mathrm{dis}i}v^nu_i). \end{cases}$$

Analysing this model, we remark that it cannot be accepted. Even if we have a term of the form $-k_{dis_i}f(t)u_i$ in the oligomer dynamics, the function $f = v^n$ increases over time in the case of low and medium concentration regimes.

We conclude that neither $\phi = nv$ nor $\phi = \{o_j\}$ can represent all the possible behaviours we observe in the experiments. We are consequently led to investigate other hypotheses, as detailed in the next section.

### 1.2.4 Two-species model

In this section, we assume a stronger hypothesis : the existence of two oligomer species. We have seen how the assumption of the disintegration process is necessary to explain a rapid loss of mass at the beginning of the experiments. We have also noticed that, in a one-species model, this same process induces too fast an oligomer disaggregation, compared to experimental data. Introducing a second species, we are able to confine the effects of the disintegration on just a part of the oligomers assuming that only one oligomer species can disintegrate. Given that we refer to *instability* as the oligomer's tendency to disintegrate, we call the disintegrating species *unstable* and the other one *stable*. We denote by :

— $w_i$ the concentration of unstable oligomers of size $i$,

— $y_i$ the concentration of stable oligomers of size $i$.

These two kinds of oligomers can be thought of as characterised by two different space structures. In other words, the stability could depend on the way monomers aggregate to one another. We are thus not assuming the existence of two different kinds of monomers. This hypothesis together with Remark 1.1.4.1 guarantees that, in our representation of the SLS data in Equation (1.8), the multiplicative coefficient $\lambda_1$ for the intensity of the light scattered off unstable oligomers is the same as the one scattered off the stable oligomers. Consequently the relation

$$z_{sls}(t) = \lambda_1 \left( v(t) + \sum_{i=i_0}^{i_1} i^2 w_i(t) + \sum_{i=i_0}^{i_1} i^2 y_i(t) + \right) \lambda_2 + \chi$$

holds true.

We are now interested in defining the possible interactions between the two oligomer species. We propose a model in which the two species only interact through exchanging monomers. Therefore, the reactions taken into account are

$$
\begin{array}{rcl}
o_i^{\text{unstb}} & \overset{k_{\text{dis}i}}{\rightarrow} & i\,m, \\
o_i^{\text{unstb}} + m & \underset{k_{\text{dep}}{}^{w}_{i+1}}{\overset{k_{\text{on}}{}^{w}_{i}}{\rightleftharpoons}} & o_{i+1}^{\text{unstb}}, \\
o_i^{\text{stb}} + m & \underset{k_{\text{dep}}{}^{y}_{i+1}}{\overset{k_{\text{on}}{}^{y}_{i}}{\rightleftharpoons}} & o_{i+1}^{\text{stb}},
\end{array}
\tag{1.26}
$$

where $o^{\text{stb}}$ and $o^{\text{unstb}}$ are the stable and unstable oligomers, respectively.
The corresponding dynamics for oligomer concentrations are given by

$$
\left\{
\begin{array}{rcl}
\dot{w}_i &=& -k_{\text{dis}i} w_i + k_{\text{on}}{}^{w}_{i-1} w_{i-1} v - k_{\text{on}}{}^{w}_{i} w_i v + k_{\text{dep}}{}^{w}_{i+1} w_{i+1} - k_{\text{dep}}{}^{w}_{i} w_i, \\
\dot{y}_i &=& k_{\text{on}}{}^{y}_{i-1} y_{i-1} v - k_{\text{on}}{}^{y}_{i} y_i v + k_{\text{dep}}{}^{y}_{i+1} y_{i+1} - k_{\text{dep}}{}^{y}_{i} y_i, \\
\dot{v} &=& \sum_{i=i_0}^{i_1} \big( - v(k_{\text{on}}{}^{w}_{i} w_i + k_{\text{on}}{}^{y}_{i} y_i) + k_{\text{dep}}{}^{w}_{i} w_i + k_{\text{dep}}{}^{y}_{i} y_i + i k_{\text{dis}i} w_i \big).
\end{array}
\right.
\tag{1.27}
$$

We call

$$u_i(t) = w_i(t) + y_i(t)$$

the total concentration of oligomers of size $i$ at time $t$. We recall that in our experiments we measure the total oligomer concentration $u$. From Equations (1.27), it follows that the differential equation solved by $u$ is

$$
\begin{aligned}
\dot{u}_i = {} & k_{\text{on}}{}^{w}_{i-1} w_{i-1} v - k_{\text{on}}{}^{w}_{i} w_i v + k_{\text{on}}{}^{y}_{i-1} y_{i-1} v - k_{\text{on}}{}^{y}_{i} y_i v \\
& + k_{\text{dep}}{}^{w}_{i+1} w_{i+1} - k_{\text{dep}}{}^{w}_{i} w_i + k_{\text{dep}}{}^{y}_{i+1} y_{i+1} - k_{\text{dep}}{}^{y}_{i} y_i - k_{\text{dis}i} w_i.
\end{aligned}
\tag{1.28}
$$

Let us introduce the *ratio of stable oligomers* of size $i$ among all the oligomers with the same size.

$$
\begin{array}{rccc}
\alpha_i : & [0,\tau] & \longrightarrow & [0,1] \\
& t & \longmapsto & \frac{y_i(t)}{u_i(t)}.
\end{array}
$$

Since $u = w + y$, we have

$$y_i(t) = \alpha_i(t) u_i(t) \quad \text{and} \quad w_i(t) = (1 - \alpha_i(t)) u_i(t).$$

By simple replacement we obtain

$$
\begin{aligned}
\dot{u}_i = {} & [k_{\text{on}}{}^{w}_{i-1}(1-\alpha_{i-1}) u_{i-1} - k_{\text{on}}{}^{w}_{i}(1-\alpha_i) u_i + k_{\text{on}}{}^{y}_{i-1} \alpha_{i-1} u_{i-1} v - k_{\text{on}}{}^{y}_{i} \alpha_i u_i] v \\
& + k_{\text{dep}}{}^{w}_{i+1}(1-\alpha_{i+1}) u_{i+1} - k_{\text{dep}}{}^{w}_{i}(1-\alpha_i) u_i + k_{\text{dep}}{}^{y}_{i+1} \alpha_{i+1} u_{i+1} - k_{\text{dep}}{}^{y}_{i} \alpha_i u_i \\
& - k_{\text{dis}i}(1-\alpha_i) u_i.
\end{aligned}
\tag{1.29}
$$

To better understand the evolution of oligomer concentration, we interpret this ODE as the first order approximation of the PDE

$$\begin{aligned}\frac{\partial u}{\partial t}(x,t) = v(t)&\frac{\partial}{\partial x}\big((k_{\text{on}}{}^{w}(x)(1-\alpha(x,t)) + k_{\text{on}}{}^{y}(x)\alpha(x,t))u(x,t)\big)\\ &- \frac{\partial}{\partial x}\big((k_{\text{dep}}{}^{w}(x)(1-\alpha(x,t)) + k_{\text{dep}}{}^{y}(x)\alpha(x,t))u(x,t)\big)\\ &- k_{\text{dis}}(x)(1-\alpha(x,t))u(x,t).\end{aligned} \tag{1.30}$$

Due to the disintegration of unstable oligomers, the rate $\alpha_i$ is a non decreasing function of time that eventually reaches the value 1 and then remains constant. We can thus say that in finite time the concentration $u$ evolves according to a first order approximation of the transport equation

$$\frac{\partial u}{\partial t}(x,t) = \frac{\partial}{\partial x}\big((k_{\text{on}}{}^{y}(x)v(t) - k_{\text{dep}}{}^{y}(x))u(x,t)\big). \tag{1.31}$$

The continuous system reaches a steady state at time $t$ if the equation $k_{\text{on}}{}^{y}(x)v(t) = k_{\text{dep}}{}^{y}(x)$ is satisfied for all $x$.

From the literature [11, 85] we know that when the system composed of monomers and stable oligomers reaches an equilibrium, the monomer concentration satisfies

$$\underline{\lim_i}\left(\frac{k_{\text{dep}}{}^{y}_{i+1}}{k_{\text{on}}{}^{y}_{i}}\right) \le v_{\text{eq}} \le \overline{\lim_i}\left(\frac{k_{\text{dep}}{}^{y}_{i+1}}{k_{\text{on}}{}^{y}_{i}}\right).$$

In particular, when the kinetic coefficients do not depend on the size, we have that both in the continuous-size and in the discrete-size model the monomer concentration at the equilibrium is given by

$$v_{\text{eq}} = \frac{k_{\text{dep}}}{k_{\text{on}}}.$$

Let us point out that this model is in good agreement with the observation of a decelerating polymerised mass loss. In fact, we would have an initial rapid mass loss due to the tendency of unstable oligomers to disintegrate. The mass loss then decreases with the increasing rate of stable oligomers. Once we are left with only stable oligomers, the system may depolymerise, stay balanced or polymerise. We can observe these three different behaviours in our three experiments of reference, at $\rho = 1,\ 3,\ 7\mu M$, respectively, confirming the interest of observing the system in three concentration regimes.

In the analysis of this model, we start from some further assumptions.

— Polymerisation and depolymerisation of unstable oligomers are negligible processes compared to the disintegration process. Consequently, for all sizes $i$ we take

$$k_{\text{dep}}{}^{w}_{i} = 0, \qquad k_{\text{on}}{}^{w}_{i} = 0. \tag{1.32}$$

— We assume that the initial stable ratio is the same for all sizes. Formally we have

$$\alpha_i(0) = \alpha. \tag{1.33}$$

For the sake of simplicity we remove the kinetic rate superscript. The final model reads

$$\begin{cases} \dfrac{\mathrm{d}w_i}{\mathrm{d}t} &= -k_{\mathrm{dis}i} w_i, \\ \dfrac{\mathrm{d}y_i}{\mathrm{d}t} &= k_{\mathrm{on}i-1} y_{i-1} v - k_{\mathrm{on}i} y_i v + k_{\mathrm{dep}_{i+1}} y_{i+1} - k_{\mathrm{dep}_i} y_i, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= \sum_{i=i_0}^{i_1} \big( - v k_{\mathrm{on}i} y_i + k_{\mathrm{dep}_i} y_i + i k_{\mathrm{dis}i} w_i \big), \\ w_i(0) &= (1-\alpha) u_i(0), \\ y_i(0) &= \alpha u_i(0), \\ v(0) &= 0, \end{cases} \tag{1.34}$$

### 1.2.5 Boundary conditions

In the previous sections we identified the main processes characterising the evolution of the oligomer system. We are now interested in defining the size range $[i_0, i_1]$ and describing the behaviour of oligomers of extreme sizes $i_0$ and $i_1$.

Let us consider SEC data in Figures 1.18a, 1.19a, 1.20a. As explained before, small elution volumes correspond to large sizes. We consider that data associated to elution volumes less than 9ml are not reliable because heavily affected by noise. By multi-wavelength static light scattering, we have associated 9ml of elution volume to the size $70mer$. We can conclude that the SEC device is able to distinguish between aggregates made of up to 70 monomers. However, we cannot assume that the maximal oligomer size is $70mer$. We set the maximal size to

$$i_1 = 150mer.$$

This size is assumed to be bigger than the size of largest oligomers in the system.

We focus now on the definition of the minimal size $i_0$. We notice that the oligomer peak and the monomer peak are always separated by a region with null concentration between 15ml and 18ml, concluding that the minimal oligomer size is the one corresponding to 15ml. Therefore, we set

$$i_0 = 25mer.$$

We define a boundary condition to guarantee that, in our model, no oligomers of size less than $i_0$ can be created. Preliminary results in the analysis of oligomer structure indicate the existence of a structural kernel to which monomers attach, see Figure 1.24. The kernel structure is such that, if a monomer detaches from it, then all the monomers are freed simultaneously. We assume that the kernel size is $25mer$.

We represent this feature by imposing a condition of full depolymerisation at size $25mer$. Specifically, oligomers of size $25mer$ depolymerise becoming of size $24mer$ and then instantaneously disintegrate into 24 isolated monomers.

These assumptions are reflected in the term

$$(i_0 - 1) k_{\mathrm{dep}_{i_0}} u_{i_0} \tag{1.35}$$

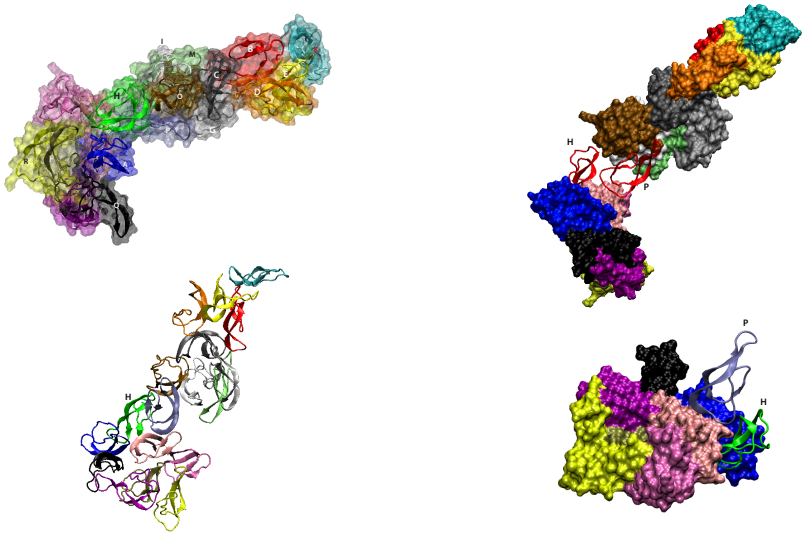


FIGURE 1.24 – Molecular dynamic simulation of O1 oligomers performed at $320°K$ using a H2H3 segment of sheep PrP. The structure shows the existence of a base composed of about 7 PrP protomers on which several other PrP molecules are grafted. The ensemble constitutes ovPrP oligomers. According to energetic considerations, the stability of the base is different than for the rest of the assembly. Therefore, the base could correspond to the minimum size of an assembly that could reach O1 before the occurrence of the disintegration process. Source :Francesca Collu and Franca Fraternali, data not yet published.

added to the dynamics of monomer concentration.

## 1.3 Inverse problem and data assimilation method

In this chapter, we have introduced a physical system of ovine prion proteins (ovPrP) and the experimental strategies designed to observe the *in vitro* evolution of this system. We have detailed how to interpret the experimental data and estimate their reliability. We have then been able to propose a mathematical model that provides a good representation of the qualitative behaviour noticed in the empirical observations. We recall that it is not possible to measure directly the kinetic parameters or the initial distribution of the two species. It is thus necessary to provide an alternative strategy to estimate these quantities. In the following, we show how we use this model to perform the estimations of the kinetic parameters and the initial conditions associated to the three experiments presented before. We solve this problem in the framework of data assimilation. Data assimilation strategies are designed to estimate the state or the parameters of a system, through the information contained in some observations of the system. A first application of data assimilation methods can be found in the fields of meteorology and oceanography [54, 75, 99]. They have then been used as a powerful tool of research in almost any applied field. In our application, we adopt a data assimilation strategy known as Extended Kalman Filter (EKF) method. This method is a natural extension of the Kalman Filter method in the non-linear case. For this reason, we first provide a practical introduction on the Kalman Filter method with the description of the general principles and the deduction of an algorithm in the linear case. We then present the Extended Kalman Filter method. Further details can be found in [61, 98, 157, 74, 21, 166, 73] and the references therein. We illustrate how we apply the EKF to our problem. We conclude by presenting the final estimations and commenting on their reliability.

### 1.3.1 Initial size distribution

In this section we focus on the initial condition of System (1.34). We notice that the initial condition is completely characterised by the total distribution $u_i(0)$ and the ratio $\alpha$.

We have already detailed how the SEC data are affected by two kinds of noise : an additive noise and a noise on the peak shape. Comparing the initial oligomer distribution in the three experimental cases, see Figure 1.7, we deduce that the peak shape is perfectly known. To have an estimation of the initial distribution, we decide to filter the additive noise in the SEC data. To do so, we refer to the SEC theory saying that the peaks in Figure 1.7 can be fitted by a Gaussian function. Furthermore, from the literature [1], we know that there is a logarithmic relationship between the elution volume ($V$) and the molecular weight. Specifically, we can write

$$\log(\text{mol weight}) = \bar{\gamma}_1 - \gamma_2 V,$$

where $\bar{\gamma}_1$, $\gamma_2 \in \mathbb{R}$. The molecular weight corresponds to the size times the monomer weight. Therefore, calling $x$ the size and $M_{\text{monomer}}$ the monomer weight, we obtain

$$\log(x) = (\bar{\gamma}_1 - \log(M_{\text{monomer}})) - \gamma_2 V = \gamma_1 - \gamma_2 V,$$

where $\gamma_1 = \bar{\gamma}_1 - \log(M_{\text{monomer}})$. Fitting the SEC data with a Gaussian function of parameters $a, m, s$, we have $z_{\text{sec}}(V) = ae^{-\frac{(V-m)^2}{s^2}} + \text{noise}$. Consequently, when we transform the data to

have the oligomer distribution with respect to the size, we obtain

$$z_{\text{sec,o}}(x) = \frac{ae^{-\frac{(\frac{\gamma_1-\log x}{\gamma_2}-m)^2}{s^2}}}{x} + \text{noise} = \frac{ae^{-\frac{(\log x+\gamma_2 m-\gamma_1)^2}{(\gamma_2 s)^2}}}{x} + \text{noise}. \tag{1.36}$$

We conclude that the oligomer size distribution data $z_{\text{sec,o}}$ can be fitted by a lognormal function.

To estimate the coefficients $\gamma_1$ and $\gamma_2$, we linearly fit the empirical calibration curve of Figure 1.4 in logscale. The resulting fit is presented in Figure 1.25. The corresponding parameters are $\gamma_1 = 5.827$, $\gamma_2 = 0.1728$. In practice, we fit the initial distribution from the

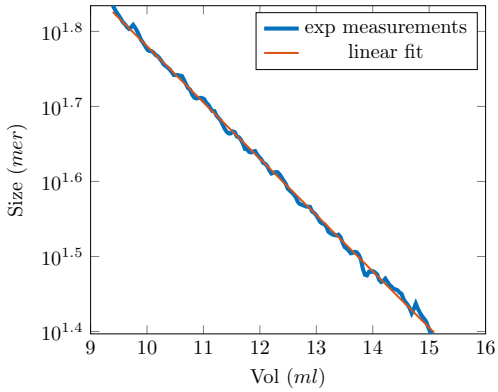



FIGURE 1.25 – Linear fit of the empirical calibration curve in log-scale.

SEC data with a Gaussian function. We then substitute the computed Gaussian parameters in the formula (1.36), obtaining an estimation of the distribution $u_i(0)$. To give an example, we show in Figure 1.26 the results of this strategy in the case of the total concentration $1\mu M$. We prefer this approach rather than fitting the size distribution with a lognormal function.

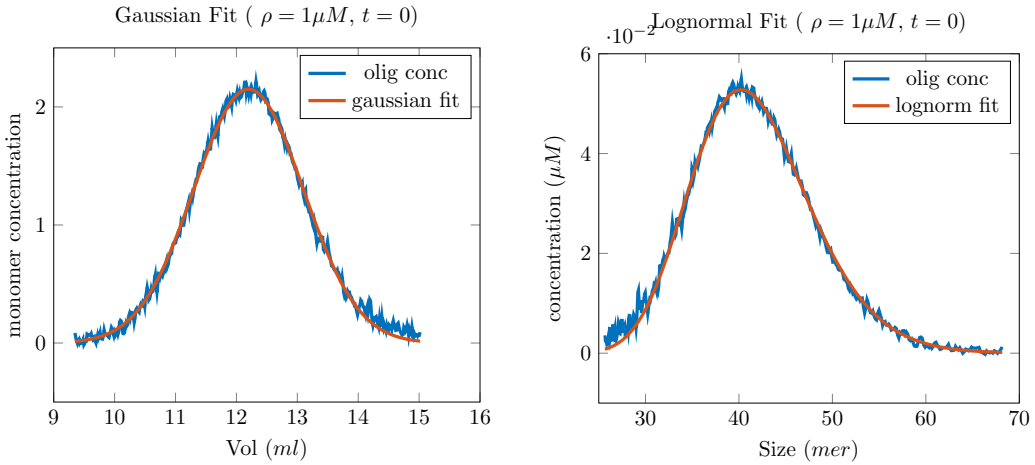



FIGURE 1.26 – Left : Gaussian fit of oligomer peak $z_{\text{sec}}$ data. Right : Fit of $z_{\text{sec,o}}$ data.

In fact, in our experience, this method results in greater accuracy and reliability.

In conclusion, using this method to estimate the distribution $u_i(0)$, we can consider the initial condition of System (1.34) fully characterised by the parameter $\alpha$.

### 1.3.2 Preliminary parameter estimation

In this section we present some strategies to estimate the system unknowns without using the system dynamics. Therefore, we use the experimental data and a qualitative description of the oligomer system. The estimations obtained in this way are, however, partial and vague. To have a more reliable estimation of these rates, it is necessary to include the model in our analysis. This task is carried out by data assimilation. We give more details about this method in the following sections.

Beforehand, however, we should point out that the kinetic rates depend on the temperature at which the experiment has been performed. The Arrhenius empirical law [7, 8] states that the kinetic coefficients are linked to the temperature $T$ in the following way

$$k_T = Ae^{-\frac{E}{RT}},$$

where $E$ and $R$ are positive constants independent of the temperature and $A$ varies slightly with temperature. Our estimations correspond to the temperature $50°C$, which was kept the same in all the experiments.

**1.3.2.a Size-reducing rates** To gain a rough idea of the order of magnitude of the kinetic rates, we consider an approach that is commonly used by biologists. We take a simplified description of the oligomer system, illustrated by the following chemical reaction

$$o \xrightarrow{k} m. \tag{1.37}$$

We represent the oligomers as one single object that we call $o$, while $m$ are the monomers. The rate $k$ stands for the velocity at which the oligomer system transforms into monomers. It can be interpreted as the sum of the depolymerisation and disintegration rates averaged in size.

The oligomer concentration, formally $o(t) = \sum_{i=25}^{70} iu_i$, evolves according to the ODE

$$\dot{o} = -ko.$$

Noticing that $o(0) = \rho$, we have $o(t) = \rho e^{-kt}$. Thanks to the SEC data we can compute the value of $o(t)$ at the measurement times. We fit these values with an exponential function to have an estimation of $k$. We can see in Figure 1.27 the results of these estimations. We do not have a good fit in the cases where $\rho = 3\mu M$ and $\rho = 7\mu M$.

If we fit the data with a sum of two exponential functions, instead, we obtain the results in Figure 1.28. We can see that in this second case we obtain better fits. The parameters of the exponential fitting functions are presented in Table 1.2.

We can use these values to estimate the order of magnitude of the kinetic parameters as being between $10^{-3}\text{min}^{-1}$ and $10^{-1}\text{min}^{-1}$.

If we want to give a biological interpretation of the fitting parameters, we need to make a step further in our still simple analysis. In writing the chemical reaction (1.37) we have treated the oligomers as one global object. Nevertheless – deriving the oligomer evolution model – we have concluded that there are at least two oligomer species. We can thus make a

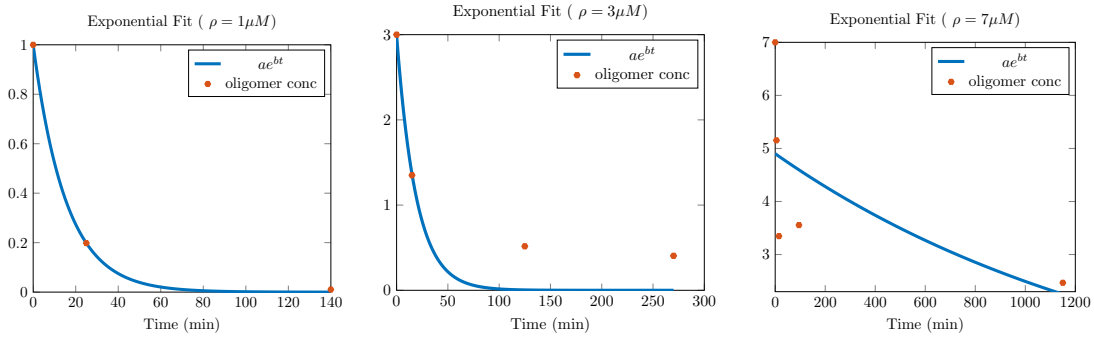



FIGURE 1.27 – Exponential fit of $\{\sum_{i=25}^{70} iu_i\}_{t_j=1,...,N\text{SEC obs}}$.

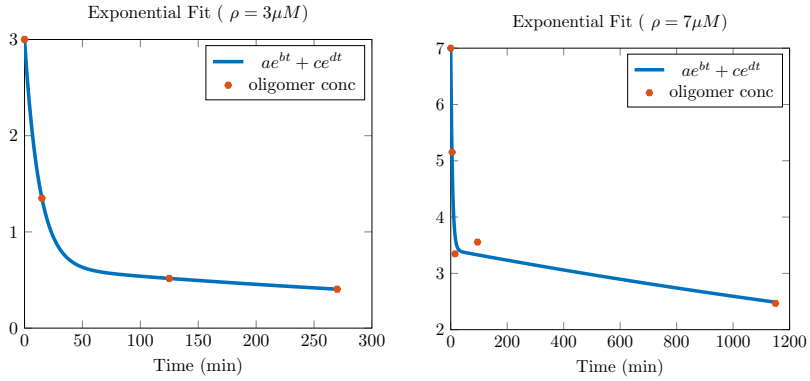



FIGURE 1.28 – Two-exponential fit of $\{\sum_{i=25}^{70} iu_i\}_{t_j=1,...,N\text{SEC obs}}$.

| $\rho$ | $a$ | $b$ | $a$ | $b$ | $c$ | $d$ |
|---|---|---|---|---|---|---|
| $1\mu M$ | 1 | -0.06476 | | | | |
| $3\mu M$ | 2.995 | -0.05244 | 2.363 | -0.07839 | 0.6371 | -0.001681 |
| $7\mu M$ | 4.901 | -0.00675 | 3.636 | -0.17 | 3.417 | -0.00276 |

TABLE 1.2 – Coefficients of the exponential fits $f(t) = ae^{bt}$ in Figure 1.27 (Columns : 1, 2) and the fits $f(t) = ae^{bt} + ce^{dt}$ in Figure 1.28 (Columns : 3, 4, 5, 6).

more accurate assumption by grouping the oligomers by their species. In this way we should consider the reactions

$$\begin{aligned} o^{\text{unstb}} &\xrightarrow{k^{\text{unstb}}} m, \\ o^{\text{stb}} &\xrightarrow{k^{\text{stb}}} m. \end{aligned} \tag{1.38}$$

For each group we can make the same considerations as in the previous case. Given that each group concentration can be represented by an exponential function, the total concentration $o(t)$ results in the sum of two exponential functions.
In conclusion, we can interpret the parameters $b$ and $d$ in the exponential fit as the estimations of the characteristic speeds at which the stable and unstable oligomers transform into monomers. However, we do not have a way to distinguish between the two parameters and thus associate them to $k^{\text{unstb}}$ or $k^{\text{stb}}$.

The results of this parameter analysis can support the existence of two kinds of oligomer species.

**1.3.2.b Size dependence of the rates** A critical point in our research has been to determine the nature of the dependency of the kinetic rates on the oligomer size. To answer this question, several hypotheses have been done and tested numerically but none of them resulted well adapted. The previous analysis on the size-reducing parameters tells us that – when we represent the kinetic parameters by their average values – we may obtain a reasonable representation of the system. We have thus considered the case of size-independent coefficients

$$k_{\text{on}i} = k_{\text{on}}, \qquad k_{\text{dep}_i} = k_{\text{dep}}, \qquad k_{\text{dis}i} = k_{\text{dis}}, \tag{1.39}$$

so that the average values correspond to the constant values $k_{\text{on}}, k_{\text{dep}}, k_{\text{dis}}$. To test this hypothesis, our collaborators designed a new biological experiment. Starting from an oligomer sample, formed with the previously detailed protocol, we perform a SEC test to separate the initial oligomer system in two groups depending on the size. The big and small aggregates are divided in two different groups. We then perform the SLS test on the two groups.

The experiment has been done twice with two different initial total concentrations : $\rho = 0.3\mu M$ and $\rho = 6\mu M$. In Figure 1.29 we compare the SLS measurements on the two groups. We notice a qualitative agreement between the SLS data on big and small sizes. We can thus assume that the two groups evolve in the same way and in particular – since the evolution is mainly ruled by the kinetic rates – that there is not a significant difference between the kinetic rates for the two groups. This experiment is thus supporting the hypothesis (1.39). In the following we show that is possible to find a set of parameter estimations in agreement with this hypothesis.

**1.3.2.c Ratio of stable oligomers** In the following, we focus on the relationship between the disintegration rate – that we assume to be size-independent – and the ratio of stable oligomers $\alpha$. In our model, the values of the oligomer distribution peak depend only on two factors. On the one hand, the dissipation – associated to the stable oligomer evolution – leads to a decrease in the peak value and to peak broadening. On the other hand, the disintegration of unstable oligomers makes the peak value decrease.

We assume that the dissipation has a just a minor contribution to the value of the peak. From Equation (1.29) and assumptions (1.39) and (1.33), we deduce that the oligomer distri-

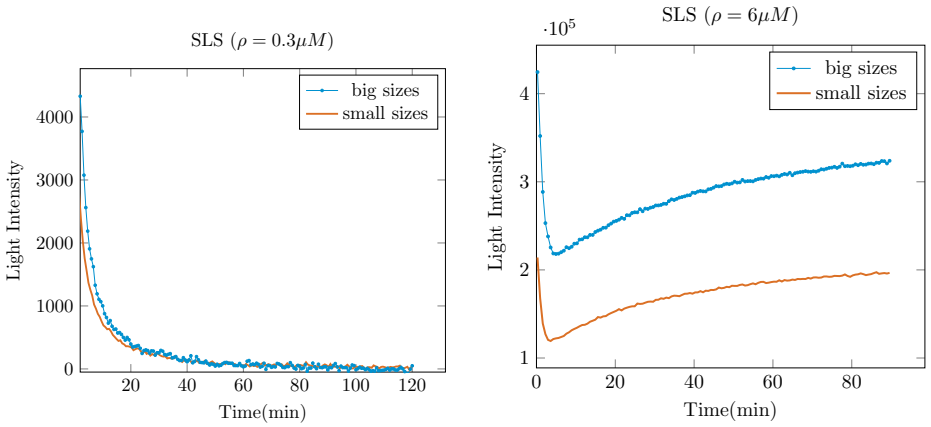



FIGURE 1.29 – SLS data. Left : total concentration $\rho = 0.3\mu M$. Right : total concentration $\rho = 6\mu M$. In blue solid-dotted line the evolutions of oligomers with sizes distributed around the size $31mer$, in orange solid line measurements on oligomers with sizes distributed around the size $42mer$.

bution peak value evolves according to the following differential equation

$$\dot{f} = -(1 - \alpha(t))k_{\text{dis}} f.$$

Consequently we have

$$\log\left(\frac{f(0)}{f(t)}\right) = k_{\text{dis}}\left(t - \int_0^t \alpha(s)\mathrm{d}s\right).$$

If the function $f$ is known, we obtain an equation linking $k_{\text{dis}}$ to $\alpha$. From the SEC data we obtain the peak value at several times. In Figure 1.30, we plot the values of $\log\left(\frac{f(0)}{f(t)}\right)$ computed from the experimental data in the three concentration cases. We notice that up

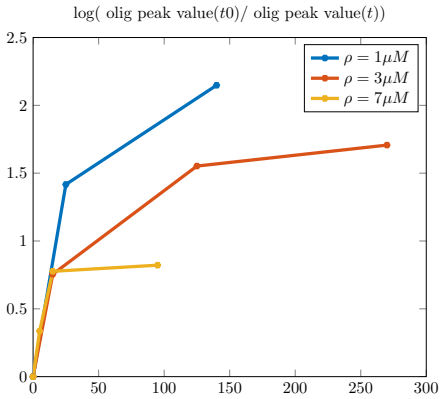



FIGURE 1.30 – Function $\log\left(\frac{f(0)}{f(t)}\right)$ obtained by linear interpolation of experimental data (dots).

to $t = 15$min the function $\log\left(\frac{f(0)}{f(t)}\right)$ grows almost linearly for all the concentrations. We

numerically estimate the slope of the function on the region $[0, 15]$min and we obtain the value 0.05. If we approximate $\log\left(\frac{f(0)}{f(t)}\right)$ by the line $0.05t$ on $[0, 15]$ min we have

$$0.05t \approx k_{\text{dis}}\left(t - \int_0^t \alpha(s)\mathrm{d}s\right), \qquad t \in [0, 15]\,\text{min}\,. \tag{1.40}$$

### 1.3.3 Inverse problem definition in a state space formalism

The definition of an inverse problem is based on the idea that we want to find some unknown value through the indirect observation of this value. For instance, to solve a linear system corresponds to solving an inverse problem. In fact, in this case we aim at finding a vector $u$ of which we can observe a linear transformation defined by the matrix $C$.
The inverse problem reads : Given $z = Cu$, find $u$.

In our case, we are interested in the case of a dynamical system on which we can take measurements. We focus on the problem of estimating the initial condition of a dynamical system using some available measurements. In the state-space formalism, we present the dynamical system in its so-called *state-space form*

$$\begin{cases} \dot{u}(t) &= A(u(t), t) + B(t)\omega(t), \qquad t \in \mathbb{R}^+, \\ u(0) &= u_\diamond + \xi, \end{cases} \tag{1.41}$$

in which we consider

— $u(t)$ the state variable at time $t$,

— $A$ the model operator,

— $B$ a model error operator

— $\omega$ some additive model noise,

— $u_\diamond$ the known part of the initial condition,

— $\xi$ the unknown part of the initial condition.

In this general formulation, we are able to take into account two sources of uncertainty : partial information on the initial condition, that has been decomposed as the sum of a known and unknown part ; a model error $B\omega$ representing the gap between the trajectory obtained with the model operator $A$ and the physical trajectory.

Furthermore – introducing the observation operator $C$ that models the process of measurement – we describe the available measurements as follows

$$z(t) = C(u, t) + \chi(t),$$

where $\chi$ denotes the observation noise. If we can directly observe our system – or part of it – the operator $C$ is simply a selection operator. However, as seen in the previous sections, in our case, we have very indirect observations albeit depending linearly on the state.

The *inverse problem* is formulated as follows

Given the observations

$$z(t) = C(\breve{u}, t) + \chi(t), \qquad t \in [0, \tau],$$

we aim at estimating $\breve{\xi}$ such that

$$\begin{cases} \dot{\breve{u}}(t) &= A(\breve{u}(t), t) + B(t)\omega(t), \qquad t \in [0, \tau], \\ \breve{u}(0) &= \breve{u}_\diamond + \breve{\xi}. \end{cases}$$

### 1.3.4 Kalman Filter theory

The Kalman filter was first designed by R.E Kalman in 1960 to solve a discrete data filtering problem [61]. The method is also known as the Kalman-Bucy filter because Bucy collaborated in the publication of the continuous-time Kalman filter in 1961 [98]. The purpose of the Kalman estimator is to approximate the state of a set of variables which are defined as the solution of a given system. By using a sequential procedure, the available observations are taken into account to improve the state estimation.

This method has a crucial role in many areas such as navigation strategy [43], target tracking [61], neuronal networks, [87], meteorology[122], oceanography [136], etc.

The Kalman estimator $\hat{u}$ is defined as the solution of a dynamical system. Its dynamics is ruled by two terms : the model of the estimating state variable and a term accounting for the distance between the experimental observations and the observations generated on the estimated state.

In a state-space formalism, we can define a sequential estimator as the solution of the following system

$$\begin{cases} \dot{\hat{u}}(t) &= A(\hat{u}(t), t) + G\big(z - C(\hat{u}(t), t)\big), \qquad t \in [0, \tau], \\ \hat{u}(0) &= u_\diamond. \end{cases}$$

The operator $G$ is called *gain operator* or *filter*. Depending on the definition of such an operator, we have different data assimilation methods.

The Kalman estimator has been designed and is only valid in a context of linear model operator and linear observation operator, see [61]. In the following we denoted the Kalman gain by $K$.

The Kalman approach was first used to solve a discrete-time problem. We started by presenting the method in this setting and then we deduce its continuous formulation. In Section 5.2.2 we then present the Kalman filter in the case of infinite-dimensional model and observation operators. For further details, we refer to [74, 157, 177].

**1.3.4.a Discrete-time Kalman estimator** The discrete-time setting is a simple framework in which to present the Kalman theory. Starting from the definition presented in this section, we will be able, in the next section, to deduce a continuous version of the Kalman estimator. Furthermore, the description of the Kalman method in a discrete setting directly provides us with a numerical algorithm to solve our estimation problem.

Let us consider a time grid $0 = t_0 < \ldots < t_N = \tau$ with constant time step $\delta t$ and a process with finite dimensional states $u_k \in \mathbb{R}^n$ at times $\{t_k\}_{k=0,\ldots,N}$. We take into account

some uncertainty on the initial condition, namely $\xi$, and the presence of some additive noise $\{\omega_k\}_{k=0,\dots,N}$ affecting the value of the process state. The stochastic states $\{u_k\}_{k=0,\dots,N}$ are defined by the following linear stochastic difference system

$$\left\{ \begin{array}{lcl} u_k & = & A_{k|k-1}u_{k-1} + B_k\omega_k, \\ u_0 & = & u_\diamond + \xi, \end{array} \right. \tag{1.42}$$

where $A_{k|k-1}$ is the linear operator that relates the state at time $t_k$ to the state at time $t_{k-1}$. Furthermore, we take into account discrete observations $z_k \in \mathbb{R}^m$ modelled by a linear observation operator $C_k$ as follows

$$z_k = C_k u_k + \chi_k. \tag{1.43}$$

We assume that the random variables $\omega_k$ and $\chi_k$ are independent, temporally uncorrelated, with zero mean and covariances $Q$ and $W$, namely *white noise*. Formally, we have

$$\mathbb{E}[\omega_k\omega_j^\intercal] = Q\delta_{k-j}, \qquad \mathbb{E}[\chi_k\chi_j^\intercal] = W\delta_{k-j},$$

where $\delta_{k-j}$ is such that $\delta_{k-j} = 1$ when $i = j$ and $\delta_{k-j} = 0$ otherwise. For the sake of simplicity, we make the assumption that the noise distribution does not change over time, since in our work we restrict ourselves to this case. However, the Kalman filter may be presented in a general setting of time-dependent covariances. Moreover, the Kalman filter can be generalised to coloured or correlated noise [32].

We introduce two state estimates for each time $t_k$

— the *a priori* state estimate $\hat{u}_k^- \in \mathbb{R}^n$, defined without using the measurement $z_k$,

— the *a posteriori* state estimate $\hat{u}_k^+ \in \mathbb{R}^n$, defined using the measurement $z_k$.

We can thus define the covariances associated to these estimates

$$P_k^- = \mathrm{Cov}(u_k - \hat{u}_k^-) = \mathbb{E}[(u_k - \hat{u}_k^-)(u_k - \hat{u}_k^-)^\intercal]$$

and

$$P_k^+ = \mathrm{Cov}(\hat{u}_k^+ - u_k) = \mathbb{E}[\hat{u}_k^+ - u_k)(\hat{u}_k^+ - u_k)^\intercal].$$

The key idea of the Kalman approach is to define the *a posteriori* state estimate as a linear combination of the *a priori* state estimate and the discrepancy between the predicted observation $C\hat{u}_k^-$ and $z_k$. Formally, we have

$$\hat{u}_k^+ = \hat{u}_k^- + K_k(z_k - C_k\hat{u}_k^-).$$

Given the multiple possible interpretations, the quantity $z_k - C\hat{u}_k^-$ has also been called *observation discrepancy*, *innovation* or *residual*. The $m \times n$ gain matrix $K_k$ is optimally defined to minimise the covariance $P_k^+$, or equivalently to minimise the *a posteriori* estimation error, over all the possible gain matrices. Formally we have

$$K_k = \arg\min_G \mathbb{E}[(\hat{u}_k^- + G(z_k - C_k\hat{u}_k^-) - u_k)(\hat{u}_k^- + G(z_k - C_k\hat{u}_k^-) - u_k)^\intercal].$$

The *a posteriori* covariance of the estimator associated to the gain $G$ may be written as follows

$$\begin{aligned} &\mathbb{E}[(\hat{u}_k^- + G(z_k - C_k\hat{u}_k^-) - u_k)(\hat{u}_k^- + G(z_k - C_k\hat{u}_k^-) - u_k)^\intercal] = \\ &\mathbb{E}[\Big((I - GC_k)(\hat{u}_k^- - u_k) + G\chi_k\Big)\Big((I - GC_k)(\hat{u}_k^- - u_k) + G\chi_k\Big)^\intercal] = \\ &(I - GC_k)P_k^-(I - GC_k)^\intercal + GWG^\intercal. \end{aligned}$$

By using the linear algebra identities

$$\frac{\partial \text{trace}[MNM^\intercal]}{\partial M} = 2MN, \qquad \frac{\partial \text{trace}[MR^\intercal]}{\partial M} = \frac{\partial \text{trace}[RM^\intercal]}{\partial M} = R,$$

with $N$ a symmetric matrix, and minimising the trace of the *a posteriori* covariance with respect to the gain, $K$ satisfies

$$-2P_k^- C_k^\intercal + 2K_k C_k P_k^- C_k^\intercal + 2K_k W = 0,$$

hence

$$K_k = P_k^- C_k^\intercal \left(C_k P_k^- C_k^\intercal + W\right)^{-1}.$$

To better understand the role of this operator, it is worth mentioning that when $W$ is small, it means that the noise in the observation is small. Moreover, given the inverse proportionality between $W$ and $K_k$, the more $W$ decreases the more $K_k$ increases. In other words, when the observation noise is small, the Kalman gain is a strong weight on the innovation term.

Furthermore, if $P_k^-$ tends to zero – which means that we have almost no error in the *a priori* state estimate – then the Kalman gain is also approaching zero. When we have a good *a priori* estimate, we do not need to correct it.

We present the Kalman estimator dynamics in a *prediction-correction* form. At each time $t_k$ we perform two steps :

— the prediction step, also called *model forecast step*, corresponds to the next step forward of our model. We compute the *a priori* state estimate $\hat{u}_k^-$ and its covariance $P_k^-$.

— In the correction step, also called *data assimilation step*, we take into account the observation $z_k$ to compute the *a posteriori* state estimate $\hat{u}_k^+$ and its covariance $P_k^+$.

The *prediction-correction* algorithm is just one of the possible forms of the Kalman filter method, we refer to [157] to have a broader overview of these algorithms.

At the initial time $t_0$, the best estimation of the initial state is given by $u_\diamond = \mathbb{E}[u_0]$. The starting condition of the Kalman filter algorithm is

$$\begin{array}{ccccc} \hat{u}_0^- & = & \hat{u}_0^+ & = & u_\diamond, \\ P_0^- & = & P_0^+ & = & \mathbb{E}[u_\diamond u_\diamond^\intercal]. \end{array}$$

At time $t_1$, the *a priori* estimation $\hat{u}_1^-$ is defined as the mean value of the state $u_1$. Using the discrete model (1.42) we can write

$$\hat{u}_1^- = \mathbb{E}[u_1] = \mathbb{E}[A_{1|0} u_0 + B_1 \omega_1] = A_{1|0} \hat{u}_0^+.$$

Consequently, the covariance $P_1^-$ is

$$\begin{aligned} P_1^- &= \mathbb{E}[(u_1 - \hat{u}_1^-)(u_1 - \hat{u}_1^-)^\intercal] \\ &= \mathbb{E}[(A_{1|0} u_0 + B_1 \omega_1 - \hat{u}_1^-)(A_{1|0} u_0 + B_1 \omega_1 - \hat{u}_1^-)^\intercal] \\ &= \mathbb{E}[(A_{1|0}(u_0 - \hat{u}_0^+) + B_1 \omega_1)(A_{1|0}(u_0 - \hat{u}_0^+) + B_1 \omega_1)^\intercal] = A_{1|0} P_0^+ A_{1|0}^\intercal + B_1 Q B_1^\intercal, \end{aligned}$$

as the $(u_0 - \hat{u}_0^+)$ and $\omega_1$ are uncorrelated.

We can extend the same argument to successive times obtaining the following formulas

$$\begin{aligned} \hat{u}_k^- &= A_{k|k-1}\hat{u}_{k-1}^+, \\ P_k^- &= A_{k|k-1}P_{k-1}^+A_{k|k-1}^\intercal + B_kQB_k^\intercal. \end{aligned} \qquad \text{(Prediction)}$$

From the *a priori* estimation we can incorporate the information of the measurement at time $t_k$ and compute the *a posteriori* state estimate $\hat{u}_k^+$ as follows

$$\begin{aligned} \hat{u}_k^+ &= \hat{u}_k^- + K_k(z_k - C_k\hat{u}_k^-), \\ K_k &= P_k^-C_k^\intercal\left(C_kP_k^-C_k^\intercal + W\right)^{-1}, \\ P_k^+ &= (I - K_kC_k)P_k^-(I - K_kC_k) + K_kWK_k^\intercal. \end{aligned} \qquad \text{(Correction)}$$

The sequential processing of measurements is one of the reasons that made this method so popular. It is, indeed, particularly well-suited to a real-time processing of the data. Moreover, since the gain operator $K$ and the covariance operators do not depend on the state of the system, they can be precomputed before the analysis of the data, saving time in real-time applications. The main drawback of this method is that it requires computing two full covariance matrices at each time step. This computation becomes prohibitive when the state dimension is too big, which would be the case, for instance, for applications in meteorology or oceanography where the state can reach millions of components [136].

**Remark 1.3.4.1**
It can be proved that the residual term $(z_k - C_k\hat{u}_k^-)$ is zero-mean white and with covariance $C_kP_k^-C_k^\intercal + W$. The Kalman filter can thus be seen as a filter that whitens the measurements extracting the maximum of the information contained [5].

**1.3.4.b Continuous-time Kalman estimator** In this section we derive the definition of the Kalman estimator and the operator $P$ in a continuous-time framework. Let us start by the relationship between the measurement error covariances in the discrete and in the continuous setting. For more details we refer to [74, 157, 97].

The continuous observations are

$$z(t) = C_{\mathrm{c}}(t)u(t) + \chi(t),$$

where the subscript 'c' stands for continuous and $\chi$ is a continuous white noise with zero mean. The noise covariance can be written as follows

$$\mathrm{Cov}(\chi(t), \chi(s)) = \delta(t - s)W(t).$$

where $\delta$ is the Dirac function with a value of $\infty$ when $t = s$, or 0 otherwise and with area equal to 1. The matrix $W(t)$ is called *spectral density matrix*. Continuous white noise is a mathematical concept that is never observed directly in any physical system. Nevertheless, it can provide an approximate description of a real process. A precise definition of continuous-time white noise is not trivial. We refer to [23] for a more complete description.

Let us take the following discrete approximation rule, for a discrete uniform time grid with time step $\delta t$,

$$u_k = \frac{1}{\delta t}\int_{t_k}^{t_k+\delta t} u(t)\mathrm{d}t.$$

We obtain

$$z_k = \frac{1}{\delta t}\int_{t_k}^{t_k+\delta t} C_{\mathrm{c}}(t)u(t) + \chi(t)\mathrm{d}t = C_k u_k + \chi_k,$$

with

$$\chi_k = \frac{1}{\delta t}\int_{t_k}^{t_k+\delta t} \chi(t)\mathrm{d}t.$$

$\chi_k$ is a discrete noise with mean

$$\mathbb{E}[\chi_k] = \frac{1}{\delta t}\int_{t_k}^{t_k+\delta t} \mathbb{E}[\chi(t)]\mathrm{d}t = 0$$

and covariance

$$\begin{aligned} W_k = \mathrm{Cov}(\chi_k) = \mathbb{E}[\chi_k\chi_k^{\intercal}] &= \frac{1}{\delta t^2}\int_{t_k}^{t_k+\delta t}\int_{t_k}^{t_k+\delta t} \mathbb{E}[\chi(t)\chi(s)^{\intercal}]\mathrm{d}t\mathrm{d}s \\ &= \frac{1}{\delta t^2}\int_{t_k}^{t_k+\delta t}\int_{t_k}^{t_k+\delta t} W(t)\delta(t-s)\mathrm{d}t\mathrm{d}s = \frac{1}{\delta t^2}\int_{t_k}^{t_k+\delta t} W(t)\mathrm{d}t. \end{aligned}$$

If, as in our case, the spectral density matrix is constant in time, we have $W(t) = W_{\mathrm{c}}$ and

$$W = W_k = \frac{W_{\mathrm{c}}}{\delta t}. \tag{1.44}$$

We consider the following discretisation scheme

$$A_{k|k-1} = I + A\delta t, \qquad B_k = B_{\mathrm{c}}, \qquad C_k = C_{\mathrm{c}}, \qquad Q = \delta t Q_{\mathrm{c}},$$

where $I$ is the identity matrix. Let us take into account the discrete-time Kalman filter and the following limit

$$\lim_{\delta t\to 0}\frac{K_k}{\delta t} = \lim_{\delta t\to 0}\frac{P_k^- C_{\mathrm{c}}^{\intercal}\left(C_{\mathrm{c}}P_k^- C_{\mathrm{c}}^{\intercal}\delta t + W_{\mathrm{c}}\right)^{-1}}{\delta t} = P_k^- C_{\mathrm{c}}^{\intercal} W_{\mathrm{c}}^{-1}.$$

We consider the equivalent expression of $P_k^+$ as $P_k^+ = (I - K_k C_k)P_k^-$. Given the discretisation scheme, the *a posteriori* covariance can be written as

$$\begin{aligned} P_{k+1}^- &= (I + A\delta t)P_k^+(I + A\delta t)^{\intercal} + \delta t B_{\mathrm{c}} Q_{\mathrm{c}} B_{\mathrm{c}} \\ &= P_k^+ + (AP_k^+ + P_k^+ A^{\intercal} + B_{\mathrm{c}}Q_{\mathrm{c}}B_{\mathrm{c}})\delta t + AP_k^+ A^{\intercal}\delta t^2 \\ &= P_k^- - K_k C_k P_k^- + (A(I - K_k C_k)P_k^- + (I - K_k C_k)P_k^- A^{\intercal} + B_{\mathrm{c}}Q_{\mathrm{c}}B_{\mathrm{c}})\delta t + O(\delta t^2). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\mathrm{d}P}{\mathrm{d}t} &= \lim_{\delta t\to 0}\frac{P_{k+1}^- - P_k^-}{\delta t} \\ &= \lim_{\delta t\to 0}\left(\frac{K_k C_k P_k^-}{\delta t} + A(I - K_k C_k)P_k^- + (I - K_k C_k)P_k^- A^{\intercal} + B_{\mathrm{c}}Q_{\mathrm{c}}B_{\mathrm{c}}\right) \\ &= -PC_{\mathrm{c}}^{\intercal}W_{\mathrm{c}}^{-1}C_{\mathrm{c}}P + AP + PA^{\intercal} + B_{\mathrm{c}}Q_{\mathrm{c}}B_{\mathrm{c}}. \end{aligned}$$

To derive the continuous version of the Kalman estimator we combine the definitions of $\hat{u}_k^-$ and $\hat{u}_k^+$, obtaining

$$\begin{aligned}\hat{u}_k^+ &= A_{k|k-1}\hat{u}_{k-1}^+ + K_k\big(z_k - C_k A_{k|k-1}\hat{u}_{k-1}^+\big)\\ &= (I + A\delta t)\hat{u}_{k-1}^+ + K_k\big(z_k - C_{\mathrm{c}}(I + A\delta t)\hat{u}_{k-1}^+\big).\end{aligned}$$

Subtracting $\hat{u}_{k-1}^+$ from both sides, dividing by $\delta t$ and taking the limit as $\delta t \to 0$ we have

$$\begin{aligned}\frac{\mathrm{d}\hat{u}}{\mathrm{d}t} &= \lim_{\delta t\to 0}\frac{\hat{u}_k^+ - \hat{u}_{k-1}^+}{\delta t}\\ &= A\hat{u} + PC_{\mathrm{c}}W_{\mathrm{c}}^{-1}\big(z - C_{\mathrm{c}}\hat{u}\big).\end{aligned}$$

In conclusion, taking out the subscripts, the Kalman estimator is defined by the following system

$$\left\{\begin{array}{rcl}\dfrac{\mathrm{d}\hat{u}(s)}{\mathrm{d}t} &=& A\hat{u}(s) + K\big(z(s) - C(s)\hat{u}(s)\big), \quad \forall s\in[0,\tau],\\ K(s) &=& P(s)C^{\intercal}(s)W^{-1}, \quad \forall s\in[0,\tau]\\ \hat{u}(0) &=& u_\diamond,\end{array}\right. \tag{1.45}$$

where $P$ is the unique solution of the Riccati differential equation

$$\left\{\begin{array}{rcl}\dfrac{\mathrm{d}P(s)}{\mathrm{d}t} &=& AP(s) + P(s)A^{\intercal} - P(s)C^{\intercal}(s)W^{-1}C(s)P(s) + B(s)QB^{\intercal}(s), \quad \forall s\in[0,\tau],\\ P(0) &=& P_0.\end{array}\right. \tag{1.46}$$

### 1.3.5 Extended Kalman Filter (EKF) theory

In most practical cases as well as in our application, the observation operator and/or the model operator are non-linear [44, 136]. As said before, the Kalman Filter cannot be applied in these cases. A widespread strategy is to consider the Extended Kalman Filter (EKF), proposed by Schmidt to solve nonlinear spacecraft navigation problems [21]. This strategy is based on the linearisation of the problem around the target trajectory using a first-order truncation of the Taylor series expansion. Strategies based on the second-order approximation of the operators exist, but the high computational costs often make them impossible to use in practice [17, 5, 109, 175].

The EKF is proved to converge only for small errors as it is based on a linearisation strategy [166, 73]. We describe how we deal with this difficulty in the application section 1.3.9.

As we did for the Kalman filter, we present the Extended Kalman filter in a discrete setting. We assume that we have a process whose states $\{u_k \in \mathbb{R}^n, k\in\mathbb{N}\}$ are governed by the non-linear model

$$\left\{\begin{array}{rcl}u_k &=& \mathbf{A_{k|k-1}}(u_{k-1}) + B_k\omega_k,\\ u_0 &=& u_\diamond + \xi\end{array}\right. \tag{1.47}$$

and can be observed through a non-linear protocol

$$z_k = \mathbf{C_k}(u_k) + \chi_k.$$

We assume that the random variables $\omega_k$ and $\chi_k$ are independent, white, with zero-mean and covariances $Q$ and $W$.

We linearise the state model around the point $u_{k-1} = \hat{u}_{k-1}^+$ obtaining

$$u_k = \mathbf{A_{k|k-1}}(\hat{u}_{k-1}^+) + B_k\omega_k + A_{k|k-1}(u_{k-1} - \hat{u}_{k-1}^+),$$

where $A_{k|k-1} = \dfrac{\mathrm{d}\mathbf{A_{k|k-1}}}{\mathrm{d}u}(\hat{u}_{k-1}^+)$ is the derivative of the state operator.

Analogously, we linearise the observation equation around the point $u_k = \hat{u}_k^-$ obtaining

$$z_k = \mathbf{C_k}(\hat{u}_k^-) + C_k(u_k - \hat{u}_k^-) + \chi_k,$$

where $C_k = \dfrac{\mathrm{d}\mathbf{C_k}}{\mathrm{d}u}(\hat{u}_k^-)$ is the derivative of the observation operator.

We have thus approximated our non-linear problem by a linear problem on which we can apply the Kalman Filter method. As for the Kalman method, the EKF can be presented in a prediction-correction form as follows

$$\begin{array}{rcll} \hat{u}_k^- &=& \mathbf{A_{k|k-1}}(\hat{u}_{k-1}^+), & \\ P_k^- &=& A_{k|k-1}P_{k-1}^+A_{k|k-1}^\intercal + B_kQB_k^\intercal & \text{(Prediction)} \end{array}$$

and

$$\begin{array}{rcll} \hat{u}_k^+ &=& \hat{u}_k^- + K_k(z_k - C_k\hat{u}_k^-), & \\ K_k &=& P_k^-C_k^\intercal\left(C_kP_k^-C_k^\intercal + W\right)^{-1}, & \text{(Correction)} \\ P_k^+ &=& (I - K_kC_k)P_k^-(I - K_kC_k) + K_kWK_k^\intercal. & \end{array}$$

We notice that the linearisation of the operators $\mathbf{A}$, $\mathbf{C}$ has been done around the best estimate of the state $u_k$ before the prediction step and before the correction step, respectively.

We should point out that the Extended Kalman Filter may be formulated to account for non-linear noise sources. We focus on the case of additive model noise and observation noise, that being the case used in the following.

For the sake of completeness, we report the continuous-time formulation of the EKF method. The EKF estimator is defined by the following system

$$\left\{\begin{array}{rcl} \dfrac{\mathrm{d}\hat{u}(s)}{\mathrm{d}t} &=& \mathbf{A}(\hat{u}(s), s) + K\big(z(s) - \mathbf{C}(\hat{u}(s))\big), \quad \forall s \in [0, \tau], \\ K(s) &=& P(s)C^\intercal W^{-1}, \quad \forall s \in [0, \tau] \\ \hat{u}(0) &=& u_\diamond, \end{array}\right. \tag{1.48}$$

where $P$ is the unique solution of the Riccati differential equation

$$\begin{cases} \dfrac{\mathrm{d}P(s)}{\mathrm{d}t} &= AP(s) + P(s)A^\intercal - P(s)C^\intercal(s)W^{-1}C(s)P(s) + B(s)QB^\intercal(s), \quad \forall s \in [0,\tau], \\ P(0) &= P_0. \end{cases} \tag{1.49}$$

Where this time $A = \dfrac{\mathrm{d}\mathbf{A}}{\mathrm{d}u}$ and $C = \dfrac{\mathrm{d}\mathbf{C}}{\mathrm{d}u}$.

### 1.3.6 Our inverse problem

Under the assumptions (1.39) and (1.33), our system is completely determined once we know the four parameters

$$k_{\mathrm{on}}, \quad k_{\mathrm{dep}}, \quad k_{\mathrm{dis}}, \quad \alpha.$$

Therefore, the solution of our inverse problem is an estimation of these parameters. In the following we define our inverse problem and more specifically

— the model operator

— the observation operator.

**1.3.6.a Model operator** The oligomer model (1.34) can be formulated in several equivalent ways. In the following we show that some formulations can meet the needs of our estimation strategy better than others.

First of all we reformulate System (1.34) including null dynamics for the kinetic parameters. In this formulation, all the unknowns of the problem appear in the initial condition and we can apply the filtering theory as presented in the previous sections. We obtain the following system

$$\begin{cases} \dfrac{\mathrm{d}y_i}{\mathrm{d}t} &= -k_{\mathrm{on}} v y_i + k_{\mathrm{on}} v y_{i-1} - k_{\mathrm{dep}} y_i + k_{\mathrm{dep}} y_{i+1}, & i_0 \le i \le i_1, \\ \dfrac{\mathrm{d}w_i}{\mathrm{d}t} &= -k_{\mathrm{dis}} w_i, & i_0 \le i \le i_1, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= (i_0 - 1) k_{\mathrm{dep}} y_{i_0} + (-k_{\mathrm{on}} v + k_{\mathrm{dep}}) \sum y_i + k_{\mathrm{dis}} \sum i w_i, & \\ \dfrac{\mathrm{d}k_{\mathrm{on}}}{\mathrm{d}t} &= 0, & \\ \dfrac{\mathrm{d}k_{\mathrm{dep}}}{\mathrm{d}t} &= 0, & \\ \dfrac{\mathrm{d}k_{\mathrm{dis}}}{\mathrm{d}t} &= 0. & \end{cases} \tag{1.50}$$

We recall that the term $(i_0 - 1)k_{\text{dep}}y_{i_0}$ comes from the boundary conditions (1.35). For the sake of simplicity, we omit the index in the sum operation, we thus consider $\sum$ instead of $\sum_{i=i_0}^{i_1}$.

Let us consider the augmented state $(y, w, v, k_{\text{on}}, k_{\text{dep}}, k_{\text{dis}})^\intercal$, where $y = (y_{i_0}, \dots, y_{i_1})^\intercal$ and $w = (w_{i_0}, \dots, w_{i_1})^\intercal$. The model operator describing the dynamics of the augmented state is nonlinear. We apply the EKF method. Once all data points have been processed, we have the estimations

$$\begin{pmatrix} \hat{y}(\tau) \\ \hat{w}(\tau) \\ \hat{v}(\tau) \\ \hat{k}_{\text{on}}(\tau) \\ \hat{k}_{\text{dep}}(\tau) \\ \hat{k}_{\text{dis}}(\tau) \end{pmatrix} \approx \begin{pmatrix} y(\tau) \\ w(\tau) \\ v(\tau) \\ k_{\text{on}}(\tau) \\ k_{\text{dep}}(\tau) \\ k_{\text{dis}}(\tau) \end{pmatrix} = \begin{pmatrix} y(\tau) \\ w(\tau) \\ v(\tau) \\ k_{\text{on}} \\ k_{\text{dep}} \\ k_{\text{dis}} \end{pmatrix}.$$

With this formulation, we would be able to estimate the kinetic rates but we would not have any information on the initial distribution of the oligomers. It would thus be necessary to estimate the initial condition with a *smoothing method*, see Section 5.3 and [157, 30, 148] for an introduction to smoothing methods and their applications.

However, by modifying the formulation of the state model we obtain a more direct solution with less computational costs and higher accuracy. We remark that the unstable oligomer concentrations can be easily expressed analytically by $w_i(t) = w_{0i}e^{-k_{\text{dis}}t}$. We decide to introduce this formula directly into the model. We consider the state $(y, w_0, v, k_{\text{on}}, k_{\text{dep}}, k_{\text{dis}})^\intercal$. The associated dynamics is

$$\begin{cases} \dfrac{\mathrm{d}y_i}{\mathrm{d}t} &= -k_{\text{on}}vy_i + k_{\text{on}}vy_{i-1} - k_{\text{dep}}y_i + k_{\text{dep}}y_{i+1}, & i_0 \le i \le i_1, \\ \dfrac{\mathrm{d}w_{0i}}{\mathrm{d}t} &= 0, & i_0 \le i \le i_1, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= (i_0 - 1)k_{\text{dep}}y_{i_0} + (-k_{\text{on}}v + k_{\text{dep}})\sum y_i + k_{\text{dis}}\sum ie^{-k_{\text{dis}}t}w_{0i}, & \\ \dfrac{\mathrm{d}k_{\text{on}}}{\mathrm{d}t} &= 0, & \\ \dfrac{\mathrm{d}k_{\text{dep}}}{\mathrm{d}t} &= 0, & \\ \dfrac{\mathrm{d}k_{\text{dis}}}{\mathrm{d}t} &= 0. & \end{cases} \tag{1.51}$$

In this case, at the end of the time window, we have

$$\begin{pmatrix} \hat{y}(\tau) \\ \hat{w}_0(\tau) \\ \hat{v}(\tau) \\ \hat{k}_{\text{on}}(\tau) \\ \hat{k}_{\text{dep}}(\tau) \\ \hat{k}_{\text{dis}}(\tau) \end{pmatrix} \approx \begin{pmatrix} y(\tau) \\ w_0(\tau) \\ v(\tau) \\ k_{\text{on}}(\tau) \\ k_{\text{dep}}(\tau) \\ k_{\text{dis}}(\tau) \end{pmatrix} = \begin{pmatrix} y(\tau) \\ w_0 \\ v(\tau) \\ k_{\text{on}} \\ k_{\text{dep}} \\ k_{\text{dis}} \end{pmatrix}.$$

As we can see, the extended Kalman estimator is able to estimate both the kinetic rates and the initial distribution of the unstable oligomers. Moreover, as the total oligomer distribution $\{u_i\}_i$ is known, we could also derive the initial distribution of stable oligomers.

Further formulations may be taken into account. For instance, by the law of mass conservation (1.2), we could substitute $v$ by $\rho - \sum i(w_i + y_i)$ and omit the differential equation for the monomer concentration $v$. However, in doing so we would introduce a higher nonlinearity into the model.

We could also treat the parameter $\alpha$ as we have done with the kinetic parameters and add the dynamics $\dfrac{d\alpha}{dt} = 0$ to the model. This choice would introduce some technical difficulty as, for example, defining the initial state covariance operator. In fact, we have seen in the previous sections that we cannot assume $\alpha$ and $k_{\text{dis}}$ to be independent but, at the same time, we do not have a complete description of their inter-relation. Furthermore, the hypothesis of size-independence of the coefficient $\alpha$ is reasonable but not as well supported as in the case of the kinetic rates. Introducing the null dynamics for $\alpha$ in the model could be too strong a condition.

**1.3.6.b Observation operator** The observation operator describes the measurement process applied to the system. We can consider two observation operators associated to the two types of measurements available : the SEC and the SLS data. The definition of these operators is given in Equations (1.3) and (1.14), respectively.

We can define several inverse problems depending on the measurements used to estimate the unknowns of the model. We have shown how the SEC data were very useful to design the model as they provide qualitative information on the system. However, the SEC data are not quantitatively reliable since the data error is difficult to model and, therefore, difficult to estimate.

The SLS data are more reliable because we can better analyse the noise and have a good estimation of its distribution. Consequently, we can derive more accurate quantitative information.

We decide to define our inverse problem taking into account only the SLS data and then validate the results with the SEC data.

A final decision to make is whether to concatenate the three sets of SLS data in a vector $z = (z_{sls,\rho=1}, z_{sls,\rho=3}, z_{sls,\rho=7})$ and thus use the data simultaneously to compute the initial condition estimation or define three independent inverse problems, one for each concentration. We prefer the second option. In fact, even if our theoretical model takes the kinetic rates to be the same in all the experiments, in practice, we more likely expect the parameters to vary within a confined range of values.

We conclude this section with the definition of the observation operator $C$ and some last remarks. We recall that

$$z_{sls}(t) = \Big(v(t) + \sum_{i=i_0}^{i_1} i^2(w_i + y_i)\Big) + \chi(t), \quad \forall t \in [0, \tau].$$

The definition of the observation operator depends on the choice of the state variable. For instance, if we consider the model formulation presented in System (1.50) and we call the state $x = (y, w, v, k_{\text{on}}, k_{\text{dep}}, k_{\text{dis}})^\intercal$, then we would have the following linear operator

$$Cx = \left(\begin{array}{ccc|ccc|c|ccc} i_0^2 & \cdots & i_1^2 & i_0^2 & \cdots & i_1^2 & 1 & 0 & 0 & 0 \end{array}\right) x(t).$$

However, we have seen that the second formulation, corresponding to System (1.51) is preferable. In this case the state is $x = (y, w_0, v, k_{\text{on}}, k_{\text{dep}}, k_{\text{dis}})^\intercal$ and observation operator reads as follows

$$C(x(t), t) = \left(\begin{array}{ccc|ccc|c|ccc} i_0^2 & \cdots & i_1^2 & i_0^2 e^{-k_{\text{dis}} t} & \cdots & i_1^2 e^{-k_{\text{dis}} t} & 1 & 0 & 0 & 0 \end{array}\right) x(t). \tag{1.52}$$

We point out that, in this last setting, the operator is dependent on the time and, more importantly, on the rate $k_{\text{dis}}$, making it non-linear.

### 1.3.7 Model operator discretisation

To discretise our model, we consider an upwind scheme. Let us define the time domain $[0, \tau] = [0, \tau]$, where $\tau$ is the experimental observation time. We consider a uniform time grid $0 = t_0 < \ldots < t_N = \tau$ with a constant time step $\delta t$. Given $f$ a function, we use the superscript $f^j$ to represent the approximation of the value $f(t_j)$. The discrete-time formulation of model (1.51) thus reads

$$\begin{cases} y_i^{n+1} &= y_i^n + \delta t\, {k_{\text{on}}}^n v^n(-y_i^n + y_{i-1}^n) + \delta t\, {k_{\text{dep}}}^n(-y_i^n + y_{i+1}^n), \\ {w_0}_i^{n+1} &= {w_0}_i^n, \\ v^{n+1} &= v^n + \delta t\,(i_0 - 1){k_{\text{dep}}}^n y_{i_0}^n + \delta t\,(-{k_{\text{on}}}^n v^n + {k_{\text{dep}}}^n)\sum y_i^n \\ &\quad + \delta t\, {k_{\text{dis}}}^n \sum i e^{-{k_{\text{dis}}}^n t_n} {w_0}_i^n, \\ {k_{\text{on}}}^{n+1} &= {k_{\text{on}}}^n, \\ {k_{\text{dep}}}^{n+1} &= {k_{\text{dep}}}^n, \\ {k_{\text{dis}}}^{n+1} &= {k_{\text{dis}}}^n. \end{cases} \tag{1.53}$$

### 1.3.8 A priori parameter estimation

So far we have built up a general understanding of the oligomer system though considerations deriving both from a qualitative description of the oligomer system and from the

information contained in the empirical data. Now, we aim at integrating this knowledge with the quantitative information provided by the model and then set the initial condition of our estimation strategy, namely $x_\diamond$.

We use the discrete model to compute the sequence $(u^n)_{0\leq n\leq N}$, approximating the state variables on the time grid. We can thus apply the observation operator on these estimations and compute the synthetic observations.

In the following, we present the steps leading to the definition of an initial estimation and its successive corrections.
We start from the following assumptions.

— Observing the experimental size distribution data at $\rho = 3\mu M$, we have designed a formula to define a first guess for the disintegrating rate $k_{\text{dis}}$.

In Figure 1.19a, we notice a pseudo steady state. In fact, the distribution at time $t = 270$min does not differ significantly from the distribution at $t = 125$min. From Equations (1.30) and (1.31), we deduce that it is possible to have a steady state only when unstable oligomers have all disintegrated. We can thus assume that, from time $t = 125$min, we do not have any unstable oligomers. Omitting the noise contribution, we have that

$$z_{sec,o}(t = 125) = u(125) = y(125), \qquad \text{for } \rho = 3\mu M.$$

Moreover, we make the hypothesis that the initial stable oligomer peak shifts over time without changing its shape. Therefore, we assume that

$$\max(y(t)) = \max(y(0)), \quad \forall t.$$

In particular, the relation holds true at $t = 125$min. We are aware that this hypothesis is not correct – since there are broadening effects and mass loss of stable oligomers– nevertheless, it is efficient enough to define an initial approximation.

Assuming that the ratio of stable oligomers at time $t = 0$ is the same at all the sizes, we have that the peak position of the initial oligomer distribution is the same as for the initial stable oligomer distribution

$$i^* = \arg\max_i u_i(0) = \arg\max_i y_i(0).$$

From the model we know that $w_i(t) = w_i(0)e^{-k_{\text{dis}}t} = (u_i(0) - y_i(0))e^{-k_{\text{dis}}t}$. Writing this relation for the size $i^*$, we obtain

$$w_{i^*}(t) = \big(\max(z_{sec,o}(0)) - \max(z_{sec,o}(125))\big)e^{-k_{\text{dis}}t}.$$

It is easy to see that we can derive the value of $k_{\text{dis}}$, if we know the value of the trajectory $w_{i^*}$ at, at least, one instant. Recalling that the kinetic coefficients are assumed size-independent, we have

$$w_{i^*}(t) = \max_i w_i(t) = \max_i(u_i(t) - y_i(t)).$$

Finally, let us call $t_{\text{stable}}$ a moment in which we have only stable oligomers and $\eta \neq 0$ a generic instant at which we have SEC measurements. We conclude the following formula

$$k_{\text{dis}} = \frac{1}{\eta}\log\left(\frac{\max(z_{sec,o}(0)) - \max(z_{sec,o}(t_{\text{stable}}))}{\max(z_{sec,o}(\eta)) - \max(z_{sec,o}(t_{\text{stable}}))}\right).$$

For the experiment at $\rho = 3\mu M$, we can take $t_{\text{stable}} = 125\text{min}$ and $\eta = 15\text{min}$.

We want to adapt the same argument to the cases $\rho = 1\mu M$, $\rho = 7\mu M$.
For the experiment at $\rho = 1\mu M$, the SEC data have been recorded three times. The natural choice is to set $t_{\text{stable}} = 140\text{min}$ and $\eta = 25\text{min}$. In fact, at the end of the experiment, it is reasonable to assume that there are only stable oligomers left.
For $\rho = 7\mu M$, we consider the experimental data in Figure 1.20. The third SEC curve, recorded at time $t = 15\text{min}$, corresponds to an increasing phase in SLS data and, thus, to a phase in which the polymerisation is the dominant process occurring. Knowing that only the stable oligomers can polymerise, we assume that at that moment the unstable oligomer concentration is negligible. We set $t_{\text{stable}} = 15\text{min}$ and $\eta = 5\text{min}$. In Table T 1.3.8.1, we report the resulting estimations of $k_{\text{dis}}$.

— Having an estimation for $k_{\text{dis}}$, we use Equation (1.40) to compute the ratio of stable oligomers. In particular, choosing a time $t_\varepsilon$ that is small enough to approximate $\alpha(t) = \alpha$ for all $t \in [0, t_\varepsilon]$, we have $k_{\text{dis}}(t - \int_0^t \alpha(s)ds) = k_{dis}(1-\alpha)t$. Consequently, we take

$$\alpha = 1 - \frac{0.05}{k_{\text{dis}}}.$$

— Equation (1.31) is useful to derive a relation between $k_{\text{on}}$ and $k_{\text{dep}}$. If the rate $k_{\text{on}}v(t) - k_{\text{dep}}$ is negative, then the stable oligomers are mainly depolymerising. On the other hand, if the rate is positive, the system is polymerising. At the equilibrium we have $k_{\text{on}}v(t) = k_{\text{dep}}$. We call $v_{\text{eq}}$ the monomer concentration that satisfies the equality.

  — We start by the case where $\rho = 3\mu M$ in which we have a pseudo steady state. In Table 1.1 we have the value of monomer concentrations at times 125 and 270 minutes, corresponding to the equilibrium phase. These values are 3.045 and 2.819 respectively. We set $v_{\text{eq},\rho=3}$ to the average value of 2.9.

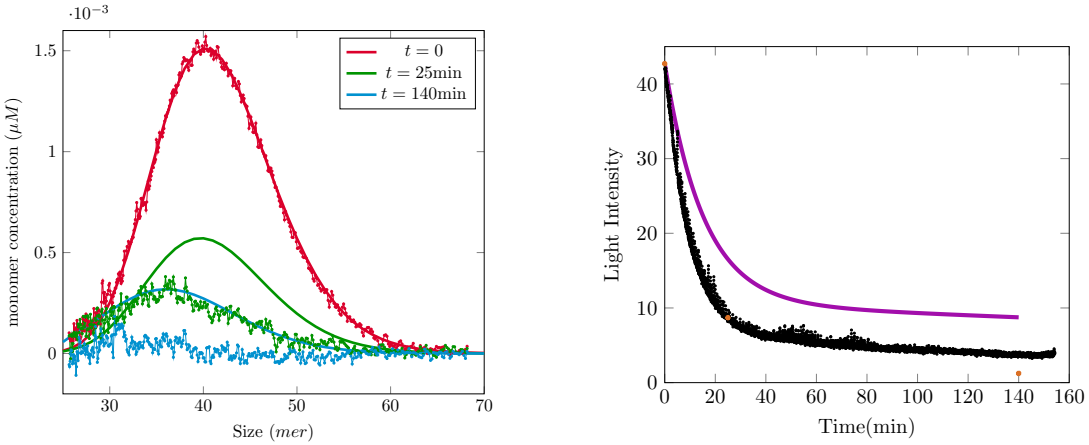
  — In the case where $\rho = 1\mu M$, the system is constantly depolymerising. Given that the final monomer concentration is of $1\mu M$, we obtain the inequality $k_{\text{on}} \cdot 1 < k_{\text{on}} v_{\text{eq}}$. We arbitrarily define $v_{\text{eq},\rho=1} = 2$.

  — At concentration $\rho = 7\mu M$, we observe that the system is depolymerising at time $t = 15\text{min}$. Given the corresponding value of monomer concentration at this time, we derive the inequality $k_{\text{on}} 3.5 < k_{\text{dep}}$. The system is, in contrast, polymerising at time $t = 1150\text{min}$ and we derive the condition $k_{\text{on}} 4.6 > k_{\text{dep}}$. We deduce that $3.5 < v_{\text{eq},\rho=7} < 4.6$. We then set $v_{\text{eq},\rho=7} = 4$.

  To conclude, we obtain the value of $k_{\text{on}}$ as follows

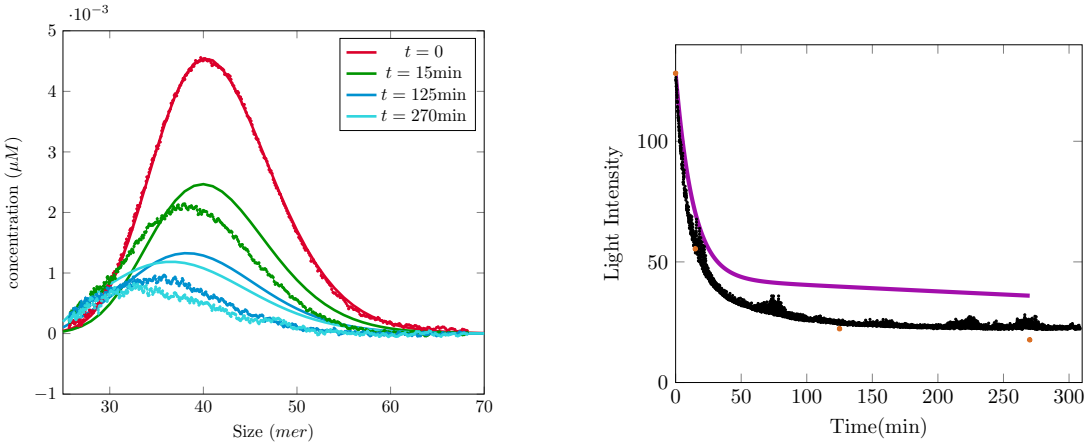
$$k_{\text{on}} = \frac{k_{\text{dep}}}{v_{\text{eq},\rho}}.$$

— The depolymerisation coefficient is arbitrarily set to $k_{\text{dep}} = 0.05$, by taking the average value of the values in Table 1.2.



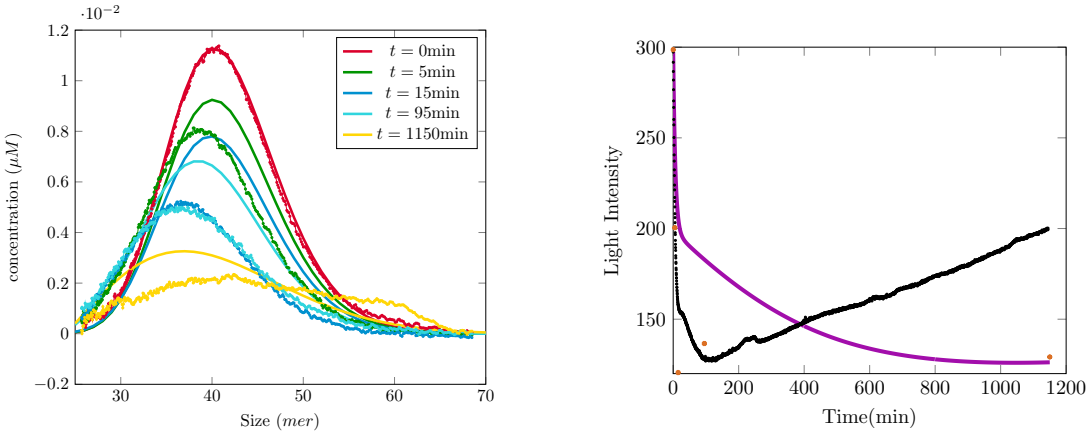

(a) SEC data (b) SLS data

FIGURE 1.31 – Fit with the coefficients of Table T 1.3.8.1 relative to $\rho = 1\mu M$.




(a) SEC data (b) SLS data

FIGURE 1.32 – Fit with the coefficients of Table T 1.3.8.1 relative to $\rho = 3\mu M$.




(a) SEC data

(b) SLS data

FIGURE 1.33 – Fit with the coefficients of Table T 1.3.8.1 relative to $\rho = 7\mu M$.

We summarise in this table our choices

| $\rho$ | $k_{\text{on}}$ | $k_{\text{dep}}$ | $k_{\text{dis}}$ | $\alpha$ | $v_{\text{eq}}$ |
|---|---|---|---|---|---|
| $1\mu M$ | 0.025 | 0.05 | 0.0653 | 0.2341 | 2 |
| $3\mu M$ | 0.0172 | 0.05 | 0.0744 | 0.3279 | 2.9 |
| $7\mu M$ | 0.0125 | 0.05 | 0.1503 | 0.6674 | 4 |

. (T 1.3.8.1)

In Figures 1.31, 1.32, 1.33, we present the associated numerical observations.
As a natural consequence of the strategy described, we notice reasonable agreement on the size distribution. Unfortunately, the SLS data – which are more trustworthy – are not well fitted. This result constitutes a good starting point for further analyses.

We recall that the experiment at $\rho = 3\mu M$ is the one in which the set of assumptions presented above are more natural. Furthermore, in this case, the SEC and SLS data are in best agreement. We have introduced in Equation (1.12) a formula to measure the distance between the SEC and SLS data : the smaller the distance, the greater the coherence between the two observations. Roughly speaking, we can say that the SEC and SLS data are in agreement if the SLS data fit the orange points well in Figures 1.31b, 1.32b, 1.33b, computed by applying the observation operator on the experimental size distribution.

Focusing on the $3\mu M$ experiment and following the guidelines presented above, we run the model several times with the set of parameters defined by slight modifications of the values in Table T 1.3.8.1. Increasing the value of $k_{\text{dep}}$ to 0.09, we find that the set of parameters

| $k_{\text{on}}$ | $k_{\text{dep}}$ | $k_{\text{dis}}$ | $\alpha$ | $v_{\text{eq}}$ |
|---|---|---|---|---|
| 0.031 | 0.09 | 0.07 | 0.2 | 2.9 |

. (T 1.3.8.2)

In Figure 1.34, we present the comparison between experimental and synthetic observations both for the experiment at $3\mu M$ and for the other concentrations. We observe a good fit of the SLS data. Moreover, as this set of parameters is in agreement with the assumptions detailed before, we still have a qualitative agreement with the SEC data. It is interesting to notice that with this choice we obtain a better fit than before for the case $1\mu M$.

It should be pointed out that it is possible to find other sets of parameters that provide a good fit of the SLS data at $\rho = 3\mu M$. For instance, we present the following parameters

| $k_{\text{on}}$ | $k_{\text{dep}}$ | $k_{\text{dis}}$ | $\alpha$ | $v_{\text{eq}}$ |
|---|---|---|---|---|
| 0.14 | 0.38 | 0.077 | 0.35 | 2.7 |

(T 1.3.8.3)

and the associated synthetic observations for the three concentration regimes in Figure 1.35.

In the following, we set the initial condition of the EKF estimator to the values given in Table T 1.3.8.2.

### 1.3.9 Extended Kalman Filter application

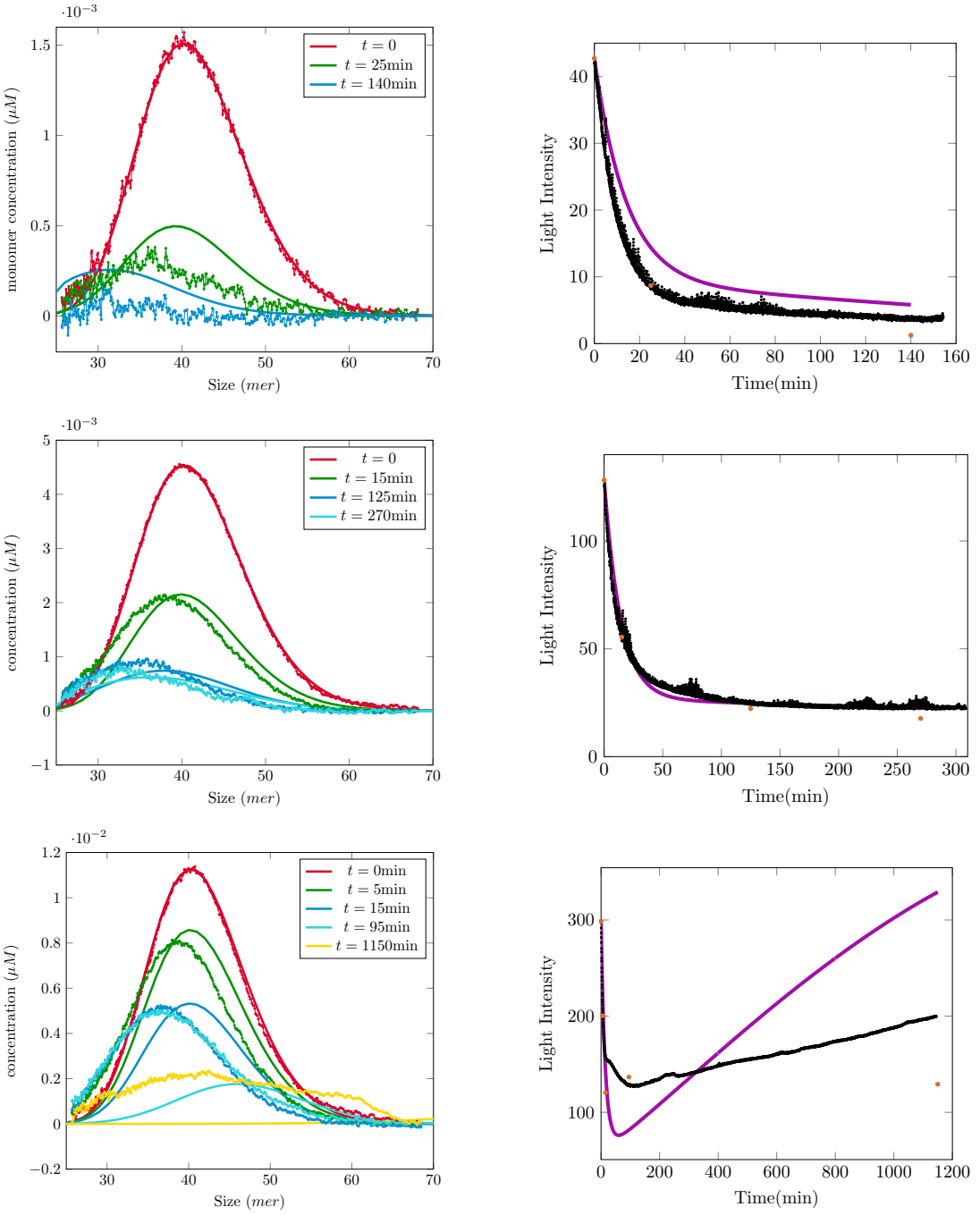
In this section we apply the EKF theory to solve our inverse problem. The method has been implemented in Matlab using the functions of the data assimilation library VerdandInMatlab, which is a module of the C++ library Verdandi [41]. The values of $k_{\text{on}}$, $k_{\text{dep}}$, $k_{\text{dis}}$, and $\alpha$ in Table T 1.3.8.2 characterise the initial condition of the Kalman estimator. We consider the model operator (1.51) and the SLS data modelled by the observation operator (1.52). In our simulations we set the time step to $\delta t = 0.1$ and the simulation time to the experimental final time.

We recall that the operators $P_0$ and $W$ represent an estimation of the amount of error in the initial state and in the observations, respectively. Consequently, they play a crucial role in the result of the estimation. Following a classical approach [21], we define these operators as the covariance of the uncertainty in the initial condition, $\xi$, and the covariance of the measurement noise $\chi$. Since both the state space and the observation space have a finite dimension, the operators can be written in a matrix form.

We recall that the discrete operator $W$ may be related to the spectral density matrix $W_{\text{c}}$ of the continuous white noise $W_{\text{c}}$, as follows [74, 157]

$$W = \delta t^{-1} W_{\text{c}}.$$

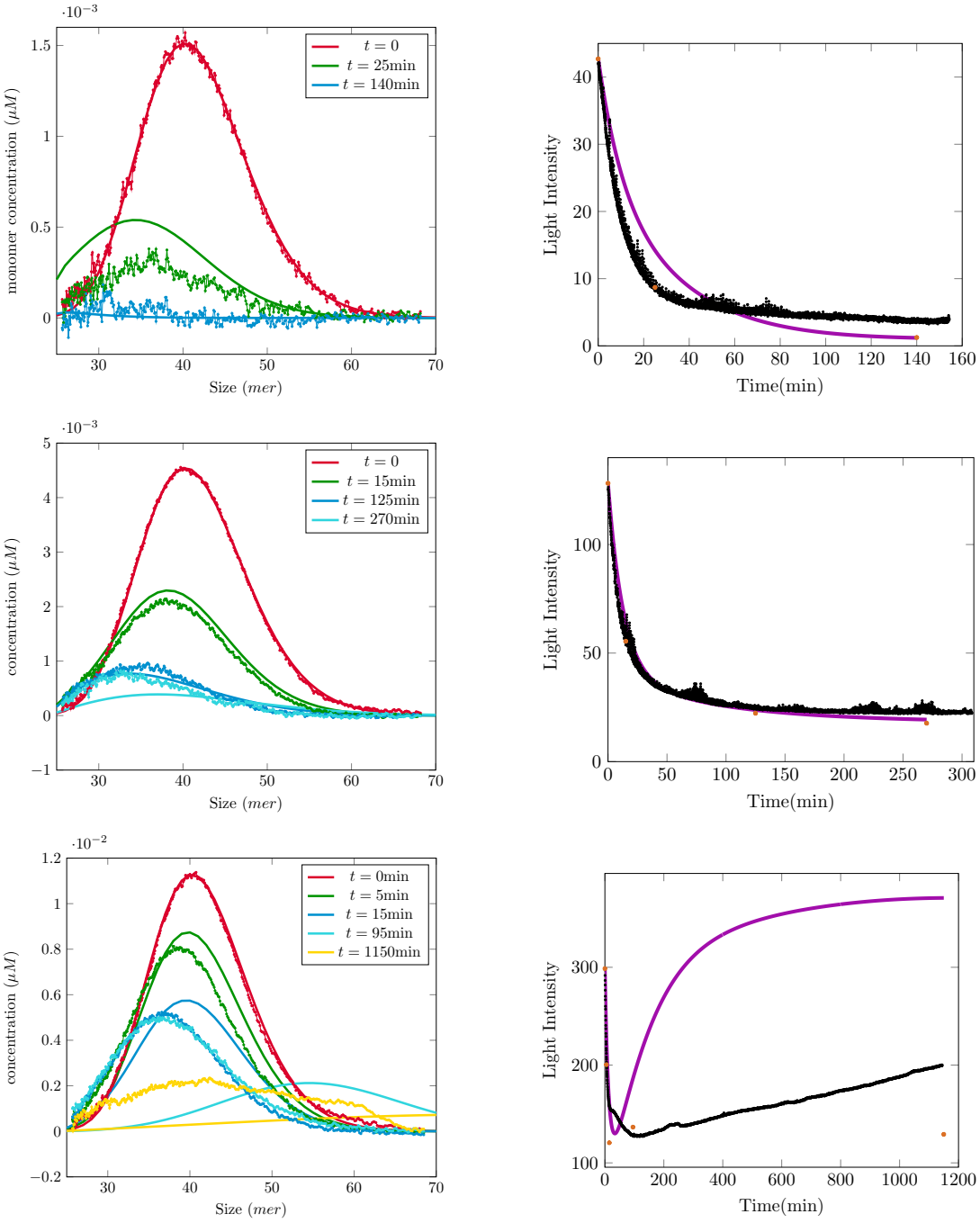
In our work, due to the small time lapse between two observations, we take $W_{\text{c}}$ as the variance of the error in our experimental data. Specifically, we compute a polynomial fit of the SLS




(a) SEC data

(b) SLS data

FIGURE 1.34 – Fit with the coefficients of Table T 1.3.8.2. From top to bottom $\rho = 1\mu M$, $\rho = 3\mu M$, $\rho = 7\mu M$.




(a) SEC data (b) SLS data

FIGURE 1.35 – Fit with the coefficients of Table T 1.3.8.3. From top to bottom $\rho = 1\mu M$, $\rho = 3\mu M$, $\rho = 7\mu M$.

data and we analyse the distribution of the residual. In practice, we set $W$ to the following values

| | $\rho = 1\mu M$ | $\rho = 3\mu M$ | $\rho = 7\mu M$ |
|---|---|---|---|
| $W$ | 1.6 | 8 | 1 |

(T 1.3.9.1)

Another possible choice would be to assume that the variance computed from the observation data $W_{\text{data}}$ satisfied $W_{\text{data}} = \delta t_{\text{data}}^{-1} W_{\text{c}}$. This would lead to the definition $W = \delta t^{-1} \delta t_{\text{data}} W_{\text{data}}$.

The operator $P_0$ is characterised by a block structure as follows

$$P_0 = \begin{pmatrix} \begin{matrix} P_y & P_{y,w} \\ P_{w,y} & P_w \end{matrix} & & & & \\ & P_v & & & \\ & & P_{k_{\text{on}}} & & \\ & & & P_{k_{\text{dep}}} & \\ & & & & P_{k_{\text{dep}}} \end{pmatrix}.$$

Entering the specifics of each block, the initial stable oligomer covariance is given by

$$P_y = \text{Cov}(y_0) = \text{Cov}(\alpha u_0) = \text{Cov}(\alpha) u_0 u_0^\intercal = P_\alpha u_0 u_0^\intercal,$$

where the definition of $P_\alpha$ is apparent from the above equation.
With analogous computations the covariance between $y_0$ and $w_0$ is

$$P_{y,w} = P_{w,y} = -P_\alpha u_0 u_0^\intercal$$

and the initial unstable oligomer covariance

$$P_w = P_\alpha u_0 u_0^\intercal.$$

The operators $P_\alpha, P_v, P_{k_{\text{on}}}, P_{k_{\text{dep}}}, P_{k_{\text{dis}}}$ are scalar values. Each of them is defined as the square of the estimated distance between the initial estimation and the target value.
We report the operator values in the following table

| $P_v$ | $P_{k_{\text{on}}}$ | $P_{k_{\text{dep}}}$ | $P_{k_{\text{dis}}}$ | $P_\alpha$ |
|---|---|---|---|---|
| $10^{-20}$ | $10^{-4}$ | $10^{-4}$ | $10^{-4}$ | $10^{-2}$ |

(T 1.3.9.2)

If we do not choose the covariance operators carefully, the Kalman estimator may exit the region of validity of the model. This would happen, for instance, when the trajectory of some parameter estimator became negative. For well-chosen covariance operator, the EKF algorithm returns the estimator $\hat{x}$. The Kalman estimator may be interpreted as the minimiser, at each time $t$, of a least square criterion given by a weighted sum of the error in the initial condition and the discrepancies $z(s) - C\hat{x}(s)$, for all times $s \leq t$. As we present more in detail in Chapter 5, the weights in the criterion are the inverse of the operators $P_0$ and $W$, respectively. In a linear setting we can prove that there exists a unique minimiser of the

least square criterion. For non-linear problems, like our application, the criterion may have several minimisers. At the end of the EKF algorithm, the Kalman estimator coincides with one of these minimisers. To avoid local minima, we adopt a multiple-run strategy, following the example of the *Iterated Extended Kalman Filter* [96, 117, 20]. More precisely, we run the algorithm again using the returned estimation as the initial condition, without changing the values of the covariance operators. We compare the output of this last run to the first estimation. If the distance between the two estimations is less than a chosen threshold, we stop. Otherwise, we iterate the EKF algorithm again and we compare the estimations. We repeat this process until the distance between two successive estimations is less than a threshold or we have reached a maximal number of EKF iterations.

We recall that the operator $P$ measures the estimation error in the state. It defines a confidence region centred on the target trajectory, in which the estimator trajectory lives. The KF, and thus the EKF, is designed to minimise the estimation error. At the end of one run of the EKF algorithm, the estimation error is less than the initial estimation error. Re-running the algorithm with the initial covariance operator instead of the final covariance allows the estimator to evolve in a "wider" region and avoid eventual local minima of the error function.

The multiple-run strategy may also be seen as a way to put a high confidence in measurements in a gradual way. In fact, using the observations at each EKF run to correct the estimation, we are implicitly placing more and more trust in these data.

Furthermore, with the multiple-run strategy we can reduce the EKF linearisation errors and thus improve the accuracy of the estimation. In fact, at each run, we linearise the model and the observation operator around a better estimation than at the iteration before.

We run the EKF algorithm taking the values in Table T 1.3.8.2 as initial conditions for the estimator and choosing the operators $P_0$ and $W$ according to Tables T 1.3.9.2 and T 1.3.9.1. We preselect the threshold on the distance between two successive estimations to $10^{-5}$ and use the distance associated to the norm $\|\cdot\|_\infty$. The maximal number of EKF iterations is set to 20. The final estimations are presented in the table below

| $\rho$ | $\hat{k}_{\text{on}}$ | $\hat{k}_{\text{dep}}$ | $\hat{k}_{\text{dis}}$ | $\hat{\alpha}$ |
|---|---|---|---|---|
| $1\mu M$ | 0.02579 | 0.10288 | 0.10151 | 0.13234 |
| $3\mu M$ | 0.24335 | 0.61940 | 0.12948 | 0.48961 |
| $7\mu M$ | 0.06447 | 0.27284 | 0.19152 | 0.50458 |

. (T 1.3.9.2)

In Figures 1.36, 1.37, 1.38 we present the results of the EKF algorithm for the concentrations $\rho = 1$, 3, $7\mu M$, respectively.

In each of these figures, we have six subfigures : two showing the comparison between the synthetic and the empirical observations and four showing the trajectories of the four parameter estimators $\hat{k}_{\text{on}}, \hat{k}_{\text{dep}}, \hat{k}_{\text{dis}}, \hat{\alpha}$. In the cases $\rho = 1$, and $7\mu M$ we have 20 parameter estimator trajectories associated to the 20 EKF iterations. In the case $\rho = 3\mu M$ we have convergence in 18 iterations. We have plotted the trajectories in a colormap varying from blue for the first iteration to red for the last iteration. Moreover, to visualise the data more clearly, the thickness of the plot line is gradually reduced from the first to the last iteration.

We can infer the reliability of the final estimation by analysing these four subfigures with parameter estimators. For instance, in Figure 1.37 – relative to $\rho = 3\mu M$ – the distance between the estimations gradually decreases until reaching a final value that triggers the stopping condition on the estimators. We can thus conclude that we have an optimal estimation.
When observing Figure 1.38 for $\rho = 7\mu M$, we notice a different behaviour. We can identify two sets of estimations that are alternatively returned by the EKF iterations. In other words if we use one of the two as the initial condition for the EKF algorithm we obtain the other one as the final condition. For example, the estimation at the end of the 19th iteration is $(\hat{k}_{\text{on}}, \hat{k}_{\text{dep}}, \hat{k}_{\text{dis}}, \hat{\alpha}) = (0.061479, 0.254831, 0.12000, 0.50316)$ while at the end of the 20th iteration we have the values shown in Table T 1.3.9.2. These two sets of parameters may be interpreted as local minima. Therefore, we cannot place much trust in these estimations.
To conclude, we comment on Figure 1.36 relative to low concentration experiment. We claim that the estimations $\hat{k}_{\text{dis}}$ and $\hat{\alpha}$ are optimal but the estimations $\hat{k}_{\text{on}}$ and $\hat{k}_{\text{dep}}$ tend to some value that has not yet been reached. We need more iterations of the EKF algorithm to reach the optimal value.

We can summarise our confidence in the estimations in Table T 1.3.9.2 as follows

| $\rho$ | $\hat{k}_{\text{on}}$ | $\hat{k}_{\text{dep}}$ | $\hat{k}_{\text{dis}}$ | $\hat{\alpha}$ |
|---|---|---|---|---|
| $1\mu M$ | ? | ? | OK | OK |
| $3\mu M$ | OK | OK | OK | OK |
| $7\mu M$ | No | No | No | No |

. (T 1.3.9.3)

Since the estimations in Table T 1.3.9.2 for $\rho = 1$, and $7\mu M$ are not satisfactory, we continue our investigation focusing on these two cases. We run the EKF algorithm with several choices of initial condition and covariance operators.

— We start from the low concentration experiment. We run the iterated EKF method starting from the initial condition given in Table T 1.3.8.2 and the covariance operators $P_v = 10^{-20}$, $P_{k_{\text{on}}} = P_{k_{\text{dep}}} = P_{k_{\text{dis}}} = P_\alpha = 10^{-2}$ and $W = 10$. This time we increase the maximal allowed number of EKF iterations to 500. We find that the convergence condition is satisfied after 222 iterations. The final estimation results in

| $\hat{k}_{\text{on}}$ | $\hat{k}_{\text{dep}}$ | $\hat{k}_{\text{dis}}$ | $\hat{\alpha}$ |
|---|---|---|---|
| 0.16960 | 0.22829 | 0.10195 | 0.14268 |

. (T 1.3.9.4)

In Figure 1.39 we present the comparison between synthetic and experimental observations as well as the parameter estimator trajectories. This time we have plotted only one trajectory every ten iterations. We can see that the estimations of $k_{\text{dis}}$ and $\alpha$ are similar to the values in Table T 1.3.9.2, reinforcing the idea that these estimations were already trustable.

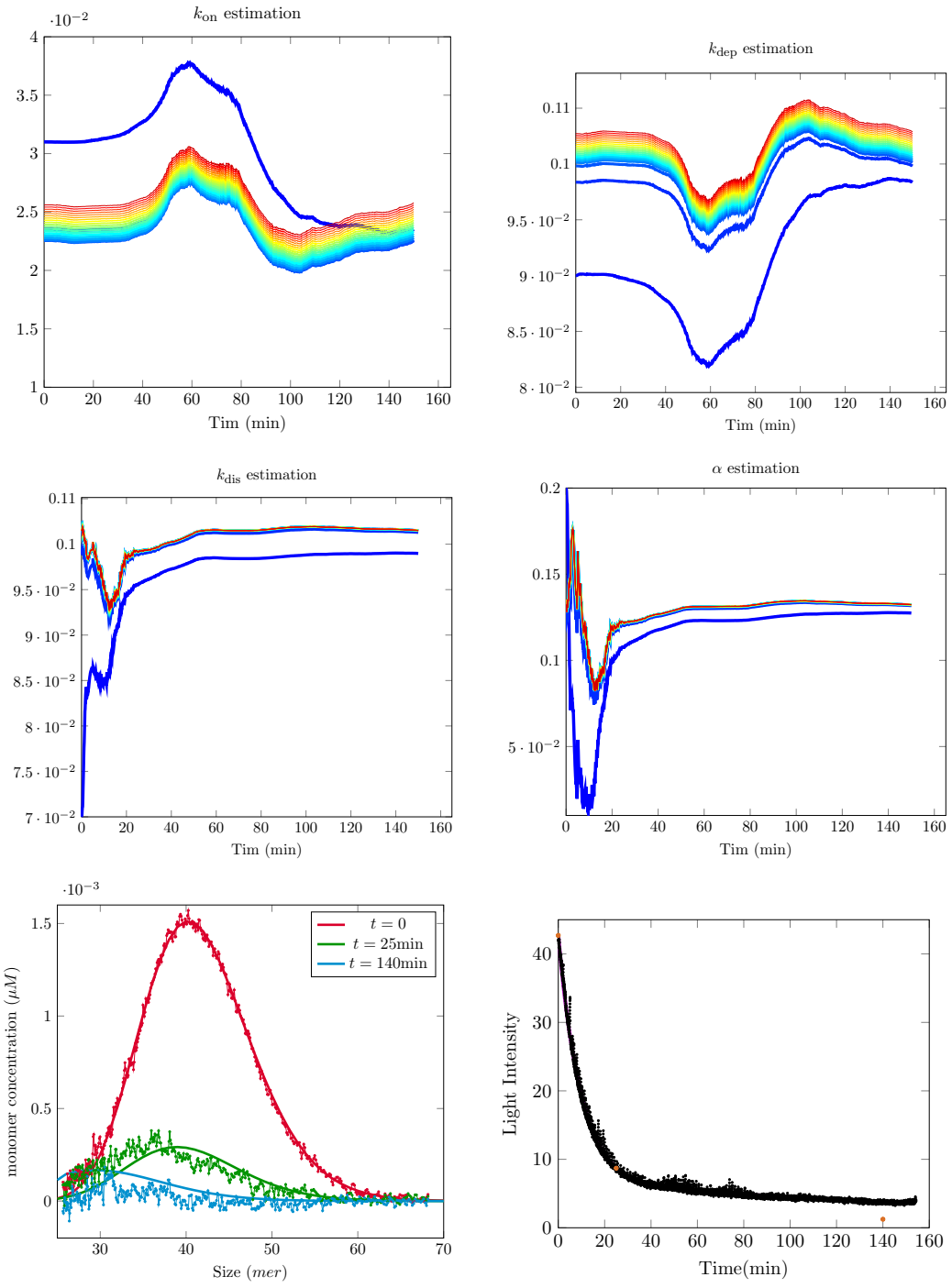



FIGURE 1.36 – Iterated EKF estimations (Table T 1.3.9.2) and observation comparison, $\rho = 1\mu M$.

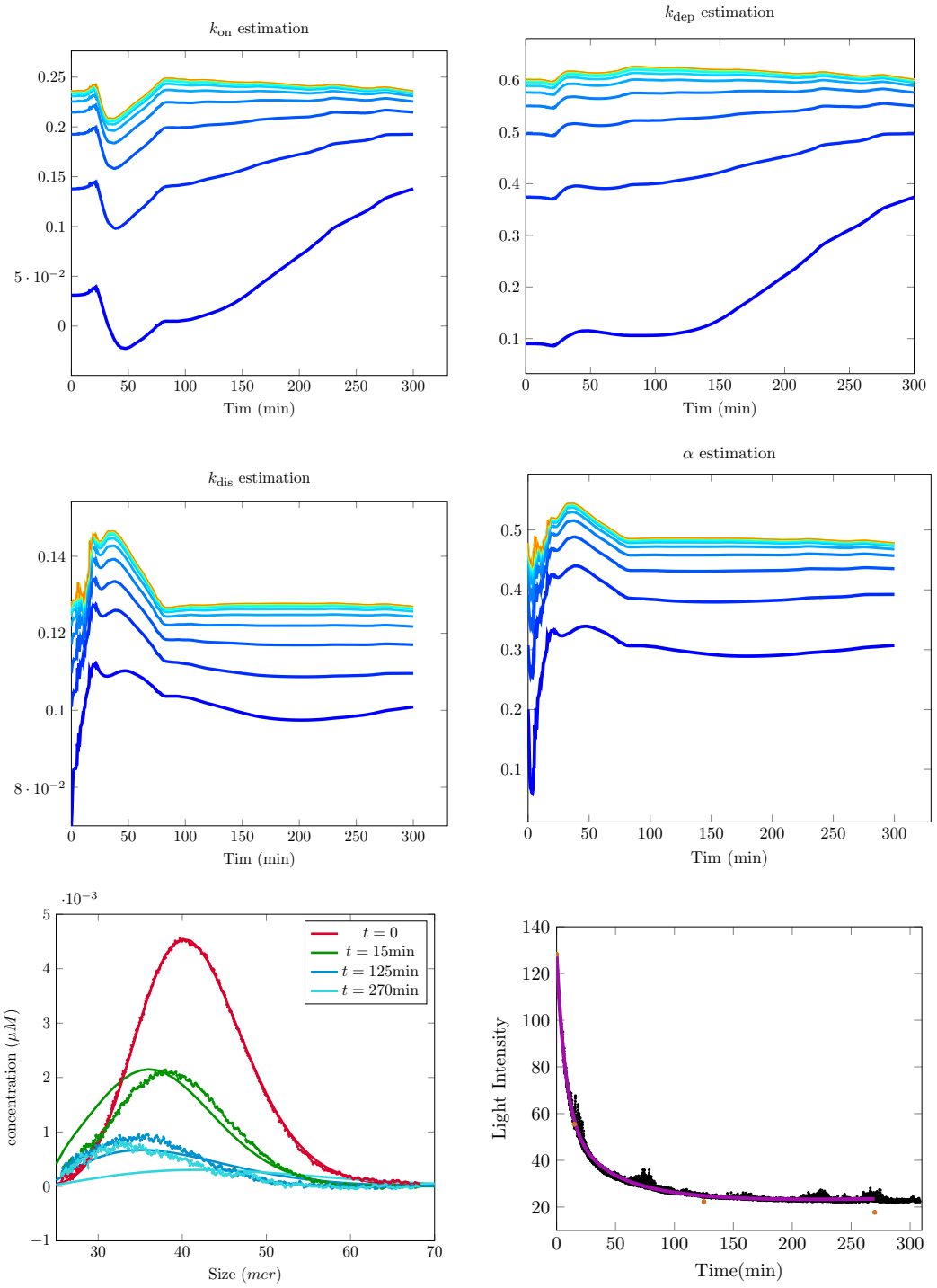



FIGURE 1.37 – Iterated EKF estimations (Table T 1.3.9.2) and observation comparison, $\rho = 3\mu M$.

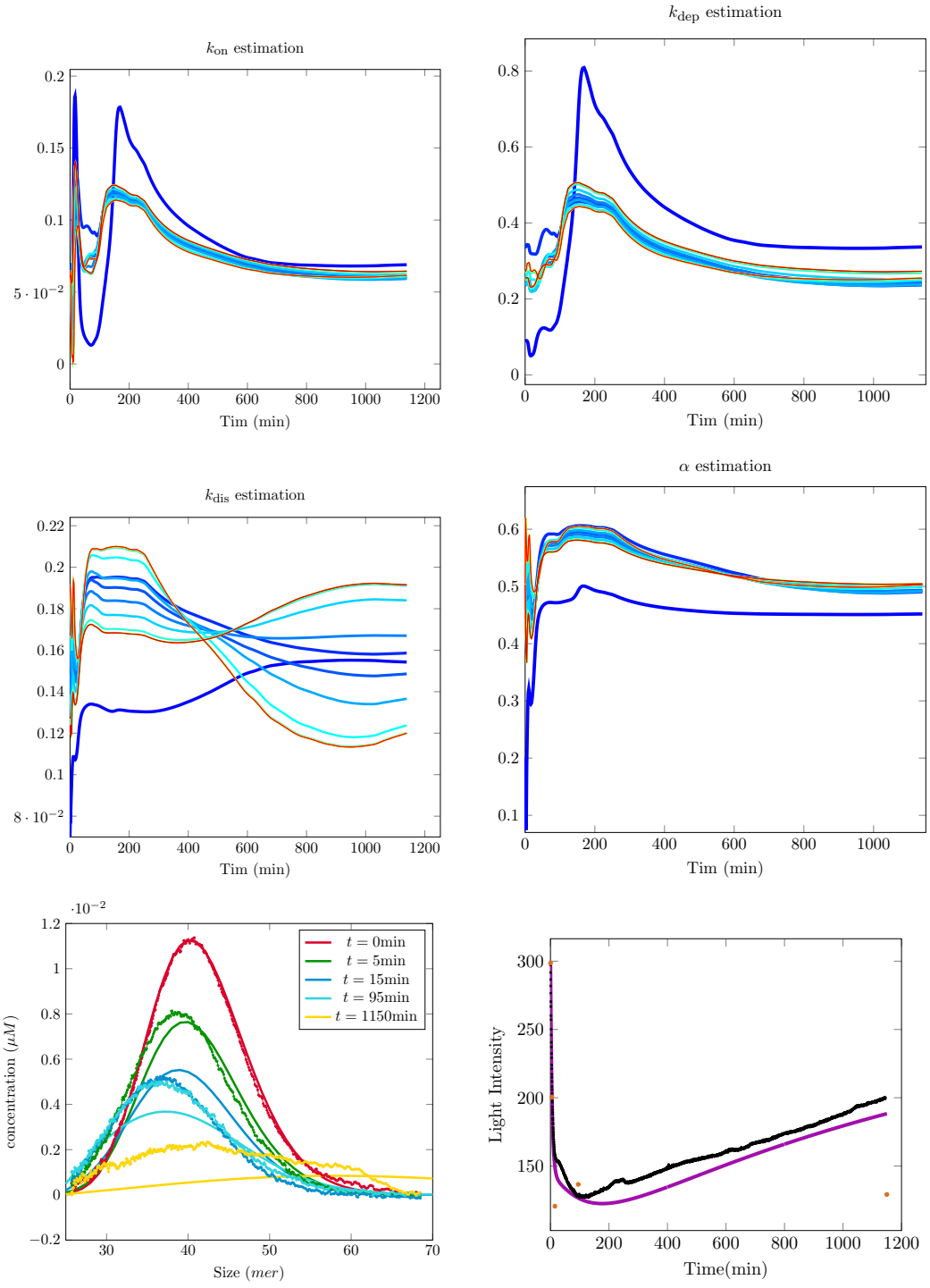



FIGURE 1.38 – Iterated EKF estimations (Table T 1.3.9.2) and observation comparison, $\rho = 7\mu M$.

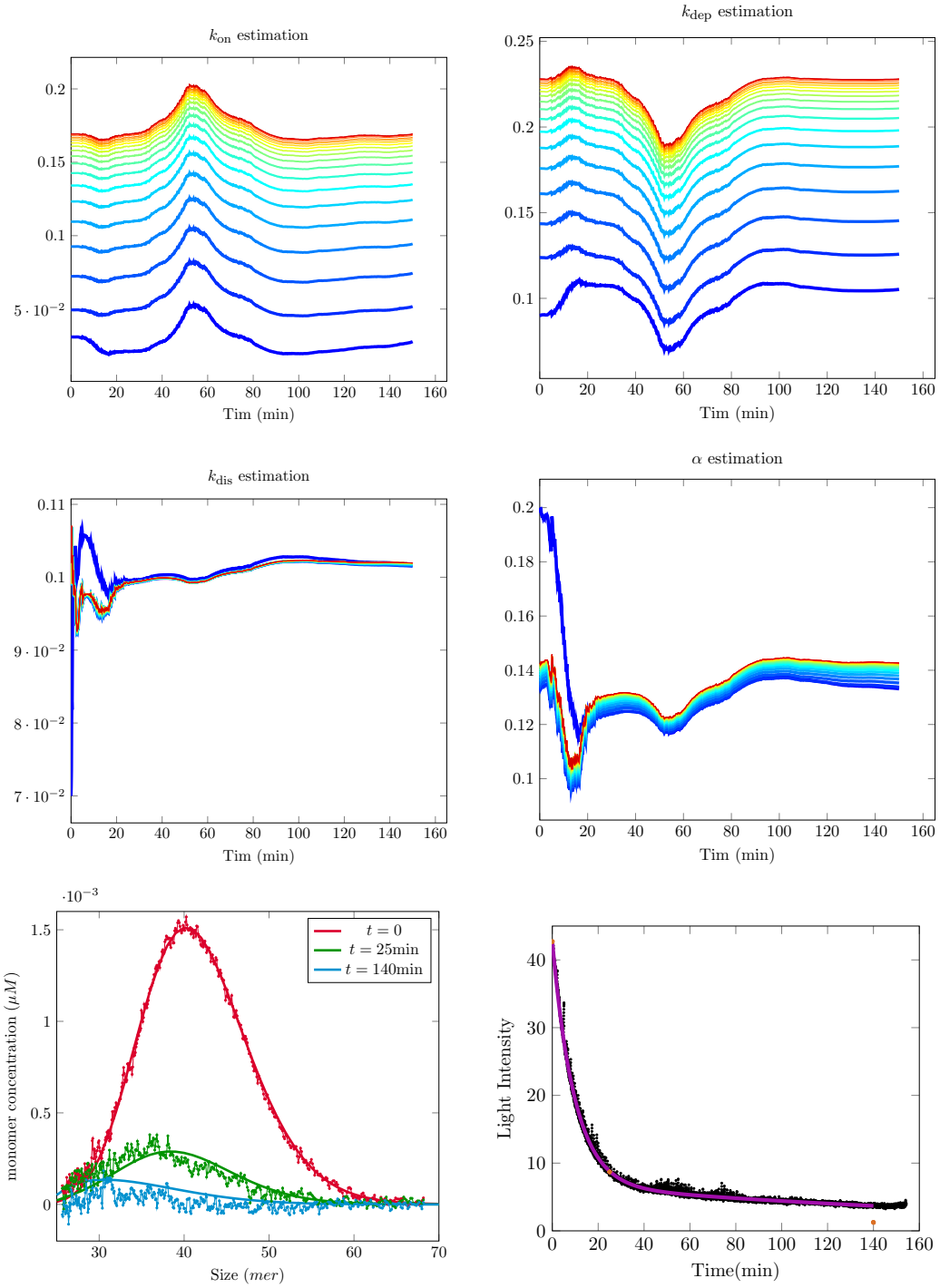



FIGURE 1.39 – Iterated EKF estimations (Table T 1.3.9.4) and observation comparison, $\rho = 1\mu M$.

— We move to the high concentration experiment. In Figure 1.40 we present the parameter estimation obtained with initial condition $(k_{\text{on}}, k_{\text{dep}}, k_{\text{dis}}, \alpha) = (0.1, 0.1, 0.1, 0.1)$ and covariance operators $P_v = 10^{-20}$, $P_{k_{\text{on}}} = P_{k_{\text{dep}}} = P_{k_{\text{dis}}} = P_\alpha = 10^{-2}$ and $W = 10$. The final estimation is

| $\hat{k}_{\text{on}}$ | $\hat{k}_{\text{dep}}$ | $\hat{k}_{\text{dis}}$ | $\hat{\alpha}$ |
|---|---|---|---|
| 0.06085 | 0.24226 | 0.15248 | 0.49153 |

. (T 1.3.9.5)

We observe that the convergence of the parameter estimations to a minimal value is reached in 9 iterations of the EKF. Furthermore, we do not see the oscillating behaviour typical of other estimations, see Figure 1.38. For these reasons we can trust these estimations.

In conclusion, here are our final estimations

| $\rho$ | $\hat{k}_{\text{on}}$ | $\hat{k}_{\text{dep}}$ | $\hat{k}_{\text{dis}}$ | $\hat{\alpha}$ |
|---|---|---|---|---|
| $1\mu M$ | 0.16960 | 0.22829 | 0.10195 | 0.14268 |
| $3\mu M$ | 0.24335 | 0.61940 | 0.12948 | 0.48961 |
| $7\mu M$ | 0.06085 | 0.24226 | 0.15248 | 0.49153 |

(T 1.3.9.6)

## 1.4 Conclusions and discussions of the chapter

Our study brought to light the presence of at least two oligomer species in the ovPrP system under investigation. We propose a qualitative description of the aggregation process, taking into account only two oligomer species. The two species are characterised as follows

— a first oligomer species, termed stable, which mostly polymerise and depolymerise

— a second oligomer species, termed unstable, mostly disintegrate.

The disintegration of the unstable oligomers generates a monomer reservoir from which stable oligomers can draw and polymerise.

We present an ODE model with the kinetic rates as parameters. Under the assumption of size-independent kinetic rates, we have shown that the model is completely determined by only three parameters. This model is able to reproduce the three typical behaviours of oligomer systems that have been identified experimentally. Specifically, we have the complete transformation of oligomers into monomers at low total concentrations, the presence of a pseudo steady-state at middle total concentrations and an increase in the average oligomer size at high total concentrations.

By means of a data assimilation strategy, we have exploited the SLS data to estimate the kinetic parameters associated to three experiments performed at three different total concentration regimes.

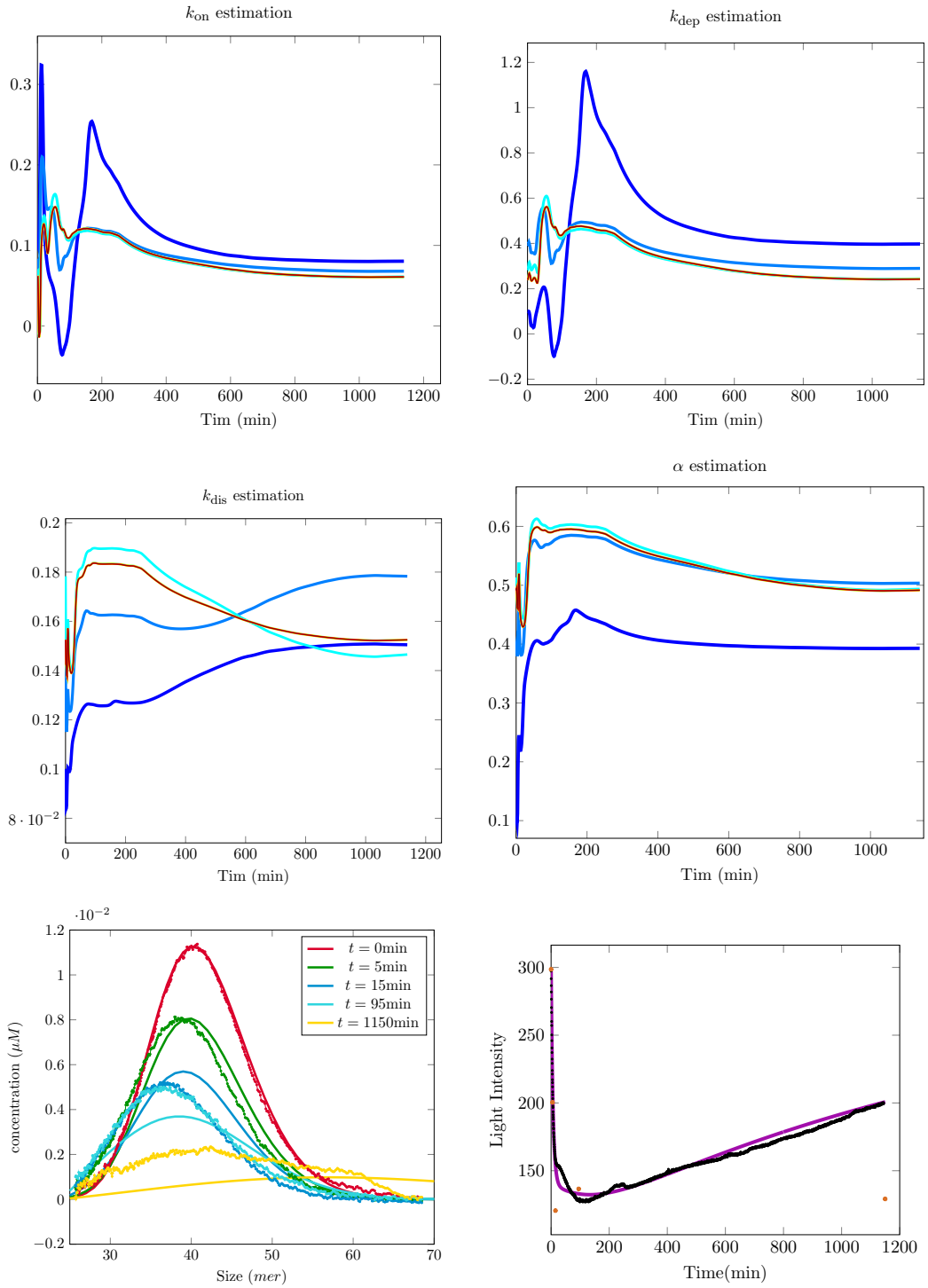



FIGURE 1.40 – Iterated EKF estimations (Table T 1.3.9.5) and observation comparison, $\rho = 7\mu M$.

The comparison between our estimation of the size distribution and the SEC data – which were not used to compute the estimations – shows a good agreement between the two. Hence, our model is able to accurately reproduce the evolution of the average size (SLS data) and the evolution of the oligomer size distribution. Taking these considerations into account we validate the model. In particular, our initial simplifying hypothesis of size-independent parameters is representative of reality. We notice small variations in the three sets of parameter estimations relative to the three experimental settings. The global agreement in their orders of magnitude further confirms the reliability of the results.

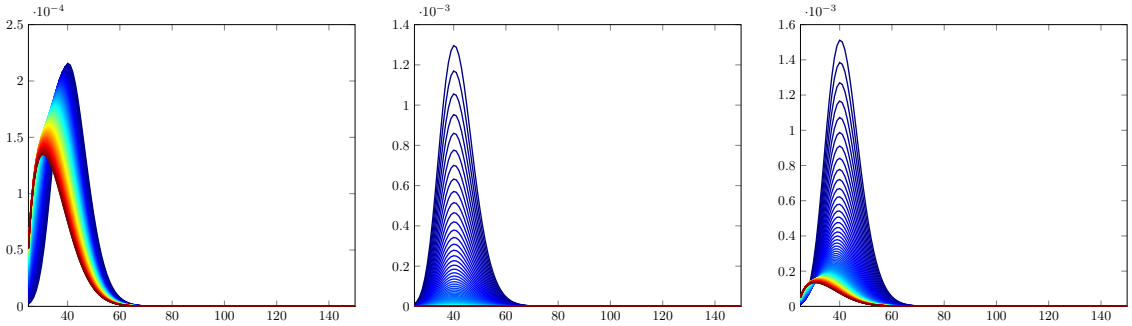
This methodology can be directly applied in future research on protein polymerisation. Two of its main advantages are the following :

— It can reduce experimental costs. We were able to compute our estimations by considering the three SLS data sets and only one of the twelve SEC data sets.

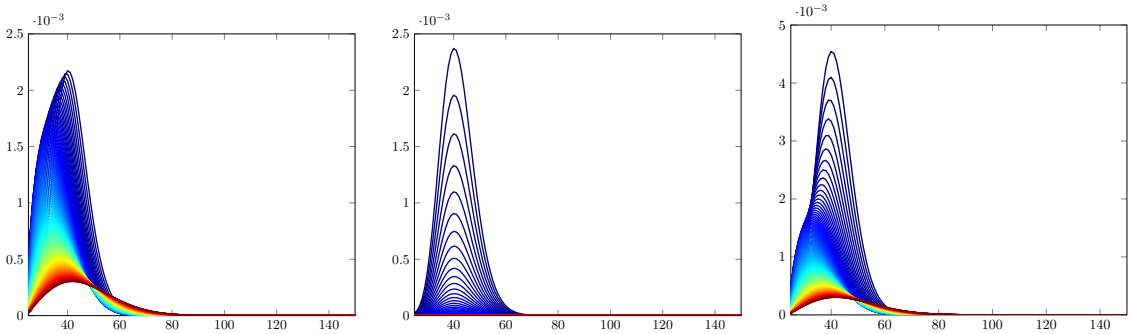
— It can provide access to new data. For instance, we can have an estimation of the oligomer size distribution, at any time. Of even more interest, is that it allows us to look at the evolutions of the two oligomer species separately. Using the estimations in Table T 1.3.9.6, we simulate the evolution of the two oligomer species for a range of sizes $[25, 150]mer$. We obtain the results illustrated in Figure 1.41a, 1.41b, 1.41c for the three concentration regimes. In particular, in the case where $\rho = 7\mu M$, we observe the presence of oligomers with sizes bigger than $75mer$. We recall that the physical features of the SEC device make it unsuitable to analyse aggregates of sizes bigger than $100mer$. These simulations could thus explain the disagreement between the SLS data and the SEC data for $t = 1150$min and $\rho = 7\mu M$, that we highlighted in our analysis.

The results of this work are gathered in Chapter 2 in the form of a pre-printed article

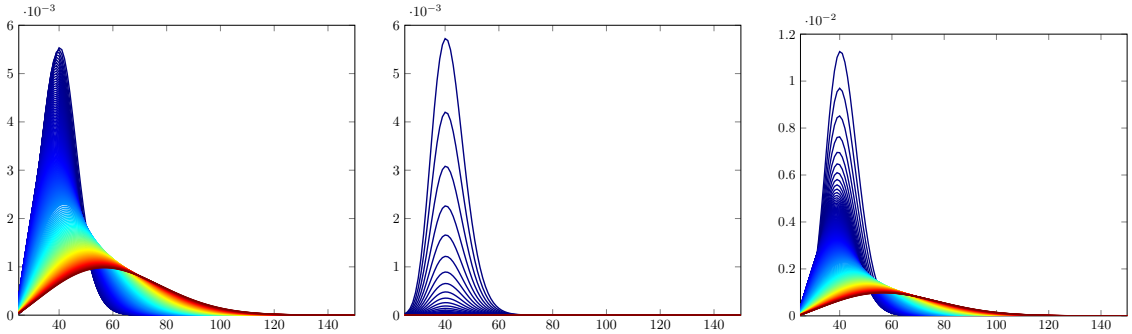
*The mechanism of monomer transfer between two structurally distinct PrP oligomers*
*A. Armiento, P. Moireau, D. Martin, N. Lepejova, M. Doumic and H. Rezaei.*



(a) Size distribution simulation with parameter in Table T 1.3.9.6, for $\rho = 1\mu M$, time step $\delta t = 0.1$, observation time $\tau = 140$min.



(b) Size distribution simulation with parameter in Table T 1.3.9.6, for $\rho = 3\mu M$, time step $\delta t = 0.15$, observation time $\tau = 270$min.



(c) Size distribution simulation with parameter in Table T 1.3.9.6, for $\rho = 7\mu M$, time step $\delta t = 0.2$, observation time $\tau = 1150$min.

FIGURE 1.41 – With a colormap going from dark blue to dark red, we represent the size distribution of stable oligomers $\{y_i(t)\}_{25\leq i\leq 150}$ (left), unstable oligomers $\{w_i(t)\}_{25\leq i\leq 150}$ (middle) and the convolution of the two $\{y_i(t) + w_i(t)\}_{25\leq i\leq 150}$ (right) on the time grid $0 = t_0 < \dots t_N = \tau$, with time step $\delta t$.

CHAPITRE 2

# Article : Mechanism of monomer transfer between two oligomer species


## Abstract

In mammals, Prion pathology refers to a class of infectious neuropathologies in which the mechanism is based on the self-perpetuation of structural information stored in the pathological conformer. The characterisation of the PrP folding landscape has revealed the existence of a plethora of pathways conducing to the formation of structurally different assemblies with different biological properties. However, the biochemical interconnection between these diverse assemblies remains unexplored. The PrP oligomerisation process leads to the formation of neurotoxic and soluble assemblies called O1 oligomers with a high size heterodispersity. By combining different size distribution estimation techniques as a function of time with kinetics modelling and data assimilation we revealed the existence of at least two structurally distinct sets of assemblies, O1a and O1b, forming O1 assemblies. These two groups exchange monomers through a disintegration process that increases the size of O1a. Our observations suggest that PrP oligomers constitute a highly dynamic population. Our results show that protein assemblies responsible for Prion diseases are a highly dynamical population, and that stable assemblies have to be explicitely targeted by drug treatments.

# The mechanism of monomer transfer between two structurally distinct PrP oligomers

Aurora Armiento[1,2], Philippe Moireau[2], Davy Martin[3], Nad'a Lepejova[3], Marie Doumic[1], Human Rezaei[3]

1: Sorbonne Universités, Inria, UPMC Univ Paris 06, Lab. J.L. Lions UMR CNRS 7598, Paris, France

2: Inria and Université Paris-Saclay, Campus de l'Ecole Polytechnique, 91128 Palaiseau, France

3: INRA, UR892, Virologie Immunologie Moléculaires, Jouy-en-Josas 78350, France

**Corresponding authors:**

Marie Doumic, Sorbonne Universités, Inria, UPMC Univ Paris 06, Lab. J.L. Lions UMR CNRS 7598, Paris, France; E-mail: marie.doumic@inria.fr

Philippe Moireau, Inria and Université Paris-Saclay, Campus de l'Ecole Polytechnique, 91128 Palaiseau, France; E-mail: philippe.moireau@inria.fr ;

Human Rezaei, Map2, VIM, INRA- Domaine de Vilvert, 78352, Jouy-en Josas, France, Tel.: (+33)1-34-65-27-89; E-mail: human.rezaei@jouy.inra.fr ;

**Running title:** Heterogeneity in Prion assemblies

**Keywords:**

Prion; oligomer; heterogeneity; quasi-species; Bekker-Döring system; Kalman filter

## Abstract

In mammals, Prion pathology refers to a class of infectious neuropathologies in which the mechanism is based on the self-perpetuation of structural information stored in the pathological conformer. The characterization of the PrP folding landscape has revealed the existence of a plethora of pathways conducing to the formation of structurally different assemblies with different biological properties. However, the biochemical interconnection between these diverse assemblies remains unexplored. The PrP oligomerization process leads to the formation of neurotoxic and soluble assemblies called O1 oligomers with a high size heterodispersity. By combining different size distribution estimation techniques as a function of time with kinetics modelling and data assimilation we revealed the existence of at least two structurally distinct sets of assemblies, $O1^a$ and $O1^b$, forming O1 assemblies. These two groups exchange monomers through a disintegration process that increases the size of $O1^a$. Our observations suggest that PrP oligomers constitute a highly dynamic population.

Transmissible spongiform encephalopathies (TSEs), or prion diseases, constitute a distinct group of fatal neurodegenerative diseases of humans and other animals. Creutzfeldt-Jakob disease (CJD), Gerstmann-Sträussler-Scheinker syndrome (GSS) and fatal familial insomnia (FFI) are the most common human prion diseases. The prion theory, which has been proposed to describe the self-perpetuation of structural information stored in prion assemblies, is now starting to be extended to a wider range of pathologies caused by protein misfolding and aggregation[1]. One of the intriguing aspects of the prion conversion process is the existence of broad panel of PrP assemblies that are highly heterogeneous in size[2]. The existence of such heterogeneity is associated to stochastic events and often to differences in the micro-environment where the conversion process occurs[3]. However, the diversity in the size of PrP assemblies could also be highly deterministic, as was observed with the oligomerisation process of recombinant PrP (recPrP) in a highly controlled environment[4]. Indeed, recPrP polymerisation at pH 4.1 and 7.2 leads to the formation of at least three structurally distinct neuro-toxic oligomers whose size and ratio are each governed by the primary structure of PrP [5]. The biochemical and biological implications of such a diversity remain unclear even if structurally different prion assemblies are claimed to be at the basis of the quasi-species phenomenon and prion adaptation to different hosts[6]. The existence of structurally different assemblies raises the question of their respective thermodynamic stability and the consequences of their coexistence in the same environment. Indeed, according to an Ostwald-like ripening phenomenon, the coexistence of assemblies structurally different could lead to a transfer phenomenon from the low stability to the high stability assemblies[7].

To address this question, we focused our study on O1 oligomers which (see Figure 1A) are highly heterogeneous regarding their size distribution, as shown by multi-wavelength static light scattering (MWLS) analysis. Two hypotheses could explain the heterogeneity in size of O1 oligomers (Figure 1B). The first hypothesis corresponds to the formation of several discrete oligomers through different polymerisation pathways. In this case, each oligomer is structurally different and could have distinct biological properties. The second hypothesis corresponds to a sequential addition of monomers to an oligomer scaffold similar to the nucleation elongation mechanism proposed for amyloid fibril formation. This second hypothesis could generate either structurally equivalent or non-equivalent objects.

In order to discriminate between these two hypotheses and explore the dynamics of different assemblies which compose the O1 peak, we adopted a strategy that consisted of inducing the depolymerisation of O1 assemblies. During the depolymerisation process, the kinetics of size variation was followed by Static Light Scattering (SLS), which reflects the variation in the

mean average molecular weight <Mw> (Figure 2, A, B and C), and by size exclusion chromatography (SEC) (Figure 2, D, E and F), which, coupled with multi-wavelength static light scattering (MWLS), gave us access to the size distribution (Figure 2, G, H and I) at several time points of the experiment. The depolymerisation of O1 assemblies at 1μM (equivalent to the monomer concentration) appears to be total and gives rise to the formation of monomeric PrP as shown by SEC and size distribution (SD) as a function of time (Figure 2 A, D and G). According to the SEC profile, the general behaviour of the shortening process, which leads to the formation of the monomers, cannot be explained by sequential depolymerisation only, but requires also a disintegration process (see SI). However, during the depolymerisation process, followed by SEC and SD, an asymmetric decrease in the O1 peak was observed. This asymmetric evolution could suggest either a faster rate of decrease of large O1 assemblies, or could result from the depolymerisation of at least two different species (Figure 1B).

As for O1 at 1 μM, the depolymerisation of O1 at 3μM, followed by the SLS signal, revealed a decrease in <Mw> of the system until a plateau was reached (Figure 2B). SEC and SD analysis as a function of time revealed, as for O1 at 1μM, a faster decrease in the amount of large assemblies. However, while for O1 at 1 μM the depolymerisation into monomers was total, at 3 μM an accumulation of small size assemblies was observed (Figure 2 E and H). Interestingly, the depolymerisation process at 7 μM led us to highlight a hidden process. First, the initial stage of the depolymerisation process at 7μM appears to be clearly *multiphasic* (Figure 2C insert). As this particular behaviour appears at a higher O1 concentration, this suggests the existence of a multi-order kinetic process such as a polymerisation process. Moreover, for O1 at 7 μM another particularity was observed. The SLS signal (i.e. <Mw> of the system) presents a minimum for t=120min. Therefore, during the first step of the process the <Mw> of the system decreases. The system is in a depolymerizing/disintegrating mode. Then for t >120, the SLS signal increases as a function of time, suggesting an increase of <Mw> of the system and a polymerizing mode. SEC and SD profiles as a function of time, which revealed the apparition of high molecular weight assemblies initially absent in the O1 peak, have confirmed this observation. Two hypotheses could explain the apparition of high molecular weight assemblies. The first hypothesis corresponds to the formation of de novo assemblies formed directly by the monomer. The second hypothesis corresponds to an uptake of monomers by thermodynamically more stable assemblies. To discriminate between these two hypotheses, monomers at concentration of 7μM were incubated in the same conditions as

O1 at 7µM and no polymerisation was observed [4]. These last observations led us to propose a mechanism in favour of the second hypothesis and to build a kinetic model that takes the overall process.

In order to build a kinetic model describing the evolution of the system, we considered different points: the first is the existence of a disintegration process highlighted at 1µM concentration of O1. The second process, which was absent at 1µM but appears at 7µM, is the occurrence of a repolymerisation reaction. The third consideration is the existence of a depolymerisation process leading to the formation of monomers that contribute to a shift to smaller sizes for all three concentrations (Figures 2 G, H and I). Finally, we exclude the spontaneous polymerisation of the monomers in our experimental conditions as was previously demonstrated [8]. Hence, any model should combine at least these three elements: disintegration, templating (i.e. recapture), and depolymerisation. However, we should also consider that if disintegration applies to all the oligomers (i.e. if we consider only structurally equivalent assemblies) then it will prevent the polymerisation process, since it would lead all polymers to disintegrate into monomers. Therefore, we should conclude on the existence of at least two structurally distinct species coexisting under the O1 peak: one unstable, subject to disintegration ($o_i^a$), the other more stable, with a disintegration rate very low ($o_i^b$). Gathering all these elements, the simplest possible model could be illustrated by Figure 3 and leads us to the following three reactions

$$O_i^a \xrightarrow[k_{dis}]{} i\, C_1,$$

$$O_i^b + C_1 \xrightarrow[k_{on}]{} O_{i+1}^b,$$

$$O_i^b \xrightarrow[k_{dep}]{} O_{i-1}^b + C_1,$$

which result in the following differential equations

$$\frac{do_i^a}{dt} = -k_{dis} o_i^a$$

$$\frac{do_i^b}{dt} = k_{on} c_1 \left(o_{i-1}^b - o_i^b\right) - k_{dep}\left(o_i^b - o_{i+1}^b\right),$$

$$\frac{dc_1}{dt} = -k_{on} c_1 \sum_{i=2}^{\infty} o_i^b + k_{dep} \sum_{i=2}^{\infty} o_i^b + k_{dis} \sum_{i=2}^{\infty} i\, o_i^a \quad ,$$

where $o_i^a$ denotes the (time-dependent) concentration of unstable oligomers of size $i$, $o_i^b$ the concentration of stable oligomers of size $i$ and $c_1$ the concentration of monomers. For the sake

of simplicity, we consider constant disintegration, polymerisation and depolymerisation rates, respectively denoted $k_{dis}$ , $k_{on}$ and $k_{dep}$, and we do not consider polymerisation and depolymerisation for the unstable species. We are thus led to estimate only four parameters: the three reaction rates, and the ratio $o_i^b/(o_i^b + o_i^a)$ at the initial time, which we also assume to be independent of the size i. To fit the model to the SLS data (not making use of the SEC data), we use a data assimilation approach by Kalman filtering (see Supporting Information for more details). Best-fit parameters are shown in Table 1 and model-data comparison in Figure 3. For such a simple model, we found a remarkable quantitative agreement, as well as parameters remaining in the same order of magnitude. Moreover, the good agreement obtained between the SEC experimental data and the curve predicted by the parameters fitted on SLS leads us to validate the monomer exchange model between $o_i^a$ and $o_i^b$ while other models failed to fit the time-dependence size evolution of oligomer assemblies (see SI).

These observations lead us to validate the monomer exchange model between the two sets of O1 assemblies. Our conclusion is thus twofold: first, a very simple two-species model is able to fit the data, whereas a one-species model, even with size-dependent coefficients, is not. Second, surprisingly, we do not need size-dependent coefficients (Table 1), so that it is possible that within a given species, it is plausible that the objects may be structurally equivalent.

| | $k_{on}$ | $k_{dep}$ | $k_{dis}$ | $\frac{o_i^b}{o_i^b+o_i^a}$ |
|---|---|---|---|---|
| $\rho = 1\mu M$ | 0.16 | 0.22 | 0.10 | 0.14 |
| $\rho = 3\mu M$ | 0.22 | 0.57 | 0.11 | 0.45 |
| $\rho = 7\mu M$ | 0.06 | 0.25 | 0.17 | 0.50 |

**Table 1:** Best-fit parameters obtained by the data assimilation method on the two-species model $o_i^a$ and $o_i^b$ (for details see also SI). $\rho$ corresponds to total oligomer concentration equivalent to monomer.

## Conclusion

Our step-by-step approach, from experimental analysis to data assimilation, leads us to a partly counter-intuitive conclusion: the existence of monomer exchange between two types of PrP oligomer assemblies. The formation of heterodisperse assemblies during the evolution of pathologies due to protein misassembly raises the question of their coexistence and their evolution. This phenomenon occurs during prion conversion for which several species could coexist and form what is also commonly called prion quasi-species[9],[10]. From a thermodynamic point of view, it is clear that not all assemblies are kinetically and energetically equivalent and some species with specific biological activities could be generated transitorily. However, the evolution of all these assemblies should follow specific thermodynamic and kinetic rules such as selection by higher stability and/or higher rate of formation. In the present work we demonstrate that from monomeric PrP at least two types of oligomers are simultaneously generated. In conditions that could biologically correspond to monomer depletion, we demonstrate that these two oligomers are able to exchange monomers. The biological consequences of such a phenomenon could be the transitory apparition of physiopathological patterns and the existence of buffer assemblies serving as monomer reservoirs to enhance and maintain more stable assemblies. It is also clear that such a phenomenon should be considered for all therapeutic purposes.

## Experimental section

**Preparation of recombinant PrP constructs.**

Full-length Ovine PrP 23-234 (Ala-136, Arg-154, Gln-171 variant) were produced in Escherichia coli and purified as described previously[11]. The O1 oligomers were generated by incubating of OvPrP at 80mM at 55°C for 6 hours and purified as previously detailed[4]. The size distribution of O1 assemblies was estimated by coupling to multi-wavelength static light scattering with size exclusion chromatography using a TSK 4000SW. The resulting data were transformed to size distribution using a custom MATLAB program. The depolymerzation of O1 assemblies was followed by light scattering by incubating O1 assemblies at 50°C.

**Kinetic simulations and data assimilation.**

Differential equations presented above have been simulated, as well as many variants using a first-order scheme in Matlab (for details see also SI). The parameters were estimated using the Extended Kalman[12] Filter Method, implemented in Matlab, warping the lines of the Verdandi

data assimilation library (http://verdandi.sourceforge.net).

## References


[1] D. C. Bolton, M. P. McKinley, S. B. Prusiner, *Science* **1982**, *218*, 1309-1311.

[2] P. Tixador, L. Herzog, F. Reine, E. Jaumain, J. Chapuis, A. Le Dur, H. Laude, V. Beringue, *PLoS Pathog* **2010**, *6*, e1000859.

[3] K. Annamalai, K. H. Guhrs, R. Koehler, M. Schmidt, H. Michel, C. Loos, P. M. Gaffney, C. J. Sigurdson, U. Hegenbart, S. Schonland, M. Fandrich, *Angew Chem Int Ed Engl* **2016**, *55*, 4822-4825.

[4] F. Eghiaian, T. Daubenfeld, Y. Quenet, M. van Audenhaege, A. P. Bouin, G. van der Rest, J. Grosclaude, H. Rezaei, *Proc Natl Acad Sci U S A* **2007**, *104*, 7414-7419.

[5] S. Prigent, H. Rezaei, *Prion* **2011**, *5*, 69-75.

[6] J. Li, S. Browning, S. P. Mahal, A. M. Oelschlegel, C. Weissmann, *Science* **2010**, *327*, 869-872.

[7] J. Zhang, M. Muthukumar, *J Chem Phys* **2009**, *130*, 035102.

[8] N. Chakroun, S. Prigent, C. A. Dreiss, S. Noinville, C. Chapuis, F. Fraternali, H. Rezaei, *FASEB J* **2010**, *24*, 3222-3231.

[9] T. Nakayashiki, K. Ebihara, H. Bannai, Y. Nakamura, *Mol Cell* **2001**, *7*, 1121-1130.

[10] C. Weissmann, J. Li, S. P. Mahal, S. Browning, *EMBO Rep* **2011**, *12*, 1109-1117.

[11] H. Rezaei, D. Marc, Y. Choiset, M. Takahashi, G. Hui Bon Hoa, T. Haertle, J. Grosclaude, P. Debey, *Eur J Biochem* **2000**, *267*, 2833-2839.

[12] R. E. Kalman, *Journal of Basic Engineering* **1960**, *82*, 35-45.

## Legends

**Figure 1: Size distribution of PrP oligomers.**

a) size exclusion chromatography (balck line) coupled to multiwavelength static light scattering lead to estimate size of oligomers generated during PrP oligomerization (in red). The O1 heterodispersity in size (i.e. moleculare weight) could result either to the formation of subpopulation of oligomer ($C_k$, $B_j$ and $A_i$) according to a multiple parallel pathways or sequential size increasing (from $C_n$ to $C_i$).

**Figure 2: Exploration of O1 oligomers stability through their depolymerisation rate.** The depolymerisation rate of O1 assemblies have been explored by using static light scattering (i.e. mean average moleculare weight ,<Mw>) as function of time (a,b and c). Arrows indicate aliquots sampling at different time for SEC (d, e and f) and MWLS (g,h and i) analysis in order to estimate size distribution analysis as function of time. Colors of arrows are reported to the curves colors and are related to the time of sampling. Left column (a,d and g) corresponds to depolymerisation experiments performed at O1 concentration of 1μM. Midel column (b,e and h) to depolymerisation experiments performed at O1 concentration of 3μM and right column (c,f and i) to depolymerisation experiments performed at O1 concentration of 7μM.

**Figure 3: Comparison between experimental data and synthetic observations**. Size distribution (a,b and c) and light scattering intensity (d,e and f) as function of time and at different O1 concentration and have been fitted (solid line) using best-fit parameters reported in Table 1. The experimental data are represented in dots. Simulation corresponding to the evolution of size distribution of $O_i^a$ (g) and $O_i^b$ (h) and the convolution of $O_i^a + O_i^b$ (i) (see Supplementary Information).

**SUPPLEMENTARY INFORMATION**

## 1. Details on the step-by-step modelling approach: how we came to the best-fit model

We detail here the approach we followed in order to explain how, from an extremely simple model, we were naturally led to our conclusion. We also believe that this methodology can be reproduced for other experiments.

To model the kinetics of oligomers, the simplest and most widespread model consists first in considering only polymerisation and depolymerisation by monomer addition. In the case where there exists only one species for each size, we model it by its time-dependent concentration $c_i(t)$, and the reactions read as follows

$$c_i + c_1 \xrightarrow{k_{on_i}} c_{i+1}$$

$$c_i \xrightarrow{k_{dep_i}} c_{i-1} + c_1.$$

This corresponds to the so-called Becker-Döring system[1]

$$\frac{dc_i}{dt}(t) = -(k_{on_i}\, c_i(t) - k_{on_{i-1}} c_{i-1}(t)) + (k_{dep_{i+1}}\, c_{i+1}(t) - k_{dep_i} c_i(t)), \quad i \geq 2,$$

$$\frac{dc_1}{dt}(t) = \sum_{i=2}^{\infty} (-c_1(t)\, k_{on_i}\, \square_i(t) + k_{dep_i} c_i(t)).$$

Here we do not take into account the spontaneous polymerisation of monomers, taking $k_{on_1} = 0$ [2]. Furthermore, the experiments start with only oligomers, or equivalently $c_1(0) = 0$, so that polymerisation does not influence the beginning of the reaction.

The first thing that we notice is that considering the SEC data (Figure 2, D, E and F) at the beginning of the reactions, the peak value both slightly shifted to the smaller sizes, lowered, and the polymerised mass decreased. At first sight, this is in line with the dynamics governed by a purely depolymerising system.

In a first approximation, the Becker-Döring system may be approximated by a transport equation – the so-called Lifshitz-Slyozov system – so that it acts mainly as a drift operator,

driving the peak either towards smaller sizes (as observed here), when depolymerisation is stronger, or towards larger sizes, when polymerisation dominates (as observed at the end of the reaction curve 7 µM, see Figure 2, F). With size-varying coefficients, the model can deform the peak, but polymerised mass can be lost only when polymers reach the smallest stable size. In a second approximation, a correction to the drift operator is given by a diffusion operator, leading the peak to be both larger and lower, as can be seen in the simulation reported in Figure 3, H.

Hence the behaviour of the peaks observed in Figure 2, D, E and F may appear qualitatively plausible at first sight – they both shifted to the left and are more diffuse. However, when simulating and trying to fit the data – only the beginning of the reactions, where polymerisation is negligible since there is only a very small number of monomers – we conclude that depolymerisation alone could not explain the curves: the correct loss of mass involves a diffusion effect too strong, and a shift towards smaller sizes that are much higher than the one observed.

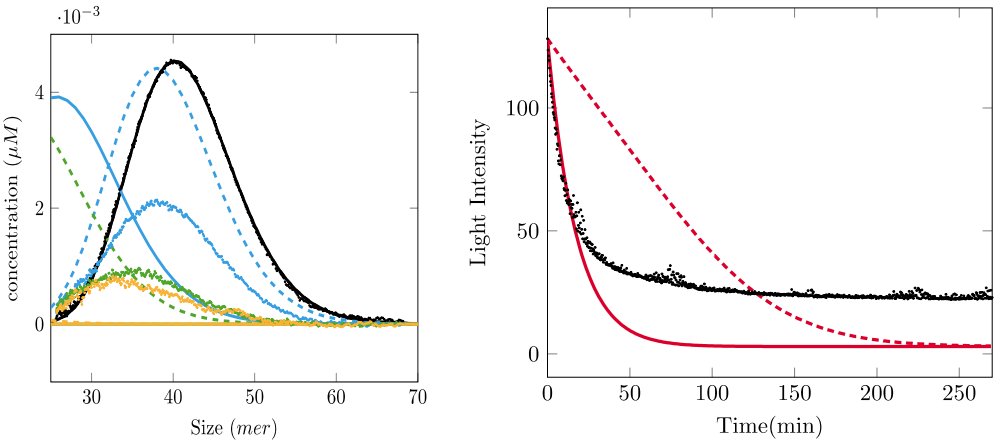



Fig. S1: comparison between experiments (dots) and simulations(dashed and solid lines) with pure depolymerisation models, for an initial concentration of 3µM. Left: SEC data at times 0 (black), 15min (blue), 125 min (green) and 270min (yellow). Right: SLS data (black dots: experiment, red dashed and solid: simulations).

Dashed curves (Left and Right) correspond to $k_{dep} = 0.16$: the position of the peak for the first time t=15min is correct for the size distribution, but its height is not and nor is the slope of the SLS data.

Solid curves (Left and Right) correspond to $k_{dep} = 1$: the slope for SLS data fits well at the beginning, but the size-distribution has shifted too much to the left.

For instance, if we take $k_{on} = 0$ and $k_{dep} = 1$, we can see in Figure S1 Right that the solid line fits the SLS data at the beginning of the experiment. However, when we compare the simulated oligomer distribution and the SEC data at $t = 15min$ we can observe a strong difference both in peak position and peak value.

To approximate the peak position of the distribution at time $t = 15min$, we consider the parameters $k_{on} = 0$ and $k_{dep} = 0.16$, resulting in the dashed lines of Figure S1: the peak position is correct, but its value is much too high, whereas the slope for the SLS data is too small.

This leads us to add a disintegration term in the system, so that we obtain

$$\frac{dc_i}{dt}(t) = -(k_{on_i}\, c_i(t) - k_{on_{i-1}} c_{i-1}(t)) + (k_{dep_{i+1}}\, c_{i+1}(t) - k_{dep_i} c_i(t)) - k_{dis_i} c_i,$$

$$i_0 \leq i \leq i\_1$$

With this additional term, the beginning of the reaction curves fits well. However, in fact, the disintegration term leads any size of polymer to vanish exponentially fast at a rate $k_{dis}$, even if there is polymerisation.

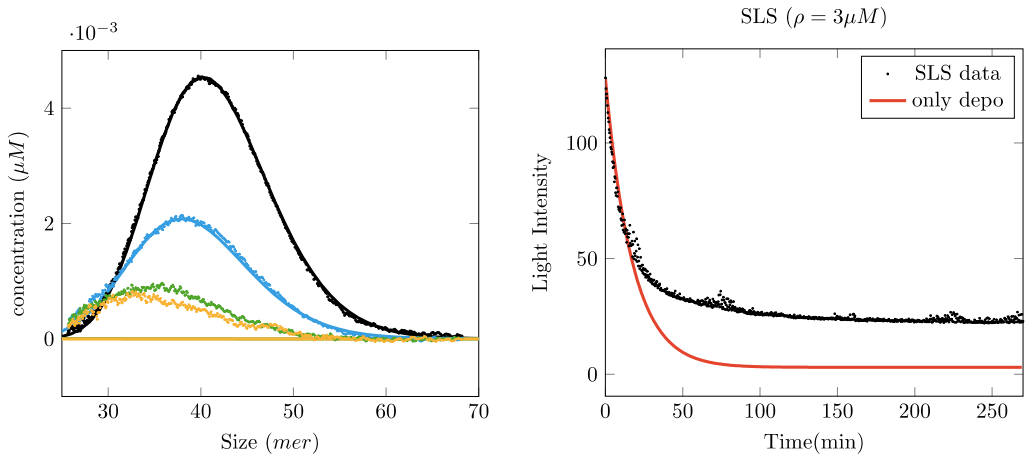



Figure S2: comparison between experiments (dots) with simulation (solid line) with a polymerisation, depolymerisation and disintegration model with one species, for an initial

concentration of 3µM. Left: SEC data times 0 (black), 15min (blue), 125 min (green) and 270min (yellow). Right: SLS data (black dots: experiment, red solid line: simulations).
We see that the size distributions and the SLS curve do not fit untill time 15min, but afterwards, due to the disintegration process, they all go to zero in the simulations, in contrast to the experimental measurements.

This behaviour is illustrated in Figure S2: choosing the kinetic parameters $k_{on} = 0, k_{dep} = 0.16$, $k_{dis} = 0.05$, we are able to fit the beginning of the 3µM experiment both in SLS data and in SEC data. The model cannot reproduce long time behaviour ($t > 20min$) because the simulated oligomer distribution tends to zero too rapidly.
With this model we can well describe the experiments at the initial concentration of 1µM, but then the recapture process observed at the end of the reaction at 7µM becomes impossible to obtain.
This leads us to dissociate the two observed phenomena:

- the disintegration corresponds to an unstable species A, denoted $o_i^a$
- the shifts of the curve, corresponding to the polymerisation/depolymerisation process of the Becker-Döring system, describe the kinetics of the species B, denoted $o_i^b$.

We then obtain the following equations, as written in the main text

$$\frac{do_i^a}{dt} = -k_{dis} o_i^a,$$

$$\frac{do_i^b}{dt} = k_{on} c_1 \left(o_{i-1}^{ab} - o_i^b\right) - k_{dep}\left(o_i^b - o_{i+1}^b\right),$$

$$\frac{dc_1}{dt} = -k_{on} c_1 \sum_{i=2}^{\infty} o_i^b + k_{dep} \sum_{i=2}^{\infty} o_i^b + k_{dis} \sum_{i=2}^{\infty} i\, o_i^a \quad .$$

We did not make the system any more complex. We claim that this simple model is, in fact, sufficient to explain the experimental results.
The entire analysis was carried out qualitatively, by iterating direct simulations. At this point, having an already qualitatively good agreement between the simulations and the experimental curves, we were ready to go further by using parameter estimation techniques.

**2. Direct simulations and comparison with experimental data**

Before using a fully quantitative parameter estimation method, we had to run simulations and visually compare them with the data, both to gain some idea about the sensitivity of each reaction rate on the model and to obtain orders of magnitude for them.

**a. Scales**

**Size Exclusion Chromatography (SEC).** The aggregates pass through a gel grid, and the device measures the concentration of molecules associated to the same elution volume. In Figure S2

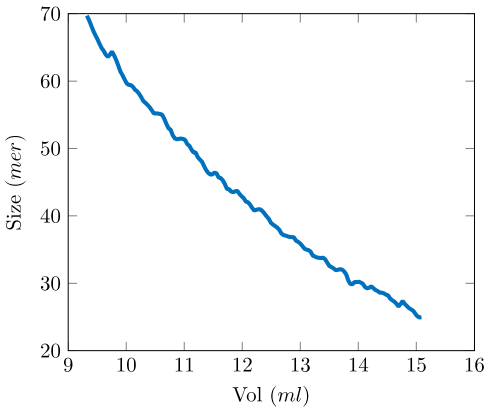



**Figure S3.**
Multi-wavelength static light scattering data, allowing us to have a correspondence between the elution volume measured by SEC (in ml) and the size of the polymers going through the SEC device (in the number of monomers, denoted *mer*).

we see the correspondence between the volume measurement and the size of a polymer. We use this scale to obtain a measurement of the size distribution, up to a constant to be determined. In Figure S4,

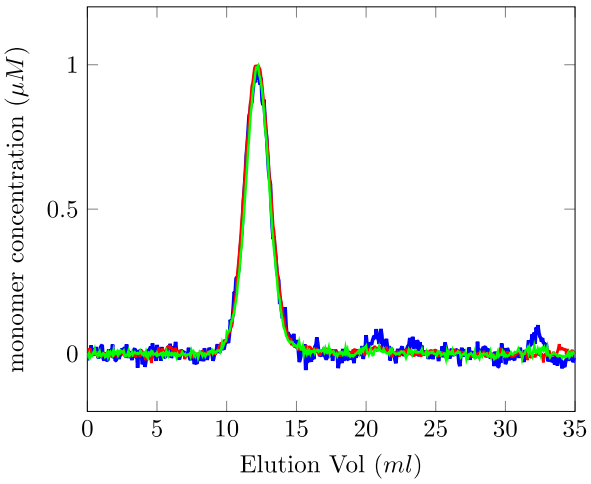



Figure S4

Normalised initial distribution for the three experiments considered, at 1µM (dark blue), 3µM (red) and 7µM (green). We can observe that they superimpose very well.

we scaled the initial size distribution of each experiment to have their peak value equal to one: they exhibited a remarkable agreement, which led us to have a high level of confidence in it. For each experiment, we then scaled this initial SEC measurement by the known initial concentration (1, 3 and 7 µM respectively), and used the same factor to scale the measurements at the following times.

**The Static Light Scattering (SLS)** measures a linear transformation of what is mathematically called the second moment of the polymer concentration, *i.e.* the quantity $\sum_{i\geq 2} i^2 o_i(t)$. Denoting *SLS(t)* the experimental measurement of the SLS at time $t$, we have, for two constants $c>0$ and $c'>0$ such that (comment: c'est peut etre peu clair si on appelle c et c' des constantes et c_i des concentrations )

$$SLS(t) = c \sum_{i\geq 2} i^2 o_i(t) + c'.$$

Here we denoted $o_i = o_i^a + o_i^b$.

To compare the simulations to the data, we thus have to estimate $c$ and $c'$. We thus proceed as follows:

- The constant $c'$ corresponds to the mean amplitude of the noise measured in a cuvette containing no protein (See Figure S5).

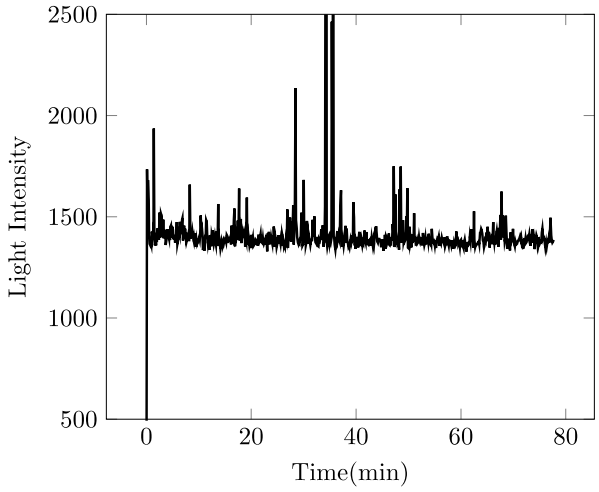



Figure S5

Noise measurement of the empty cuvette for the SLS device.

We measured *c'=1393* (in Light Intensity).

- For the constant *c*, there are several methods are possible methods to estimate it. After testing several, and evaluating the confidence we may have in each, the best appeared to be to use the initial SEC measurements to estimate $\sum_{i\geq 2} i^2 o_i\,(0)$, and then take this value to calculate *c* such that

$$SLS(0) = c\sum_{i\geq 2} i^2 o_i\,(0) + c'.$$

This gave us three values for c, namely 93 (1 µM), 109 (3 µM) and 114 (7 µM). We chose the mean of these values, c=105, which moreover gave a good time-dependent agreement between SLS and SEC data (see Figure S6 )

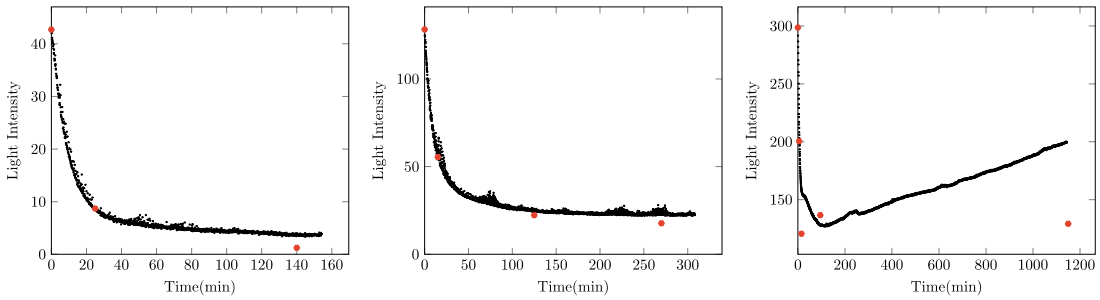



Figure S6

Comparison between SLS (black dots) data and second moment computed from SEC data (red points) for the three experiments, from left to right: 1μM, 3μM and 7μM.

**b. Simulations**

We ran the simulations in Matlab, using a simple first-order scheme to solve the ODE system. All sizes are simulated from size 25 to size 150. We assumed the size limit 24mer to disintegrate instantaneously, since none is experimentally observed, representing an unstable oligomer structure. The upper bound of 150mer has been arbitrarily chosen to encompass all possible oligomer sizes. We remark that SEC data give us the oligomer distribution for sizes between 25 and 70mer.

We simulated the model with various values for the four parameters $k_{on}, k_{dep,}\ k_{dis}$ and the ratio $\theta = o_i^b/(o_i^b + o_i^a)$.

**2. Parameter estimation: Kalman filtering approach**

To estimate the four parameters $k_{on}$, $k_{dep,}\ k_{dis}$ and $\theta = o_i^b/(o_i^b + o_i^a)$, we used the Extended Kalman Filter Method[3]. The codes have been implemented in Matlab.

The state variables $(o_i^a, o_i^b, c_1)$ are extended with the parameters $k_{on}, k_{dep,} k_{dis}$ with dynamics

$$\frac{dk_{on}}{dt} = 0, \frac{dk_{dep}}{dt} = 0, \frac{dk_{dis}}{dt} = 0.$$

To apply this method we need to define an estimation of the initial state that would be the inital condition of the estimator built by the method.

Several simulations of the model allowed us define the *a priori* estimation $k_{on_\circ} = 0.1$, $k_{dep_\circ} = 0.1$, $k_{dis_\circ} = 0.1038$, $\theta_\circ = 0.3$ of the parameters $k_{on}, k_{dep,}\ k_{dis}$ and θ respectively.

The initial condition is thus fixed to $((1-\theta_\circ)o_i(0), \theta_\circ o_i(0), 0, k_{on_\circ}, k_{dep_\circ}, k_{dis_\circ})$.

The dynamics of the Kalman estimator results in the contribution of two terms :

1) the model -- that summarizes our knowledge on the oligomer system
2) 2) a corrective term exploiting the availability of some observation on the system.

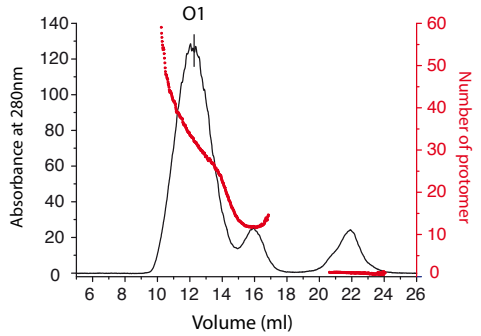
In our case we have used only the SLS data - the observation operator being the second moment as in [4] - to obtain the estimation. We then use the SEC data to validate the estimations. The final estimations are given in Table 1 in the manuscript.

References

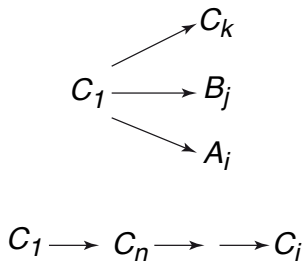

[1] R. Becker, W. Döring, *Ann. Phys* **1935**, *24*, 719-752.
[2] F. Eghiaian, T. Daubenfeld, Y. Quenet, M. van Audenhaege, A. P. Bouin, G. van der Rest, J. Grosclaude, H. Rezaei, *Proc Natl Acad Sci U S A* **2007**, *104*, 7414-7419.
[3] R. E. Kalman, *Journal of Basic Engineering* **1960**, *82*, 35-45.
[4] A. Armiento, M. Doumic, P. Moireau, H. Rezaei, *J Theor Biol* **2016**, *397*, 68-88.

Figure 1

a

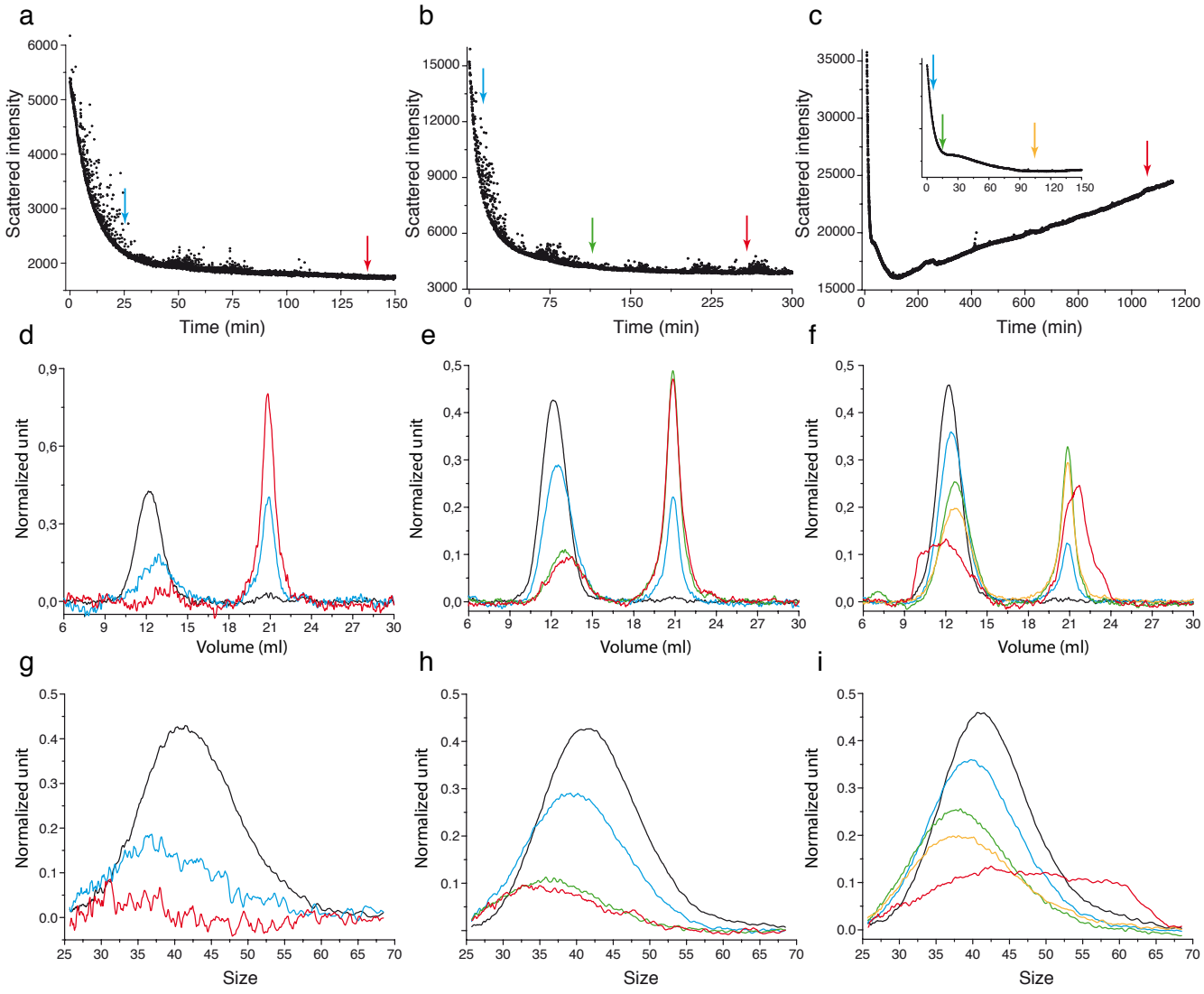

O1
Absorbance at 280nm
Number of protomer
Volume (ml)
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60
6
8
10
12
14
16
18
20
22
24
26

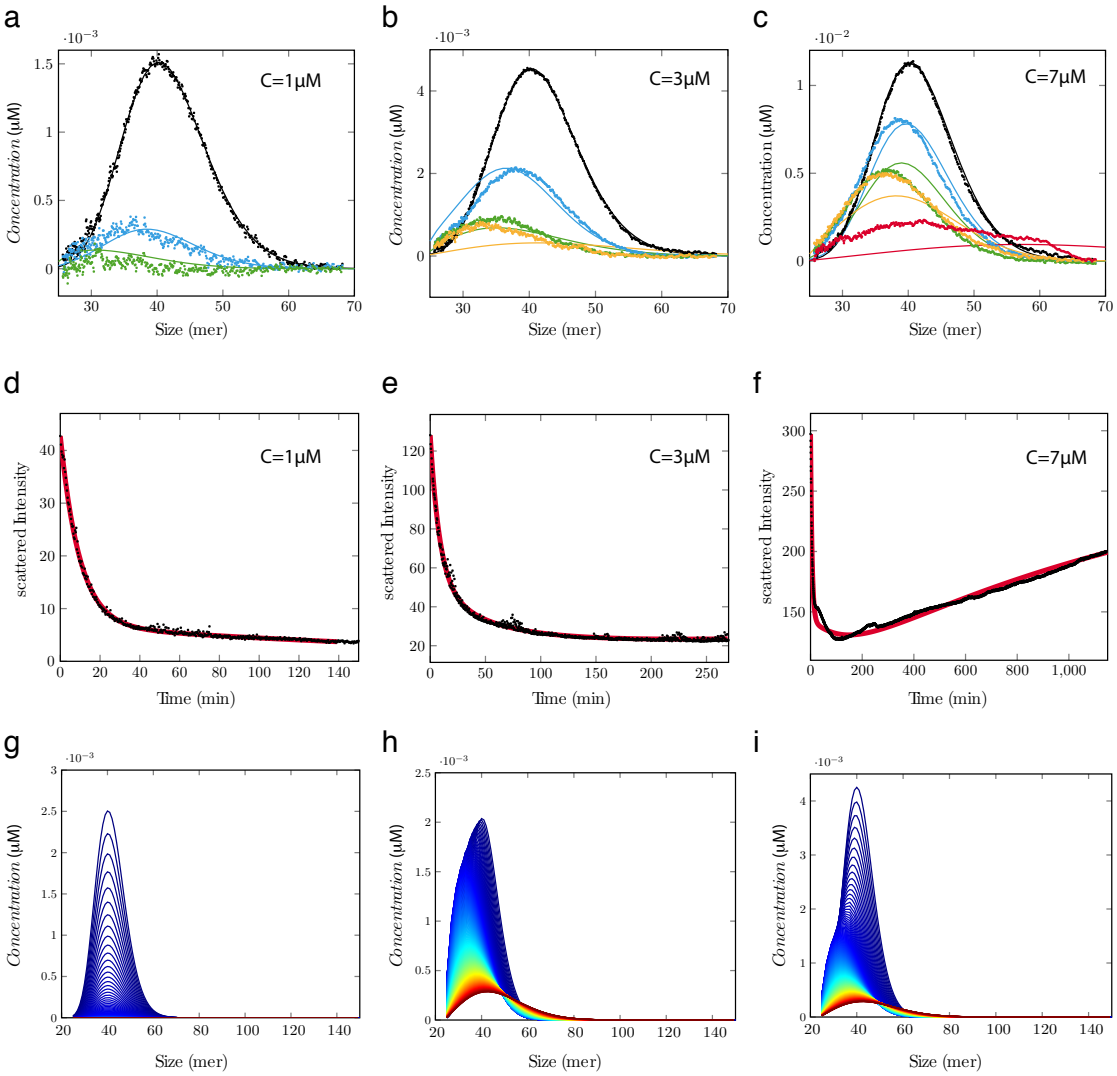
b


$C_k$
$C_1$
$B_j$
$A_i$


$C_1$
$C_n$
$C_i$

Figure 2


a
b
c
Scattered intensity
Time (min)
d
e
f
Normalized unit
Volume (ml)
g
h
i
Size


a
b
c
C=1μM
C=3μM
C=7μM
Concentration (μM)
Size (mer)
d
e
f
scattered Intensity
Time (min)
g
h
i
Concentration (μM)
Size (mer)

## Deuxième partie

# Data assimilation on a PDE model of polymerisation

CHAPITRE 3

# The transport model as a simple prion model

In the second part of this thesis we present an overview of the strategies that can be used to solve the following inverse problem

> Given the observation of some moments of a state function $u$, the solution of a transport equation, we want to estimate the initial condition and/or the transport velocity.

To the best of our knowledge, very few studies on this subject are available in the literature [135, 59, 14]. Interestingly, it finds an application on prion protein modelling. In this introduction we briefly explain how the phenomenon of protein aggregation can be modelled by a transport equation. To do so, we introduce two classical models for phase transition phenomena :

— the infinite ODE system proposed by Becker and Döring [19], which reads as follows

$$\begin{cases} \dot{u}_i = J_{i-1}(u) - J_i(u), & i > 1 \\ \dot{u}_1 = -J_1(u) - \sum_{i=1}^{\infty} J_i(u), & \\ J_i(u) = a_i u_1 u_i - b_{i+1} u_{i+1}, \quad u = (u_i)_{i \geq 1}, & \end{cases}$$

— and the integro-differential system proposed by Lifshitz-Slyozov [111]

$$\begin{cases} \partial_t u(x,t) + \partial_x(V(x,t)u(x,t)) &= 0, & \forall (x,t) \in \mathbb{R}^+ \times \mathbb{R}^+ \\ V(x,t) &= a(x)v(t) - b(x), & \\ v(t) + \int_0^\infty xu(t,x)\mathrm{d}x &= \rho > 0 & \forall t \in \mathbb{R}^+ \\ u_{|t=0} &= u_0, & \\ v_{|t=0} &= v_0. & \end{cases}$$

Collet, Goudon, Poupaud and Vasseur – introducing a scaling parameter – demonstrated in [47] that the solution of the Becker-Döring system converges to the solution of a Lifshitz-Slyozov system when this parameter tends to zero.

At end of this chapter, we formalise the inverse problem studied in the rest of this thesis work. For this second part no information from the previous part is required. Chapter 4 takes the form of an article gathering the work done in collaboration with M. Doumic, P. Moireau and H. Rezaei [6]. It presents two inverse problem solutions : one given by a kernel regularisation method and the other by a data assimilation method called 4d-Var. The two strategies are analysed, compared and illustrated by means of a practical example.
Chapter 5 provides an overview of data assimilation methods.

## Becker-Döring theory and discrete-size coagulation-fragmentation model overview

A very classical approach to describe the evolution of a system of molecules or cells was presented by Smoluchowski in 1917 [162, 163]. In this study the author focuses on the phenomenon of coagulation. Coagulation is the binding of two clusters to form a bigger cluster. This phenomenon is also known in the literature as *aggregation*. By denoting $x_i$ the clusters of size $i$, the coagulation can be represented by the following chemical equation

$$x_i + x_j \xrightarrow{a_{i,j}} x_{i+j}$$

where $a_{i,j}$ are positive kinetic coefficients.

The Becker-Döring theory, formulated for the first time in 1935 [19], models the behaviour of clusters that can just gain one monomer [1] or lose one monomer at a time according to the chemical equations

$$\begin{aligned} x_i + x_1 &\xrightarrow{a_i} x_{i+1} \qquad i > 1, \\ x_i &\xrightarrow{b_i} x_{i-1} + x_1. \end{aligned}$$

In the original version, the quantity of monomers was assumed to be constant over time. Later, in 1979, this model was modified by Penrose and Lebowitz [134], considering the case of monomers used to form larger clusters. When we refer to the Becker-Döring system we usually consider this second formulation.

Let us call $u_i(t)$ the expected number of $i$-particle clusters per unit of volume at time $t$ and $u = (u_i)_{i\geq 1}$. By the law of mass action, the cluster concentrations are ruled by the following equations

1. As in the first part of this thesis, we call *monomers* the clusters of size one.

$$\begin{cases} \dot{u}_i = J_{i-1}(u) - J_i(u), & i > 1 \\ \dot{u}_1 = -J_1(u) - \sum_{i=1}^{\infty} J_i(u), \end{cases} \tag{3.1}$$

where, $J_i(u) = a_i u_1 u_i - b_{i+1} u_{i+1}$, and $a_i, b_i$ are positive kinetic coefficients.

Furthermore, in this model the mass of the system is conserved over time. Formally it yields

$$M_{\text{monomer}} \sum_{i=1}^{\infty} i u_i(t) = \tilde{\rho} = M_{\text{monomer}}\, \rho,$$

where $M_{\text{monomer}}$ is the mass of one monomer. Therefore, the density is a conserved quantity for the solutions.

A complete study of the Becker-Döring system, for finite total mass and positive initial condition is provided by Ball, Carr and Penrose in [11] (1986) and [10] (1988). In particular, an existence and uniqueness theorem for the solution of the Becker-Döring system has been demonstrated.

**Theorem 1** *i) Consider the problem* (3.1) *with initial data* $u_{|t=0} = u^0$ *verifying*

$$\sum_{i=1}^{\infty} i u_i^0 = \rho < \infty.$$

*If the coefficients satisfy* $a_i, b_i = \mathcal{O}(\sqrt{i})$*, then there exists one and only one solution.*

*ii) Let* $c$ *be a solution of* (3.1)*, and* $\phi_i$ *be a nonnegative sequence satisfying*

$$\begin{cases} \int_{t_1}^{t_2} \sum_{i=1}^{\infty} |\phi_{i+1} - \phi_i| a_i u_i(t) \mathrm{d}t < \infty, \\ \sup_t \sum_{i=1}^{\infty} \phi_i u_i(t) < \infty, \\ \phi_{i+1} - \phi_i \geq 0 \text{ for } i \text{ big enough.} \end{cases}$$

*Then for any* $m \geq 2$ *the following relation holds true*

$$\sum_{i=m}^{\infty} \phi_i u_i(t_2) + \int_{t_1}^{t_2} \sum_{i=m}^{\infty} (\phi_{i+1} - \phi_i) b_{i+1} u_{i+1} \, \mathrm{d}s =$$
$$\sum_{i=m}^{\infty} \phi_i u_i(t_1) + \int_{t_1}^{t_2} \sum_{i=m}^{\infty} (\phi_{i+1} - \phi_i) a_i + u_i u_1 \, \mathrm{d}s + \int_{t_1}^{t_2} \phi_m (a_{m-1} u_1 u_{m-1} - b_m u_m) \, \mathrm{d}s.$$

We notice that the second point of the theorem is useful to study the asymptotic behaviour of the solution $u$ and its moments. In fact, when the kinetic coefficients are bounded, the hypothesis is satisfied and we can take $\phi_i = 1$ or $\phi_i = i^\alpha$ for $\alpha \geq 1$, see [11].

To conclude this overview of Becker-Döring models, we cite the work of Simha in 1941, who modelled the fragmentation of long-chain polymers [156]. The Becker-Döring theory has been extended to the *discrete coagulation-fragmentation model* formulated by Spouge in 1984 [167]

as follows

$$\begin{cases} \dot{u}_i &= \frac{1}{2}\sum_{s=1}^{i-1} a_{s,i-s}u_s u_{i-s} - u_i \sum_{s=1}^{\infty} a_{i,s}u_s + \sum_{s=i+1}^{\infty} b_{s,i}u_s - \frac{u_i}{i}\sum_{s=1}^{i-1} s b_{i,s}\ , \quad i>1 \\ \dot{u}_1 &= -u_1 \sum_{s=1}^{\infty} a_{1,s}u_s + \sum_{s=2}^{\infty} b_{s,1}u_s. \end{cases} \tag{3.2}$$

This set of ODEs models the coagulation of two clusters – of size $i$ and $s$ respectively – forming an $i+s$-cluster at rates

$$\begin{array}{ll} a_{i,s}u_i u_s & \text{if } i \neq s, \quad \text{with } a_{i,s} = a_{s,i}, \\ \frac{1}{2}a_{i,i}u_i^2 & \text{if } i = s \end{array}$$

and the fragmentation of clusters without mass loss. If we consider a binary fragmentation, each cluster of size $s$ is decomposed into two clusters of sizes $i$ and $s-i$ at rates $b_{s,i}u_s$. Consequently, the fragmentation coefficients are such that $b_{i,s} = b_{i,i-s}$.

In the discrete coagulation-fragmentation models, depending on the definition of the kinetic rates, the mass conservation can break down in finite time. This phenomenon is known as *gelation*, for more details we refer to [27, 110, 88].

We point out that we retrieve the Becker-Döring system, by setting the parameters to

$$\begin{array}{lll} a_i = a_{i,1} = a_{1,i} & b_{i+1} = b_{i+1,1} = b_{i+1,i} & i>1 \\ 2a_1 = a_{1,1}, & 2b_2 = b_{2,1}, & \\ a_{i,s} = 0, & b_{i,s} = 0 & \text{otherwise.} \end{array}$$

For other models of coagulation-fragmentation we cite the works of Friedlander in 1960 [72], Binder in 1977 [25] and the book by Drake [60]. For more results about the Becker-Döring theory, we recall the studies [38, 94, 124, 132, 133, 176, 171].

## 3.1 Lifshitz-Slyozov theory

The Lifshitz-Slyozov model describes, in continuous time, the removal or the addition of monomers to the clusters. In this model, the cluster size is a continuous variable $x \in \mathbb{R}_+$ because the monomer size is assumed to be infinitesimally small compared to the cluster size. It was originally designed to model the formation of a new phase in solid solution. The system proposed in [111] consists of a transport equation coupled with an integro-differential equation as follows

$$\begin{cases} \frac{\partial u}{\partial t}(x,t) + \frac{\partial}{\partial x}\Big(\big(a(x)v(t) - b(x)\big)u(x,t)\Big) = 0, & x \geq 0,\ t \geq 0, \\ v(t) + \int_0^{\infty} x u(x,t)\mathrm{d}x = \rho > 0, & t \geq 0 \\ v(0) = v_0, & \\ u(x,0) = u_0(x), & \end{cases} \tag{3.3}$$

where $u(x,t) \geq 0$ denotes the concentration of aggregates of size $x$ at time $t$ and $v$ is the monomer concentration. The second line of this system comes straightforward by the conservation of total mass and hence the conservation of the total concentration. In fact,

the quantity $\int_0^\ell xu(x,t)\mathrm{d}x$ at the right hand side of the equation may be interpreted as the concentration of monomers in the polymerised form. For every size $x$, the concentration of monomers in aggregates of size $x$ is given by the concentration $u(x,t)$ times the number of monomers in each aggregate, namely $x$. In conclusion, we can naturally read the second line as the sum of isolated monomer concentration and polymerised monomer concentration, which is the total mass concentration.

The first equation in System (3.3) is a transport equation. The positive functions $a$ and $b$ are associated to the kinetic rates at which the aggregates take or lose monomers, respectively. Therefore, the quantity $a(x)v(t) - b(x)$ corresponds to the growth rate of clusters of size $x$ at time $t$.

The expressions of $a$ and $b$ depend on the polymer size and the mechanism of exchange during the reactions, more details are provided in the review [161]. For instance, in the original work of Lifshitz and Slyozov [111], the authors assume that the mass transfer is driven by monomer diffusion, hence they obtain the coefficients

$$a(x) = x^{\frac{1}{3}}, \qquad b(x) = 1.$$

The Lifshitz-Slyozov system has been used to investigate Ostwald ripening, a phenomenon commonly described as "*large grains are growing at the expense of smaller ones*".
Starting by analysing the size of a single cluster in a bath of monomers, they conclude that the evolution of an $x$-cluster is determined by the ratio between the monomers concentration $v(t)$ and an equilibrium concentration $v_{\mathrm{eq}}(x)$, characterised by the size $x$. Generally, $v_{\mathrm{eq}}(x)$ is a decreasing function of the size. Let us assume that there exists a unique critical size $x_{v(t)}$, which splits the size domain into

$$\begin{aligned} a(x)v(t) - b(x) < 0, \quad &\text{for} \quad 0 < x < x_{v(t)}, \\ a(x)v(t) - b(x) > 0, \quad &\text{for} \quad x > x_{v(t)}. \end{aligned}$$

Then, if $v(t) < v_{\mathrm{eq}}(x)$, the cluster of size $x$ shrinks. Otherwise, the cluster of size $x$ expands. Consequently, there is an energetic advantage making the small grains dissolve and transfer their mass to the large clusters.

The well-posedness of the Lifshitz-Slyozov system is studied by Collet and Goudon in [46]. We present here the existence and uniqueness theorem formulated in their paper.

**Theorem 2**
*Assume that $a, b$ are $C^1$ functions on $[0, +\infty[$ satisfying*

$$\begin{cases} a(x) \geq 0, \quad b(x) \geq 0, \\ a(0)\rho - b(0) \leq 0, \\ |a'(x)| + |b'(x)| \leq K. \end{cases}$$

*Let the initial data $u_0$ be nonnegative and satisfy*

$$\int_0^\infty u_0(x)\mathrm{d}x < \infty, \qquad \int_0^\infty xu_0(x)\mathrm{d}x \leq \rho.$$

*Then System (3.3) has a unique solution*

$$(v, u) \in C^0([0,\tau]) \times C^0([0,\tau], \omega - L^1(\mathbb{R}^+)).$$

The condition $a(0)\rho - b(0) \leq 0$, guarantees that at any time $a(0)v(t) - b(0) \leq 0$. Therefore, the characteristics of the transport equation are directed outside the domain at $x = 0$ and we do not need a supplementary boundary condition at $x = 0$.

Let us consider the evolution of the total cluster concentration

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_0^\infty u(x,t)\mathrm{d}x\right) &= \int_0^\infty \frac{\partial u}{\partial t}(x,t)\mathrm{d}x = -\int_0^\infty \frac{\partial}{\partial x}\Big(\big(a(x)v(t) - b(x)\big)u(x,t)\Big)\mathrm{d}x \\ &= \big(a(0)v(t) - b(0)\big)u(0,t),\end{aligned}$$

where the last equation is true for solution $u$ vanishing at infinity. This assumption is for instance true, when we treat systems with finite total mass $\rho$. We deduce that total cluster concentration increases or remains the same if $a(0)v(t) \geq b(0)$ and decreases otherwise.

Further results can be found in [103, 125, 123, 126, 85, 172].

## 3.2 The Lifshitz-Slyozov system as an asymptotic limit of the Becker-Döring system

In [47], the authors show that the Lifshitz-Slyozov system can be obtained as an asymptotic limit of the Becker-Döring system. The leading idea to demonstrate the asymptotic equivalence is to consider the functions $\{u_i(t)\}_{i>1}$, the solution of the Becker-Döring system, as a discretisation in space of a function $u(x,t)$, that, with a function $v$, solves the Lifshitz-Slyozov system.

In the following, we briefly describe the main steps to get this result. We start by rewriting System (3.1) in a dimensionless form. We rescale every variable by its characteristic value – denoted by capital letters

$$\begin{aligned}&\bar{t} = \frac{t}{T}, \quad \bar{u}_1 = \frac{u_1(\bar{t}T)}{U_1}, \quad \bar{u}_i = \frac{u_i(\bar{t}T)}{U}, \quad \bar{\rho} = \frac{\rho}{M}, \\ &\bar{a}_i = \frac{a_i}{A} \text{ for }, \quad \bar{a}_1 = \frac{a_1}{A_1}, \quad \bar{b}_i = \frac{b_i}{B}, \qquad \text{for } i \geq 2.\end{aligned}$$

The dimensionless form of System (3.1) is then (taking out the overlines)

$$\begin{cases} \dfrac{\mathrm{d}u_i}{\mathrm{d}t} &= \alpha(a_{i-1}u_1u_{i-1} - a_iu_1u_i) + \beta(b_{i+1}u_{i+1} - b_iu_i) \qquad \text{for } i > 2, \\ \dfrac{\mathrm{d}u_2}{\mathrm{d}t} &= \alpha_1 a_1 u_1^2 - \alpha a_2 u_1 u_2 + \beta(b_3u_3 - b_2u_2), \\ \dfrac{\mathrm{d}u_1}{\mathrm{d}t} &= -\gamma\left[2(\alpha_1 a_1 u_1^2 - \beta b_2 u_2) + \sum_{i=2}^\infty (\alpha a_i u_1 u_i - \beta b_{i+1} u_{i+1})\right] \end{cases}$$

and the equation of mass conservation becomes

$$u_1 + \gamma \sum_{i=2}^\infty i u_i = \mu\rho,$$

where

$$\gamma = \frac{U}{U_1}, \quad \mu = \frac{M}{M_{\text{monomer}}U_1}, \quad \alpha = ATU_1, \quad \alpha_1 = \frac{TA_1U_1^2}{U}, \quad \beta = BT.$$

We then introduce a scaling factor $\varepsilon > 0$ and the function $u^\varepsilon(x,t)$ such that $u^\varepsilon$ is piecewise constant on the space grid $\{x_i = i\varepsilon\}$ and it is defined as follows

$$\begin{cases} u^\varepsilon(x,t) &= u_i^\varepsilon(t) \quad \text{for } x \in [x_i, x_{i+1}), \ \ t > 0 \text{ and } i \geq 2, \\ u^\varepsilon(x,t) &= 0 \qquad \text{for } x \in [0, 2\varepsilon), \end{cases}$$

where $u_i^\varepsilon$ is the solution of the dimensionless system with a suitable scaling of the parameters with respect to $\varepsilon$. In particular, the scaling proposed in the article [47] is

$$\gamma = \varepsilon^2, \quad \mu = 1, \quad \alpha = \beta = \frac{1}{\varepsilon}, \quad \alpha_1 \leq \frac{1}{\varepsilon},$$

leading to the system

$$\begin{cases} \dfrac{\mathrm{d}u_i^\varepsilon}{\mathrm{d}t} &= \frac{1}{\varepsilon}(a_{i-1}u_1^\varepsilon u_{i-1}^\varepsilon - a_i u_1^\varepsilon u_i^\varepsilon) + \frac{1}{\varepsilon}(b_{i+1}u_{i+1}^\varepsilon - b_i u_i^\varepsilon) \quad \text{for } i > 2, \\ \dfrac{\mathrm{d}u_2^\varepsilon}{\mathrm{d}t} &= \alpha_1 a_1 (u_1^\varepsilon)^2 - \frac{1}{\varepsilon} a_2 u_1^\varepsilon u_2^\varepsilon + \frac{1}{\varepsilon}(b_3 u_3^\varepsilon - b_2 u_2^\varepsilon), \\ \dfrac{\mathrm{d}u_1^\varepsilon}{\mathrm{d}t} &= -2\varepsilon^2 \alpha_1 a_1 (u_1^\varepsilon)^2 + \varepsilon b_2 u_2^\varepsilon - \varepsilon \textstyle\sum_{i=2}^\infty (a_i u_1^\varepsilon u_i^\varepsilon - b_{i+1} u_{i+1}^\varepsilon) \end{cases}$$

and mass conservation equation

$$u_1^\varepsilon + \varepsilon^2 \sum_{i=2}^\infty i u_i^\varepsilon = \rho.$$

Finally, as $\varepsilon \to 0$, we obtain the Lifshitz-Slyozov system as the following theorem states [47].

**Theorem 3 ([47])**
*Assume the kinetic coefficients $a_i, b_i$ satisfy*

$$a_i, b_i \leq K, \qquad |a_{i+1} - a_i| \leq \frac{K}{i}, \qquad |b_{i+1} - b_i| \leq \frac{K}{i}$$

*for some constant $K$. Then, there exists a subsequence and two functions $a, b \in W^{1,\infty}((0,\infty)) \cap L^\infty(\mathbb{R}^+)$ s.t.*

$$\lim_{\varepsilon \to 0} \sup_{h/\varepsilon < i < H/\varepsilon} (|a_i - a(i\varepsilon)| + |b_i - b(i\varepsilon)|) = 0 \quad \forall\, 0 < h < H < \infty.$$

*Assume that there exist constants $0 < s < 1$, $0 < \rho, M_0, M_s < \infty$ for which $\forall \varepsilon > 0$*

$$\varepsilon \sum_{i=2}^\infty u_i^\varepsilon(0) \leq M_0, \qquad u_1^\varepsilon(0) + \varepsilon^2 \sum_{i=2}^\infty i u_i^{0,\varepsilon} = \rho, \qquad \varepsilon \sum_{i=2}^\infty (\varepsilon i)^{1+s} u_i^\varepsilon(0) \leq M_s.$$

*Then, as $\varepsilon \to 0$, up to a subsequence, we have*

$$\begin{cases} u^\varepsilon \rightharpoonup u, \qquad xu^\varepsilon \rightharpoonup xu \qquad \text{in } C^0([0,T]; \mathcal{M}^1((0,\infty))\text{-weak-}^*), \\ \\ \qquad u_1^\varepsilon(t) \to v(t) \qquad\quad \text{uniformly in } C^0([0,T]), \end{cases}$$

*where $(v,u)$ is the solution of* (3.3).

The space $\mathcal{M}^1(0,\infty)$ is the space of bounded measures on $(0,\infty)$. This space is the dual of the space of continuous functions vanishing at infinity and for $x = 0$, namely $C_0^0((0,\infty))$. The function $u$ is thus such that $u(\cdot,t) \in \mathcal{M}^1(0,\infty)$, see [46].

A modified version of the Lifshitz-Slyozov model is presented in [86], with an analysis of the steady state. The asymptotic limit of the discrete coagulation-fragmentation model in Equations (3.2) has also been investigated in the studies [60, 3, 104].

## 3.3 Prion replication model

We find the first application of aggregation-fragmentation models to the protein polymerisation in [129]. Fragmentation was then taken into account in later studies such as [26, 138, 179] and recently [68, 154].

A first model for *in vivo* prion replication was the model by Nowak *et al.* in 1998 [127] given by the ODE system

$$\begin{cases} \dot{u}_i &= u_1 a_{i-1} u_{i-1} - u_1 a_i u_i - d u_i + \sum_{j=i+1}^{\infty} (b_{i,j} + b_{i,i-j}) u_j - \sum_{j=1}^{i-1} b_{i,j} u_i, \quad i > 1 \\ \dot{u}_1 &= \lambda - \gamma u_1 - \sum_{i=1}^{\infty} a_i u_1 u_i, \end{cases}$$

where, with the notation introduced for the Becker-Döring system, $u_i$ is the concentration of polymers containing $i$ monomers. Each prion polymer of size $i$ gains a monomer at reaction rate $a_i$. Polymers can split into smaller aggregates. Any polymer of size $i$ can break into two pieces of sizes $j$ and $i-j$ at rate $b_{i,j}$ or degrade at rate $d$. Monomeric $\mathrm{PrP}^{\mathrm{sc}}$ is produced at rate $\lambda$ and metabolically removed at rate $\gamma$.

One year later Masel *et al.* [116] proposed the more general model

$$\begin{cases} \dot{u}_i &= u_1 a_{i-1} u_{i-1} - u_1 a_i u_i - b_i u_i + 2 \sum_{j=i+1}^{\infty} b_j u_j k_{i,j}, \qquad i \geq i_0 \\ \dot{u}_1 &= \lambda - \gamma u_1 - \sum_{i=1}^{\infty} a_i u_1 u_i + 2 \sum_{j \geq i_0} \sum_{i < i_0} i k_{i,j} b_j u_j, \end{cases} \tag{3.4}$$

in which, in contrast to the Nowak model, the polymers of size $i$ can break at rate $b_i$ into two pieces of sizes $j$ and $i-j$ at probability $k_{j,i} = k_{i-j,i}$. Furthermore, there exists a minimal size $i_0$ such that all polymers with fewer than $i_0$ monomers disintegrate instantaneously into monomers.

Greer *et al.* formulated a continuous model to describe a population of polymers evolving by nucleation, polymerisation, fragmentation and degradation of monomers as well as production of monomers by the cell [78]. The model reads as follows

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) + \dfrac{\partial}{\partial x}\big(a(x)v(t)u(x,t)\big) + b(x)u(x,t) = 2 \displaystyle\int_x^{\infty} k(x,y)b(y)u(y,t)\mathrm{d}y, \quad x \geq x_0 \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} = \lambda - \gamma v - v \displaystyle\int_{x_0}^{\infty} a(x)u(x,t)\mathrm{d}x + 2 \int_0^{x_0} \int_{x_0}^{\infty} x k(x,y) b(y) u(y,t) \mathrm{d}y \mathrm{d}x \\ a(x_0)u(x_0,t) = 0. \end{cases} \tag{3.5}$$

Several authors analysed this model. For instance, we recall the studies [65] or [146] in which – under the assumption of size-independent kinetic rates – the authors proposed a model of three differential equations : one for monomers $v$, one for the total number of polymers $U(t) = \displaystyle\int_{x_0}^{\infty} u(x,t)\mathrm{d}x$ and one for the total number of polymerised monomers $P(t) = \displaystyle\int_{x_0}^{\infty} x u(x,t)\mathrm{d}x$. In [37, 36], the authors analyse the case of size-dependent kinetic rates. In [58] it has been proved that, under assumptions on the coefficients, Greer's model (3.5) can be obtained as an asymptotic limit of the Masel model (3.4). Following a similar approach to the one described in the previous section, a scaling factor $\varepsilon$ is introduced. The authors proposed the choice

$$\varepsilon = \frac{1}{\langle i \rangle},$$

with $\langle i \rangle$ being the average polymer length. Therefore, the condition $\varepsilon \to 0$ corresponds to $\langle i \rangle \to \infty$.

**Remark 3.3.0.1**
The continuous models are suitable for cases of large polymers like *prion fibrils*.

The Masel model has been extended to a model that accounts for *nucleation*, which is the spontaneous aggregation of monomers into an oligomer structure [139]. For a more complete review of prion models, we refer to [107, 120].

## 3.4 Inverse problem

In our work, we consider the *in vitro* evolution of a population of prion proteins. The *in vitro* condition allows us to simplify the model and consider only two main reactions : the *polymerisation* and the *depolymerisation*. Under such assumptions, the Lifshitz-Slyozov model [111] or the Becker-Döring model [19] can describe, in a continuous or discrete way, the behaviour of the system observed.

A set of experiments was performed to study prion behaviour. In these experiments, an initial population of $\mathrm{PrP}^{\mathrm{SC}}$ fibrils evolves in a liquid solution. The system is observed by an SLS device, which measures the intensity of the light scattered by the system and provides information on its average molecular weight. Prion fibrils are structures made up of thousands of monomers and when we consider other kinds of proteins we may have polymers made of millions of monomers. An ODE model requires one differential equation for each polymer size and, for large polymers, results in high computational costs. In the following, we consider the continuous-size formulation. We point out that, contrarily to the construction of the asymptotic limit presented in Section 3.2, we do not consider a dimensionless writing of the system. In this way we can consider the physical order of magnitudes of polymer sizes and directly compare the model solution to the experimental observations on the system. As the total mass of the system is finite and conserved during the *in vitro* experiments, we can assume the polymer sizes to be in the finite interval $[0, \ell]$. We refer to [13] for a discussion and theoretical justification of the use of a continuous size variable rather than a discrete one. In this paper the authors treat large polymers models with efficient numerical schemes, reducing the system dimension with respect to the ODE system. Furthermore, the authors consider continuous models in the physical size range, proposing size discretisation steps $\delta x \geq 1$ and a scaling parameter $\varepsilon = O(\frac{\delta x}{x})$.

The experimental data are recorded at discrete times. The time lapse between two observations is of about 1.6 seconds, while the observation time scale is in hours. We consider a linear interpolation of the data and we treat them as a continuous function of time. We refer to [45] for a more precise analysis of the possible strategies to treat discrete-time data. For more details about the SLS technique we refer to Section 1.1.4 or [55, 165]. We recall that there is no size-scaling in data returned by SLS devices. We thus choose to not rescale the sizes and describe our model on the physical rage of sizes $[0, \ell]$.

In a continuous-size model, the SLS measurements formally read

$$z(t) = \lambda_1 \Big( v(t) + \int_0^\ell x^2 u(x,t) \mathrm{d}x \Big) + \lambda_2 + \chi(t), \tag{3.6}$$

where $\lambda_1 > 0$, $\lambda_2 \in \mathbb{R}$ and $\chi$ is an additive centred white noise.

By the law of mass conservation, the concentration of monomers is comparable to the first moment of the polymer concentration function, namely $\int_0^\ell xu(x,t)\mathrm{d}x$. Denoting as $\langle x \rangle$ the average polymer size, we notice that the moment of order $n$ of the polymer concentration is of the order of magnitude of $\langle x \rangle^{n-1} v$. Since we are not considering a scaling in the sizes, $\langle x \rangle$ may be very large when we study polymers. For these reasons, in our case, we can make the following assumption

$$v \ll \int_0^\ell x^n u(x,t)\mathrm{d}x. \tag{3.7}$$

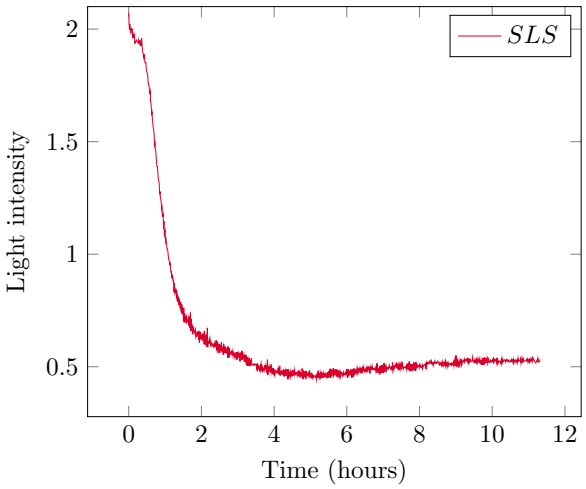



FIGURE 3.1 – Static Light Scattering measurements on PrP fibrils.

Let us now focus on the model. In the experiments of reference, we have only polymers at the beginning. In Figure 3.1, we show an example of the experimental data. We can identify two phases : a first phase in which the average polymer size is mostly decreasing (up to 4 hours), followed by a second phase in which it is mostly increasing. As $v(0) = 0$, the growth rate $a(0)v(0) - b(0)$ is negative. Consequently, the polymerised mass decreases over a certain time domain $[0, \bar{t})$.

Setting opportune experimental conditions – like for instance a low total mass concentration $\rho$ – the aggregation process may be considered a minor process. In these cases, we can take the following assumption

$$a(x)v(t) \ll b(x), \qquad \forall t, \forall x. \tag{3.8}$$

Under these two assumptions we approximate System (3.3) and observations (3.6) by

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) + \dfrac{\partial}{\partial x}\Big(b(x)u(x,t)\Big) = 0, \qquad x \in [0,\ell],\ t \geq 0, \\ u(\ell,t) = 0, \\ u(x,0) = u_0(x) = u_\diamond + \xi, \end{cases} \tag{3.9}$$

$$z(t) = \lambda_1 \int_0^\ell x^n u(x,t)\mathrm{d}x + \lambda_2 + \chi(t), \tag{3.10}$$

Setting $n = 2$, the equation (3.10) corresponds to (3.6). We have chosen to set a general framework and to consider the observations as the moment of order $n \in \mathbb{N}$. For example, it is possible to observe the first moment of $u$ by Thioflavin T (ThT) fluorescence which provides the measurement of the total polymerised mass [18].

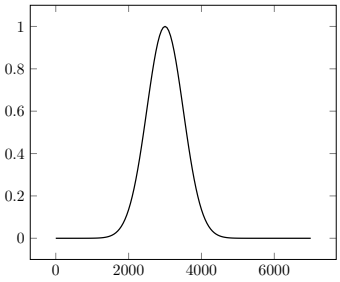
Let us recall that, from the mass conservation law in System (3.3), we have $v(t) = \rho - \int_0^\ell xu(x,t)\mathrm{d}x$. Therefore, the transport equation in System (3.3) results in

$$\frac{\partial u}{\partial t}(x,t) + \frac{\partial}{\partial x}\Big(a(x)\left(\rho - \int_0^\ell xu(x,t)\mathrm{d}x\right) - b(x)u(x,t)\Big) = 0, \quad x \in [0,\ell],\ t \geq 0$$

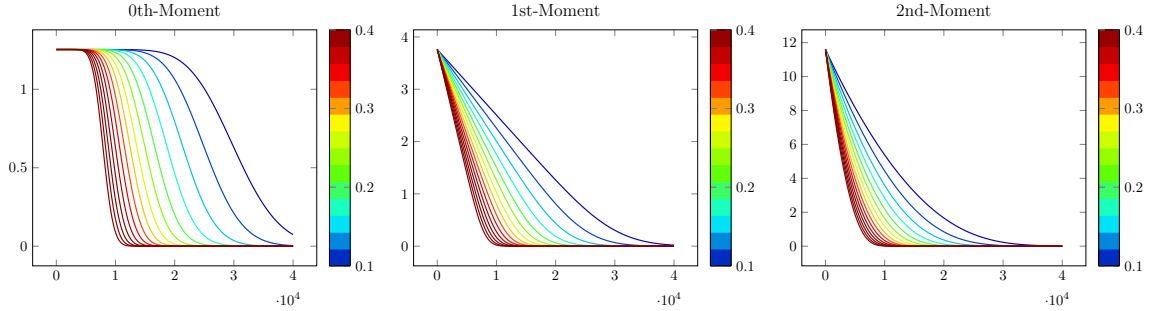
which is non linear in $u$. Assumption (3.8) is particularly useful since it allows us to set up an inverse problem methodology in a simpler linear setting. This study constitutes a first necessary step to, in the future, solve the initial condition estimation problem for the Lifshitz-Slyozov model.

To illustrate the nature of the observations we are going to work with, in Figure 3.2 we present the 0-th, 1st and 2nd moments of the solution of System (3.9) with the Gaussian function in Figure 3.2a, as initial condition $u_0(x)$. Moreover, to highlight the sensitivity of the observations with respect to the transport velocity, we plot the observations associated to the choices of $b(x) = b$ varying between the values 0.1 and 0.4.

In conclusion, we introduce the inverse problem that we tackle in the rest of this work.




(a) initial condition




(b) moment observation

FIGURE 3.2 – (a) Gaussian function $u_0(x) = e^{\frac{(x-3000)^2}{2\cdot 10^6}}$. (b) From left to right the 0-th, 1st and 2nd moments of the solution of System (3.9) with the initial condition $u_0$. In a colormap from blue to red, we show the moments associated with transport rates from 0.1 to 0.4.

**Inverse Problem**

Given the observations

$$z(t) = \lambda_1 \int_0^\ell x^n \breve{u}(x,t)\mathrm{d}x + \lambda_2 + \chi(t),$$

to find

$$\breve{\xi}$$

such that

$$\begin{cases} \dfrac{\partial \breve{u}}{\partial t}(x,t) - \dfrac{\partial}{\partial x}\big(b(x)\breve{u}(x,t)\big) = 0, \qquad x \in [0,\ell],\ t \geq 0, \\ \breve{u}(\ell,t) = 0, \\ \breve{u}(x,0) = u_\diamond + \breve{\xi}, \end{cases} \tag{3.11}$$

In the next chapter, we present two strategies to solve this inverse problem. The first belongs to the class of kernel methods, while the second is a variational data assimilation method called 4d-Var. In particular, the first is designed for the specific case of constant transport velocity. In Chapter 5 we provide a more complete overview of data assimilation methods.

CHAPITRE 4

# Article : Estimation from moments measurements for amyloid depolymerisation


## Abstract

Estimating reaction rates and size distributions of protein polymers is an important step for understanding the mechanisms of protein misfolding and aggregation, a key feature for amyloid diseases. This study aims at setting this framework problem when the experimental measurements consist in the time-dynamics of a moment of the population (*i.e.* for instance the total polymerised mass, as in Thioflavine T measurements, or the second moment measured by Static Light Scattering). We propose a general methodology, and we solve the problem theoretically and numerically in the case of a depolymerising system. We then apply our method to experimental data of depolymerising oligomers, and conclude that smaller aggregates of ovPrP protein should be more stable than larger ones. This has an important biological implication, since it is commonly admitted that small oligomers constitute the most cytotoxic species during prion misfolding process.

# Estimation from Moments Measurements for Amyloid Depolymerisation

Aurora Armiento * Marie Doumic † Philippe Moireau ‡ H. Rezaei §

February 23, 2016

**Abstract**

Estimating reaction rates and size distributions of protein polymers is an important step for understanding the mechanisms of protein misfolding and aggregation, a key feature for amyloid diseases. This study aims at setting this framework problem when the experimental measurements consist in the time-dynamics of a moment of the population (*i.e.* for instance the total polymerised mass, as in Thioflavine T measurements, or the second moment measured by Static Light Scattering). We propose a general methodology, and we solve the problem theoretically and numerically in the case of a depolymerising system. We then apply our method to experimental data of depolymerising oligomers, and conclude that smaller aggregates of ovPrP protein should be more stable than larger ones. This has an important biological implication, since it is commonly admitted that small oligomers constitute the most cytotoxic species during prion misfolding process.

**Keywords:** Amyloid, prion, protein stability, oligomer, transport equation, state estimation, inverse problem, data assimilation

## Introduction

Protein aggregation is a key feature of a large range of diseases, called *amyloid* diseases, among which we can quote Alzheimer's, Parkinson's, Huntington's, transmissible spongiform encephalopaties (or prion diseases - *e.g.* Creutzfeldt–Jakob's, Kuru, bovine spongiform encephalopathy/madcow), etc [20, 23].

This category of diseases takes its name from the protein fibrils, called *amyloids*, which are formed during the disease and accumulate into the tissue. Their formation arise from

* Univ Paris Diderot, Sorbonne Paris Cité, Lab. J.L. Lions, UMR CNRS 7598, Inria , Paris, France
†Sorbonne Universités, Inria, UPMC Univ Paris 06, Lab. J.L. Lions UMR CNRS 7598, Paris, France
‡Inria and Université Paris-Saclay, Campus de l'Ecole Polytechnique, 91128 Palaiseau, France
§Virologie et Immunologie Moléculaires, Institut National de la Recherche Agronomique, F-78352 Jouy-en-Josas, France

misfolded versions of proteins present naturally in the body, each disease having its specific precursor protein (*e.g.* APP for Alzheimer's, PrP for Prion, $\beta_2$m for haemodialysis-associated amyloidosis). While their accumulation in organs is characteristic for the disease, the reason for their association as well as their role in tissue damages are still unclear. Moreover, their aggregation mechanisms - most probably specific for each protein involved - are at the moment largely unknown.

The main reasons for so many open questions to remain, despite both the longstanding interest raised in the biological, biophysical and biochemical communities, and the major importance of amyloid diseases for public health, are twofold. First, the number of possible chain-reactions involved is huge, possibly infinite - as the size of aggregates is. Hence model design and discrimination is very complex, and conclusions made on a specific protein are hardly translatable to another one. Second, the most common experimental devices can measure averaged quantities on the polymerised proteins, such as the total polymerised mass (Thioflavine T measurements [5]) or the average size of polymers (Static Light Scattering (SLS) [27]). How such measurements may be used to estimate reaction rates (which may also be an infinity) and size distribution of aggregates, and thus to select the major mechanisms, is an emerging field of inverse problems with few theoretical progress [1] and positive results on experimental data [29, 30].

To contribute to this new field, this article focus on one of the major concerns in pathologies due to protein misassembly and aggregation: the determination of oligomer size distribution. It has been reported that – while amyloid fibrils present low biological activity – oligomers and small assemblies are the cytopathogenic elements [26, 13]. Depending on the type of pathology and the protein involved, oligomers could either be involved into the pathway of amyloid fibrils formation or be associated to an independent pathway, which only leads to the formation of oligomers. Oligomer size characterisation can play a key role in distinguishing between these pathways. Therefore, the investigation on size distribution remains the first step to understand how oligomers are formed, their biological activity and their biophysical characterisation to finally design therapeutic strategies.

This question - how to estimate size distributions - leads us to setting a framework problem and studying it, both theoretically and numerically, in one of its simplest possible version. We then apply our method to experimental data, using the time-dependent average size of polymers (measured by SLS) to reconstruct the oligomer initial size distribution. We compare our estimation to the experimental estimation obtained by chromatography and discuss the implications of our results. Eventually, we discuss the new problems and possibilities opened-up by these results, and how this methodology could easily be adapted to other models and experiments.

## Mathematical Setting

Since protein aggregates can reach extremely large average sizes, we adopt here a continuous framework [22] and denote $x \in (0, \infty)$ the *size* of an aggregate, *i.e.* $x$ represents the (rescaled) quantity of monomers contained in a given polymer. We thus call $u(x, t)$ the concentration of polymers of size $x$ at time $t$ (see [2] for a discussion and theoretical

justification of the use of a continuous size variable rather than a discrete one).

One of the techniques most widely used is the measurement of Thioflavin T (ThT) fluorescence, [5], which provides measurements of the total polymerised mass, *i.e* a linear transformation of the first moment of the concentration function

$$z_{\text{tht}}(t) = c_1 \int_0^\infty xu(x,t)dx.$$

The Static Light Scattering (SLS) technique, [27], could give us an affine transformation of the second moment

$$z_{\text{sls}}(t) = c_1 \int_0^\infty x^2 u(x,t)dx + c_2,$$

where $c_1 \geq 0,\ c_2 \in \mathbb{R}$.
The framework problem we want to contribute stands: Under which assumptions (and limitations) is it possible to estimate the reaction rates and/or the initial size distribution, from a time measurement $z_{\text{tht}}(t)$ or $z_{\text{sls}}(t)$?

As a first simplifying assumption, we model the primary reactions involved in the evolution of polymers with the Lifshitz-Slyozov system, that is one of the most common polymerising/depolymerising model. In this system, polymers (or clusters, in another application context) can only grow by monomer addition, with a size-dependent reaction rate $a(x)$, and depolymerise by monomer loss, with a reaction rate $b(x)$. This results in the following system

$$\begin{cases} \dfrac{\partial}{\partial t}u(x,t) + \dfrac{\partial}{\partial x}\Big(\big(a(x)v(t) - b(x)\big)u(x,t)\Big) = 0, \qquad x \in [0,\ell],\ t \geq 0, \\ u(\ell,t) = 0, \\ u(x,0) = u_0(x), \end{cases} \tag{1}$$

where $\ell \in (0,\infty]$ is the upper bound of polymer sizes. We assume here $\ell < \infty$, in contrast with the initial Lifshitz-Slyozov model [14]. The function $v(t)$ is the concentration of monomers in the cuvette and is directly related to polymer concentration from the following mass conservation law

$$v(t) + \int_0^\ell xu(x,t)\mathrm{d}x = v(0) + \int_0^\ell xu_0(x)\mathrm{d}x > 0 \quad \forall\, t \geq 0. \tag{2}$$

When applied to amyloid formation, this model may be seen as a qualitative model taking into account what biologists call *primary pathway* and neglecting, as a first approach, *secondary pathways* such as fragmentation or coalescence [8]. Note that there are many other possible applications of this model, such as phase transition, which was the original application for which it had been designed [14].

The problem now stands: Measuring $z_{\text{sls}}(t)$ or $z_{\text{tht}}(t)$, or more generally the time-dependence of a $n$-th moment defined by $\int_0^\ell x^n u(x,t)dx$, with $u(x,t)$ solution of System (1)(2), what may be possibly estimated among the unknown quantities, *i.e.* the initial state $u_0(x)$ and the parameter functions $a(x)$ and $b(x)$?

This problem in its full generality is both nonlinear and highly ill-posed. Hence, we proceed to further simplifications and study the state estimation of a model of pure depolarisation. Assuming to start with no monomers, *i.e.* $v(0) = 0$, we can neglect the polymerisation term, at least during the beginning of the reaction – see Figure 12 for measurements of such an experiment. The model then becomes

$$\begin{cases} \dfrac{\partial}{\partial t}u(x,t) - \dfrac{\partial}{\partial x}\Big(b(x)u(x,t)\Big) = 0, \qquad x \in [0,\ell],\ t \geq 0, \\ u(\ell,t) = 0, \\ u(x,0) = u_0(x). \end{cases} \tag{3}$$

The state estimation problem may be formulated as follows:

(IP) How to estimate $u_0$ – the initial condition of System (3) – from the given *a priori* knowledge of $b(x)$ and measurement $\int_0^\ell x^n u(x,t)dx$?

In order to settle a general framework, easy to adapt to more complex problems in the future, we introduce below the notations for the standard state-space formalism used for dynamical systems.

### State space formalism

We introduce the state space $\mathcal{U} = \boldsymbol{L}^2([0,\ell])$ equipped with its natural norm and introduce the state variable $u$ standing for the function

$$u : \left| \begin{array}{ccc} [0,\tau] & \longrightarrow & \mathcal{U} \\ t & \longmapsto & u(t) : \left| \begin{array}{ccc} [0,\ell] & \longrightarrow & \mathbb{R} \\ x & \longmapsto & u(x,t). \end{array} \right. \end{array} \right.$$

Then we rewrite System (3) in the state-space form

$$\begin{cases} \dfrac{du}{dt} = Au, \\ u(0) = u_0, \end{cases} \tag{4}$$

where $A$ is the linear functional operator – called model operator –

$$A : \left| \begin{array}{ccc} \mathcal{D}(A) \subset \boldsymbol{L}^2([0,\ell]) & \longrightarrow & \boldsymbol{L}^2([0,\ell]), \\ f & \longmapsto & \partial_x(bf), \end{array} \right.$$

of domain

$$\mathcal{D}(A) = \{f \in \boldsymbol{H}^1([0,\ell]) \mid f(\ell) = 0\}.$$

Assuming $b' \in \boldsymbol{L}^\infty([0,\ell])$, we easily prove that there exists $\lambda$ such that the operator $A - \lambda Id$ is dissipative, hence $A$ is the generator of a strongly continuous semigroup $\mathbb{T}_t$ – see for instance [3] for an introduction to such concepts.

We formalise our measurement procedure by introducing the observation space $\mathcal{Z} = \mathbb{R}$ and a so-called observation operator associated in our case to the $n$-th momentum of a given state variable

$$C : \left| \begin{array}{ccc} \mathcal{U} & \longrightarrow & \mathcal{Z}, \\ u & \longmapsto & \int_0^\ell x^n u(x) \mathrm{d}x, \end{array} \right. \tag{5}$$

which is a time-invariant linear bounded operator with $\|C\| \leq \ell^{n+\frac{1}{2}}$. In the following sections ,we will use the notation $C^n$ when we need to stress the dependence on the order of the moment.

Note that the observation operator is defined independently of the model $A$. However – by taking into account the model dynamics – we easily write the relation between observations and the initial condition of polymer concentration. To do so, we introduce the operator $\Psi_\tau \in \mathcal{L}(\mathcal{U}, \boldsymbol{L}^2([0,\tau], \mathcal{Z}))$

$$\Psi_\tau : \left| \begin{array}{ccc} \mathcal{U} & \longrightarrow & \boldsymbol{L}^2([0,\tau]), \\ u_0 & \longmapsto & C\mathbb{T}_\tau u_0, \end{array} \right.$$

since in our case $\boldsymbol{L}^2([0,\tau], \mathcal{Z}) = \boldsymbol{L}^2([0,\tau])$.

Let us now denote by $z(t)$ the observations at our disposal. We can say that the observations are related to a target solution $\breve{u}$ of System (3) up to some measurement errors – *i.e.* observation noise – $\chi$. Formally, we have

$$z = C\breve{u} + \chi.$$

Using the various operators introduced, we formulate our inverse problem in two equivalent forms. In the more classical inverse-problem formulation, our objective appears as

| |
|---|
| Inverting $\Psi_\tau$ to reconstruct $\breve{u}_0$ from the given measurement $z$ generated through time $t \in [0,\tau]$. |

In a more data assimilation form we aim at

| |
|---|
| Estimating $\breve{u}_0$ from given measurements $z$ generated through time $t \in [0,\tau]$, knowing the model dynamics $A$ and the model of observation operator $C$. |

In Section 1, we consider the specific case where the depolymerisation rate $b(x)$ is constant: we show that the problem is equivalent to the estimation of the $(n+1)$-th derivative of the measurement, so that we can use (for instance) a kernel regularisation method for which we recall the standard convergence results. This gives us some light on what we could expect for convergence in more general cases. In Section 2, we turn to the variational formulation, recall its intrinsic links with the previous regularisation method, and extend it to non constant $b(x)$. We then illustrate our results by numerical simulations in Section 3. We apply our method, together with a statistical study for the measurement noise, to analyse the experimental data in Section 4. All this exploratory study leads us to sketch perspectives for future work and open problems.

# 1 First Approach: kernel regularisation

In this section we assume to have a constant depolymerisation rate $b(x) = b > 0$. We know that in this case, the solution of System (3) is given by $u(x,t) = u_0(x+bt)$. We have, by a simple change of variable, that

$$\forall t > 0, \quad C^n u(t) = \int_0^\ell x^n u(x,t)dx = \int_{bt}^\ell (x'-bt)^n u_0(x')\mathrm{d}x',$$

and therefore

$$\Psi_\tau^n : \left| \begin{array}{ccc} \boldsymbol{L}^2([0,\ell]) & \longrightarrow & \boldsymbol{L}^2([0,\tau]), \\ u_0 & \longmapsto & \left(t \to \displaystyle\int_{bt}^\ell (x-bt)^n u_0(x)\mathrm{d}x\right), \end{array} \right. \tag{6}$$

where $n$ can be avoided when not necessary. We easily see that

$$\mathrm{Ran}\Psi_\tau^n = \left\{ u \in \boldsymbol{H}^{n+1}([0,\tau]), u_0(\tau) = \cdots = u_0^{(n)}(\tau) = 0 \right\}.$$

Deriving recursively $\Psi_\tau u_0$, we obtain

$$\frac{d^{n+1}}{dt^{n+1}}(\Psi_\tau u_0) = (-b)^{n+1} n! u_0(bt) \quad \text{for } n \geq 0,$$

so that we have the following explicit formula for $u_0$

$$\boxed{u_0(x) = \frac{1}{n!(-b)^{n+1}} \frac{d^{n+1}}{dt^{n+1}}(\Psi_\tau u_0)\left(\frac{x}{b}\right), \quad \text{for } n > 0.} \tag{7}$$

In the previously seen formalism, we model an additive noise as follows: we call $\varepsilon$ the upper bound for the noise level measured in a Sobolev space $\boldsymbol{W}^{-s,p}([0,\tau])$-norm, and we assume

$$\|\chi\|_{\boldsymbol{W}^{-s,p}([0,\tau])} \leq \varepsilon. \tag{8}$$

The choice for the parameters $s$ and $p$ depends on the kind of noise ($s = 0$ for a deterministic noise, $s = \frac{1}{2}$ and $p = 2$ for a deterministic equivalent of a gaussian white noise [17]). The ill-posedness of the problem comes from the fact that the noisy measurement $z$ is in general not differentiable, so that we cannot use directly Equality (7) to solve our problem. This is a classical linear ill-posed problem of order $\delta_{IP} = n+1$ in the scale $\boldsymbol{W}^{k,p}$, see [12]. Before applying Formula (7), we need to regularise our measurement $z$. A classical regularisation method consists in convolving the measurement with a mollifier sequence, method called kernel density estimation for the statistical problem of estimating the density from an i.i.d. sample [28]. Thanks to classical results, we know that the regularity of the convolution depends on the regularity of both the measurement and the kernel. Let us take a kernel function $\rho \in \mathcal{C}_c^\infty(\mathbb{R})$, such that

$$\int_{\mathbb{R}} \rho(x)dx = 1, \quad \int_{\mathbb{R}} x^k \rho(x)dx = 0, \quad \text{for } 1 \leq k \leq m. \tag{9}$$

We define the family of mollifiers $\rho_\alpha$ by

$$\rho_\alpha = \frac{1}{\alpha}\rho\left(\frac{x}{\alpha}\right), \tag{10}$$

depending on the parameter $\alpha > 0$. Our estimation of the initial condition is carried out by the function

$$\hat{u}_0^{\varepsilon,\alpha} = \frac{d^{n+1}}{dx^{n+1}}\rho_\alpha * \left(\frac{1}{n!(-b)^{n+1}} z\left(\frac{x}{b}\right)\right)$$

where the convolution operator $*$ is defined by $f * g(x) = \int_{\mathbb{R}} g(x')f(x-x')dx'$. The accuracy of our approximation shall depend on the noise level $\varepsilon$, on the regularity of the kernel family, on the parameter $\alpha$ and on the order of the derivative, that is $n+1$. Classically, we obtain an optimal upper bound for the accuracy of the estimation as stated in the following proposition.

**Proposition 1**
*Let $1 \leq p < \infty$, $n \in \mathbb{N}$, $0 \leq s < 1$ and let $\breve{u}_0 \in \boldsymbol{W}^{m+1,p}([0,\ell])$ with $m$ defined as in Equation (9). Let $\Psi_\tau \breve{u}_0 \in \boldsymbol{W}^{m+n+2,p}([0,\tau])$ defined in Equation (6). Let $z \in \boldsymbol{W}^{-s,p}([0,\tau])$ a measurement of the $n$-th momentum $\Psi_\tau \breve{u}_0$ such that $\tau \geq \frac{\ell}{b}$ and*

$$\|z - \Psi_\tau \breve{u}_0\|_{\boldsymbol{W}^{-s,p}([0,\tau])} \leq \varepsilon.$$

*Let us define*

$$\breve{u}_0(x) = \frac{1}{n!(-b)^{n+1}} \frac{d^{n+1}}{dt^{n+1}} \Psi_\tau \breve{u}_0 \left(\frac{x}{b}\right), \tag{11}$$

*Let $\rho$ defined by Equation (9) and $\rho_\alpha$ by Equation (10), with $\alpha \in (0,1)$. We define*

$$\hat{u}_0^{\varepsilon,\alpha} = \frac{d^{n+1}}{dx^{n+1}}\rho_\alpha * \left(\frac{1}{n!(-b)^{n+1}} z\left(\frac{x}{b}\right)\right) \tag{12}$$

*as an approximation of $\breve{u}_0$. Then the following estimation is of optimal order in the sense of [12]*

$$\|\hat{u}_0^{\varepsilon,\alpha} - \breve{u}_0\|_{\boldsymbol{L}^p([0,\ell])} \leq \Theta\left(\frac{\varepsilon}{\alpha^{n+s+1}} + \alpha^{m+1}\right) = F_\varepsilon(\alpha), \tag{13}$$

*where the constant $\Theta$ depends on $\|\Psi_\tau \breve{u}_0\|_{\boldsymbol{W}^{m+n+2,p}([0,\tau])}$, $\|x^{m+1}\rho\|_{\boldsymbol{L}^1(\mathbb{R})}$, $\|\rho^{(n)}\|_{\boldsymbol{L}^1(\mathbb{R})}$, $\|\rho^{(n+1)}\|_{\boldsymbol{L}^1(\mathbb{R})}$.*

For the sake of completeness, the proof of this proposition is recalled in Appendix A. This gives us an *a priori* method to choose the parameter $\alpha$: Aiming at the smallest approximation error – we select the $\alpha$ that minimises $F_\varepsilon(\alpha)$. The *a priori* optimal choice for $\alpha$ is the minimiser of the convex function $F_\varepsilon(\alpha)$

$$\alpha_{opt} = O\left(\varepsilon^{\frac{1}{n+m+2+s}}\right). \tag{14}$$

By this choice, we obtain an estimation $\hat{u}_0^{\varepsilon,\alpha_{\text{opt}}}$ such that

$$\boxed{\|\hat{u}_0^{\varepsilon,\alpha_{\text{opt}}} - \breve{u}_0\|_{\boldsymbol{L}^p([0,\ell])} = O\left(\varepsilon^{\frac{m+1}{n+m+2+s}}\right).} \tag{15}$$

In the case of a variable depolymerisation rate $b(x)$, computations are not so easy and in general we do not have such an explicit relation between measurements and initial condition. This is part of the reasons why we now turn to data assimilation approaches.

## 2 Second Approach: a data assimilation variational approach

In this section, we propose to base our inverse problem solving strategy on the so-called 4d-Var approach as named by [15]. The principle consists in minimising – hence the variational designation – with respect to the initial condition a least-square criterion $\mathcal{J}$ combining the discrepancy between the actual data and the simulation, with additional regularisation terms accounting for the confidence in the model.

The advantage of this method lies in its very general formalism that leads to high flexibility in the choice of the model operator or the observation operator.

Typically, we decompose $\breve{u}_0$ as the sum of a known *a priori* $u_\circ$, and an unknown variation $\breve{\xi}$ representing the uncertain part of our initial concentration

$$\breve{u}_0 = u_\circ + \breve{\xi}. \tag{16}$$

As $\breve{\xi}$ is unknown, the trajectory $\{\breve{u}(t),\ t \in [0,\tau]\}$ cannot be obtained directly. However, we can parametrise the dynamics (4) with respect to any guess $\xi$ of $\breve{\xi}$. We denote by $\{u_{|\xi}(t),\ t \in [0,\tau]\}$ the resulting state trajectory knowing the guess $\xi$

$$\begin{cases} \dot{u}_{|\xi} = A u_{|\xi} \\ u_{|\xi}(0) = u_\circ + \xi. \end{cases} \tag{17}$$

We then write the criterion to minimise

$$\mathcal{J}_\tau(\xi) = \frac{1}{2}\langle \xi, P_0^{-1}\xi\rangle_{\mathcal{U}}^2 + \frac{1}{2}\int_0^\tau \gamma |z - C(u_{|\xi})|^2 dt. \tag{18}$$

The isomorphism on $\mathcal{U}$, namely $P_0$, and the scalar $\gamma$ are weights on the natural norm on $\mathcal{U}$ and $\mathcal{Z}$, respectively. These weights are defined in accordance with the level of confidence into our *a priori* on the initial condition and the *measurement* – typically based on an *a priori* evaluation of the noise $\chi$. Note that contrarily to the kernel regularisation method, if the space for the noise is less regular than $L^2$, this method cannot be used directly: prior regularisation on the measurement is needed. On the contrary, this method provides a unique minimiser for general rates $b(x)$ as soon as the direct problem is well-posed.

Our objective is to minimise $\mathcal{J}_\tau$ under the constraint of the model dynamics (17). We thus introduce the so-called *adjoint variable* $q_{|\xi,\tau}$ as the Lagrange multiplier associated

with the dynamical constraint (17). The adjoint variable is then solution – see for instance [7] – of the dynamics

$$\begin{cases} \dot{q}_{|\xi,\tau} + A^* q_{|\xi,\tau} = -\gamma C^* \left( z - C u_{|\xi} \right), \quad t \in [0,\tau] \\ q_{|\xi,\tau}(\tau) = 0, \end{cases} \tag{19}$$

where $A^*$ is the adjoint of the model operator defined by

$$\begin{array}{rccc} A^* : & \mathcal{D}(A^*) \subset \boldsymbol{L}^2([0,\ell]) & \longrightarrow & \boldsymbol{L}^2([0,\ell]) \\ & f & \longmapsto & -b(x)\partial_x f \end{array}$$

with domain

$$\mathcal{D}(A^*) = \{ f \in \boldsymbol{H}^1([0,\ell]) \mid f(0) = 0 \}$$

and $C^*$ is the adjoint of the observation operator $C$ defined by (5), hence

$$\begin{array}{rccc} C^* : & \mathbb{R} & \longrightarrow & \boldsymbol{L}^2([0,\ell]) \\ & r & \longmapsto & f_r : x \mapsto x^n r. \end{array}$$

Therefore, the adjoint system reads in strong formulation

$$\begin{cases} \dfrac{\partial}{\partial t} q_{|\xi,\tau}(x,t) - b(x)\partial_x q_{|\xi,\tau}(x,t) \\ \qquad\qquad = -\gamma x^n \Big( z - \displaystyle\int_0^\ell x'^n \bar{u}(x',t) dx' \Big), \qquad x \in [0,\ell],\ t \in [0,\tau] \\ q_{|\xi,\tau}(0,t) = 0, \\ q_{|\xi,\tau}(x,\tau) = 0. \end{cases} \tag{20}$$

Using the adjoint variable, a standard computation allows to characterise $\bar{\xi}_{|\tau} = \arg\min_\xi \mathcal{J}_\tau$ as

$$\bar{\xi}_{|\tau} = P_0 \bar{q}_{|\tau}(0),$$

where $\bar{q}_{|\tau}$ is the adjoint variable associated to the $\bar{u}_{|\tau} = u_{|\bar{\xi}_{|\tau}}$, hence leading to a famous both-end problem formulation [7]

$$\begin{cases} \dot{\bar{u}}_{|\tau} = A \bar{u}_{|\tau}, & t \in [0,\tau] \\ \dot{\bar{q}}_{|\tau} + A^* \bar{q}_{|\tau} = -\gamma C^* \left( z - C \bar{u}_{|\tau} \right), & t \in [0,\tau] \\ \bar{u}_{|\tau}(0) = u_\circ + P_0 \bar{q}_{|\tau}(0), \\ \bar{q}_{|\tau}(\tau) = 0. \end{cases} \tag{21}$$

## 2.1 Equivalence with the kernel regularisation method

According to a classical interpretation, we can read the second term of the criterion as the ordinary least-square data fitting term, while the first term is often considered as a regularisation term by choosing

$$P_0 = \frac{1}{\beta} \mathrm{Id},$$

with $\beta$ small enough so that

$$\beta\|\bar{\xi}_\tau\|^2 \ll \gamma \int_0^\tau \|\chi\|^2 \, dt.$$

Minimising the criterion $\mathcal{J}_\tau$ is then equivalent to minimise for $\alpha^2 = \frac{\beta}{\gamma}$

$$\min_{\xi} \left\{ \alpha^2 \|\xi\|^2_{\boldsymbol{L}^2([0,\ell])} + \|z(t) - \Psi_\tau(\xi)\|^2_{\boldsymbol{L}^2([0,\tau])} \right\},$$

where clearly appears the classical Tikhonov regularisation.

Moreover, we can consider different criteria by changing the state space $\mathcal{U}$ or considering different $P_0$. For instance, when choosing

$$P_0 = \frac{1}{\beta}\mathrm{Id}, \quad \mathcal{U} = \boldsymbol{H}^s([0,\ell]),$$

the variational method is equivalent to the *generalized* Tikhonov method where we minimise

$$\min_{\xi} \left\{ \alpha \|\xi\|^2_{\boldsymbol{H}^s([0,\ell])} + \|z(t) - \Psi_\tau(\xi)\|^2_{\boldsymbol{L}^2([0,\tau])} \right\}.$$

Note that $s > -\frac{n+1}{2}$ is necessary for this minimisation to be a regularising method - see for instance the analysis of Tikhonov's regularisation in Hilbert scales in [6], and below the comments on the observability condition.

To give some insight into the links between the two regularisation methods, let us take the case of classical Tikhonov regularisation, $b$ constant, $u_\circ = 0$ with $\tau \geq \frac{\ell}{b}$.

We recall that $\Psi_\tau(\xi)(t) = \int_{bt}^{\ell} (y - bt)^n \xi(y) dy$. The operator $\Psi_\tau$ is injective, compact, and with dense image when taken from $L^2([0,\ell])$ to $L^2([0,\tau])$ with $\tau = \ell/b$. Its adjoint operator is

$$\Psi^*_\tau(v)(x) = \int_0^{\frac{y}{b}} (y - bt)^n v(t) dt.$$

This provides us with the following result.

**Proposition 2**
*For any $z \in L^2([0,\tau])$, there exists a unique minimiser $\bar{\xi}$ for $J(\xi)$ defined by*

$$J(\xi) = \frac{\alpha^2}{2}\|\xi\|^2_{L^2([0,\ell])} + \frac{1}{2}\int_0^\tau |z(t) - \Psi_\tau(\xi)|^2 dt,$$

*and $\bar{\xi} \in \boldsymbol{H}^{n+1}([0,\ell])$. If moreover $\breve{\xi} \in \boldsymbol{H}^{n+1}([0,\ell])$ with $\breve{\xi}(0) = \cdots = \breve{\xi}^{(n)}(0) = 0$, we have the following estimate*

$$\|\breve{\xi} - \bar{\xi}\|_{\boldsymbol{L}^2([0,\ell])} \leq \frac{1}{\alpha}\|z - \Psi_\tau(\breve{\xi})\|_{\boldsymbol{L}^2([0,\tau])} + \alpha\|\breve{\xi}\|_{\boldsymbol{H}^{n+1}([0,\ell])}.$$

We recognise here the case $s = 0$ and $n = m$ of Proposition 1, by denoting $\alpha = \tilde{\alpha}^{n+1}$. For the sake of completeness, we sketch out the proof in Appendix A.

## 2.2 Observability Condition

In data assimilation, the well-posedness or ill-posedness of the inverse problems is characterised by a so-called observability condition translating that there is enough information in the data to reconstruct the initial condition. Typically, this condition is of the form

There exists a time $\tau_0$ and a constant $\Theta > 0$ such that, for all $u \in C\big([0,\tau],\mathcal{U}\big)$ solution of System (4), we have

$$\forall \tau > \tau_0, \quad \int_0^\tau |Cu(t)|^2 dt > \Theta \|u_0\|^2_{\mathcal{U}}. \tag{22}$$

Following our previous computation, we remark that when $\mathcal{U} = \boldsymbol{L}^2([0,\ell])$ equipped with its natural norm Inequality (22) cannot be satisfied. Let us see this in the $b$-constant case.

$$\int_0^\tau |Cu|^2 dt = \int_0^\tau \left[\int_0^\ell x^n u(x,t) dx\right]^2 dt = \int_0^\tau \left[\int_{bt}^\ell x^n u_0(x+bt) dx\right]^2 dt$$

If $n = 0$, we call $F(bt) = \int_{bt}^\ell u_0(x+bt)dx$. The observability condition then reads

$$\int_0^\tau F(s)^2 ds \geq \Theta \int_0^\ell F'(s)^2 ds.$$

Counter examples proving that this cannot be uniformly the case for any $F$ are well-known, take for instance any mollifier sequence $\rho_\alpha = \frac{1}{\alpha}\rho(\frac{x}{\alpha})$, where $\rho \in C_c^\infty\big((0,\min\{\ell,\tau\})\big)$.

However, it would have been possible to have the observability condition if we would have chosen different metrics. For example, let us consider the case of very regular observations in $\boldsymbol{H}^{n+1}([0,\tau],\mathcal{Z})$ with the seminorm

$$|f|_{\boldsymbol{H}^{n+1}([0,\tau])} = \int_0^\tau \left|\frac{d^{n+1}f}{dt^{n+1}}\right|^2 dt$$

and state space $\mathcal{U} = \boldsymbol{L}^2([0,\ell])$. Thanks to Equation (7), we can easily find a constant $\Theta$ that satisfies, $\forall\, \tau > \tau_0 = \frac{\ell}{b}$, the observability condition

$$\int_0^\tau \left|\frac{d^{n+1}\Psi_\tau u_0(t)}{dt^{n+1}}\right|^2 dt \geq \Theta \|u_0\|_{\boldsymbol{L}^2([0,\ell])},$$

associated to the criterion

$$J(\xi) = \frac{\gamma_\xi}{2}\|\xi\|^2_{\boldsymbol{L}^2([0,\ell])} + \frac{\gamma_z}{2}\|z(t) - \Psi_\tau u_0(t)\|^2_{\boldsymbol{H}^{n+1}([0,\tau])}.$$

Alternatively, we can satisfy the observability condition in the case of less regular initial condition in $\mathcal{U} = \boldsymbol{H}^{-(n+1)}([0,\ell])$ and observations in $\boldsymbol{L}^2([0,\tau])$. The criterion thus reads

$$J(\xi) = \frac{\gamma_\xi}{2}\|\xi\|^2_{\boldsymbol{H}^{-(n+1)}([0,\ell])} + \frac{\gamma_z}{2}\|z(t) - \Psi_\tau u_0(t)\|^2_{\boldsymbol{L}^2([0,\tau])}.$$

Therefore, the inequality of the observability condition becomes

$$\int_0^\tau |\Psi_\tau u_0(t)|^2 \geq \Theta \|u_0\|^2_{\boldsymbol{H}^{-(n+1)}([0,\ell])}.$$

According to Equation (7), we can rewrite this inequality as

$$\int_0^\tau |\Psi_\tau u_0(t)|^2 \geq \frac{\Theta}{n!(-b)^{n+1}} \left\| \frac{d^{n+1}}{dt^{n+1}}(\Psi_\tau u_0) \right\|^2_{\boldsymbol{H}^{-(n+1)}([0,\tau])}.$$

It is easy to prove that

$$\left\| \frac{d^{n+1}}{dt^{n+1}} \Psi_\tau u_0 \right\|_{\boldsymbol{H}^{-(n+1)}([0,\tau])} \leq \|\Psi_\tau u_0\|_{\boldsymbol{L}^2([0,\tau])},$$

and we can satisfy the observability condition with $\Theta = n!(-b)^{n+1}$ and $\tau_0 = \frac{\ell}{b}$.

In these two cases (the space $[0,\tau] \to \mathcal{Z}$ equipped with a very regular norm, or on the contrary the very weak assumption on the regularity of the initial state $\mathcal{U}$), the observability condition shows that the problem is well-posed for $\tau \geq \frac{\ell}{b}$ and so there is no need for either a regularisation or an *a priori* information. However they cannot be used for real applications since we do not observe $z(t)$ in such a regular space, and we want to reconstruct regular initial states.

# 3 Numerical Analysis

## 3.1 Model discretisation

In this section we describe the numerical implementation and the comparison between the two approaches.

We set the space domain to $[0,\ell] = [0,\, 200]mer$. We define the notation $mer$ for monomer which is the fundamental unit aggregating into oligomers. We set the time domain to $[0,\tau] = [0,\, 100]$min and the transport velocity to $b = 2\text{min}^{-1}$.
We present two cases associated to two initial concentration conditions: the gaussian function $u_{0\text{g}} = e^{\frac{1}{2}\frac{-(x-100)^2}{20^2}}\mu M$ and the characteristic function $u_{0\text{ch}} = I_{[70,130]}\mu M$. For the sake of simplicity, in the following of this section we omit the units.
We consider a uniform space grid $0 = x_0 < \ldots < x_{N_x} = \ell$, with a constant space step $\delta x$. By evaluating the continuous initial conditions on this grid, we obtain the vector

$$\breve{u}_0 = (\breve{u}_{0,j})_{0\leq j\leq N_x} = \breve{u}_0(x_j).$$

We fix a time discretisation $t_0 < \ldots < t_{N_t}$ of the time domain $[0, \tau]$ with a constant time step $\delta t$. We call $u_j^k$ the approximation of $u(x_j, t_k)$. The cluster concentrations, at time $t_k$, are approximated by the vector $u^k = (u_j^k)_j$. To compute these quantities, we refer to the discrete model

$$\begin{cases} u^{k+1} = A_{k+1|k} u^k, & \text{for } k \in \mathbb{N} \\ u^0 = u_0. \end{cases} \tag{23}$$

The expression of the discrete model operator $A_{k+1|k}$ depends on the numerical scheme which is adopted to discretise the transport equation of System (3). For the upwind scheme it is

$$A_{k+1|k} = \mathbb{1}_{N_x} + \delta t \, bD, \tag{24}$$

where the discrete differential operator $D$ is such that

$$(Du^k)_j = \frac{u_{j+1}^k - u_j^k}{\delta x} \qquad \text{if } b > 0 \qquad\qquad (Du^k)_j = \frac{u_j^k - u_{j-1}^k}{\delta x} \qquad \text{if } b < 0.$$

We can also use a numerical scheme with higher approximation order such as a Lax-Wendroff scheme. The discrete model operator associated to this scheme is

$$A_{k+1|k} = \mathbb{1}_{N_x} + \frac{b\delta t}{\delta x} D_x^c + \frac{b^2 \delta t^2}{2\delta x^2} D_{xx},$$

where

$$(D_x^c u^k)_j = \frac{u_{j+1}^k - u_{j-1}^k}{2\delta x} \qquad \text{and} \qquad (D_{xx} u^k)_j = \frac{u_{j+1}^k - 2u_j^k + u_{j-1}^k}{2\delta x^2}.$$

We choose space and time steps satisfying the *Courant–Friedrichs–Levy (CFL) condition* $\left|\frac{b\delta t}{\delta x}\right| \leq 1$ that ensures the stability of the schemes [16].

### 3.2 Synthetic data generation

To test our inversion strategies, we generate synthetic observations. In this respect we fix uniform grids on $[0, \tau]$ and $[0, \ell]$ with discretisation steps much smaller than the ones considered solving the inverse problem. Specifically, we take the time step $\delta t = 10^{-3}$ and space step $\delta x = 2 \cdot 10^{-3}$. We use the discrete model (23) with $\breve{u}_0$ as initial condition to compute the sequence $(\breve{u}^k)_{1 \leq k \leq N_{\text{obs}}}$. Consequently, we compute the observations thanks to the discrete observation operator

$$C_k^{(n)} = \delta x \begin{pmatrix} \frac{x_0^n}{2} & x_1^n & \ldots & x_{N_x-1}^n & \frac{x_{N_x}^n}{2} \end{pmatrix}$$

obtained by using the trapezoidal rule to approximate the space integral appearing in the continuous definition. We remark that – since the continuous observation operator $C$ is time independent – $C_k$ does not depend on $k$.

We consider synthetic observations of the form

$$z_k = C_k \breve{u}_k + \chi_k, \quad k = 0, \ldots, N_{\text{obs}}, \tag{25}$$

where $\chi_k = \varepsilon\omega_k$ and the values $\omega_k$ are randomly generated according to the standard gaussian distribution. As we can see in [18], this construction produces a white gaussian noise on the observations such that heuristically

$$\|z - C\breve{u}\|_{\boldsymbol{H}^{-\frac{1}{2}}([0,\tau])} \leq \varepsilon.$$

Consequently, we can take the $\varepsilon$ as the noise level in $\boldsymbol{H}^{-\frac{1}{2}}([0,\tau])$.

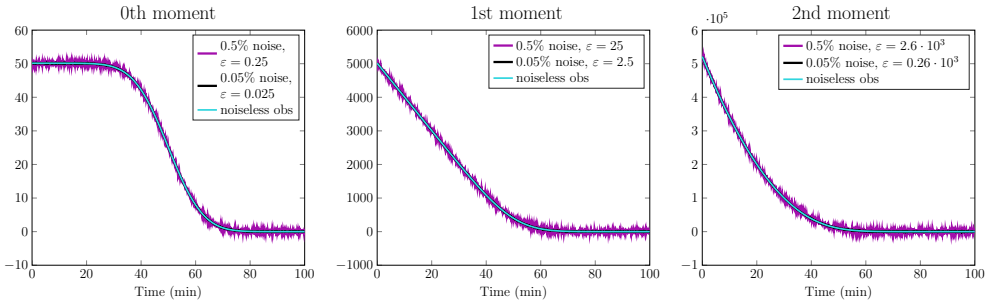



Figure 1: Three moments of the state function $u$ having dynamics (23) where $u_{0\mathrm{g}} = e^{\frac{1}{2}\frac{-(x-100)^2}{20^2}}$ and $b = 2$.

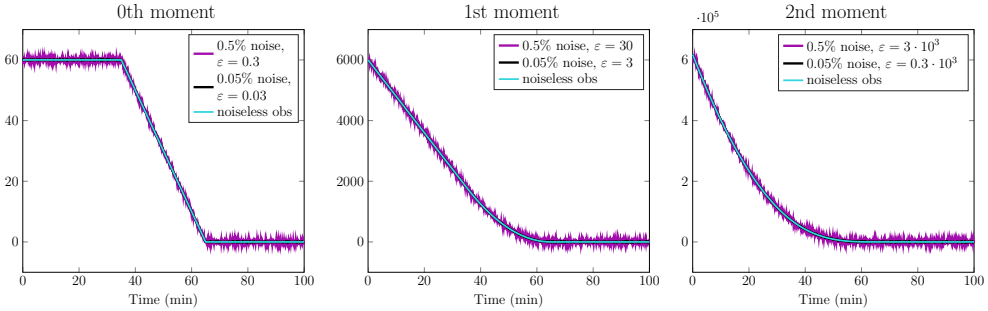



Figure 2: Three moments of the state function $u$ having dynamics (23) where $u_{0\mathrm{ch}} = I_{[70,130]}$ and $b = 2$.

In Figures 1 - 2, we present the synthetic observations associated to the gaussian function $u_{0\mathrm{g}}$ and the characteristic function $u_{0\mathrm{ch}}$, respectively. In both cases we have computed the first three moments of the state function $u$. Moreover, we consider noised observations. The noise corresponds to errors of 0.05% and 0.5%. The corresponding values of the noise level $\varepsilon$ have been reported in figure legends.

### 3.3 Numerical Simulations: kernel regularisation method

In this section we present some examples of numerical initial condition estimation by the kernel regularisation method. We recall that the estimation is given by the function $\hat{u}_0^{\varepsilon,\alpha}$

defined in Equation (12). The parametrised kernel family $\rho_\alpha$ is defined by $\rho_\alpha(x) = \frac{1}{\alpha}\rho(\frac{x}{\alpha})$. The kernel $\rho$ is chosen as the gaussian kernel, $\rho = \frac{1}{0.3\sqrt{2\pi}} e^{-\frac{x^2}{2(0.3)^2}}$. According to these choices, the coefficient $m$ – defined in Equation (9) – is equal to 1. We compute the $(n+1)$-th derivative of the convolution between $z$ and the regularisation kernel $\rho_\alpha$ as the convolution between $z$ and the $(n+1)$-th derivative of $\rho_\alpha$. The derivative can be either analytically computed or approximated by finite differences. In the examples of this section we have considered the analytic expression of kernel derivatives.

To compute the discrete convolution we need two vectors. One vector is the set of measurements $z$. The other vector is obtained by evaluating the derivative of the kernel function, $\frac{d^{n+1}\rho_\alpha}{dx^{n+1}}$, on a discrete grid. First of all we approximate the kernel $\rho$ by $\tilde{\rho} = \rho I_{[-2,2]}$, where $I_{[-2,2]}$ is the characteristic function for the domain $[-2, 2]$. We remark that when we numerically compute the integral of $\tilde{\rho}$ we obtain 1, which is the same value as the integral of $\rho$ over $\mathbb{R}$. Hence, approximating $\rho$ by $\tilde{\rho}$ we make an error smaller than machine-precision. Consequently, we consider the support of $\rho_\alpha$ and its derivatives to be included in $[-2\alpha, 2\alpha]$. We evaluate $\frac{d^{n+1}\rho_\alpha}{dx^{n+1}}$ over the grid $-2\alpha = x_1 < \ldots < x_r = 2\alpha$ with the same discretisation step, $\delta t$, as the measurement time grid.

The discrete convolution of two vectors $x$, $y$ of lengths $m$ and $h$, respectively, is the vector $w$ such that $w_k = \sum_j x_j y_{k-j+1}$, for $k = 1, \ldots m+h+1$. For every component $k$ the index sum $j$ varies between $\max(1, k+1-h)$ and $\min(k, m)$. This algorithm is equivalent to extend with zeros the vectors for indices $j$ outside the range $[\max(1, k+1-h), \min(k, m)]$.

We remark that zero-padding the kernel derivative vector is equivalent to evaluating the function $\frac{d^{n+1}\rho_\alpha}{dx^{n+1}}$ outside the domain $[-2\alpha, 2\alpha]$. Moreover, extending with zeros the measurements for times bigger than the observation time is coherent with the biological interpretation: a depolymerising system in which all polymers have been reduced into monomers cannot change its state.

On the other side, adding zeros for negative times would lead to a bad reconstruction of the initial condition on the left border. Our idea is to extend the observation data for negative times in $[-\tau_\alpha, 0]$. The positive value $\tau_\alpha$ is such that the component $w_k$ – corresponding to the left border value of the estimation $\hat{u}_0^{\varepsilon,\alpha}$ – is computed as a complete sum. To this purpose, we fix $\tau_\alpha$ bigger than the length of the kernel domain, specifically $4\alpha$. To consider negative times, we extend the definition of the initial condition for negative sizes by $u_0(x) = 0$ if $x < 0$. The $n$-th moment for negative times reads

$$C^n u(t) = \int_{bt}^{\ell} (x - bt)^n u_0(x) \mathrm{d}x = \int_0^{\ell} (x - bt)^n u_0(x) \mathrm{d}x.$$

We notice that the $n$-th moment is a polynomial of degree $n$ in $t$. Assuming there is a size $x_{min} > 0$ such that the support of $u_0$ is included in $[x_{min}, \ell]$, we obtain that, for every $t \leq \frac{x_{min}}{b}$, $C^n u(t) = \int_{x_{min}}^{\ell} (x - bt)^n u_0(x) \mathrm{d}x$. To conclude, we assume $x_{min} = 10$ in our numerical examples. We fit the observations relative to times in the range $[0, 5]$ with an $n$ degree polynomial. We present in Figure 3 an example of extension of observation data

for the first three moments in the case of gaussian initial condition $u_{0g}$. We discretise the arbitrarily chosen negative domain $[-10, 0]$ with a time step $\delta t$. We evaluate the polynomial fit on this grid and we use these data to extend our observations.

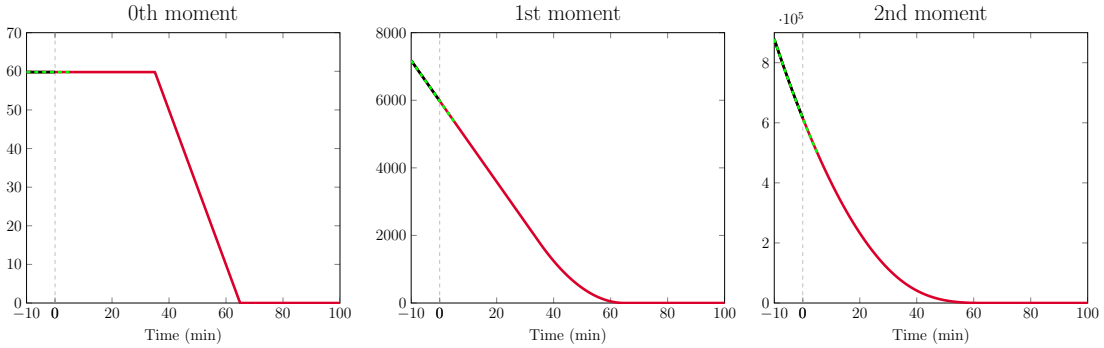



Figure 3: From the left to the right the 0th, 1st and 2nd moment relative to the initial condition $u_{0g}$ and depolymerisation rate $b = 2$. In red the moments for positive times, in black the moments for negative times in $[-10, 0]$, in green the polynomial function fitting the red curve in $[0, 5]$ and used to extend the data on $[-10, 0]$.

We compute the discrete convolution between the extended observation and the kernel derivative vector. We multiply the resulting vector by $\delta t$ – to approximate the continuous integral of the convolution – and by $\frac{1}{(-b)^{n+1}n!}$ in accordance with Equation (12).

We present in Figure 4 and Figure 6 the estimation of the initial conditions $u_{0\mathrm{g}}$ and $u_{0\mathrm{ch}}$, respectively. We remark that the quality of the estimation decreases when the order of the moment and the noise level increase.

### 3.4 Numerical Simulations: the data assimilation method

We now turn to the variational approach detailed in Section 2. In order to discretise and simulate the two-end problem (21), we rely on a discretised version of the optimal criterion (18) to be minimised under the constraint of the discretised model (23). The resulting time-discretised optimal system can then be proved to converge to time-continuous solution of (21) [7]. Therefore, we decompose the initial condition by defining $\breve{\xi} \in \mathbb{R}^{N_x}$ such that $\breve{u}_0 = u_\circ + \breve{\xi}$ and seek an estimate of $\breve{\xi}$ given by

$$\bar{\xi} = \arg\min_{\xi} J_{N_t}(\xi) = \arg\min_{\xi} \left( \frac{1}{2}\|\xi\|^2_{P_0^{-1}} + \frac{1}{2}\sum_{k=0}^{N_t} \|z_k - C_k u^k\|^2_{M_k} \right). \tag{26}$$

The matrix $M_k$ is the discrete approximation of the operator $\gamma Id_{\mathcal{Z}}$ and it depends on the quadrature rule chosen to approximate the integral in time. We fix $M_k = \delta t \gamma I_{N_t}$.

Furthermore – if we assume that the initial *a priori* approximates the unknown initial condition with the same error on every cluster size – we can take $P_0 = \dfrac{1}{\delta x \beta} I_{N_x}$.

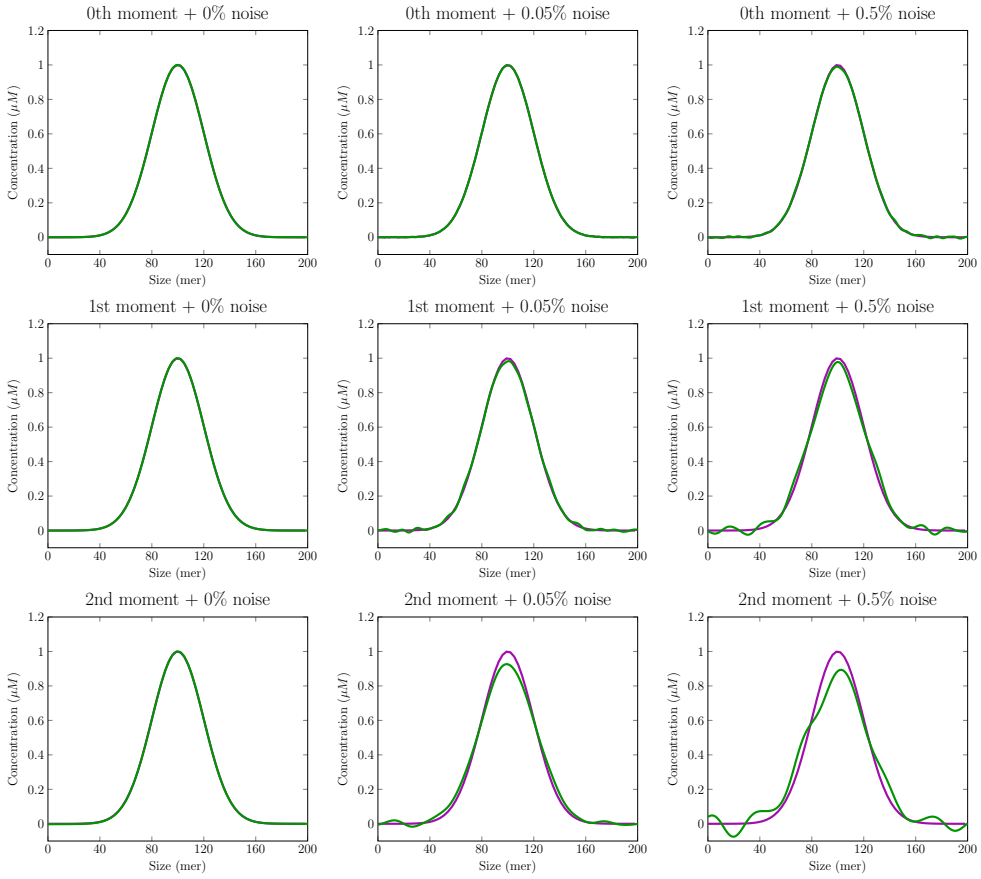



Figure 4: Comparison between the exact gaussian initial condition $u_{0\mathrm{g}}$ (purple line) and the approximations $\hat{u}_0^{\varepsilon,\alpha}$ (green line) provided by the kernel regularisation method. Each estimation is associated with the measurements in Figure 5.

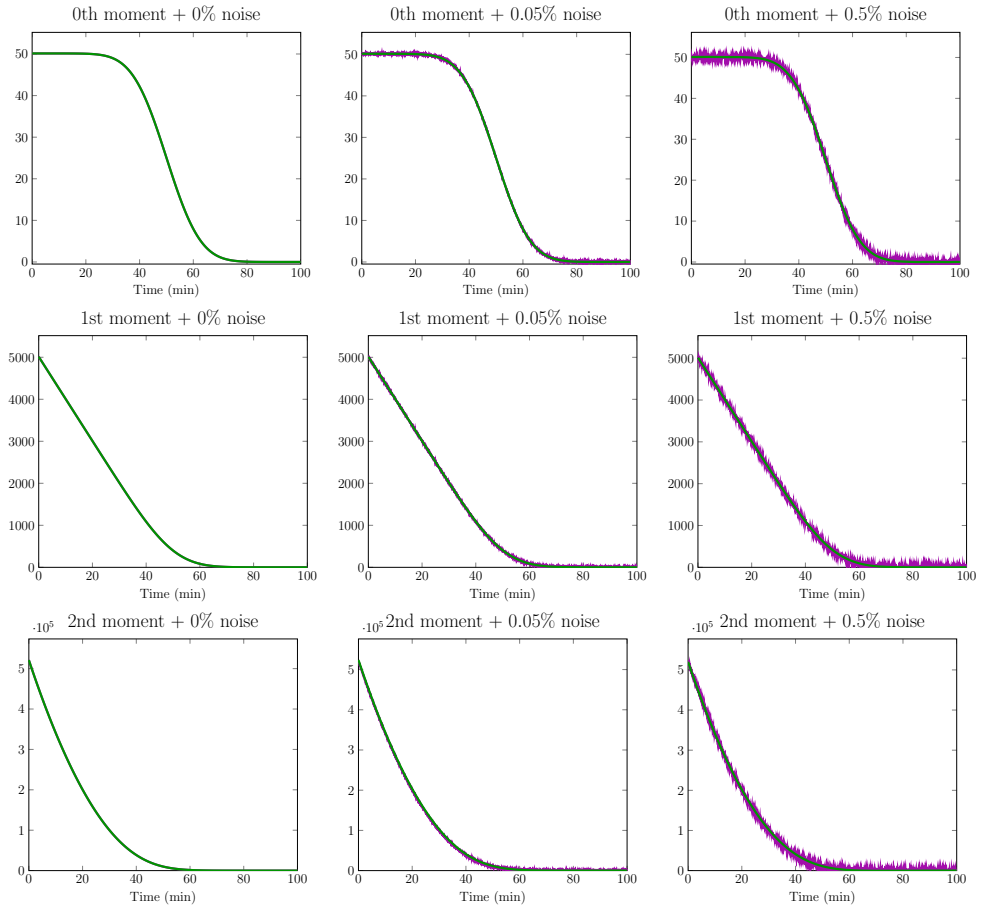



Figure 5: Synthetic observation in the case of gaussian initial condition $u_{0g}$ (purple line) and relative fit given by the observations generated from the data assimilation estimator. From the top to the bottom by rows, we see the 0th-moment, 1st-moment, 2nd-moment. From the left to the right by columns the noise corresponds to a 0%, 0.05%, 0.5% error.

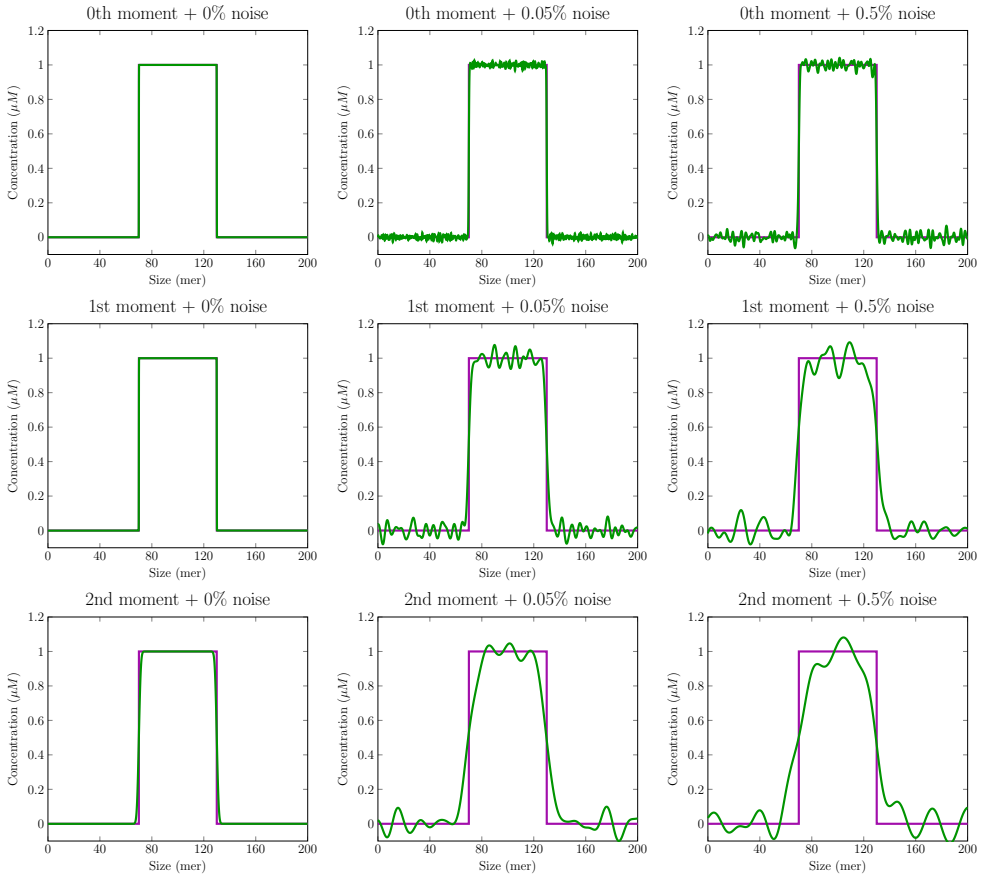



Figure 6: Comparison between the exact initial condition $u_{0\text{ch}}$ (purple line) and the approximations $\hat{u}_0^{\varepsilon,\alpha}$ (green line) provided by the data assimilation method. Each estimation is associated with the measurements in Figure 7.

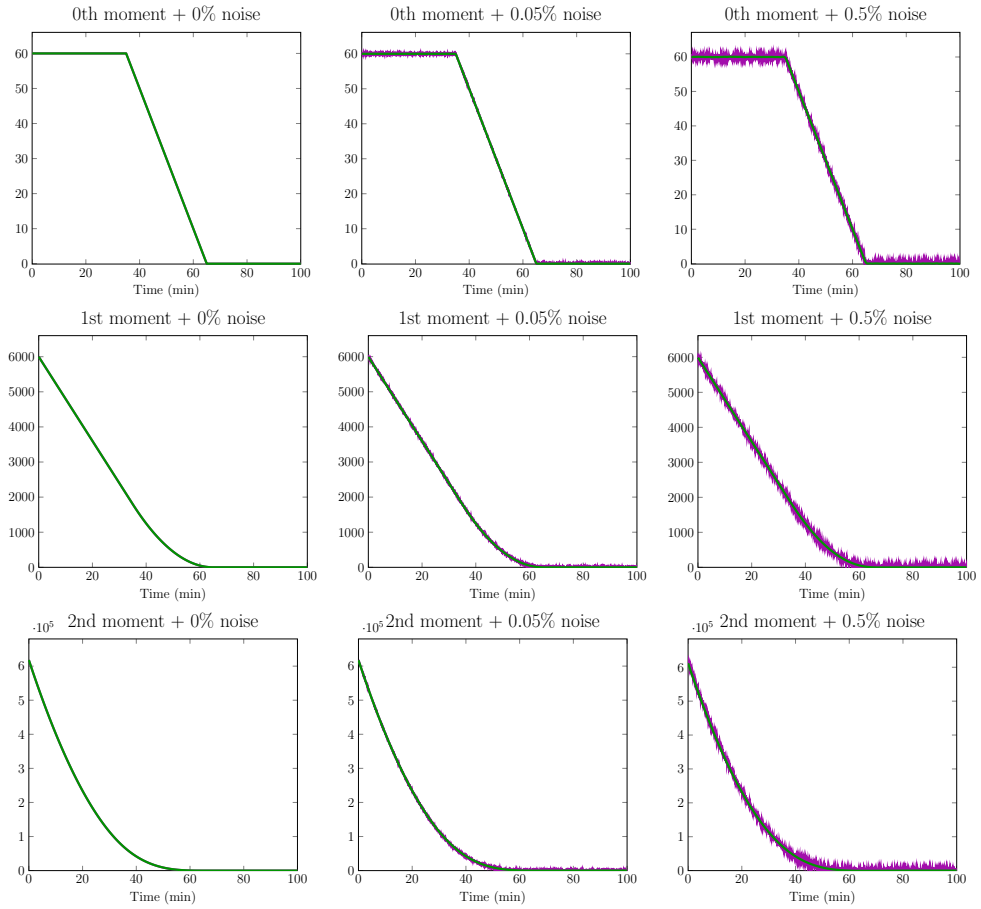



Figure 7: Synthetic data in the case of $u_{0\text{ch}}$ as initial condition (purple line) and relative fit given by the observations generated from the data assimilation estimator. From the top to the bottom by rows, we see the 0th-moment, 1st-moment, 2nd-moment. From the left to the right by columns the noise corresponds to a 0%, 0.05%, 0.5% error.

These two matrices are classically defined as $P_0 = \mathrm{Cov}(\breve{u}_0 - u_\circ)$ and $M_k = \mathrm{Cov}(\chi_k)^{-1}$. As well explained in [25] – if $W(t)$ is the time-independent covariance of the continuous white gaussian process $\chi(t)$ – the covariance of measurement noise in discrete time is $W_k = \frac{W(t_k)}{\delta t}$. Since $M_k = W_k^{-1}$, we find the relation $M_k = \delta t M(t) = \delta t \gamma Id_{\mathcal{Z}}$.
Consequently, we define $\gamma$ as $\gamma = (\sigma^2)^{-1}$, where $\sigma^2$ is the variance of the white gaussian process $\chi(t)$. The parameter $\beta$ is analogously defined as $\beta^{-1} = \|\breve{\xi}\|^2_{\boldsymbol{L}^2([0,\ell])}$. In practice, we can numerically estimate the parameter $\gamma$ – by analysing the noise on the data – while the value of the parameter $\beta$ reflects the confidence that we have on the *a priori* information on the initial condition.

To minimise the criterion we use a gradient-descent based optimisation method that – starting from the initial guess $\xi = 0$ – iteratively attempts to estimate the minimum from the criterion gradient evaluated on the current guess

$$\nabla J_{N_t}(\xi) = P_0^{-1}\xi - (q^0_{|\xi}),$$

where $q^0_{|\xi}$ is the time-discrete adjoint variable at time 0 solution of the time-discrete system

$$\begin{cases} q^k_{|\xi} = A_{k+1|k} q^{k+1}_{|\xi} - C_k^\intercal M_k (z_k - C_k u^k_{|\xi}), \quad 0 \leq k \leq N_t \\ q^{N_t+1}_{|\xi} = 0. \end{cases} \tag{27}$$

Note that the time-discrete adjoint variable is the Lagrange multiplier associated with the dynamical constraint (23) in the minimisation of (26). Besides, it is also a time-discretisation of the time-continuous adjoint variable (19).

As numerical synthetic test cases, we present in Figure 8 the estimation of a gaussian initial condition by the variational data assimilation method. Respectively in Figure 10, we estimate a characteristic function. As previously done with the kernel method, the nine estimation curves correspond to the nine observation curves presented in Figure 1 or Figure 2.

# 4 Application on experimental data

Having presented, theoretically investigated and numerically tested our mathematical approach, we are now ready to apply our method to experimental data.

## 4.1 Presentation of the experimental protocol and noise analysis

The data to analyse consist in observations on ovine prion protein oligomers (PrP oligomers), in depolymerising conditions. PrP oligomers are a kind of amyloid deposit generated by the concatenation of monomers forming chains of a few tens of proteins. These structures are relatively small compared to protein polymers, that could be composed by up to thousands of proteins [24].

We refer to Appendix B for details on the protocol used to form and make measurements on oligomer systems. We present in Figure 12 an example of Static Light Scattering

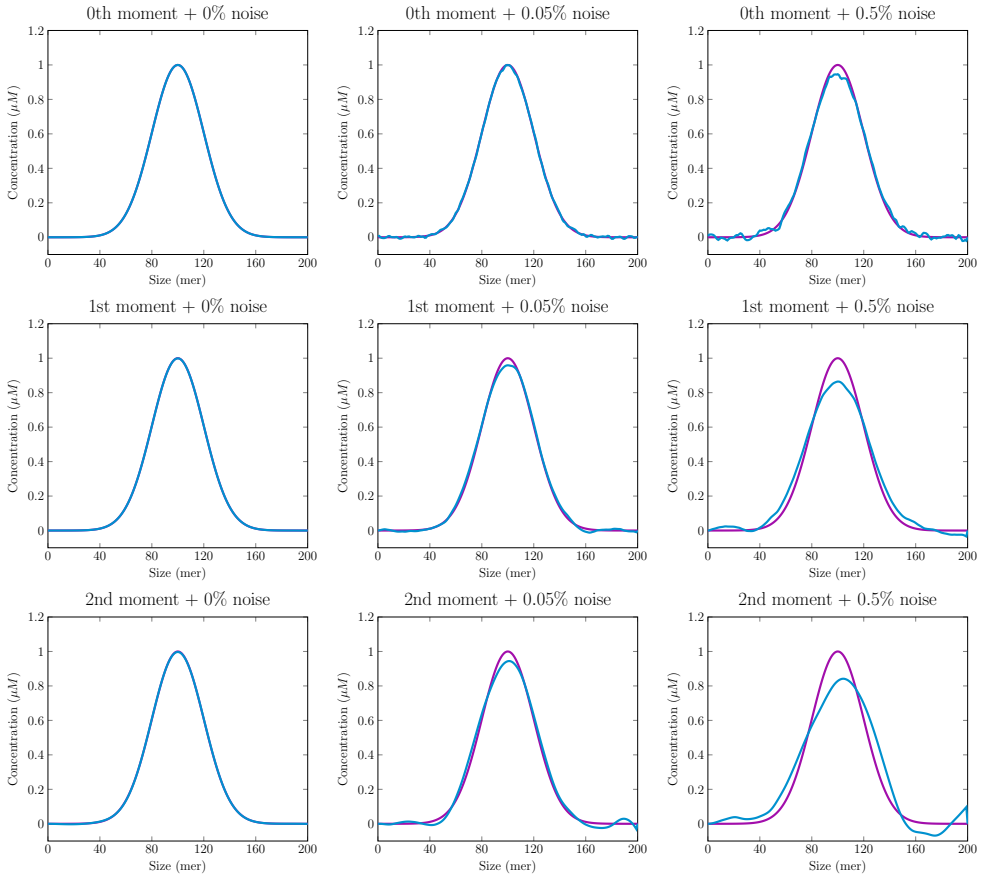



Figure 8: Comparison between the exact gaussian initial condition $u_{0\mathrm{g}}$ (purple line) and the approximations $\hat{u}_0^\alpha$ (blue line) provided by the data assimilation method. Each estimation is associated with the measurements in Figure 9.

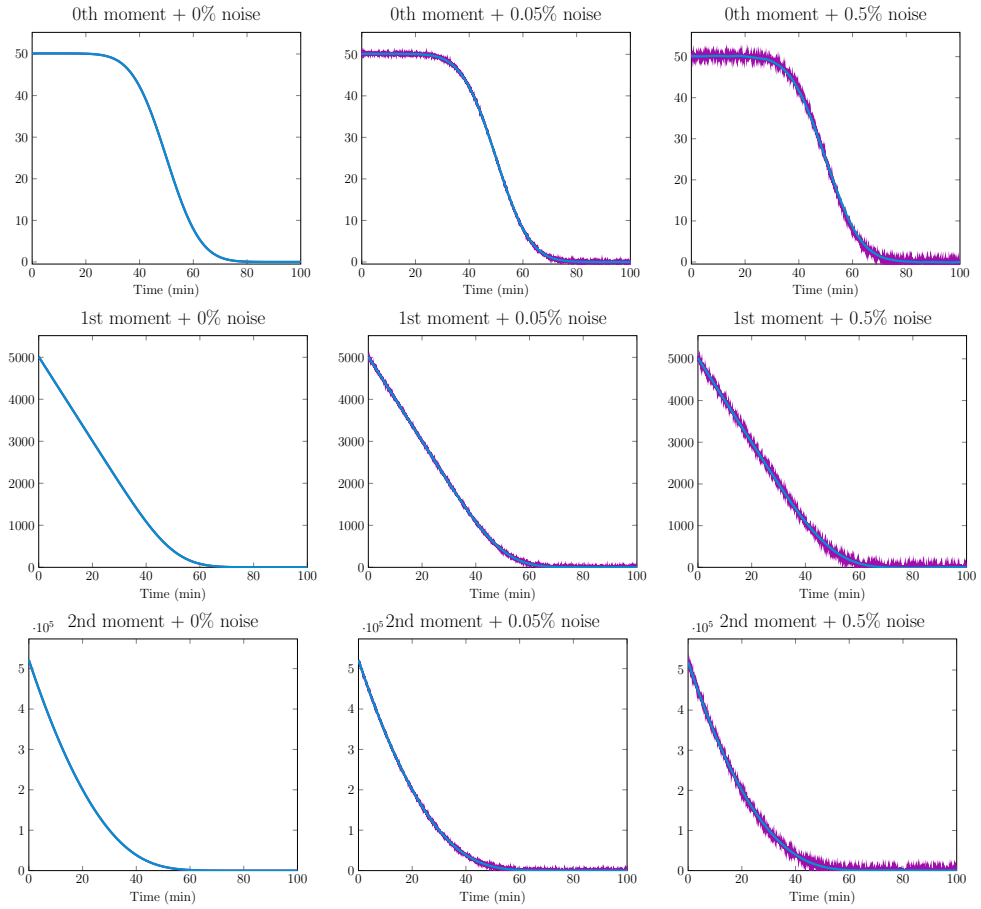



Figure 9: Synthetic observation in the case of gaussian initial condition $u_{0g}$ (purple line) and relative fit given by the observations generated from the data assimilation estimator. From the top to the bottom by rows, we see the 0th-moment, 1st-moment, 2nd-moment. From the left to the right by columns the noise corresponds to a 0%, 0.05%, 0.5% error.

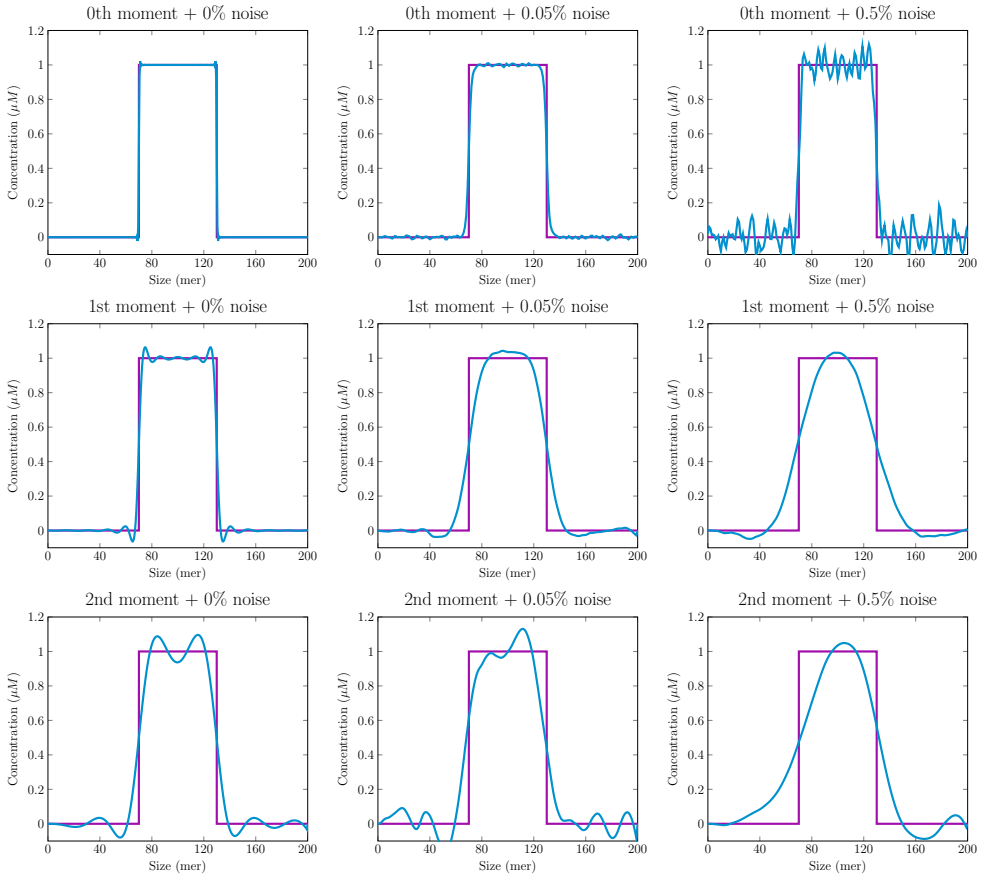



Figure 10: Comparison between the exact initial condition $u_{0\text{ch}}$ (purple line) and the approximations $\hat{u}_0^\alpha$ (blue line) provided by the data assimilation method. Each estimation is associated with the measurements in Figure 11.

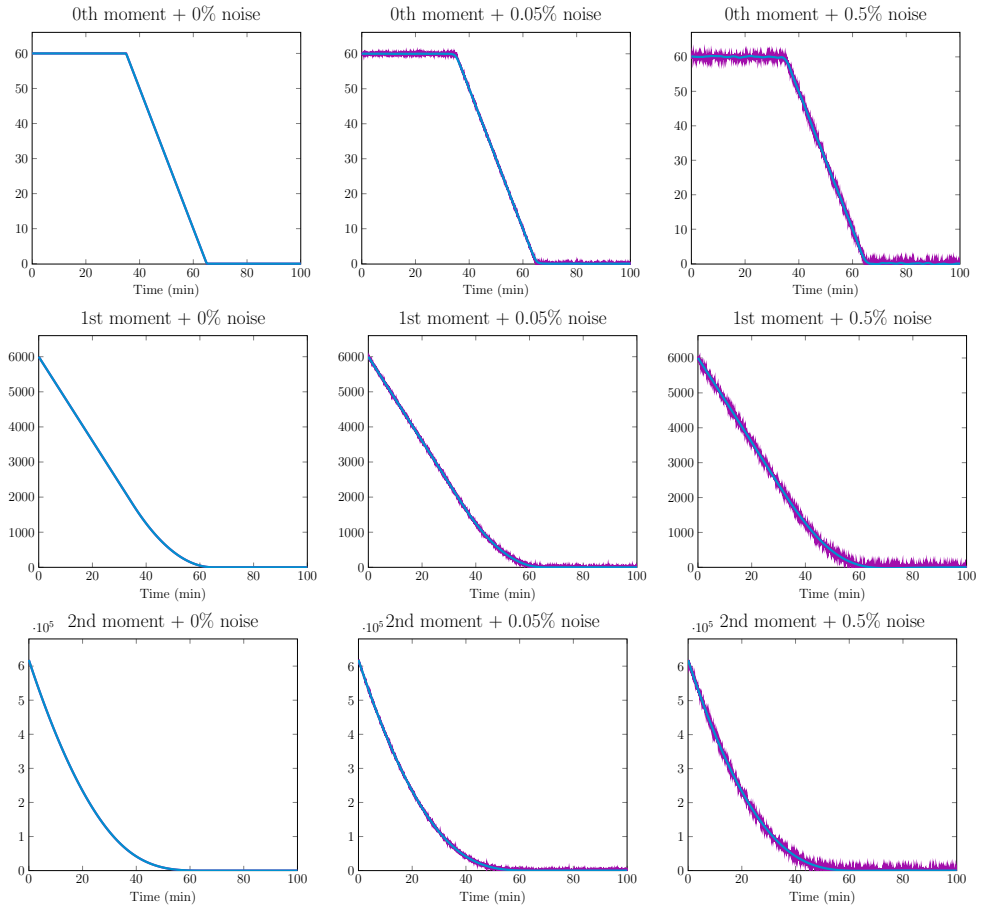



Figure 11: Synthetic data in the case of $u_{0\text{ch}}$ as initial condition (purple line) and relative fit given by the observations generated from the data assimilation estimator. From the top to the bottom by rows, we see the 0th-moment, 1st-moment, 2nd-moment. From the left to the right by columns the noise corresponds to a 0%, 0.05%, 0.5% error.

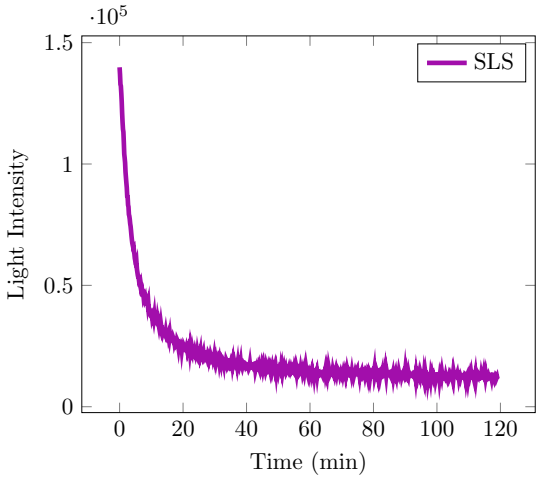


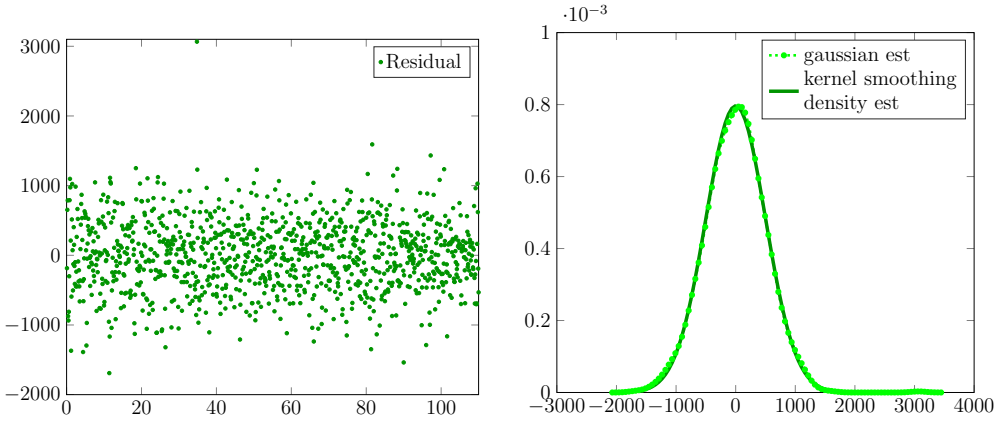

Figure 12: Depolymerisation kinetics of ovPrP oligomers monitored by static light scattering.

measurements on a depolymerising system of PrP oligomers. We recall that the SLS measurement is a linear transformation of the second moment $z_{SLS}(t) = c_1 Cu(t) + c_2$, with unknown parameters $c_1, c_2$. We assume to observe the experiment until all the oligomers are depolymerised into monomers. We fix the parameter $c_2$ such that the mean of the measurements at end of the observation domain is zero. We thus consider the shifted data $z_{SLS} - c_2$ as measurements. Solving the inverse problem with this observation data, we estimate the function $c_1 u_0$. In the following, we assume that $c_1 = 1$.

In order to analyse the measurement noise, we assume it to be a white gaussian additive noise and test this hypothesis. Since we assume that the initial size distribution is a regular function, we expect the corresponding second moment to be a regular function with a smooth graph. For this reason, we start by filtering the data. We use a cubic Savitzky-Golay filter. For more details about this filter see [19].

The difference between the empirical data and the fit gives us an estimation of the noise contribution, see Figure 13a. We run a $\chi^2$ numerical test to test the null hypothesis of residual points following a gaussian distribution. The test accepts the null hypothesis at the 5% significance level. We then estimate the mean and the standard deviation of the gaussian distribution generating the residual. We estimate the mean at 0 and the standard deviation $\sigma = 501$. The purple dotted line in Figure 13b shows the estimated gaussian density function: this leads us to accept our noise model and keep this value of $\sigma$ as a reasonable estimation of the noise level.

The experimental protocol also includes the separation of oligomers by size, using the Size Exclusion Chromatography (SEC) device. Thanks to this technique we can measure the initial oligomer distribution. We point out that it would not have been possible to make these measurements on fibrils, due to the large size of the aggregates and the limits of the device. We can see in Figure 14 the measurement of the initial oligomer distribution




(a) (b)

Figure 13: Analysis of the noise on SLS data. We present (left) the residuals obtained as the difference between the SLS data and the cubic Savitzky-Golay filter of the data. In the right figure we present two estimations of the density function associated to the residual data.

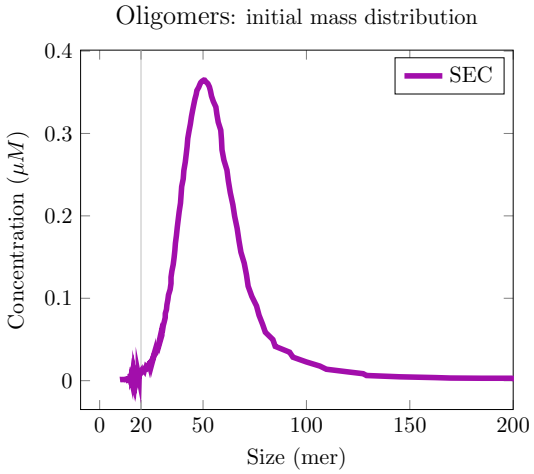



Figure 14: Size-exclusion chromatogram of purified ovPrP oligomers.

associated to the data in Figure 12. We notice that the biggest oligomer size taken into account is $200mer$, with an almost zero relative concentration, while the smallest detectable size is $20mer$. In this set of data, the most present oligomers have sizes between $30mer$ and $100mer$. We also remark that this distribution has only one peak centred around the size $50mer$. Evaluating the noise level on the SEC device is a complex subject, going beyond the scope of this study.

In the following we set up the inverse problem of estimating the initial size distribution by using the SLS measurements only. We then discuss the results obtained when we also take into account the SEC measurement.

### 4.2 Initial state estimation without *a priori*

Oligomer dynamics can be modelled by System (1), [9]. We remark that in this model there are two unknowns: the depolymerisation rate $b$ and the initial condition $u_0$. The approaches that we have presented in this paper are designed to estimate only the initial condition. We thus start our analysis with the simple model of constant backward transport

$$\begin{cases} \dfrac{\partial}{\partial t}u(x,t) - b\dfrac{\partial}{\partial x}u(x,t) = 0, \\ u(L,t) = 0, \\ u(x,0) = u_0(x). \end{cases} \tag{28}$$

Our strategy is to fix an arbitrary value for $b$ and then perform the initial condition estimation. The resulting estimation depends explicitly on $b$, as we have seen in Equation (7). For example, consider two models associated to the rates $b_1 \neq b_2$. Equation (7), in the case of a noiseless second moment observation, reads

$$u_{0|_{b_i}}(x) = \frac{1}{2(-b_i)^3}\frac{d^3}{dt^3}z\left(\frac{x}{b_i}\right),$$

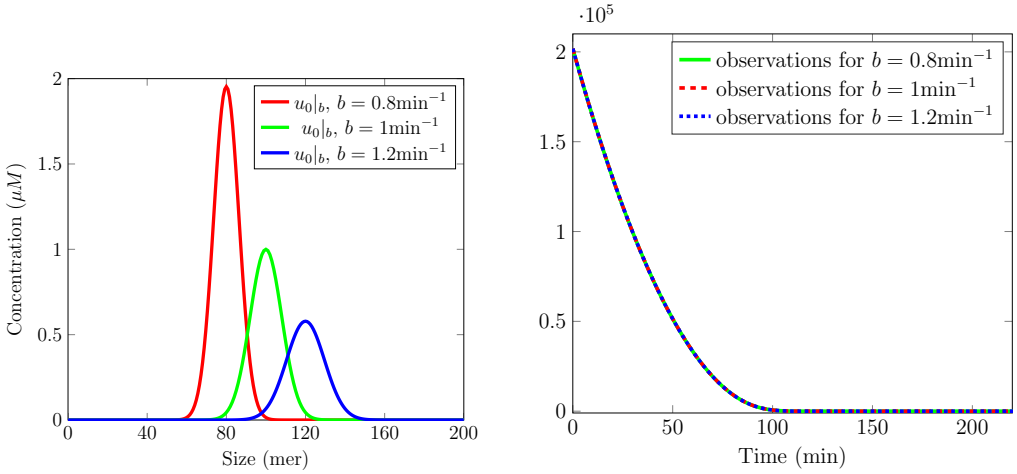
for $i = 1, 2$. We use the notation $u_{0|_{b_i}}$ to indicate the solution of the inverse problem when we consider the transport velocity $b_i$ in the model (28). Eventually, we can notice that

$$u_{0|_{b_1}}(x) = \left(\frac{b_2}{b_1}\right)^3 u_{0|_{b_2}}\left(\frac{b_2}{b_1}x\right). \tag{29}$$

This relation leads us to the conclusion that, when we fix a depolymerisation rate, we obtain a function that differs from the exact one in a linear change of variables and a scaling factor. To illustrate this relation, we show in Figure 15a an example of distributions which produce the same second moment observation, see Figure 15b, evolving with different rates.

We thus fix the depolymerisation rate to the arbitrary value $b = 2\text{min}^{-1}$. We consider the experimental time domain $[0, \tau] = [0, 110]$min. Biological considerations lead us to the definition of the size domain $[0, \ell] = [0, 200]mer$.

To apply the data assimilation method we need to define the least square criterion. We choose the isomorphism $P_0$ of the form $P_0 = \frac{1}{\beta}\text{Id}$. Consequently, we only need to fix




(a) Initial condition

(b) Second moment observations

Figure 15: Several choices of initial condition and depolymerisation rate can lead to the same second moment observation.

a value for the regularisation parameters $\beta$ and $\gamma$. As explained before, these parameters are linked by an inversely proportional relation to the confidence on the *a priori* on the initial condition and the noise level on measurements, respectively.

Since for the moment we do not consider additional information on the initial condition, our *a priori* is the zero constant function. In particular, we do not know whether this *a priori* is far or not from the target initial condition. Therefore, we assume to have low confidence on the *a priori* or equivalently we allow the estimations to be far from the *a priori*. This assumption corresponds to the choice of a small value for $\beta$. In the following we fix $\beta = 10^{-2}$.

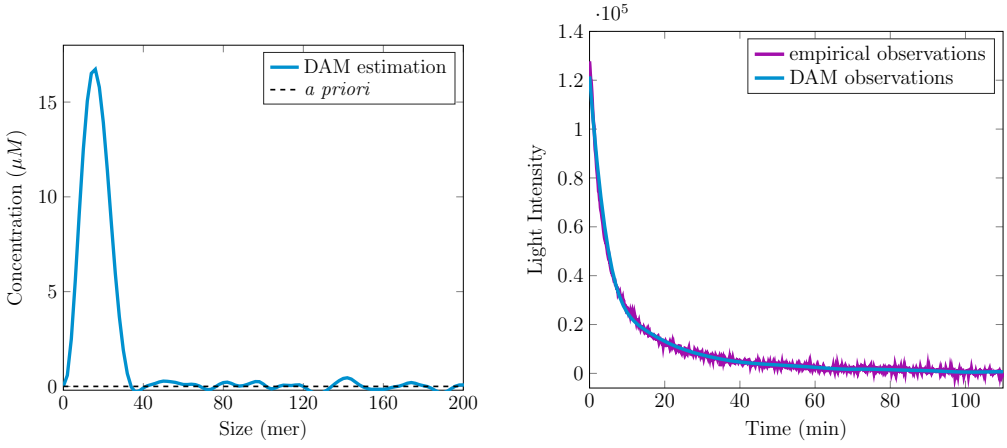
We consider $\gamma \simeq \frac{1}{\sigma^2}$, where $\sigma$ is the standard deviation of noise distribution. Consequently, we take $\gamma = 10^{-6}$.

We show in Figure 16 the results of data assimilation estimation. We see in Figure 16b that we obtain a good fit of the experimental data. In Figure 16a we have the initial state estimation. We recall that this estimation is associated to the arbitrary choice of $b$ and the real initial distribution can be a transformation of this function, according to Formula (29). Nevertheless, we can infer interesting features such as the presence of one main peak and the fact that the peak starts from small sizes.

### 4.3 Estimation with *a priori*

In this section we take into account the SEC measurement of Figure 14 to discuss the result of our initial state estimation of Figure 16a.

A first possibility is to admit that the chromatography technique cannot trustfully measure the variation of the distribution but it can nevertheless find the position of the




(a) (b)

Figure 16: Left: Initial condition estimation by variational approach when we choose $b = 2\,\mathrm{min}^{-1}$, $\beta = 10^{-2}$, $\gamma = 10^{-6}$. Right: comparison between the SLS measurements and the observations generated by the observation operator on the state estimation.

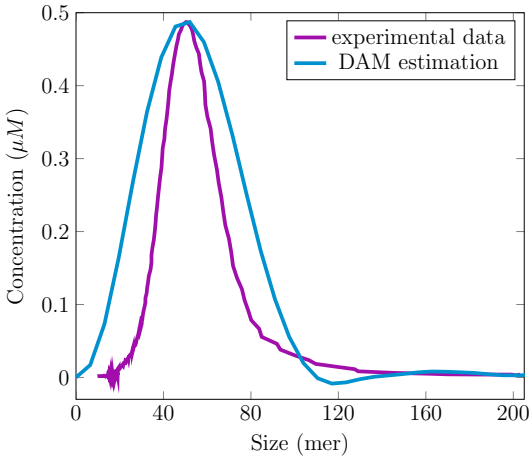



Figure 17: Comparison between the initial condition estimation – associated to the rate $b = 6.5\mathrm{min}^{-1}$ – and the experimental SEC data rescaled to have the same maximum as the estimation.

peak. Hence, we transform the estimation according to Formula (29) to move the peak to the same value as in the chromatography measurements. Since experimental data have been normalised to have integral equal to one, we can define a coefficient to make the two curves have the same maximum. In this way, we could use the SEC to identify the depolymerisation rate – that in this case corresponds to $b = 6.5\text{min}^{-1}$. Consequently, we consider the blue curve in Figure 17 as the estimation of the initial condition. However, according to the SEC specification and methodology the difference between the two distributions - the experimental one and the estimated one, see Figure 17 - seems too important to correspond to a noise on the measurement.

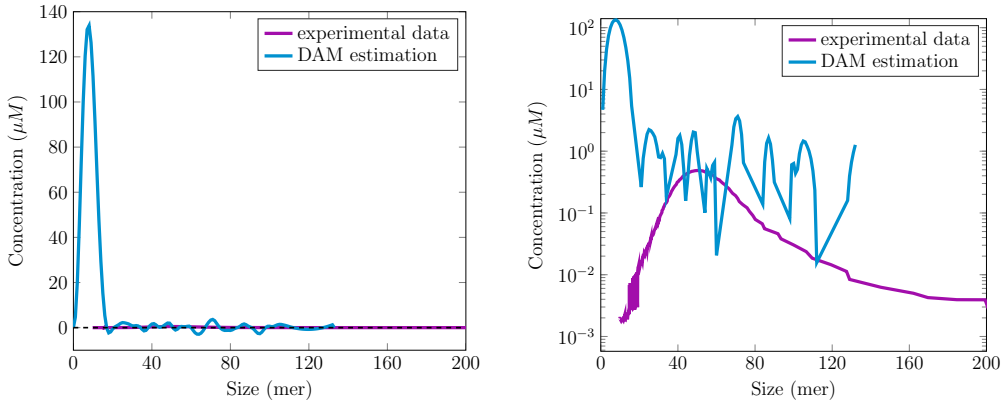



Figure 18: Initial condition estimation associated to the rate $b = 1\text{min}^{-1}$. The peak corresponds to sizes less than $20mer$. Second Figure in semi-log scale, to point out the difference of magnitude between the two regions $[0, 20]mer$ and $[20, 200]mer$.

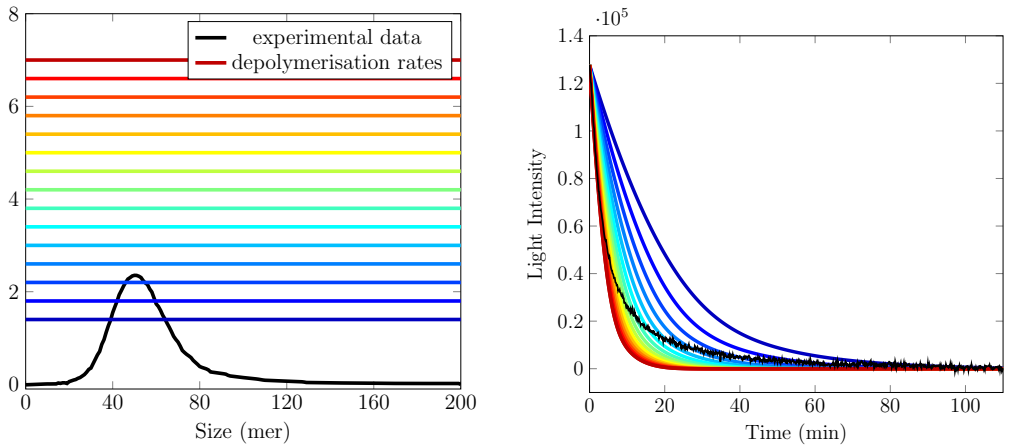
A second possible interpretation is to think that the peak of the estimator does not correspond to the peak we can see in SEC measurements and it may concern sizes up to $20mer$, instead. We recall that the SEC device cannot detect aggregates composed by less than 20 monomers. For instance, if we take $b = 1\text{min}^{-1}$ we would have an initial condition that can illustrate this case.

We present in Figure 18 such an initial condition. We can notice that the maximum of oligomer concentrations for sizes bigger than $20mer$ is much smaller than the value of the peak. This would imply that the concentration of oligomers measured by SEC is negligible compared to a high concentration of (hidden) small oligomers: this is barely plausible.

On the contrary, let us assume that we trust completely the SEC data and that those data represent the overall distribution, i.e., we extend the data by zero in the region $[0, 20]mer$. Having fixed the initial condition, our problem can thus be seen as a parameter identification problem. This problem can be presented as

| Estimating the depolymerisation rate $b$, appearing in the definition of the model dynamics $A$, from given measurements $z$ generated through time $t \in [0, \tau]$, knowing the initial condition $\breve{u}_0$ and the model of observation operator $C$. |
|---|

To have a rough approximation of $b$, we run the direct model for several choices of $b$




(a) (b)

Figure 19: Influence of the depolymerisation rate on the measurement. Left: the initial condition (black line), given by the SEC data, and several constant functions $b$. Right: experimental SLS data (black line) and the second moment observations associated to the $b$ of the same colour on the left figure.

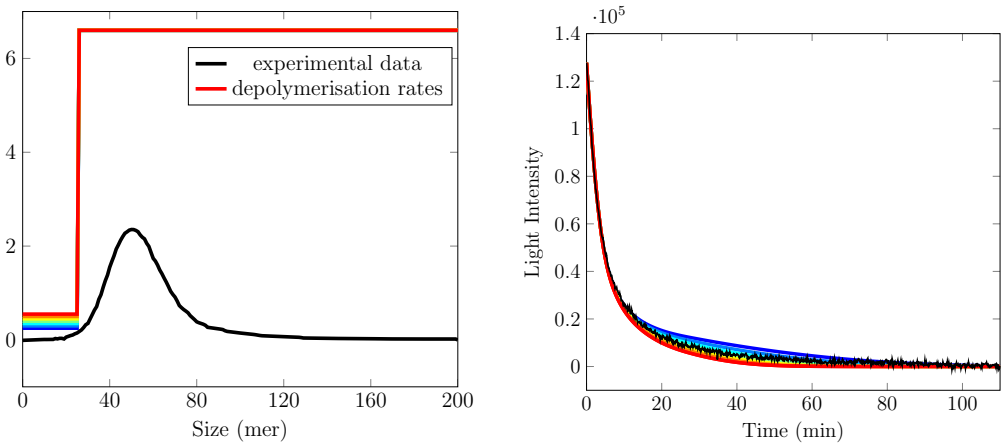
and then, for each of these choices, we compare the second moment generated with the operator $C$ to the SLS data $z$. We can see in Figure 19 the result of this analysis when we choose $b$ as a constant function with values varying between $1\text{min}^{-1}$ and $7\text{min}^{-1}$. We came to the conclusion that, if we assume the initial condition in Figure 19a, the solution of the parameter identification problem is not a constant function.

This conclusion follows from the fact that – whenever we consider two rates $b_1 < b_2$, – we have $(y - b_1 t)^n - (y - b_2 t)^n > 0$ and consequently

$$z_2 = Cu_2 = \int_{b_2 t}^{\ell} (y - b_2 t)^n u_0(y) dy \leq \int_{b_1 t}^{\ell} (y - b_1 t)^n u_0(y) dy = Cu_1 = z_1. \tag{30}$$

Specifically, all the curves start from the same value and then they do not cross anymore. Anyway we can see in Figure 19b that experimental data intersect all the synthetic observations. We deduce that it is not possible to define a constant parameter $b \in [1, 7]\text{min}^{-1}$, solution of the parameter identification problem. A sensitivity analysis could also be carried out to gain more insights, see [4, 1]. Furthermore, Inequality (30) implies that any value $b \in (0, 1) \cup (7, \infty)\text{min}^{-1}$ would lead to an observation far from the experimental data.

We have so concluded that, if we take the curve in Figure 19a as the initial oligomer distribution, we need to consider a size-dependent depolymerisation rate. Let us discuss a simple case in which we distinguish two depolymerisation rates, one for small aggregates and one for big aggregates. For instance, let us define $b$ as a piecewise function that takes only two values, $b(x) = b_1 I_{x<a} + b_2 I_{x>a}$. Running our direct model, we notice that the parameters $b_1$ and $b_2$ are associated to the slopes at the end and at the beginning of the




(a) (b)

Figure 20: On the left we have the initial condition (black line), given by the SEC data, and several piecewise constant functions $b(x)$. On the right we present the experimental SLS data (black line) and the second moment observations associated to the $b$ of the same colour.

data $z$, respectively. Furthermore, the size $a$ is associated to the point of slope change. After testing several possibilities, we have chosen $a = 25mer$ and $b_2 = 6.6\text{min}^{-1}$. In particular, we can see in Figure 20 that if we consider $b(x) = b_1 I_{x<25} + 6.6 I_{x>25}$ and $b_1$ varying between $0.25\text{min}^{-1}$ and $0.55\text{min}^{-1}$ we can well approximate the experimental measurements. This simple model of parameter variation therefore illustrates the idea that a low depolymerisation rate for small aggregates and a high depolymerisation rate for big aggregates allows a better agreement with the data.

To conclude, this modelling choice for the variation of $b$ gives a first insight of the expected behaviour of the parameter and will be refined in future works by fully solving a complete parameter identification problem. In this respect, our approach allows easily to consider an initial condition estimation problem associated to the parameter identification problem and then apply the variational approach on this new problem. Specifically, we consider an augmented state $u^a$ composed by the state function $u$ and the parameters $u^a = (u; b)$. The dynamics of this new variable is given by

$$\dot{u^a} = \begin{pmatrix} \dot{u} \\ \dot{b} \end{pmatrix} = \begin{pmatrix} A(u,b) \\ 0 \end{pmatrix}$$

and the initial condition of the system is $u^a(0) = (u(0); b(0))$. We can use the SEC data to fix an *a priori* on the initial condition so that only the parameters remain unknown. Keeping the same notation as before, we call our target $\breve{\xi}^a = (0; b)$. We remark that this new model presents the additional difficulty of the nonlinearity. A common strategy to overcome this difficulty is to approximate this model replacing the model operator by its tangent.

### 4.4 Discussion

Departing from our general methodology, we were able to estimate the *shape* of the initial size distribution of ovPrP oligomers under the assumption of a constant depolymerisation rate. We then compared it to the experimental distribution obtained with the SEC device (Figure 17), and the discrepancy between these two distributions led us to revisit our previous assumptions.

To explain this discrepancy, two hypotheses could be evocated. The first one is a very important underestimation of the amount of small oligomers assemblies by SEC techniques (Figure 18). The second possibility is a higher depolymerisation rate for large PrP assemblies compared to smaller one (Figure 20). According to SEC specification and methodology the first hypothesis can be excluded. Therefore, lower stability of large assemblies appears as the best possible explanation.

The fact that large assemblies present lower stability than small oligomers could have an important biological implication. Indeed it is commonly admitted that small oligomers constitute the most cytotoxic species during prion misfolding process [26]. Therefore the low stability of large assemblies could make an accumulation of lower molecular weight assemblies and contribute to increase the toxicity level.

## Conclusion

In this paper, we have defined the inverse problem of estimating the initial condition of a dynamical system – whose evolution is described by a transport equation – given the measurements of the second moment over time. We have described two possible approaches to solve this problem. The first belongs to the family of kernel regularisation methods, and allowed us to define a good strategy that exploits all the features of the model and deals precisely with the regularity of the functions. We have introduced this strategy in a case of constant transport velocity both to provide the method guidelines and to give an insight into the properties of the model and the relations between the state function, its moments and the transport velocity. However, since it has been designed for this very specific case, this method lacks of flexibility.

The second approach belongs to the family of data assimilation methods. The inverse problem is written in terms of operators and we obtain a general formalism that can be applied directly on a variety of models. We have seen how this approach is equivalent to the first approach in the case of constant transport velocity and we refer to this second approach to address the more general case of variable velocity and a priori information on the data. The pure depolymerisation problem that we have presented in this paper is a very specific case with relatively narrow possible applications. However, the highly flexible second approach is fundamental and easily adaptable to any more complex situations, and to begin with, the full polymerisation-depolymerisation system given by the Liftshitz-Slyozov model (1). We have also briefly explained in the previous section how we can use the same strategy to solve a problem of parameter identification, see also [21]. With no more effort we can treat the case of multiple measurements. It would be enough to define

an observation operator that – applied on the state function – returns the concatenation of the measurements. This perspective is particularly interesting since we have seen above an example of how by SEC and by SLS it is possible to get several measurements on the same system and the more observations we have the better we can estimate the solution of the inverse problem.

The two methods have been numerically tested and compared on synthetic data. The results of these numerical estimations are in agreement with theoretical estimates. We remark that the estimations we get can take negative values. This is an expected behaviour because we look for an estimation in $\boldsymbol{L}^p$-norm and we do not enforce positivity constraints. A possible improvement for the variational approach would be to either do a constrained optimisation or parametrise the state function to guarantee its positivity at all times.

In the last section we have presented and discussed our inverse problem methodology applied to experimental data of ovPrP oligomers, and this study exemplified the flexibility of the data assimilation framework. We were led to the conclusion that most probably the smaller polymers are more stable than the larger ones. To support this conclusion, further experiments have to be carried out. The simultaneous measurement of the first moment, *i.e.* of the total polymerised mass (by ThT), and of the second moment (by SLS) should be much more informative, and would lead to interesting extensions of our approach. This is a direction for future work.

## Acknowledgments

The research of M. Doumic and part of the research of A. Armiento are supported by the ERC Starting Grant SKIPPER$^{AD}$ (number 306321).

## Appendices

## A Mathematical proofs

In order to prove Proposition 1 we first recall a classical lemma on convolution products.

**Lemma 3**
*Let $n \in \mathbb{N}$, $p \geq 1$, $\alpha \in (0,1)$, the function $\rho \in \mathcal{C}_c^\infty(\mathbb{R})$ and the coefficient $m$ satisfying Assumptions (9). We define the function $\rho_\alpha(x) = \frac{1}{\alpha}\rho(\frac{x}{\alpha})$.*

*i) If the function $f$ is in $\boldsymbol{W}^{1,p}(\mathbb{R}_+)$, we have*

$$\|f - \rho_\alpha * f\|_{\boldsymbol{L}^p(\mathbb{R}_+)} \leq c_1 \alpha \|f\|_{\boldsymbol{W}^{1,p}(\mathbb{R}_+)}, \tag{31}$$

*where $c_1 = \|x\rho\|_{\boldsymbol{L}^1(\mathbb{R})}$*

*ii) Let $n \leq m$, if the function $f$ is in $\boldsymbol{W}^{n+1,p}(\mathbb{R}_+)$, we have*

$$\|f - \rho_\alpha * f\|_{\boldsymbol{L}^p(\mathbb{R}_+)} \leq c_2 \alpha^{n+1} \|f\|_{\boldsymbol{W}^{n+1,p}(\mathbb{R}_+)}, \tag{32}$$

*where $c_2 = \frac{1}{n!}\|x^{n+1}\rho(x)\|_{\boldsymbol{L}^1(\mathbb{R})}$.*

*iii) Furthermore, we have*

$$\|\rho_\alpha * f^{(n)}\|_{\boldsymbol{L}^p(\mathbb{R}_+)} = \|\rho_\alpha^{(n)} * f\|_{\boldsymbol{L}^p(\mathbb{R}_+)} \leq c_3 \alpha^{-n} \|f\|_{\boldsymbol{L}^p(\mathbb{R}_+)}, \tag{33}$$

*where* $c_3 = \|(\rho^{(n)})_\alpha\|_{\boldsymbol{L}^1(\mathbb{R})}$.

*iv) Given* $s \in [0,1)$, *if the function* $f$ *is in* $\boldsymbol{W}^{-s,p}(\mathbb{R}_+)$ *and* $\rho$, $\rho^{'} \in \boldsymbol{L}^1(\mathbb{R})$, *we have*

$$\|\rho_\alpha * f\|_{\boldsymbol{L}^p(\mathbb{R}_+)} \leq c_4 \alpha^{-s} \|f\|_{\boldsymbol{W}^{-s,p}(\mathbb{R}_+)}, \tag{34}$$

*where* $c_4$ *depends on* $\|\rho\|_{\boldsymbol{L}^1(\mathbb{R})}$, $\|\rho^{'}\|_{\boldsymbol{L}^1(\mathbb{R})}$.

*v) Given* $s \in [0,1)$, *if the function* $f$ *is in* $\boldsymbol{W}^{-s,p}(\mathbb{R}_+)$ *and* $\rho^{(n)}$, $\rho^{(n+1)} \in \boldsymbol{L}^1(\mathbb{R}_+)$, *we have*

$$\|\rho_\alpha * f^{(n)}\|_{\boldsymbol{L}^p(\mathbb{R}_+)} \leq c_5 \alpha^{-(n+s)} \|f\|_{\boldsymbol{W}^{-s,p}(\mathbb{R}_+)}. \tag{35}$$

*where* $c_5$ *depends on* $\|\rho^{(n)}\|_{\boldsymbol{L}^1(\mathbb{R})}$, $\|\rho^{(n+1)}\|_{\boldsymbol{L}^1(\mathbb{R})}$.

Let us now state and prove Proposition 1.

**Proposition 4 (Proposition 1)**
*Let* $1 \leq p < \infty$, $n \in \mathbb{N}$, $0 \leq s < 1$ *and the function* $\Psi_\tau u_0$ *defined in Equation* (6). *Let* $\Psi_\tau \breve{u}_0 \in \boldsymbol{W}^{m+n+2,p}([0,\tau])$, *with* $m$ *defined as in Equation* (9). *Let* $z \in \boldsymbol{W}^{-s,p}([0,\tau])$ *a measurement of the* $n$*-th momentum* $\Psi_\tau \breve{u}_0$ *such that* $\tau \geq \frac{\ell}{b}$ *and*

$$\|z - \Psi_\tau \breve{u}_0\|_{\boldsymbol{W}^{-s,p}([0,\tau])} \leq \varepsilon.$$

*The following relation holds true*

$$\breve{u}_0(x) = \frac{1}{n!(-b)^{n+1}} \frac{d^{n+1}}{dt^{n+1}} \Psi_\tau \breve{u}_0 \left(\frac{x}{b}\right). \tag{36}$$

*Let* $\rho$ *defined by Equation* (9) *and* $\rho_\alpha$ *by Equation* (10), *with* $\alpha \in (0,1)$. *We consider*

$$\hat{u}_0^{\varepsilon,\alpha} = \frac{d^{n+1}}{dx^{n+1}} \rho_\alpha * \left( \frac{1}{n!(-b)^{n+1}} z\left(\frac{x}{b}\right)\right) \tag{37}$$

*as approximation of* $\breve{u}_0$. *Then the following estimation is of optimal order in the sense of [12]*

$$\|\hat{u}_0^{\varepsilon,\alpha} - \breve{u}_0\|_{\boldsymbol{L}^p([0,\ell])} \leq \Theta\left(\frac{\varepsilon}{\alpha^{n+s+1}} + \alpha^{m+1}\right) = F_\varepsilon(\alpha), \tag{38}$$

*where the constant* $\Theta$ *depends on* $\|\Psi_\tau \breve{u}_0\|_{\boldsymbol{W}^{m+n+2,p}([0,\tau])}$, $\|x^{m+1}\rho\|_{\boldsymbol{L}^1(\mathbb{R})}$, $\|\rho^{(n)}\|_{\boldsymbol{L}^1(\mathbb{R})}$, $\|\rho^{(n+1)}\|_{\boldsymbol{L}^1(\mathbb{R})}$.

*Proof.* We start by defining the function

$$\hat{u}_0^\alpha = \rho_\alpha * u_0.$$

By the triangle inequality for $\boldsymbol{L}^p$-norm we have

$$\|\hat{u}_0^{\varepsilon,\alpha} - u_0\|_{\boldsymbol{L}^p([0,\ell])} \leq \|\hat{u}_0^{\alpha} - u_0\|_{\boldsymbol{L}^p([0,\ell])} + \|\hat{u}_0^{\varepsilon,\alpha} - \hat{u}_0^{\alpha}\|_{\boldsymbol{L}^p([0,\ell])}.$$

Let us consider the two terms on the right-hand separately.

The first term is $\|\hat{u}_0^{\alpha} - u_0\|_{\boldsymbol{L}^p([0,\ell])} = \|\rho_\alpha * u_0 - u_0\|_{\boldsymbol{L}^p([0,\ell])}$. By using Inequality (32), we obtain

$$\|\hat{u}_0^{\alpha} - u_0\|_{\boldsymbol{L}^p([0,\ell])} \leq \gamma_1 \alpha^{(m+1)} \|u_0\|_{\boldsymbol{W}^{m+1,p}([0,\ell])},$$

with $\gamma_1 = \frac{1}{m!}\|x^{(m+1)}\rho\|_{\boldsymbol{L}^1(\mathbb{R})}$. While the second term is

$$\begin{aligned}
\|\hat{u}_0^{\varepsilon,\alpha} - \hat{u}_0^{\alpha}\|_{\boldsymbol{L}^p([0,\ell])} &= \|\rho_\alpha * \hat{u}_0^{\varepsilon} - \rho_\alpha * u_0\|_{\boldsymbol{L}^p([0,\ell])} \\
&= \|\rho_\alpha * (\hat{u}_0^{\varepsilon} - u_0)\|_{\boldsymbol{L}^p([0,\ell])} \\
&= \frac{1}{n!(-b)^{n+1}} \left\| \rho_\alpha * \frac{d^{n+1}}{dt^{n+1}} (z - \Psi_\tau \breve{u}_0) \right\|_{\boldsymbol{L}^p([0,\ell])}.
\end{aligned}$$

By recalling Inequality(35) , with $f = z - \Psi_\tau \breve{u}_0$, we have

$$\|\hat{u}_0^{\varepsilon,\alpha} - \hat{u}_0^{\alpha}\|_{\boldsymbol{L}^p([0,\ell])} \leq \frac{1}{n!(-b)^{n+1}} \gamma_2 \alpha^{-(n+1+s)} \|z - \Psi_\tau \breve{u}_0\|_{\boldsymbol{W}^{-s,p}([0,\ell])}.$$

In conclusion, we obtain

$$\begin{aligned}
\|\hat{u}_0^{\varepsilon,\alpha} - u_0\|_{\boldsymbol{L}^p([0,\ell])} &\leq \gamma_1 \alpha^{(m+1)} \|u_0\|_{\boldsymbol{W}^{m+1,p}([0,\ell])} + \frac{1}{n!(-b)^{n+1}} \gamma_2 \alpha^{-(n+1+s)} \|z - \Psi_\tau \breve{u}_0\|_{\boldsymbol{W}^{-s,p}([0,\tau])} \\
&\leq \gamma_1 \frac{1}{n!(-b)^{n+1}} \alpha^{(m+1)} \|\Psi_\tau \breve{u}_0\|_{\boldsymbol{W}^{n+m+2,p}([0,\tau])} \\
&\quad + \frac{1}{n!(-b)^{n+1}} \gamma_2 \alpha^{-(n+1+s)} \|z - \Psi_\tau \breve{u}_0\|_{\boldsymbol{W}^{-s,p}([0,\tau])} \\
&\leq \Theta \left( \alpha^{m+1} + \frac{\varepsilon}{\alpha^{n+1+s}} \right),
\end{aligned}$$

where $\Theta = \frac{1}{n!(-b)^{n+1}} \max\{\gamma_1 \|\Psi_\tau \breve{u}_0\|_{\boldsymbol{W}^{m+n+2,p}([0,\tau])}, \gamma_2\}$. ■

We now turn to Proposition 2.

**Proposition 5 (Proposition 2)**
*For any $z \in L^2([0,\tau])$, there exists a unique minimiser $\bar{\xi}$ for $J(\xi)$ defined by*

$$J(\xi) = \frac{\alpha^2}{2} \|\xi\|^2_{L^2([0,\ell])} + \frac{1}{2} \int_0^\tau |z(t) - \Psi_\tau(\xi)|^2 dt,$$

*and $\bar{\xi} \in \boldsymbol{H}^{n+1}([0,\ell])$. If moreover $\breve{\xi} \in \boldsymbol{H}^{n+1}([0,\ell])$ with $\breve{\xi}(0) = \cdots = \breve{\xi}^{(n)}(0) = 0$, we have the following estimate*

$$\|\breve{\xi} - \bar{\xi}\|_{\boldsymbol{L}^2([0,\ell])} \leq \frac{1}{\alpha} \|z - \Psi_\tau(\breve{\xi})\|_{\boldsymbol{L}^2([0,\tau])} + \alpha \|\breve{\xi}\|_{\boldsymbol{H}^{n+1}(0,\ell)}.$$

*Proof.* This result is based on general inequalities for Tikhonov method, that we recall below.

**Lemma 6 (Estimates for Tikhonov Regularisation)**
*Let $K : \mathcal{U} \to \mathcal{Y}$ a compact injective operator between two Hilbert spaces $\mathcal{U}$ and $\mathcal{Y}$, with norms $\|\cdot\|_{\mathcal{U}}$ and $\|\cdot\|_{\mathcal{Y}}$, $K^* : \mathcal{Y} \to \mathcal{U}$ its adjoint (relatively to these norms) and $K^\dagger$ its Moore-Penrose pseudo-inverse. Let $y_\varepsilon \in \mathcal{Y}$. Let $u_{\varepsilon,\alpha} \in \mathcal{U}$ the unique solution of*

$$(K^*K + \alpha^2)u_{\varepsilon,\alpha} = K^* y_\varepsilon. \tag{39}$$

*Then $u_{\varepsilon,\alpha}$ is also the unique minimiser of the following functional*

$$J_K(u) := \frac{1}{2}\|Ku - y_\varepsilon\|_{\mathcal{Y}}^2 + \frac{\alpha^2}{2}\|u\|_{\mathcal{U}}^2.$$

*Moreover, if $u_{\varepsilon,\alpha} \in Ran(K^*)$ and we have*

$$\|Ku_{\varepsilon,\alpha}\|_{\mathcal{Y}} \leq \|y_\varepsilon\|_{\mathcal{Y}}, \qquad \|x_{\varepsilon,\alpha}\|_{\mathcal{U}} \leq \frac{1}{\alpha}\|y_\varepsilon\|_{\mathcal{Y}}.$$

*If moreover $y \in Ran(K)$, denoting $Ku = y$ and $u_\alpha$ the solution to (39) with $y_\varepsilon = y$, we have $u_\alpha \in Ran(K^*K)$ and*

$$\|u_\alpha\|_{\mathcal{U}} \leq \|x\|_{\mathcal{U}},$$

*If moreover $u \in Ran(K^*)$, we have*

$$\|u_\alpha - u\|_{\mathcal{U}} \leq \alpha\|K^{*\dagger}u\|_{\mathcal{Y}},$$

*if moreover $x \in Ran(K^*K)$, we have*

$$\|u_\alpha - u\|_{\mathcal{U}} \leq \alpha^2\|(K^*K)^\dagger u\|_{\mathcal{U}}.$$

We now apply this result to $\mathcal{U} = \boldsymbol{L}^2([0,\ell])$, $\mathcal{Y} = \boldsymbol{L}^2([0,\tau])$ and $K = \Psi_\tau$. We take $u = \breve{\xi}$ and $y_\varepsilon$ to be the measurement function $t \in [0,\tau] \to z(t)$, and $y = \Psi_\tau\breve{\xi}$.

$$\mathrm{Ran}\Psi_\tau^n = \left\{u \in \boldsymbol{H}^{n+1}([0,\tau]), u(\tau) = \cdots = u^{(n)}(\tau) = 0\right\},$$

and

$$\mathrm{Ran}\Psi_\tau^{*n} = \left\{u \in \boldsymbol{H}^{n+1}([0,\ell]), u(0) = \cdots = u^{(n)}(0) = 0\right\}.$$

Lemma 6 gives us that $u_{\varepsilon,\alpha} = \bar{\xi}$ is the unique minimizer for $J$. We decompose as is well-known

$$\|\breve{\xi} - \bar{\xi}\|_{\boldsymbol{L}^2([0,\ell])} \leq \|\breve{\xi} - u_\alpha\|_{\boldsymbol{L}^2([0,\ell])} + \|u_\alpha - u_{\alpha,\varepsilon}\|_{\boldsymbol{L}^2([0,\ell])}.$$

In Proposition 2, the assumptions on $\breve{\xi}$ mean that $\breve{\xi} \in Ran(K^*)$, hence

$$\|u_\alpha - \breve{\xi}\|_{\mathcal{U}} \leq \alpha\|K^{*\dagger}\breve{\xi}\|_{\mathcal{Y}} \leq \alpha\|\breve{\xi}\|_{\boldsymbol{H}^{n+1}([0,\ell])}.$$

Concerning the term $\|\breve{\xi} - u_\alpha\|_{\boldsymbol{L}^2([0,\ell])}$, we apply the second inequality of Lemma 6 to Equation (39) with $y_\varepsilon$ replaced by $y_\varepsilon - y$, for which $u_{\varepsilon,\alpha} - u_\alpha$ is a solution, and find

$$\|\breve{\xi} - u_\alpha\|_{\boldsymbol{L}^2([0,\ell])} \leq \frac{1}{\alpha}\|y_\varepsilon - y\|_{\mathcal{Y}} = \frac{1}{\alpha}\|z - \Psi_\tau(\breve{\xi})\|_{\boldsymbol{L}^2([0,\tau])}.$$

This ends the proof. ∎

# B Oligomer formation protocol

The protocol to form and make measurements on oligomer systems – previously described in [10] – consists in inducing a partial unfolding of full-length ovine PrP protein by thermal treatment. This partial unfolding leads to generation of three distinct oligomers that can be purified by size exclusion chromatography, for further investigation we refer to [11]. The conversion of PrP into the oligomeric form is performed in $20mM$ sodium citrate buffer (pH 3.40). The PrP – at a final concentration of $50\mu M$ – is incubated in a Perkin Elmer GenAmp2400 thermocycler at $65°C$ for two hours. Homogeneous fractions of oligomers are then collected after separation by size exclusion chromatography (SEC), as first described in [10]. The SEC is performed at $20°C$ using a TSK 4000SW ($7mm * 600mm$) gel-filtration column (Interchim, Montluçon, France) with $20mM$ sodium citrate (pH 3.35). Protein elution is monitored by UV absorption at $280nm$. The size distribution of oligomer assemblies has been determined by the SEC device coupled with the static light scattering device. Depolymerisation kinetics are performed with an in-lab device using $407nm$ laser beams in a $2mm$-path-length quartz cuvette. Kinetic experiments are performed according to a standardise methodology, as reported in [11]: $72°C$ in $20mM$ sodium citrate buffer (pH 3.40). The oligomer concentration has been fixed at $3\mu M$.

# References


[1] H. T. Banks, M. Doumic, C. Kruse, S. Prigent, and H. Rezaei. Information content in data sets for a nucleated-polymerisation model. *Journal of Biological Dynamics*, 9(1):172—197, January 2015.

[2] H. T. Banks, Marie Doumic-Jauffret, and Carola Kruse. Efficient numerical schemes for nucleation-aggregation models: Early steps. Mar 2014.

[3] H.T. Banks. *A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering*. CRC Press, 2012.

[4] H.T. Banks and D.M. Bortz. A parameter sensitivity methodology in the context of hiv delay equation models. *Journal of Mathematical Biology*, 50(6):607–625, 2005.

[5] Kiersten M. Batzli and Brian J. Love. Agitation of amyloid proteins to speed aggregation measured by ThT fluorescence: A call for standardization. *Materials Science and Engineering: C*, 48(0):359 – 364, 2015.

[6] Johann Baumeister and Antonio Leitão. *Topics in inverse problems*. Publicações Matemáticas do IMPA. [IMPA Mathematical Publication]. Instituto Nacional de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, 25º colóquio brasileiro de matemática. [25th brazilian mathematics colloquium] edition, 2005.

[7] Alain Bensoussan. *Filtrage Optimal des Systèmes Lineaires*. Méthodes Mathématiques de l'informatique. Dunod, 1971.

[8] M. F. Bishop and F. A. Ferrone. Kinetics of nucleation-controlled polymerization. a perturbation treatment for use with a secondary pathway. *Biophys J.*, 46(5):631–644, November 1984.

[9] M. Doumic, T. Goudon, and T. Lepoutre. Scaling limit of a discrete prion dynamics model. *Communications in Mathematical Sciences*, 7(4):839–865, 2009.

[10] F. Eghiaian, T. Daubenfeld, Y. Quenet, M. van Audenhaege, AP. Bouin, G. van der Rest, J. Grosclaude, and H. Rezaei. Diversity in prion protein oligomerization pathways results from domain expansion as revealed by hydrogen/deuterium exchange and disulfide linkage. *Proc Natl Acad Sci U S A.*, 104(18):7414–7419, may 2007.

[11] F. Eghiaian, T. Daubenfeld, Y. Quenet, M. van Audenhaege, AP. Bouin, G. van der Rest, J. Grosclaude, and H. Rezaei. The oligomerization properties of prion protein are restricted to the H2H3 domain. *The FASEB Journal*, 24(9):3222–3231, sep 2010.

[12] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, 1996.

[13] Barnabas James Gilbert. The role of amyloid $\beta$ in the pathogenesis of Alzheimer's disease. *J Clin Pathol*, 66:362–366, March 2013.

[14] Lifshitz I.M. The kinetics of precipitation from supersaturated solid solutions. *Journal of physics and chemistry of solids*, 19:35–50, 1961.

[15] Francois-Xavier Le Dimet and O Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus A*, 38(2):97–110, 1986.

[16] Randall J. LeVeque. *Finite-Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[17] M. Nussbaum. Asymptotic equivalence of density estimation and white noise. *Ann. Statist.*, 24:2399–2430, 1996.

[18] Michael Nussbaum and Sergej V Pereverzev. *The Degree of Ill Posedness in Stochastic and Deterministic Noise Model*. WIAS, 1999.

[19] Sophocles J. Orfanidis. *Introduction to Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[20] Sian-Yang Ow and Dave E. Dunstan. A brief overview of amyloids and alzheimer's disease. *Protein Science*, 23(10):1315–1331, 2014.

[21] Antoine Perasso, Béatrice Laroche, Yacine Chitour, and Suzanne Touzeau. Identifiability analysis of an epidemiological model in a structured population. *Journal of Mathematical Analysis and Applications*, 374(1):154 – 165, 2011.

[22] S. Prigent, A. Ballesta, F. Charles, N. Lenuzza, P. Gabriel, L. M. Tine, H. Rezaei, and M. Doumic. An Efficient Kinetic Model for Assemblies of Amyloid Fibrils and Its Application to Polyglutamine Aggregation. *PLoS ONE*, 7(11), 11 2012.

[23] Marina Ramìrez-Alvarado, Jane S. Merkel, and Lynne Regan. A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro. *Proceedings of the National Academy of Sciences*, 97(16):8979–8984, 2000.

[24] Human Rezaei, Frédéric Eghiaian, Javier Perez, Bénédicte Doublet, Yvan Choiset, Thomas Haertle, and Jeanne Grosclaude. Sequential Generation of Two Structurally Distinct Ovine Prion Protein Soluble Oligomers Displaying Different Biochemical Reactivities. *Journal of Molecular Biology*, 347:665–679, apr 2005.

[25] D Simon. *Optimal State Estimation: Kalman, H Infinity, And Nonlinear Approaches*. Springer, 2006.

[26] S. Simoneau, H. Rezaei, N. Salès, G. Kaiser-Schulz, M. Lefebvre-Roque, C. Vidal, JG. Fournier, J. Comte, F. Wopfner, J. Grosclaude, H. Schätzl, and C. Lasmézas. In vitro and in vivo neurotoxicity of prion protein oligomers. *PLoS Pathog*, 3(8), Aug 2007.

[27] Daniel Some. Light-scattering-based analysis of biomolecular interactions. *Biophysical Reviews*, 5(2):147–158, 2013.

[28] M.P. Wand and M.C. Jones. *Kernel Smoothing*. CRC Press, 1994.

[29] W-F Xue, S W Homans, and S E Radford. Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly. *PNAS*, 105:8926–8931, 2008.

[30] W-F Xue and S E Radford. An imaging and systems modeling approach to fibril breakage enables prediction of amyloid behavior. *Biophys. Journal*, 105:2811–2819, 2013.

CHAPITRE 5

# Complements on data assimilation strategies for infinite-dimensional operators

In this chapter we focus on *data assimilation methods* for PDE models. These methods can be divided into two main classes : variational methods and sequential methods. The former are based on the minimisation of a variational criterion, the latter define an estimator that is able to correct its trajectory sequentially, thereby reducing the estimation error when a new observation is available.

Two of the best known methods in these classes are the variational *4d-Var* method and the sequential *Kalman Filter* method. In Chapter 4, we presented the 4d-Var method without accounting for a model error, that is considered in this chapter. An introduction to the Kalman method was provided in Chapter 1 in the case of ODE models considering a stochastic setting. In this chapter we present the method for PDE models in a deterministic setting and we provide its stochastic interpretation.

We start by recalling the state-space formalism used in Chapter 4. We present the 4d-Var method more in detail and we conclude by introducing the Kalman approach and its application to initial condition estimation problems.

## 5.1 State-space formalism and model error

The state-space formalism is particularly useful to describe data assimilation methods in a compact way. In fact, we can handle a variety of applications with slight differences in the formalism.

To start with, let us briefly recall the formalism introduced in the last chapter. The key idea is to take the state of the model as a function of time with values in the state space. Formally, we consider the state space $\mathcal{U} = \mathbb{L}^2([0,\ell])$, which is the usual Hilbert space of square integrable functions. Consequently, the state function is defined as follows

$$u : \left| \begin{array}{ccc} [0,\tau] & \longrightarrow & \mathcal{U} \\ t & \longmapsto & u(t) : \left| \begin{array}{ccc} [0,\ell] & \longrightarrow & \mathbb{R} \\ x & \longmapsto & u(x,t) \end{array} \right. \end{array} \right.$$

### 5.1.1 Model operator

In the state-space formalism, the model is given by the *model operator*. In our case, the model operator is the following unbounded linear operator

$$A : \left| \begin{array}{ccc} \mathcal{D}(A) \subset \mathbb{L}^2([0,\ell]) & \longrightarrow & \mathbb{L}^2([0,\ell]) \\ f & \longmapsto & \dfrac{\mathrm{d}(bf)}{\mathrm{d}x}, \end{array} \right.$$

where $b \in \mathbb{H}^1([0,\ell])$ is a known function. The *domain of* $A$ is defined as follows

$$\mathcal{D}(A) = \{f \in \mathbb{H}^1([0,\ell]) \mid f(\ell) = 0\}.$$

It is easy to prove that $\mathcal{D}(A)$ is dense in $\mathbb{L}^2([0,\ell])$. It is indeed sufficient to note that it contains $C_0^\infty([0,\ell])$, which is dense in $\mathbb{L}^2([0,\ell])$ [31].

Our model (3.11) can thus be written in the state-space formalism as

$$\left\{ \begin{array}{l} \dfrac{\mathrm{d}u}{\mathrm{d}t} = Au, \\ u(0) = u_0. \end{array} \right. \tag{5.1}$$

As mentioned in the last chapter, the 4d-Var estimator can be defined by introducing the adjoint variable. The dynamics of the adjoint variable are determined by the adjoint of the model operator. The operator $A$ being densely defined, it is indeed possible to consider its *adjoint model operator* $A^*$. In a general formulation, the adjoint model operator is an unbounded operator

$$A^* : \mathcal{D}(A^*) \subset \mathbb{L}^2([0,\ell]) \longrightarrow \mathbb{L}^2([0,\ell]),$$

characterised by the following relation

$$\langle Au, v\rangle = \langle u, A^*v\rangle \quad \forall u \in \mathcal{D}(A), \forall v \in \mathcal{D}(A^*),$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product and $\|\cdot\|$ the $\mathbb{L}^2$ norm. The domain of the adjoint model operator is defined by

$$\mathcal{D}(A^*) = \{v \in \mathbb{L}^2([0,\ell]) | \exists c \geq 0 \text{ s. t. } |\langle Au, v\rangle| \leq c\|u\| \quad \forall u \in \mathcal{D}(A)\}.$$

Let us consider a function $v \in \mathbb{L}^2([0,\ell])$ and

$$\phi_v : u \in \mathcal{D}(A) \longmapsto \langle Au, v\rangle.$$

If $v \in \{f \in \mathbb{H}^1([0,\ell]) \mid f(0) = 0\}$, we can write

$$\phi_v(u) = \langle Au, v\rangle = \int_0^\ell \frac{\mathrm{d}(bu)}{\mathrm{d}x} v \mathrm{d}y = [buv]_0^\ell - \int_0^\ell bu\frac{\mathrm{d}v}{\mathrm{d}x}\mathrm{d}y = -\int_0^\ell ub\frac{\mathrm{d}v}{\mathrm{d}x}\mathrm{d}x.$$

Therefore, the adjoint model operator is defines by

$$A^* : \left| \begin{array}{ccc} \mathcal{D}(A^*) \subset \mathbb{L}^2([0,\ell]) & \longrightarrow & \mathbb{L}^2([0,\ell]) \\ f & \longmapsto & -b\dfrac{\mathrm{d}f}{\mathrm{d}x} \end{array} \right.$$

with

$$\mathcal{D}(A^*) = \{f \in \mathbb{H}^1([0,\ell]) \mid f(0) = 0\}.$$

We conclude this section by noticing that the operator $A$ is closed. In fact, the domain $\mathcal{D}(A^*)$ is dense in $\mathbb{L}^2([0,\ell])$, since it contains the dense subspace $C_0^\infty([0,\ell])$. It is then enough to verify that $(A^*)^* = A$ and recall that the adjoint of an operator is always closed [31].

### 5.1.2 Observation operator

To model the measurement process, we consider the functional operator $C$ that, at each function in the state space, associates its moment of order $n$. Let $\mathcal{Z} = \mathbb{R}$ be the observation space, we have

$$C : \left| \begin{array}{ccc} \mathcal{U} & \longrightarrow & \mathcal{Z}, \\ u & \longmapsto & \int_0^\ell x^n u(x)\mathrm{d}x, \end{array} \right. \tag{5.2}$$

which is a time-invariant linear bounded operator such that $\|C\| \leq \ell^{n+\frac{1}{2}}$.

The *adjoint observation operator* $C^*$ is the operator satisfying, for all $v \in \mathbb{R}$,

$$\langle Cu, v\rangle_{\mathbb{R}} = v\int_0^\ell x^n u\mathrm{d}x = \int_0^\ell (vx^n)u\mathrm{d}x = \langle u, vx^n\rangle.$$

Consequently, it is defined as follows

$$C^* : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{L}^2([0,\ell]) \\ r & \longmapsto & f_r : x \mapsto x^n r. \end{array}$$

### 5.1.3 Modelling uncertainties

When we consider a mathematical description of a physical application, we need to take into account several sources of uncertainties. In this chapter, we will see how to treat these uncertainties in a deterministic setting and we provide a description of the stochastic setting. We consider three kinds of error :

1. We can have a *model error*. This kind of error is commonly due to necessary simplifications and to a partial knowledge of the physics observed. The model uncertainty may be modelled as an additive noise $\omega \in \mathbb{L}^2([0,\tau],\mathcal{W})$, affecting the state dynamics as follows
$$\frac{\mathrm{d}u}{\mathrm{d}t} = Au(t) + B\omega(t).$$
For the sake of simplicity we assume $B$ to be a known bounded linear operator, $B \in \mathcal{L}(\mathcal{W},\mathcal{U})$. In Chapter 5, we have performed our analysis in the case $B = 0$.
2. We consider a *noise on the initial condition*. For this reason, we decompose the initial state as
$$u_0 = u_\diamond + \xi,$$
where $u_\diamond \in \mathcal{U}$ – also called state *a priori* – and $\xi \in \mathcal{U}$ denote its known and unknown part, respectively.
3. We take into account some additive *measurement noise* $\chi \in \mathbb{L}^2([0,\tau],\mathcal{Z})$. In our case, the observation space is $\mathcal{Z} = \mathbb{R}$. The noised measurements are given by
$$z(t) = Cu(t) + \chi(t), \qquad t \in [0,\tau].$$

Incorporating these uncertainties in the model, it results
$$\mathcal{M}(\omega,\xi) = \begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}t} = Au + B\omega, \\ u(0) = u_\diamond + \xi. \end{cases}$$

In the following we use the name $\mathcal{M}(\omega,\xi)$ to highlight the dependence of the model on the unknowns $\omega$ and $\xi$.

Let us call $\breve{u}$ the trajectory we want to estimate. The inverse problem, in this more general setting, is

**Inverse Problem** : Estimating $\breve{u}$, the solution of the system $\mathcal{M}(\breve{\omega},\breve{\xi})$, given the measurements
$$z = C\breve{u} + \chi,$$
generated through time $t \in [0,\tau]$.

In the following we show how to treat this problem in a data assimilation framework.

## 5.2 Data assimilation methods

In this section, we present two data assimilation strategies. The first, called 4d-Var, belongs to the class of variational methods since it relies on the minimisation of a quadratic functional. This strategy provides an optimal estimation of the target trajectory, based on the observations collected over a given time window.

The second strategy, called Kalman filtering, belongs to the class of sequential methods. Sequential methods seek to estimate the target trajectory, by filtering the discrepancy between empirical data and theoretical prediction. The Kalman approach is also called optimal filtering, since it builds the estimator that at any time $t$ corresponds to the minimiser of a variational criterion based on the observations up to time $t$.

### 5.2.1 Variational method : 4d-Var

The variational data assimilation method 4d-Var has been widely used in the literature, see for instance [99, 63, 105]. The name stands for "four-dimensional variational method" since it considers a physical state in 3d reality plus observations distributed in time. It provides an estimator that best fits the observations taken for a given time window.

In our paper [6] (Chapter 4) we have shown how it is possible to use either a kernel regularisation method or the 4d-Var data assimilation method to estimate the initial condition of System (3.9).

Now, we want to present the 4d-Var method in the more general setting of a non-null modelling error, $B \neq 0$. This case often occurs since we may have incomplete information on the process or we need to consider a simplified version of the reality in which only the principal features are taken into account.

The method provides the estimations of $\breve{\omega}, \breve{\xi}$, namely $\bar{\omega}$ and $\bar{\xi}$, by minimising a least square criterion. The criterion contains three terms weighting the amount of error associated to the uncertainties $\xi$, $\omega$, $\chi$. The weights in the criterion are given by the self-adjoint, non-negative and invertible operators

$$P_0 \in \mathcal{L}(\mathcal{U}), \quad W \in \mathcal{L}(\mathcal{Z}), \quad Q \in \mathcal{L}(\mathcal{W}).$$

Let us define the following norms and scalar products

$$\begin{array}{ll} \forall u \in \mathcal{U} & \|u\|^2_{P_0^{-1}} = \langle u, u\rangle_{P_0^{-1}} = \langle P_0^{-1}u, u\rangle_{\mathcal{U}}, \\ \forall z \in \mathcal{Z} & \|z\|^2_{W-1} = \langle z, z\rangle_{W^{-1}} = \langle W^{-1}z, z\rangle_{\mathcal{Z}}, \\ \forall \omega \in \mathcal{W} & \|\omega\|^2_{Q^{-1}} = \langle \omega, \omega\rangle_{Q^{-1}} = \langle Q^{-1}\omega, \omega\rangle_{\mathcal{W}}. \end{array}$$

The least square criterion is given by

$$\mathcal{J}_\tau(\omega, \xi) = \frac{1}{2}\|\xi\|^2_{P_0^{-1}} + \frac{1}{2}\int_0^\tau \Big(\|z - C(u)\|^2_{W^{-1}} + \|\omega\|^2_{Q^{-1}}\Big)\mathrm{d}t, \tag{5.3}$$

under the constraint that $u$ is the solution of $\mathcal{M}(\omega, \xi)$. We define

$$(\bar{\omega}, \bar{\xi}) = \underset{(\omega,\xi)}{\arg\min}\, \mathcal{J}_\tau(\omega, \xi). \tag{5.4}$$

The isomorphisms $P_0, W, Q$ weight the natural norms of the state, observation and model noise spaces, respectively. These operators play a crucial role in the estimation since they reflect the confidence we have in the initial condition approximation $u_\diamond$, in the observations $z$ and in the model $A$. In an ideal noiseless and errorless case, $\bar{u} = \breve{u}$. In a more general case, $\bar{u}$ is an estimation of the target solution. As we have shown in the last chapter, in the simple case of scalar weights, the resulting estimation depends on the relative weights of each term with respect to the others. These operators may be associated with a stochastic interpretation [149, 33] as the covariance of the uncertainties. More precisely, $P_0 = \mathrm{Cov}(\xi)$ and, for all times $t \in [0, \tau]$, $W(t) = \mathrm{Cov}(\chi(t))$, $\quad Q(t) = \mathrm{Cov}(\omega(t))$.

In the case of linear model operators and linear observation operators, it is possible to prove that there exists a unique minimiser of $\mathcal{J}_\tau$ under the constraint that $u$ satisfies the model dynamics. To characterise this minimiser, we introduce the *adjoint variable*, namely $q$.

Given the notation $d_u A$, standing for the derivative of the model operator with respect to the $u \in \mathcal{D}(A)$, we define the adjoint of $u$ as the solution of the following dynamical system

$$\begin{cases} \dfrac{\mathrm{d}q}{\mathrm{d}t} + d_u A^* q = -C^* W^{-1} \left(z - Cu\right), \quad t \in [0, \tau] \\ q(\tau) = 0. \end{cases} \tag{5.5}$$

This variable may be seen as the Lagrangian multiplier associated to the minimisation constraint that $u$ is the solution of $\mathcal{M}(\omega, \xi)$.

Let us denote by $\bar{u}$ the solution of $\mathcal{M}(\bar{\omega}, \bar{\xi})$ and $\bar{q}$ its adjoint variable. We can easily define the optimal estimation $\bar{\xi}$ as follows

$$\bar{\xi} = P_0 \bar{q}(0). \tag{5.6}$$

For the sake of completeness, let us detail the derivation of this result. Using the notation $\mathrm{d}_\xi$ and $\mathrm{d}_\omega$ for the derivatives with respect to the initial unknown $\xi$ and the model noise $\omega$, respectively, we have

$$\mathrm{d}_\xi \mathcal{J}_\tau \cdot \delta\xi = \langle P_0^{-1}\xi,\, \delta\xi\rangle + \frac{1}{2}\int_0^\tau d_\xi \langle z - Cu,\, z - Cu\rangle_{W^{-1}} \cdot \delta\xi \mathrm{d}t,$$

We recall that the uncertainties $\xi$ and $\omega$ are assumed to be independent. Let us focus on the second term of this sum. For all $f \in \mathcal{U}$, we denote by $u_f$ the solution of $\mathcal{M}(\omega, f)$. Consequently, we have

$$\begin{aligned} d_\xi \langle z - Cu_\xi,\, z - Cu_\xi\rangle_{W^{-1}} \cdot \delta\xi &= \lim_{h\to 0} \frac{\langle z - Cu_{\xi+h\delta\xi},\, z - Cu_{\xi+h\delta\xi}\rangle_{W^{-1}} - \langle z - Cu_\xi,\, z - Cu_\xi\rangle_{W^{-1}}}{h} \\ &= -2\langle z - Cu_\xi,\, Cd_\xi u(\delta\xi)\rangle_{W^{-1}}, \end{aligned}$$

where $\mathrm{d}_\xi u(\delta\xi)$ is the sensitivity of the trajectory with respect to the initial condition $\xi$. We can notice that $\mathrm{d}_\xi u \cdot \delta\xi$ solves the following system

$$\begin{cases} \dot{y} = d_u A y \\ y(0) = \delta\xi. \end{cases}$$

Substituting this result into the computation of the criterion derivative and taking out the subscript, we have

$$\begin{aligned} \mathrm{d}_\xi \mathcal{J}_\tau \cdot \delta\xi &= \langle P_0^{-1}\xi,\, \delta\xi\rangle - \int_0^\tau \langle C^* W^{-1}(z - Cu),\, d_\xi u(\delta\xi)\rangle \mathrm{d}t \\ &= \langle P_0^{-1}\xi,\, \delta\xi\rangle + \int_0^\tau \langle \dot{q} + d_u A^* q,\, d_\xi u(\delta\xi)\rangle \mathrm{d}t \\ &= \langle P_0^{-1}\xi,\, \delta\xi\rangle + \left[\langle q,\, d_\xi u(\delta\xi)\rangle\right]_0^\tau = \langle P_0^{-1}\xi - q(0),\, \delta\xi\rangle. \end{aligned}$$

This identity being satisfied for all choices of $\delta\xi$, we can conclude Equation (5.6).

Analogously, we can consider the derivative of the criterion with respect to the model noise $\omega$. Let us notice that $\mathrm{d}_\omega u(\delta\omega)$ – which is the sensitivity of the trajectory with respect to the noise $\omega$ – solves the system

$$\left\{\begin{array}{rcl} \dot{y} & = & d_u A y + B\delta\omega, \\ y(0) & = & 0. \end{array}\right.$$

We have

$$\begin{aligned}
\mathrm{d}_\omega \mathcal{J}_\tau \cdot \delta\omega &= \int_0^\tau \Big(\frac{1}{2} d_\omega \langle z - Cu,\, z - Cu\rangle_{W^{-1}} \cdot \delta\omega + \langle Q^{-1}\omega,\, \delta\omega\rangle\Big)\mathrm{d}t \\
&= \int_0^\tau \Big( - \langle C^* W^{-1}(z - Cu),\, d_\omega u(\delta\omega)\rangle + \langle Q^{-1}\omega,\, \delta\omega\rangle\Big)\mathrm{d}t \\
&= \int_0^\tau \Big(\langle \dot{q} + d_u A^* q,\, d_\omega u(\delta\omega)\rangle + \langle Q^{-1}\omega,\, \delta\omega\rangle\Big)\mathrm{d}t \\
&= \langle q,\, d_\omega u(\delta\omega)\rangle_0^\tau + \int_0^\tau \Big( - \langle q,\, d_u A\, d_\omega u(\delta\omega) + B\delta\omega\rangle + \langle d_u A^* q,\, d_\omega u(\delta\omega)\rangle + \langle Q^{-1}\omega,\, \delta\omega\rangle\Big)\mathrm{d}t \\
&= \int_0^\tau \langle -B^* q + Q^{-1}\omega,\, \delta\omega\rangle \mathrm{d}t.
\end{aligned}$$

Therefore, we conclude

$$\bar{\omega}(t) = QB^* q(t), \quad \forall t \in [0, \tau]. \tag{5.7}$$

The variational estimator $\bar{u}$ of $\breve{u}$ is the solution of the *both-ends problem*

$$\left\{\begin{array}{rcl} \dfrac{\mathrm{d}\bar{u}}{\mathrm{d}t} & = & A\bar{u} + BQB^*\bar{q}, \\ \dfrac{\mathrm{d}\bar{q}}{\mathrm{d}t} + A^*\bar{q} & = & -C^* W^{-1}\left(z - C\bar{u}\right), \\ \bar{u}(0) & = & u_\diamond + P_0\bar{q}(0), \\ \bar{q}(\tau) & = & 0. \end{array}\right. \tag{5.8}$$

In practical applications, this system is decoupled and solved iteratively. Let us start by setting $(u^0, q^0) = (u_\diamond, 0)$. Then, for all iterations $k \geq 1$, we solve the following systems. First, we solve the state initial condition system

$$\left\{\begin{array}{rcl} \dfrac{\mathrm{d}u^k}{\mathrm{d}t} & = & Au^k + BQB^* q^{k-1}, \\ u^k(0) & = & u_\diamond + P_0 q^{k-1}(0). \end{array}\right.$$

Then, we solve the adjoint final condition system

$$\left\{\begin{array}{rcl} \dfrac{\mathrm{d}q^k}{\mathrm{d}t} + A^* q^k & = & -C^* W^{-1}\left(z - Cu^k\right), \\ q^k(\tau) & = & 0. \end{array}\right.$$

At each iteration we go forward in time with the state variable and come backward with the adjoint variable. We iterate until convergence of the method, namely until a stopping condition is met. This method provides very good estimations, but it commonly requires a high number of iterations to reach convergence. For an overview of the 4d-Var strategy in a non-linear setting we refer to [42].

### 5.2.2 Kalman Filter

In this section we introduce a sequential data assimilation method known as Kalman filtering. The method was first introduced by R.E. Kalman in 1960 [61]. It aims at estimating the target trajectory $\breve{u}$ in a sequential way. Starting from the known state *a priori*, the Kalman estimator $\hat{u}$ is designed to correct its trajectory in an optimal way and, eventually, reach the target state function. The advantage of using this strategy, rather than 4d-Var, consists in a reduced computational cost. In fact, as we detail in the following, the Kalman estimator coincides with the solution of 4d-Var at the end of the time domain, but it needs just one simulation run.

In Chapter 1, we have already presented this method in a stochastic framework, in the case of finite-dimensional spaces. Now, we treat the method in a deterministic framework for infinite-dimensional spaces. The stochastic framework for infinite-dimensional spaces is in fact not trivial. A formal presentation of it is provided in Section 5.2.3.

We define the Kalman estimator as the variable that, at any time $t \in [0, \tau]$, corresponds to the optimal estimation obtained from the observations up to time $t$. Formally, we define the Kalman estimator as follows

$$\hat{u}(t) = \bar{u}_t(t), \quad \forall t \in [0, \tau], \tag{5.9}$$

where $\bar{u}_t$ is the estimator obtained with the 4d-Var method, by minimising the criterion

$$\mathcal{J}_t(\omega, \xi) = \frac{1}{2}\|\xi\|^2_{P_0^{-1}} + \frac{1}{2}\int_0^t \Big(\|z(s) - Cu(s)\|^2_{W^{-1}} + \|\omega(s)\|^2_{Q^{-1}}\Big)\mathrm{d}s, \tag{5.10}$$

defined over the time window $[0, t]$. We take

$$P_0 \in \mathcal{L}(\mathcal{U}), \quad W \in \mathcal{L}(\mathcal{Z}), \quad Q \in \mathcal{L}(\mathcal{W}) \tag{5.11}$$

self-adjoint, non-negative and invertible operators. As seen in the previous section, the variational minimiser $\bar{u}_t$ together with its adjoint $\bar{q}_t$ solve the both-ends problem

$$\begin{cases} \dfrac{\mathrm{d}\bar{u}}{\mathrm{d}t} &= A\bar{u} + BQB^*\bar{q}, \quad \forall s \in [0, t] \\ \dfrac{\mathrm{d}\bar{q}}{\mathrm{d}t} + A^*\bar{q} &= -C^*W^{-1}\left(z - C\bar{u}\right), \quad \forall s \in [0, t] \\ \bar{u}(0) &= u_\diamond + P_0\bar{q}(0), \\ \bar{q}(t) &= 0. \end{cases}$$

However, in practical applications, this definition cannot be used, since it would require solving a minimisation problem at each time $t$. A second characterisation is thus provided by the following theorem.

**Theorem 4 ([23])**
*For all $\tau > 0$, let us consider the following Riccati differential equation*

$$\begin{cases} \dfrac{\mathrm{d}P(s)}{\mathrm{d}t} &= AP(s) + P(s)A^* - P(s)C^*W^{-1}CP(s) + BQB^*, \quad \forall s \in [0,\tau], \\ P(0) &= P_0. \end{cases} \tag{5.12}$$

*and the dynamical system*

$$\begin{cases} \dfrac{\mathrm{d}\hat{u}(s)}{\mathrm{d}t} &= A\hat{u}(s) + P(s)C^*W^{-1}\big(z(s) - C\hat{u}(s)\big), \quad \forall s \in [0,\tau] \\ \hat{u}(0) &= u_\diamond. \end{cases} \tag{5.13}$$

*i) Let us assume that System*(5.12) *has a solution $P \in C^1([0,\tau];\mathcal{L}(\mathcal{U}))$. Let $(\bar{u},\bar{q})$ be the solution of System* (5.8)*. Then, the variable $\hat{u}$ defined by the following equation*

$$\hat{u}(s) = \bar{u}(s) - P(s)\bar{q}(s) \quad \forall s \in [0,\tau], \tag{5.14}$$

*is the unique solution $\hat{u} \in C^1([0,\tau];\mathcal{U})$ of Equation* (5.13)*.*

*ii) The variable $\hat{u}$, defined by Equation* (5.14)*, is the Kalman estimator.*

*Idea of the proof i)* Let $\widetilde{u}$ be defined as $\widetilde{u} = \bar{u} - P\bar{q}$. Let us compute its dynamics. We will use the notation $\dot{\widetilde{u}}$ to indicate the derivative. By a simple calculation, we obtain

$$\begin{aligned} \dot{\widetilde{u}} &= A\bar{u} + BQB^*\bar{q} - \Big(AP + PA^* - PC^*W^{-1}CP + BQB^*\Big)\bar{q} - P\Big( - A^*\bar{q} - C^*W^{-1}\left(z - C\bar{u}\right)\Big) \\ &= A(\bar{u} - P\bar{q}) + PC^*W^{-1}\big(z - C(\bar{u} - P\bar{q})\big) = A\widetilde{u} + PC^*W^{-1}\big(z - C\widetilde{u}\big). \end{aligned}$$

Furthermore,

$$\widetilde{u}(0) = \bar{u}(0) - P(0)\bar{q}(0) = u_\diamond.$$

*ii)* To prove the second result, we notice that Equation (5.14), at time $s = \tau$, reads $\widetilde{u}(\tau) = \bar{u}_\tau(\tau)$. As Equation (5.14) is true for all choices of $\tau \in \mathbb{R}_+$, we obtain that $\widetilde{u}$ satisfies Definition (5.9).

To define the Kalman estimator we thus need that Systems (5.12) and (5.13) admit a solution. In the following, we show that it is possible to define a mild solution of System (5.12).

Let us define, for all $t \in [0,\tau]$, the operator $P(t)$ as the bounded operator given by

$$\bar{u}_{t,\lambda}(t) = P(t)\lambda, \qquad \forall \lambda \in \mathcal{U}. \tag{5.15}$$

The variable $\bar{u}_{t,\lambda}(t)$ is the 4d-Var estimator associated to the following criterion

$$\mathcal{J}_{t,\lambda}(\omega,\xi) = \frac{1}{2}\|\xi\|^2_{P_0^{-1}} + \frac{1}{2}\int_0^t \Big(\|Cu\|^2_{W^{-1}} + \|\omega\|^2_{Q^{-1}}\Big)\mathrm{d}s - \langle\lambda,\, u(t)\rangle_{\mathcal{U}}, \tag{5.16}$$

with the constraint that $u$ solves the system

$$\begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}t} = Au + B\omega, \quad \forall s \in [0,t], \\ u(0) = \xi. \end{cases}$$

Similarly to what was done for the definition of the 4d-Var method, we can find the minimiser of the criterion $\mathcal{J}_{t,\lambda}$ by differentiation. We obtain

$$\bar{\xi}_{t,\lambda} = P_0 \bar{q}_{t,\lambda}(0) \quad \text{and} \quad \bar{\omega}_{t,\lambda}(s) = QB^* \bar{q}_{t,\lambda}(s), \quad \forall s \in [0,t].$$

The 4d-Var estimator $\bar{u}_{t,\lambda}$ and its adjoint $\bar{q}_{t,\lambda}$ solve the following both-ends system over the time domain $[0,t]$

$$\begin{cases} \dfrac{\mathrm{d}\bar{u}_{t,\lambda}}{\mathrm{d}t} &= A\bar{u}_{t,\lambda} + BQB^*\bar{q}_{t,\lambda}, & \forall s \in [0,t], \\ \dfrac{\mathrm{d}\bar{q}_{t,\lambda}}{\mathrm{d}t} + A^*\bar{q}_{t,\lambda} &= C^*W^{-1}C\bar{u}_{t,\lambda}, & \forall s \in [0,t], \\ \bar{u}_{t,\lambda}(0) &= u_{\diamond t,\lambda} + P_0\bar{q}_{t,\lambda}(0), \\ \bar{q}_{t,\lambda}(t) &= \lambda. \end{cases} \tag{5.17}$$

In particular, by System (5.17), we see that

$$\bar{q}_{t,\lambda}(t) = \lambda. \tag{5.18}$$

Therefore, Equation (5.15) is equivalent to

$$\bar{u}_{t,\lambda}(t) = P(t)\bar{q}_{t,\lambda}(t). \tag{5.19}$$

In the following, we briefly sketch the steps to prove that this operator is the mild solution of System(5.12). Let us recall that the *mild solution* of the Riccati differential equation (5.12) is an operator $P : [0,\tau] \to \mathcal{L}(\mathcal{U})$ such that

i) $\forall t \in [0,\tau]$, $P(t)$ is self-adjoint.

ii) $\forall t \in [0,\tau]$ and $\forall q \in \mathcal{U}$, $P$ satisfies

$$\begin{aligned} P(t)q =& T(t)P_0T(t)^*q + \int_0^t T(t-s)BQB^*T(t-s)^*q\mathrm{d}s - \\ & \int_0^t T(t-s)P(s)C^*W^{-1}CP(s)T(t-s)^*q\mathrm{d}s, \end{aligned} \tag{5.20}$$

where $T(t)$ is the strongly continuous semigroup generated by $A$ [1].

iii) $\forall q \in \mathcal{U}$, the map $t \in [0,\tau] \mapsto P(t)q$ is continuous.

We refer to [52, 51, 23] for further details.

1. The family of operators $(T(t))_{t\geq 0}$ on $\mathcal{U}$ satisfies : 1) $T(0) = I$, 2) $T(t+h) = T(t) \circ T(h)$, for all $t \geq 0, h \geq 0$, 3) $\lim_{t\to 0^+} \|S(t)u - u\|_{\mathcal{U}} = 0, \quad \forall u \in \mathcal{U}$. Furthermore, $Au = \lim_{t\to 0^+} \dfrac{\|T(t)u - u\|_{\mathcal{U}}}{t}$.

**i) $P(t)$ is a self-adjoint and non-negative operator** In the following we prove that $P$ verifies the first condition to be a mild solution of System(5.12). From Equations (5.18), (5.19) and (5.17), we have $\forall q_1, q_2 \in \mathcal{U}$

$$\begin{aligned}
\langle q_1,\, P(s)q_2\rangle_{\mathcal{U}} &= \langle \bar{q}_{s,q_1}(s),\, \bar{u}_{s,q_2}(s)\rangle_{\mathcal{U}} \\
&= \langle \bar{q}_{s,q_1}(0),\, P(0)\bar{u}_{s,q_2}(0)\rangle_{\mathcal{U}} + \int_0^s \Big(\langle \dot{\bar{q}}_{s,q_1}(r),\, \bar{u}_{s,q_2}(r)\rangle_{\mathcal{U}} + \langle \bar{q}_{s,q_1}(r),\, \dot{\bar{u}}_{s,q_2}(r)\rangle_{\mathcal{U}}\Big)\mathrm{d}r \\
&= \langle \bar{q}_{s,q_1}(0),\, P(0)\bar{u}_{s,q_2}(0)\rangle_{\mathcal{U}} + \int_0^s \langle -A^*\bar{q}_{t,q_1} + C^*W^{-1}C\bar{u}_{t,q_1},\, \bar{u}_{s,q_2}(r)\rangle_{\mathcal{U}}\mathrm{d}r + \\
&\quad \int_0^s \langle \bar{q}_{s,q_1}(r),\, A\bar{u}_{t,q_2} + BQB^*\bar{q}_{t,q_2}\rangle_{\mathcal{U}}\mathrm{d}r \\
&= \langle \bar{q}_{s,q_1}(0),\, P(0)\bar{u}_{s,q_2}(0)\rangle_{\mathcal{U}} + \int_0^s \langle W^{-1}C\bar{u}_{t,q_1}(r),\, C\bar{u}_{s,q_2}(r)\rangle_{\mathcal{Z}}\mathrm{d}r + \\
&\quad \int_0^s \langle B^*\bar{q}_{s,q_1}(r),\, QB^*\bar{q}_{t,q_2}(r)\rangle_{\mathcal{W}}\mathrm{d}r.
\end{aligned}$$

As the operators $P_0$, $W^{-1}$, $Q$ are self-adjoint, we can conclude that $P(s)$ is self-adjoint, for all $s \in [0,t]$. Furthermore, from the positivity of $P_0$, $W^{-1}$, $Q$, by taking $q_1 = q_2 = q$ we have

$$\langle q,\, P(s)q\rangle_{\mathcal{U}} \geq 0.$$

**ii) $P(t)$ solves Equation** (5.20) Let us introduce the operator

$$D : t \in [0,\tau] \mapsto -P(t)C^*W^{-1}C \in \mathcal{L}(\mathcal{U}). \tag{5.21}$$

We can prove that $P(t)$ solves Equation (5.20) taking into account the following two results, relying on the mild evolution operator theory [51] :

— For all $t \in [0,\tau]$ and for all $q \in \mathcal{U}$

$$P(t)q = T(t)P_0S_D(t,0)^*q + \int_0^t T(t-s)BQB^*S_D(t,s)^*q\mathrm{d}s. \tag{5.22}$$

— For all $q \in \mathcal{U}$, the operator $S_D(t,s) : \{(t,s) \in [0,\tau]^2 | s \leq t\} \to \mathcal{L}(\mathcal{U})$ is the unique solution, in the set of strongly continuous bounded linear operators on $\mathcal{U}$, of the following system

$$\left\{\begin{array}{rcl} S_D(t,s)q &=& T(t-s)q + \int_s^t T(t-r)D(r)S_D(r,s)q\mathrm{d}r, \\ S_D(t,s)q &=& T(t-s)q + \int_s^t S_D(t,r)D(r)T(r-s)q\mathrm{d}r. \end{array}\right. \tag{5.23}$$

From Equations (5.23), $S_D(t,s)^*$ satisfies

$$S_D(t,s)^*q = S(t,s)^*q - \int_s^t S_D(r,s)^*\widetilde{q}(r)\mathrm{d}r,$$

with $\widetilde{q}(r) = C^*W^{-1}CP(r)S(t,r)^*q$. Therefore, from Equation (5.22) we have

$$\begin{aligned}
P(t)q =& S(t,0)P_0S(t,0)^*q - \int_0^t S(t,0)P_0S_D(r,0)^*\widetilde{q}(r)\mathrm{d}r \\
&+ \int_0^t S(t,s)BQB^*S(t,s)^*q\mathrm{d}s - \int_0^t S(t,s)BQB^*\Big(\int_s^t S_D(r,s)^*\widetilde{q}(r)\mathrm{d}r\Big)\mathrm{d}s.
\end{aligned}$$

The second and the last integral may be written as follows

$$\int_0^t S(t,0)P_0S_D(r,0)^*\widetilde{q}(r)\mathrm{d}r + \int_0^t S(t,s)BQB^*\Big(\int_s^t S_D(r,s)^*\widetilde{q}(r)\mathrm{d}r\Big)\mathrm{d}s =$$
$$\int_0^t S(t,r)S(r,0)P_0S_D(r,0)^*\widetilde{q}(r)\mathrm{d}r + \int_0^t S(t,r)\int_0^r S(r,s)BQB^*S_D(r,s)^*\widetilde{q}(r)\mathrm{d}s\mathrm{d}r =$$
$$\int_0^t S(t,r)P(r)\widetilde{q}(r)\mathrm{d}r = \int_0^t S(t,r)P(r)C^*W^{-1}CP(r)S(t,r)^*q\mathrm{d}r.$$

Then, we can easily conclude.

**iii) The mapping $t \in [0,\tau] \mapsto P(t)q$ is continuous** For all $q \in \mathcal{U}$, let us consider $P(t)q$ as defined in Equation (5.22). From the strong continuity of $T$ and $S_D$, we have the function $t \in [0,\tau] \mapsto T(t)P_0S_D(t,0)^*q$ is continuous. Given the following upper-bound for $S_D(t,0)$ [51]

$$\|S_D(t,s)\|_{\mathcal{L}(\mathcal{U})} \leq m_1 e^{m_2(t-s)}, \tag{5.24}$$

with $m_1$, $m_2 \in \mathbb{R}_+$, we deduce that for all $s \in [0,\tau]$

$$\sup_{t\in[s,\tau]} \|S(t,s)BQB^*S_D(t,s)^*\|_{\mathcal{L}(\mathcal{U})} \leq m_1 e^{m_2\tau}\|B\|^2_{\mathcal{L}(\mathcal{W},\mathcal{U})}\|Q\|_{\mathcal{L}(\mathcal{W})}.$$

Therefore, the function $t \mapsto I_{[s,\tau]}S(t,s)BQB^*S_D(t,s)^*$ is continuous. Moreover, taking into account the inequality (5.24), we also have the bound

$$\sup_{(t,s)\in\Delta_\tau} \|S(t,s)BQB^*S_D(t,s)^*q\|_{\mathcal{L}(\mathcal{U})}$$
$$\leq \tilde{\alpha}e^{\tilde{m}\tau}\|B\|^2_{\mathcal{L}(\mathcal{W},\mathcal{U})}\|Q\|_{\mathcal{L}(\mathcal{W})}\Big(\sup_{\Delta_\tau}\|S_D(t,s)^*\|_{\mathcal{L}(\mathcal{U})}\Big)\|q\|_{\mathcal{U}} < \infty$$

From Lebesgue dominated convergence we have the continuity of the function $t \mapsto \int_0^t S(t,s)BQB^*S_D(t,s)^*q\mathrm{d}s$. We can thus conclude that the mapping $t \mapsto P(t)q$ is continuous on $[0,\tau]$.

With this final result, we conclude that $P$ is a mild solution of System (5.12). We refer to [52] for a proof of its uniqueness. A mild solution of System (5.13) is given by

$$\hat{u}(t) = S_D(t,0)u_\diamond + \int_0^t S_D(t,s)P(s)C^*W^{-1}z(s)\mathrm{d}s.$$

Finally, to have a classical solution of System (5.13), we need some further condition as the following

$$T(t)P_0 \text{ and } T(t)BQB^* : \mathcal{U} \to \mathcal{D}(A) \quad \forall t \in [0,\tau] \tag{5.25}$$

and

$$\sum_{j=0}^{\infty}\mu^2\int_0^\tau \|AT(t)\phi_j\|^2\mathrm{d}t < \infty \tag{5.26}$$

where $\{(\mu_j,\phi_j)\}_{j=0}^{\infty}$ are the eigenvalues and the eigenvectors of $P_0 \in \mathcal{L}(\mathcal{U})$ [52].

### 5.2.3 Stochastic deduction of the Kalman Filter

As we have seen in Section 1.3.4 for finite dimensional problems, the Kalman estimator has a natural probabilistic interpretation as the minimiser at each time step of the covariance of the estimation error. Here, we also want to provide a formal statistical framework of the Kalman estimator for infinite-dimensional problems. We refer to [53, 66] for more details, since a complete presentation would go beyond the scope of this thesis. Note, however, that it is well known that the Kalman estimator can be equivalently presented in a deterministic or stochastic framework, see for instance [178, 91].

In this chapter we have considered the following model

$$\left\{\begin{array}{rcl} \dfrac{\mathrm{d}u(t)}{\mathrm{d}t} & = & Au(t) + B(t)\omega(t), \quad \forall t \in [0,\tau], \\ u(0) & = & u_0 = u_\diamond + \xi. \end{array}\right. \tag{5.27}$$

When we perform an experiment and we collect some observations on the system, we aim at finding the trajectory $\breve{u}$ solution of

$$\left\{\begin{array}{rcl} \dfrac{\mathrm{d}u(t)}{\mathrm{d}t} & = & Au(t) + B(t)\breve{\omega}(t), \quad \forall t \in [0,\tau], \\ u(0) & = & \breve{u}_0 = u_\diamond + \breve{\xi}. \end{array}\right. \tag{5.28}$$

We can thus interpret System (5.28) as a deterministic model describing the trajectory of the state function relative to one experiment, while System (5.27) is a stochastic model describing all possible experiments. Therefore, $\xi$ and $\omega(t)$ are random variables, and $\breve{\xi}$ and $\breve{\omega}(t)$ are their realisations. We assume $\xi$ to have zero mean. Let us consider $\Omega$ the space of the realisations of $\omega$. We recall that a Wiener process is a random process $\mu : (t,\nu) \in [0,\tau] \times \Omega \mapsto \nu(t)$ such that for all $t$, $\mu(t,\cdot)$ is centred and almost everywhere continuous, with orthogonal increments and such that

$$\mathrm{Cov}(\mu(t,\cdot),\mu(s,\cdot)) = \min(t,s)Q.$$

By differentiation in time, this process can be associated to a white noise $\omega$ as follows

$$\mathrm{d}\mu = \omega \mathrm{d}t.$$

The covariance of $\omega$ is then given by

$$\mathrm{Cov}(\omega(t),\omega(s)) = \delta(t-s)Q.$$

where $\delta$ is the Dirac function. Therefore, System (5.27) should be written in the form of a stochastic linear differential equation as follows

$$\mathrm{d}u = Au\mathrm{d}t + B(t)\mathrm{d}\mu, \quad u(0) = u_\diamond + \xi.$$

Let us consider the finite time interval $[0,\tau]$ and $T(t)$ the strongly continuous semigroup on $\mathcal{U} = \mathbb{L}^2([0,\ell])$, generated by the deterministic operator $A$. The solution of the stochastic System (5.27) can thus be written by means of stochastic integration as follows

$$u(t) = T(t)u_0 + \int_0^t T(t-s)B(s)\mathrm{d}\mu, \quad \forall t \in [0,\tau].$$

Let us now move to the observations. In this chapter, we have modelled the observations as follows

$$z(t) = Cu(t) + \chi(t), \quad t \in [0, \tau]. \tag{5.29}$$

We assume $\chi$ to be a white noise, see for instance [23] for a complete presentation of white noise in infinite-dimensional settings. The observations used to estimate $\breve{u}$ are a realisation of this stochastic model and can be written as follows

$$z(t) = C\breve{u}(t) + \breve{\chi}(t), \quad t \in [0, \tau].$$

To define the Kalman estimator at time $t$, as mentioned in the deterministic setting, we process the observations sequentially and $\hat{u}(t)$ corresponds to the optimal estimation obtained through the information in the observations up to time $t$. Therefore, let us consider $y(t)$ the total observation up to time $t$, it can be described by the following stochastic differential equation

$$\mathrm{d}y = Cu\mathrm{d}t + \mathrm{d}\eta, \quad y(0) = z(0).$$

where $\eta$ is the Wiener process associated to the white noise $\chi$.

For the sake of simplicity, let us consider for the moment $u_\diamond = 0$. Let us now consider the estimators

$$\hat{u}_\Psi(t) = \int_0^t \Psi(t, s)\mathrm{d}y(s),$$

with

$$\begin{array}{c} \Psi(t,\cdot) \in \mathcal{L}(\mathcal{Z},\mathcal{U}) \text{ such that } \int_0^h \|\Psi(t,\cdot)\|^2 \mathrm{d}s < \infty \text{ and} \\ \langle \Psi(t,\cdot)z,\, u\rangle \text{ is measurable } \forall z \in \mathcal{Z}, \forall u \in \mathcal{U}. \end{array} \tag{5.30}$$

These estimators correspond to the family of sequential estimators. As in the deterministic setting the Kalman estimator is the sequential estimator associated to an optimal gain, in the stochastic setting it corresponds to an optimal choice of $\Psi$. Specifically, let $u$ be the solution of System (5.27), the Kalman estimator is the minimiser of the following criterion

$$\mathcal{J}_t(v) = \mathbb{E}[\|u(t) - v\|_{\mathcal{U}}^2] = \mathbb{E}[\langle u(t) - v,\, u(t) - v\rangle_{\mathcal{U}}], \qquad \forall t \in [0, \tau]. \tag{5.31}$$

In the following we show a relation between this criterion and the covariance of the estimation error. We notice that, as $u_\diamond = 0$, for any time $t \in [0, \tau]$, the mean of the state $u(t)$ is

$$\mathbb{E}[u(t)] = T(t)\mathbb{E}[u_0] + \int_0^t T(t - s)B\mathbb{E}[\omega(s)]\mathrm{d}s = 0$$

and the mean of the estimation error is

$$\mathbb{E}[\hat{u}_\Psi(t)] = \int_0^t \Psi(t, s)\Big(C\mathbb{E}[u(s)] + \mathbb{E}[\chi(s)]\Big)\mathrm{d}s = 0.$$

Therefore, the mean of the estimation error, namely $e_\Psi(t) = \hat{u}_\Psi(t) - u(t) \in \mathcal{U}$, is

$$\mathbb{E}[e_\Psi(t)] = \mathbb{E}[\hat{u}_\Psi] - \mathbb{E}[\breve{u}] = 0.$$

Let us call its variance

$$P_\Psi(t) = \mathrm{Cov}(e_\Psi(t)).$$

Let $\{\psi_j\}_j = 0, \dots, \infty$ be an orthonormal basis of $\mathcal{U}$. We write $e_\Psi$ on the element of the basis $e_\Psi = \sum_{j=0}^{\infty} \langle e_\Psi, \phi_j \rangle \psi_j$. Consequently, $\langle e_\Psi, e_\Psi \rangle = \sum_{j=0}^{\infty} \langle e_\Psi, \phi_j \rangle^2$. The criterion is linked to the error covariance in the following way

$$\mathcal{J}_t(\hat{u}_\Psi) = \mathbb{E}[\sum_{j=0}^{\infty} \langle e_\Psi(t), \phi_j \rangle^2] = \sum_{j=0}^{\infty} \mathbb{E}[\langle e_\Psi(t), \phi_j \rangle^2] = \sum_{j=0}^{\infty} \langle \mathrm{Cov}(e_\Psi(t))\phi_j, \phi_j \rangle = \sum_{j=0}^{\infty} \langle P_\Psi(t)\phi_j, \phi_j \rangle.$$

Consequently,

$$\mathcal{J}_t(\hat{u}_\Psi) = \|P_\Psi(t)\|_1,$$

where, for all self-adjoint non-negative operators $L$, the norm $\|L\|_1 = \sum_{j=0}^{\infty} \langle L\phi_j, \phi_j \rangle$, for any choice of the orthonormal basis $\{\psi_j\}_{j=0,\dots,\infty}$.

To conclude, we need to characterise the optimal $\Psi$. By introducing the operator

$$\Lambda(r, s) = \mathbb{E}[u(r) \circ u(s)] \quad r, s \geq 0,$$

where the operation denoted by $\circ$ is defined by $(a \circ b)c = \langle b, c \rangle a$, for all $a, b, c, \in \mathcal{U}$, we can define the optimal $\Psi$, namely $\hat{\Psi}$, as the solution of

$$\int_0^t \Psi(t, r) C \Lambda(r, s) C^* \mathrm{d}r + \Psi(t, s) W = \Lambda(t, s) C^* \quad \text{for almost all } s \in [0, t]. \tag{5.32}$$

Furthermore, we can characterise $\hat{\Psi}$ as

$$\hat{\Psi}(t, s) = S_D P(s) C^* W^{-1}$$

where $S_D$ is the operator defined in Equation 5.23 and $P$ is the weak solution of System (5.12).

In conclusion, the Kalman estimator is defined as

$$\hat{u}(t) = S_D(t, 0) u_\diamond + \int_0^t S_D(t, s) P(s) C^* W^{-1} \mathrm{d}y(s) \quad t \in [0, \tau]$$

satisfies

$$\mathcal{J}(t, \hat{u}) = \min_\Psi \mathcal{J}(t, \hat{u}_\Psi), \quad t \in [0, \tau].$$

Furthermore,

$$P = \mathrm{Cov}(\hat{u} - \breve{u}),$$

and consequently

$$\|P\|_1 = \min_\Psi \|P_\Psi\|_1.$$

As in the finite-dimensional case, we can thus interpret the operator $P$ as the covariance of the Kalman estimator.

## 5.3 Smoothing methods

In Section 5.2.2, we have shown how the Kalman estimator, starting from the *a priori* state estimation $u_\diamond$, takes into account the observations, sequentially, to correct its trajectory over time. When all data have been processed, the Kalman estimator satisfies the condition

$$\hat{u}(\tau) = \bar{u}(\tau). \tag{5.33}$$

Therefore, the method is not designed to retrieve the initial condition but rather to pursue the target trajectory. The application of some further strategy is needed, in order to complete the initial condition estimation.

These strategies are known as *smoothing methods* or *retrodiction filter methods* [12]. Smoothing methods aim at estimating the model state at some time $t$ in the past, given some measurements. In our work, we apply these methods to estimate the state at time 0, from the observations over the time interval $[0, \tau]$.

In the following we introduce some smoothing methods, we comment on their applicability in our case and we conclude by presenting the one we used in practice.

### 5.3.1 Forward-backward filtering

The first two methods we want to illustrate are based on the idea of computing the classical Kalman estimator forward in time and then go backward in time by means of a second estimator. The strategy of combining two estimators in a *forward-backward method* was first suggested in [71].

**5.3.1.a Rauch-Tung-Striebel smoothing** A first approach is one of the most commonly used strategies. It was presented by Rauch-Tung-Striebel [148]. The method consists in associating the forward Kalman estimator to the following backward estimator

$$\begin{cases} \dot{u}_{\text{rts}}(t) = A u_{\text{rts}}(t) + BQB^* P^{-1}(t)\big(u_{\text{rts}}(t) - \hat{u}(t)\big) \quad t \in [0, T] \\ \\ u_{\text{rts}}(\tau) = \hat{u}(\tau) \end{cases}$$

with associated covariance

$$\begin{cases} \dot{P_{\text{rts}}}(t) = \big(A + BQB^* P^{-1}(t)\big) P_{\text{rts}}(t) + P_{\text{rts}}\big(A(t) + BQB^* P^{-1}(t)\big)^* - BQB^*, \qquad t \in [0, \tau], \\ \\ P_{\text{rts}}(T) = P(T). \end{cases}$$

This strategy, however, cannot be applied in our case since it would require the inversion of the operator $P$. In the following we discuss the invertibility of $P$ in our application. Let us consider a uniform time grid $0 = t_0 < \ldots < t_N = \tau$ and a uniform size grid $0 = x_0 < \ldots < x_M = \ell$, with relative discretisation steps $\delta t$ and $\delta x$.

Let us consider the prediction-correction Kalman algorithm as presented in Section 1.3.4.a. At time $t_k$, the *a priori* covariance operator $P_k^-$ is given by

$$P_k^- = A_{k|k-1} P_{k-1}^+ A_{k|k-1}^\intercal + B_k Q B_k^\intercal. \tag{5.34}$$

where $A_{k|k-1}$ is the discrete model operator, $P_{k-1}^+$ is the *a posteriori* covariance operator, $B_k$ is the discrete model noise operator and $Q$ the covariance of the discrete noise associated to the continuous noise $\omega$.

To study the invertibility of $P_k^-$ let us start by defining the discrete model operator $A_{k|k-1}$. We consider two discretisation schemes. The first is the upwind scheme that gives

$$A_{k+1|k} = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} + \frac{\delta t\, b}{\delta x} \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & -1 \end{pmatrix}.$$

To take into account a more accurate scheme we can also consider the second-order Lax-Wendroff scheme, which results in

$$A_{k+1|k} = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} + \frac{\delta t\, b}{2\delta x} \begin{pmatrix} 0 & 1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & -1 & 0 \end{pmatrix} + \frac{\delta t^2\, b^2}{2\delta x^2} \begin{pmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix}.$$

In both the cases, when we set the CFL is equal to one [108], namely $\dfrac{\delta t\, b}{\delta x} = 1$, we have the nilpotent operator

$$A_{k+1|k} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}. \tag{5.35}$$

When there is no model noise, namely $B = 0$, from Equations (5.34) and (5.35), we have

$$P_k^- = \left(\begin{array}{ccc|c} & \tilde{P} & & 0 \\ & & & \vdots \\ & & & 0 \\ \hline 0 & \cdots & 0 & 0 \end{array}\right),$$

where $\tilde{P}$ is the $M-1 \times M-1$ submatrix composed by the last $M-1$ rows and columns of $P_{k-1}^+$. We conclude that, in order to invert $P_k^-$ we need to introduce some "noise" in the model. This may be done in two ways : either as a numerical error, by setting $\frac{\delta t\, b}{\delta x} < 1$, or by considering some model noise. The first way, however, would result in an error affecting all the state components and it makes the matrix ill-conditioned. Therefore, a better choice would be to design and consider some non-null model error $B$. The minimum necessary error is in the form

$$B_k Q B_k^\intercal = \left(\begin{array}{ccc|c} & 0 & & 0 \\ & & & \vdots \\ & & & 0 \\ \hline 0 & \cdots & 0 & \gamma \end{array}\right),$$

where $\gamma \neq 0$. In order to obtain this result, a possible choice is

$$B_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

This choice corresponds to some noise at the boundary $x_M = \ell$. In this way, we are accounting for some error in the Dirichlet condition $u(\ell, t) = 0$. Given the model noise space $\mathcal{W} = \mathbb{R}$ and $Q = \gamma$, the continuous model noise operator is given by the unbounded operator

$$\begin{array}{rccc} B: & \mathbb{R} & \longrightarrow & \mathbb{L}^2([0,\ell]) \\ & r & \longmapsto & r\delta_\ell, \end{array}$$

where $\delta_\ell(x) = 0$ for $0 \leq x < \ell$ and $\delta_\ell(\ell) = 1$. We recall that in this thesis we have presented the Kalman method in the case of bounded model error operators. For more details on the method for bounded model error operators we refer to [48]. However, in practice, we have considered a smoothing method that does not require the inversion of $P$.

**5.3.1.b Adjoint smoothing** A second method can be derived straightforward by the definition of the adjoint variable as the solution of the both-ends problem (5.8). Solving the coupled system of equations in (5.8) represents the bottleneck of the 4d-Var method. As previously mentioned, an iterative decoupling strategy can be taken into account. However, it would require a high number of iterations, each iteration involving the solution of the state model and the adjoint model.

The Kalman estimator may be used to decouple the both-ends system. In fact, from Equation (5.14), we can write the dynamics of the adjoint variable $\bar{q}$ independently of the 4d-Var estimator $\bar{u}$ as follows

$$\frac{\mathrm{d}\bar{q}}{\mathrm{d}t} + A^*\bar{q} = -C^*W^{-1}\left(z - C\big(\hat{u} + P\bar{q}\big)\right) = -C^*W^{-1}\left(z - C\hat{u}\right) - C^*W^{-1}CP\bar{q}.$$

An initial condition estimation method can thus consist in :

1. computing the Kalman estimator $\hat{u}$ and the Riccati operator $P$, by solving the Systems (5.13) and (5.12) forward in time,
2. solving the backward system

$$\begin{cases} \dfrac{\mathrm{d}\bar{q}}{\mathrm{d}t} + \left(A^* + C^*W^{-1}CP\right)\bar{q} = -C^*W^{-1}\left(z - C\hat{u}\right), \quad t \in [0,\tau], \\ \\ \bar{q}(\tau) = 0. \end{cases}$$

3. estimating the initial condition by

$$u_\diamond + P(0)\bar{q}(0).$$

In numerical applications, this method requires storing both the matrixes $P_k$ and the the Kalman estimator state $\hat{u}_k$ for each time $t_k$. The consequently cumbersome computational memory costs, make it impossible to use in our application.

In the following we present some strategies to bypass this problem.

## 5.3.2 Augmented-state method

The augmented-state strategy is the one that we have preferred in our applications. For all times $t \in [0,\tau]$, let us define the augmented-state as

$$\mathfrak{u}(t) = \big(u(t)\,,\, u_0\big)^\intercal = \big(u(t)\,,\, \theta(t)\big)^\intercal,$$

having the both the state and the initial condition $u_0$ in its components.

The initial condition being constant over time, the evolution of the augmented-state is modelled as follows

$$\dot{\mathfrak{u}} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \mathfrak{u} + \begin{pmatrix} B \\ 0 \end{pmatrix} \omega = \mathbb{A}\mathfrak{u} + \mathbb{B}\omega.$$

with initial condition $\mathfrak{u}(0) = \big(u_0\,,\, u_0\big)^\intercal = \big(u_\diamond\,,\, u_\diamond\big)^\intercal + \big(\xi\,,\, \xi\big)^\intercal = \mathfrak{u}_\diamond + \xi$.

Given the augmented observation operator $\mathbb{C} = (C, 0)$, the observation data can be modelled in the augmented-state framework as follows

$$z(t) = \mathbb{C}\mathfrak{u}(t) + \chi(t), \qquad t \in [0, \tau].$$

Let us denote the Kalman estimator of the augmented-state as

$$\hat{\mathfrak{u}} = \big(\hat{u}\,,\, \hat{\theta}\big)^\intercal$$

and the associated Riccati operator as

$$\mathbb{P}(t) = \begin{pmatrix} P_u(t) & P_{u,\theta}(t) \\ P_{\theta,u}(t) & P_\theta(t) \end{pmatrix}.$$

Replacing the operators in Systems (5.13) and (5.12) by their augmented versions, we obtain the following differential systems

$$\begin{cases} \dot{\hat{u}} = A\hat{u} + P_u C^* \Big(z - C\hat{u}\Big), \\ \dot{\hat{\theta}} = P_{\theta,u} C^* \Big(z - C\hat{u}\Big) \end{cases} \tag{5.36}$$

and

$$\begin{cases} \dot{P}_u = AP_u + P_u A^* - P_u C^* W^{-1} C P_u + BQB^*, \\ \dot{P_{u,\theta}} = AP_{u,\theta} - P_u C^* W^{-1} C P_{u,\theta}, \\ \dot{P}_\theta = -P_{\theta,u} C^* W^{-1} C P_{u,\theta}. \end{cases}$$

We point out that $P_u$ solve the differential equation (5.12). By setting $P_u(0) = P_0$, we have $P_u = P$. Consequently, the estimator appearing in the first components of the augmented-state $\hat{\mathfrak{u}}$, denoted by $\hat{u}$, is the state Kalman estimator described by Systems (5.13).

To define the initial value of the Riccati operator, namely $\mathbb{P}(0)$, we consider Equation (5.14) in this setting. We have

$$\hat{\mathfrak{u}}(t) = \bar{\mathfrak{u}}(t) - \mathbb{P}(t)\bar{\mathfrak{q}}(t),$$

where $\bar{\mathfrak{u}}$ and $\bar{\mathfrak{q}}$ are the state estimator and its adjoint, defined by the 4d-Var for the augmented-state problem.

The equation above at time $t = 0$ gives

$$\begin{aligned} \bar{\mathfrak{u}}(0) = \begin{pmatrix} \bar{u}(0) \\ \bar{\theta}(0) \end{pmatrix} = \mathfrak{u}_\diamond + \mathbb{P}(0)\bar{\mathfrak{q}}(0) &= \begin{pmatrix} u_\diamond \\ u_\diamond \end{pmatrix} + \begin{pmatrix} P_u(0) & P_{u,\theta}(0) \\ P_{\theta,u}(0) & P_\theta(0) \end{pmatrix} \begin{pmatrix} \bar{q}_u(0) \\ \bar{q}_\theta(0) \end{pmatrix} \\ &= \begin{pmatrix} u_\diamond \\ u_\diamond \end{pmatrix} + \begin{pmatrix} P_u(0)\bar{q}_u(0) \\ P_{\theta,u}(0)\bar{q}_u(0) \end{pmatrix}. \end{aligned} \tag{5.37}$$

In the last equivalence, we have used the fact that the components of the adjoint variable associated to the initial condition $\theta$ solve the system

$$\begin{cases} \dot{\bar{q}}_\theta = 0 \\ \bar{q}_\theta(\tau) = 0 \end{cases} \quad \Longrightarrow \quad \bar{q}_\theta = 0.$$

From what we saw before, $\bar{u}$ and $\bar{q}_u$ are the 4d-Var estimator of $u$ and its adjoint. Therefore, from Equation 5.37, we have

$$\bar{\theta}(0) = \bar{u}(0) = u_\diamond + P_0\bar{q}(0) \Longrightarrow P_{\theta,u}(0) = P_0.$$

The operator $\mathbb{P}$ being self-adjoint, we obtain $P_{u,\theta}(0) = P_0$. The initial value of $P_\theta$ can be easily obtained from the stochastic interpretation of $P$ as state covariance, resulting in $P_\theta = P_0$.

To conclude, the initial state estimation can be obtained by selecting the components $\theta$ of the augmented-state Kalman estimator at the end of the time window. In fact, we would have

$$\hat{\theta}(\tau) = \bar{\theta}(\tau) = \bar{\theta}(0) = \bar{u}(0).$$

This approach is particularly interesting, since it can also be applied to perform a joint state-parameter estimation, as we illustrated in the first part of this thesis.

In our study, we apply this last method to estimate the initial condition. In agreement with the theoretical equivalence in Equation (5.33), we found the same numerical estimations as the ones computed by 4d-Var and shown in Chapter 4.

## 5.4 Conclusion of Chapter 5

In this chapter we have presented two of the most important and commonly used data assimilation methods : the variational 4d-Var method and the sequential Kalman filtering. We have briefly illustrated how the estimators built by these two methods are liked to each other. Both the methods provide accurate solutions to the inverse problem, but they also have weakness. The 4d-Var requires the solution of the state model and adjoint model over the entire time domain at each iteration of a gradient descendent algorithm resulting in potentially high computational costs. The Kalman method requires the computation of the operator $P$ that for high dimensional problems corresponds to high computational memory and time costs. In low-dimensional inverse problems like the one presented in Chapter 4 we have compared the two methods on synthetic data. Both the methods are able to well estimate the initial condition, but the computational time is drastically reduced when using the Kalman method going from the order of tens of minutes to the order of minutes.

# Appendix A : Data assimilation for model validation

Let us consider a physical system that can be experimentally studied by collecting some measurements on it. We gather in a model the physical information of the system. In this thesis we have shown that, when we have a model and some data, we can put these elements together and build an estimation of the real system. In particular, in the case of a linear state model and a linear observation model, the existence of the minimum of the variational criterion for the 4d-Var method or the convergence of the Kalman estimator for the Kalman filter method are guaranteed.

In Chapter 5, we present data assimilation methods as useful strategies to estimate the initial condition of a system. In Chapter 1, we apply a data assimilation method to estimate some model parameters. In this appendix, we illustrate on a specific example that data assimilation methods may be a valid instrument to analyse the validity of a model.

Let us take the concrete example of a system of ovPrP fibrils, as presented in Section 3.4. We observe these fibrils in an experiment of depolymerisation. Measuring the static light scattering intensity of a system with total concentration $\rho = 0.25\mu M$, we obtain the SLS data in Figure 5.1. We want to apply a data assimilation method to estimate the initial condition of this system. We describe its evolution by the linear model

$$\begin{cases} \dfrac{\partial \breve{u}}{\partial t}(x,t) + \breve{b}\dfrac{\partial \breve{u}}{\partial x}(x,t) = 0, \qquad x \in [0,\ell],\ t \in [0,5]h, \\ \breve{u}(\ell,t) = 0, \\ \breve{u}(x,0) = \breve{u}_0(x), \end{cases} \tag{5.38}$$

in which we consider a constant depolymerisation rate $\breve{b}$. The SLS data are then modelled as

$$z(t) = \int_0^\ell x^2 \lambda_1 \breve{u}(x,t)\mathrm{d}x + \chi(t), \quad t \in [0,5]h, \tag{5.39}$$

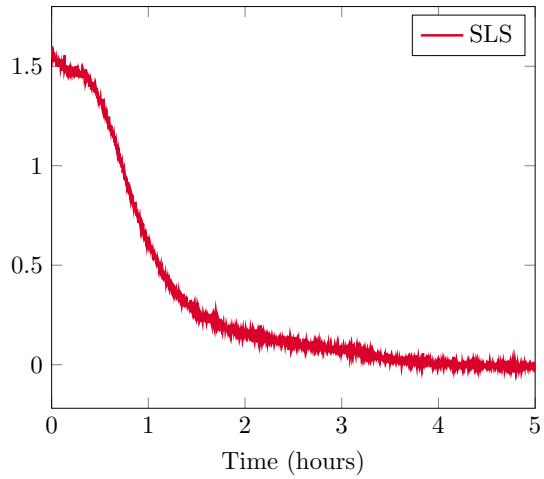



FIGURE 5.1 – Static Light Scattering measurements on ovPrP fibrils.

with $\lambda_1 > 0$ and $\chi$ some additive noise on the measurements. We notice that, in the case of prion fibrils, we cannot perform the SEC test and it is thus not possible to estimate the coefficient $\lambda_1$. However, thanks to the linearity of the model, we know that $\lambda_1 u$ evolves with the dynamics in System (5.38) and the initial condition $\lambda_1 u_0$. When we apply our strategy to estimate the initial condition, we retrieve $\lambda_1 u_0$.

Another difficulty comes from the unknown depolymerisation rate $\breve{b}$. As we explain in [6], when we apply our estimation method with a different choice of depolymerisation rate $b$, we estimate a transformation of the target initial condition $\breve{u}_0$. Considering the depolymerisation rate $b$ in the model, we denote as $u_0$ the initial condition estimator. We have

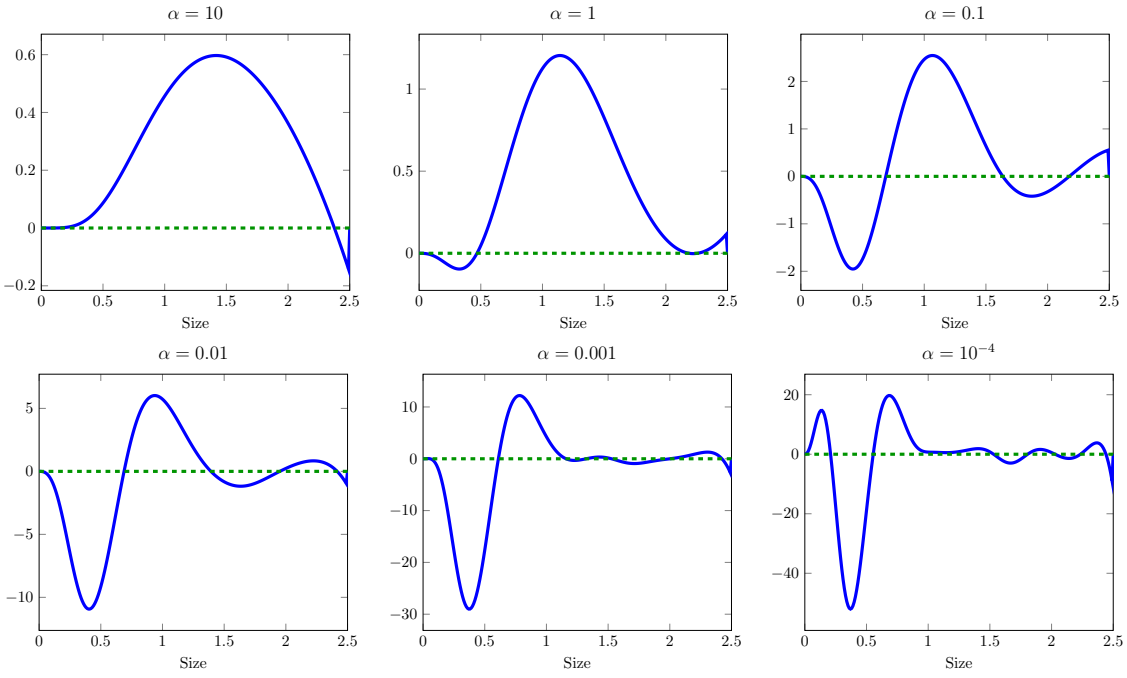
$$u_0(x) = \left(\frac{\breve{b}}{b}\right)^3 \breve{u}_0\left(\frac{\breve{b}}{b}x\right).$$

When we know neither the multiplicative coefficient $\lambda_1$ nor the depolymerisation rate, our estimator is

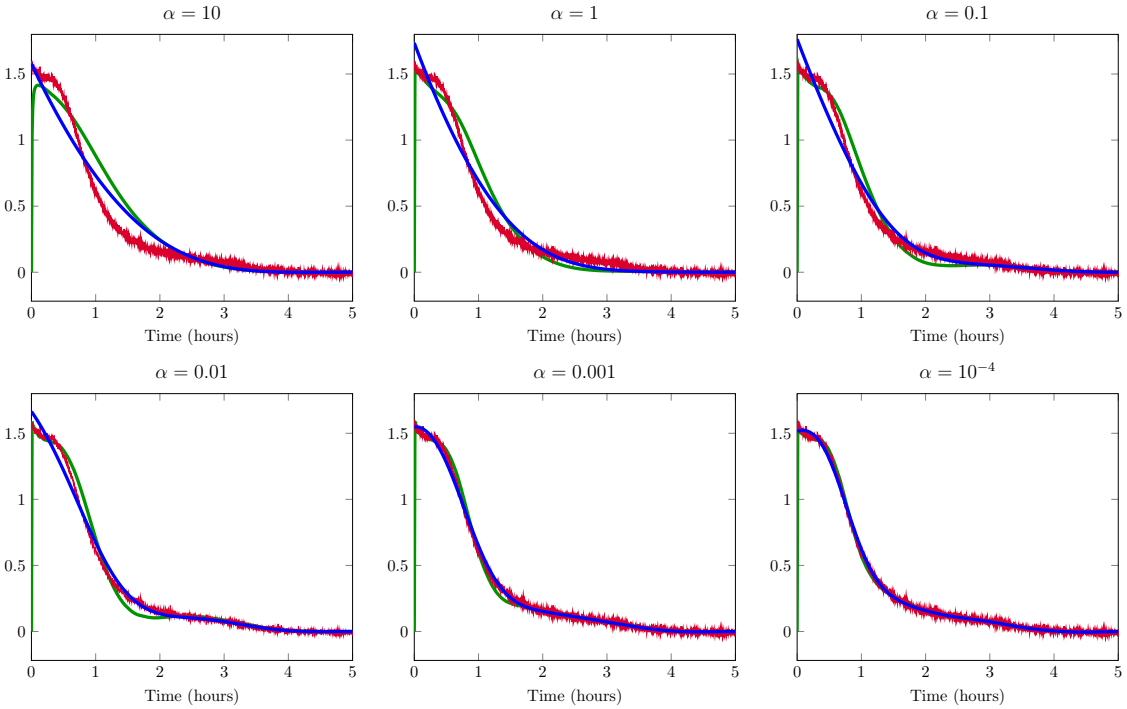
$$u_0(x) = \lambda_1 \left(\frac{\breve{b}}{b}\right)^3 \breve{u}_0\left(\frac{\breve{b}}{b}x\right) = c_1 \breve{u}_0\left(c_2 x\right), \tag{5.40}$$

with $c_1$ and $c_2$ positive coefficients. We can be interested in computing $u_0$, since it contains precious information about the profile of the target initial condition.

Let us consider the Kalman estimator starting with a zero *a priori* with several choices of operators $W$ and $P_0$. We set the covariance matrices to $W = \beta I_{N_t}$, $P_0 = \gamma I_{N_x}$, where $I$ is the identity matrix. We obtain the results shown in Figure 5.2a with associated fits of the observations in Figure 5.2b. We define $\alpha = \frac{\beta}{\gamma}$. The smaller $\alpha$ is, the more we trust the observations over the *a priori* estimation on the initial condition. We notice that for low values of $\alpha$ we have good fits of the data. However, when we have a good fit of the observations, the corresponding estimation of the initial condition becomes negative. From the formula (5.40), we deduce that $\breve{u}_0$ also takes negative values. This result cannot be admissible since $\breve{u}$ describes a concentration function, which is by definition non-negative. We conclude that System (5.38) is not suitable for describing the experimental data in Figure 5.1.




(a) Kalman estimator of $\breve{u}_0(x)$ (blue line), starting from a zero *a priori* (dotted green line).




(b) Observations fit. Comparison between the experimental data (red line), the observations of the Kalman estimator trajectory (green line) and the observations of the solution of System (5.38) with the estimation provided by the Kalman filter method (blue line) as initial condition.

FIGURE 5.2 – Kalman initial condition estimator and observation fit. The estimations were performed setting a discretisation time step to $\delta t = 10^{-4}$, the discretisation space step $\delta x = 10^{-4}$, the depolymerisation rate to $b = 0.5$ and the covariances $W = \beta I$, $P_0 = \gamma I$. The value of $\alpha = \frac{\beta}{\gamma}$ is given for each of the plots.

# Appendix B : Reminders on the transport equation

The simplest transport equation is given by the following linear constant transport equation

$$\left\{ \begin{array}{rcll} \dfrac{\partial u}{\partial t}(x,t) + \nu \dfrac{\partial u}{\partial x}(x,t) & = & 0, & \forall (x,t) \in \mathbb{R} \times \mathbb{R}^+ \\ u(x,0) & = & u_0, & \forall x \in \mathbb{R}, \end{array} \right. \tag{5.41}$$

where $\nu \in \mathbb{R}$. In this case we can easily compute the solution.

**Theorem 5**
*If $u_0 \in C^1(\mathbb{R})$, there exists a unique solution of the Cauchy problem* (5.41) *given by*

$$u(x,t) = u_0(x - \nu t) \qquad \forall (x,t) \in \mathbb{R} \times \mathbb{R}^+.$$

*Démonstration.* We introduce the characteristic curves in $\mathbb{R}^2$ given by $(X(t), t)$ where $X(t)$ is the solution of

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \nu.$$

Let $(y, s)$ be a point such that $X(s) = y$, then we have

$$X(t) = y + \nu(t - s).$$

Along the characteristics the function $u$, the solution of System (5.41), satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(u(X(t),t)\big) = \frac{\mathrm{d}X}{\mathrm{d}t}(t)\frac{\partial u}{\partial x}(X(t),t) + \frac{\partial u}{\partial t}(X(t),t) = 0.$$

We can therefore say that the solutions are constant along the characteristics. For all points $(x, t)$ we can find the unique characteristic function passing through that point, namely $X_{x,t}$, and thus write

$$u(x,t) = u(X_{x,t}(t),t) = u(X_{x,t}(0),0) = u(x - \nu t, 0) = u_0(x - \nu t).$$

■

Let us consider now the more general situation of variable transport velocity.

$$\left\{ \begin{array}{rcll} \dfrac{\partial u}{\partial t}(x,t) + \nu(x,t)\dfrac{\partial u}{\partial x}(x,t) & = & 0, & \forall (x,t) \in ]a,b[\times\mathbb{R}^+ \\ u(x,0) & = & u_0,, & \forall x \in ]a,b[, \end{array} \right. \tag{5.42}$$

with $\nu(a,t) = \nu(b,t) = 0$. As done in the constant case, we look for the characteristic curves, solutions of

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t) = \nu(X(t),t). \tag{5.43}$$

The variation of the solution of System (5.42) along the characteristics is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(u(X(t),t)\big) = \frac{\mathrm{d}X}{\mathrm{d}t}(t)\frac{\partial u}{\partial x}(X(t),t) + \frac{\partial u}{\partial t}(X(t),t) = 0.$$

Therefore, we have again that the solutions are constant along the characteristics defined by equation (5.43).

We conclude by discussing the system

$$\left\{ \begin{array}{rcll} \dfrac{\partial u}{\partial t}(x,t) + \dfrac{\partial \nu u}{\partial x}(x,t) & = & 0, & \forall (x,t) \in ]a,b[\times\mathbb{R}^+ \\ u(x,0) & = & u_0, & \forall x \in ]a,b[, \end{array} \right. \tag{5.44}$$

which is the system studied in the second part of this thesis.

We notice that in this last setting the PDE can be written as

$$\frac{\partial u}{\partial t}(x,t) + \frac{\partial \nu u}{\partial x}(x,t) = \frac{\partial u}{\partial t}(x,t) + \frac{\partial \nu}{\partial x}(x,t)u(x,t) + \nu(x,t)\frac{\partial u}{\partial x}(x,t) = 0.$$

Given the characteristics defined by the ODE (5.43), we can repeat the same calculations as before and compute the variation of the solution along the characteristics :

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(u(X(t),t)\big) = \frac{\mathrm{d}X}{\mathrm{d}t}(t)\frac{\partial u}{\partial x}(X(t),t) + \frac{\partial u}{\partial t}(X(t),t) = -\frac{\partial \nu}{\partial x}(X(t),t)u(X(t),t).$$

We thus conclude that the solutions vary along the characteristics.

Let us introduce

$$w(t) = u(X(t),t)e^{\int_0^t \frac{\partial \nu}{\partial x}(X(s),s)\mathrm{d}s}.$$

Its derivative is given by

$$\frac{\mathrm{d}w(t)}{\mathrm{d}t} = \left(\frac{\mathrm{d}u}{\mathrm{d}t}(X(t),t) + u(X(t),t)\frac{\partial \nu}{\partial x}(X(t),t)\right) e^{\int_0^t \frac{\partial \nu}{\partial x}(X(s),s)\mathrm{d}s} = 0.$$

The function $w$ is hence constant. If characteristics do not cross, for each point $(x,t)$ we can consider the unique characteristic function passing through it and define

$$u(X(t),t)e^{\int_0^t \frac{\partial \nu}{\partial x}(X(s),s)\mathrm{d}s} = u(X(0),0) = u_0(X(0)).$$

We conclude that the solution of System (5.44) satisfies

$$u(X(t),t) = u_0(X(0))e^{-\int_0^t \frac{\partial \nu}{\partial x}(X(s),s)\mathrm{d}s}.$$

# Troisième partie

# Conclusion

CHAPITRE 6

# Conclusions and perspectives

The work presented in this thesis is a contribution to the study of the mechanisms governing the aggregation of proteins. Studying these mechanisms experimentally is a real challenge, as few experimental tests are available to observe protein aggregation and, in the majority of cases, only very indirect measurements can be obtained. However, in the study of neurodegenerative prion diseases, it is possible to measure the polymerised mass or the average aggregate size over time, and for small aggregates we can also measure the size distribution accurately at certain (few) instants. At present, there are no experimental strategies to measure the division rates or aggregation rates.

For these reasons, a mathematical – direct and inverse – approach of the problem modelling is essential. In this thesis, we apply data assimilation methods to estimate the kinetic parameters and the initial size distribution through the observations over time of the average size of a system of ovPrP oligomers.

In the following, we highlight three of our main contributions to the study of PrP oligomers :

1. We set up a data analysis methodology in order to study the amount of noise in the experimental observations and to estimate the unknown scaling factors in the data. This methodology applies to all kinds of molecules observed by SLS and SEC.
2. We identified three main chemical reactions governing the evolution of the ovPrP : polymerisation, depolymerisation and disintegration. The main novelty of this study is that we take into consideration the disintegration process. Moreover, our analysis indicated the presence of at least two different oligomer species obtained by aggregation

of the monomeric ovPrP. An explanation of this heterogeneity may lie in the physical structures of the oligomers. The two species show very different behaviours. One of the species, termed stable, evolves by gaining or losing monomers through polymerisation and depolymerisation. The other species, termed unstable, is mostly disintegrating. The two species interact through the exchange of monomers : the monomers freed by the disintegration of unstable oligomers may subsequently be used by the stable oligomers to grow in size. We gathered our conclusions on the dynamics of the oligomer system to propose the following ODE model

$$\begin{cases} \dfrac{\mathrm{d}w_i}{\mathrm{d}t} &= -k_{\mathrm{dis}} w_i, \\ \dfrac{\mathrm{d}y_i}{\mathrm{d}t} &= k_{\mathrm{on}} y_{i-1} v - k_{\mathrm{on}} y_i v + k_{\mathrm{dep}} y_{i+1} - k_{\mathrm{dep}} y_i, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} &= \sum_{i=i_0}^{i_1} \big( - v k_{\mathrm{on}} y_i + k_{\mathrm{dep}} y_i + i k_{\mathrm{dis}} w_i \big), \\ w_i(0) &= (1-\alpha) u_i(0), \\ y_i(0) &= \alpha u_i(0), \\ v(0) &= 0, \end{cases} \tag{6.1}$$

It provides a simple representation of the most important features of the physical system.

3. By means of the extended Kalman approach, we estimated the kinetic parameters $k_{\mathrm{on}}$, $k_{\mathrm{dep}}$, $k_{\mathrm{dis}}$ and the initial ratio of stable oligomers $\alpha$ in three relevant experimental conditions, through the information contained in SLS data. The reliability of these estimations has been tested by comparing our oligomer size estimation to the SEC data. A possible improvement in this validation step would be possible by considering a more precise description of the error affecting the peak shapes in SEC data. To this end, repeating the experiments would allow us to determine an average peak shape and estimate the variability.

We believe that the methodology proposed in this work is a promising approach that can be extended to the study of various kinds of prions and prion-like proteins. It is worth mentioning that – even if only the average size measurements were used to solve the estimation problem – the observation of the size distribution played a crucial role in the success of this study. The simultaneous analysis of SLS and SEC data gave us a general understanding of the dynamics of the oligomer system, which would have been impossible to obtain with the SLS data alone. Moreover, SEC data enabled us to determine a good initial condition *a priori* estimation. We see the measurement of both the average polymer size and the size distribution as a key strategy to obtain important progress in the study of protein aggregation.

In order to analyse large protein aggregates, considering an ODE model may require analysing a very large number of equations. To bypass this difficulty, PDE models are commonly preferred as they may provide a better understanding of the population dynamics, as for instance we have shown in Section 1.2. To set up a methodology for the study of fibrils, we considered the Lifshitz-Slyozov theory in the case of a depolymerising system, thereby

obtaining the following backward transport model

$$\begin{cases} \frac{\partial}{\partial t}u(x,t) - \frac{\partial}{\partial x}(b(x)u(x,t)) = 0, \\ u(+\infty,t) = 0, \\ u(x,0) = u_0(x). \end{cases} \tag{6.2}$$

We investigated the inverse problem of estimating the initial condition of System (6.2) when observing the $n$-th moment of the state function $u$ over time. We propose three possible strategies to solve this linear problem.

1. The first one applies in the case of constant transport velocity. This strategy is based on an explicit formula linking the initial condition to the measurements in the noiseless case. To apply this relation on noisy measurements, we regularise the data through convolution with a regularisation kernel. The advantage of this method is that we can obtain a very precise estimation of the approximation error. However, it is difficult to extend it to other cases because any modification in the problem would require the definition of a new explicit relation.

2. To treat the more general case of a variable transport velocity we consider data assimilation methods. These methods can be presented in a general formalism that does not depend on the specific problem. We start by investigating the 4d-Var variational method. This method is based on the minimisation of a least square criterion. The method takes into account all the available data at once. The minimisation is in practice performed by a gradient descent-based method. At each iteration of this method, we need to integrate the state model forward in time and the adjoint model backward in time. This method provides very good estimations, but, as our experience confirmed, the minimisation method requires hundreds of iterations before convergence.

3. To obtain the initial condition estimation in lower computational times, we consider the Kalman method. This method integrates the data points sequentially to correct the trajectory of the estimator. The Kalman method in a linear setting is proved to converge to the solution of the 4d-Var method when all the data have been processed. This method is faster than the variational method since it goes through the time domain only once, and, as mentioned, it produces an equally good estimation. The main drawback of this method lies in the computation, at each time step, of the full covariance matrix $P$. This computation may require prohibitive computational costs and therefore be impossible when treating two-dimensional or three-dimensional problems. However, it can be done in the case of one-dimensional problems, as in our application. The maybe counter-intuitive use of the Kalman approach to estimate the initial condition of a system is explained by the fact that it can be applied on an augmented state having the initial condition in its components. The initial condition, being constant in time, is estimated at the end of the time domain.

This work opens up new questions and perspectives. In the following, we start by presenting some of the mathematical perspectives for each of the two parts of this thesis. We conclude by proposing a selection of the possible biological perspectives.

# Perspectives for Part I

## Identifiability

In Chapter 1, we presented an inverse problem consisting in the estimation of the initial condition of an ODE model through the observation over time of the second moment of the solution. Assuming size-independent kinetic rates and a size-independent initial ratio of stable oligomers, our problem can be seen as a parameter identification problem. The well-posedness of the inverse problem can be tested by analysing the identifiability of the system. A system is identifiable if distinct parameters should give distinct observations. Formally, let us consider the dynamical system

$$\mathcal{M}(\theta) = \left\{ \begin{array}{rcll} \dot{u}(t) & = & A(u(t,\theta),\theta,t), & t \in \mathbb{R}^+, \\ u(0,\theta) & = & u_0(\theta), & \end{array} \right.$$

where the components of $\theta \in \Theta \subset \mathbb{R}^d$ are the parameters of the model. In our case, $d = 4$. In the noiseless case, the observations are given by

$$z(\theta,t) = C(u(t,\theta),\theta,t), \qquad t \in [0,\tau].$$

We say that the system is *identifiable in* $\theta$ if

$$\forall\, \widetilde{\theta} \in \Theta, \qquad z(\theta,\cdot) = z(\widetilde{\theta},\cdot) \Rightarrow \theta = \widetilde{\theta}.$$

The system is *identifiable* if it is identifiable for all $\theta \in \Theta$.

A classical method to study identifiability for ODE systems is based on Taylor series [137]. A variety of other approaches also exists. Among them we can cite methods based on the state isomorphism theorem [56], on algebro-differential elimination [112, 70], on the Kalman filter [24, 130, 35] and on 4d-Var [42].

To illustrate the problem, we briefly outline in our case the method based on Taylor series. We assume that the model operator $A$ and the observation operator $C$ are infinitely differentiable. We consider the Taylor series expansion of $z(\theta,\cdot)$ around $t = 0$. A sufficient condition for identifiability is that the following equations

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}\left(C(u(t,\theta),\theta,t)\right)(0) = \lambda_k, \quad k = 1,\dots,\infty \tag{6.3}$$

have a unique solution $\theta \in \Theta$. The values $\lambda_k$ are knowns and correspond to the value measured in 0 of the derivative of order $k$ of the observation function $z(\theta,\cdot)$.

The main drawback of this approach is that it may lead to difficult computations. However, our model belongs to one of the few classes of non-linear problems in which this approach can be used in practice. In fact, System (6.1) can be written in the state-space form as follows

$$\left\{ \begin{array}{rcll} \dot{u}(t) & = & A(u(t,\theta),\theta,t)u(t,\theta), & t \in \mathbb{R}^+, \\ u(0,\theta) & = & u_0(\theta) & \end{array} \right.$$

and the observation operator $C$ associated to the SLS measurements is linear.

In this case, System (6.3) can be written as follows

$$\left\{ \begin{array}{l} Cu_0(\theta) = \lambda_0, \\ C\Big(\sum_{i=1}^{k} \dfrac{(k-1)!}{(k-i)!(i-1)!} \dfrac{\mathrm{d}^{k-i}A}{\mathrm{d}t^{k-i}}(0) \dfrac{\mathrm{d}^{i-1}u}{\mathrm{d}t^{i-1}}(0)\Big) = \lambda_k, \qquad \forall\, k \geq 1. \end{array} \right.$$

Studying the solutions of the system of equations obtained with these formulas, we can ascertain the identifiability of our system. In our application, finding an analytical solution of this system seems extremely challenging.

### Size-continuous model

Another interesting mathematical question is the study of the asymptotic limits of the ODE system (6.1), when the average aggregate size is large. To this end, one can follow the approach described briefly in Section 3.2 and detailed in [47, 58]. The size-continuous model, obtained with this strategy, can then be taken into account to describe the evolution of large protein aggregates such as fibrils.

## Perspectives for Part II

In the second part of this thesis, we focus on the study of an estimation problem in the case of a transport model (6.2) and the observation of the moment of order $n$ of the state function $u(x,t)$.

In Chapter 3 we saw that this model describes a depolymerising system well when the monomer size is asymptotically small. When we treat the problem numerically, we have to discretise the size-domain. To ensure the validity of the discrete model, we need to set a very small size-discretisation step $\delta x$.

Representing the function $u(x,t)$ by its values on a uniform size-grid with size-step $\delta x$, we obtain the state vector $(u(i\delta x,t))_{i=0,\dots,[\ell\delta x^{-1}]}$. Unfortunately, when the length of this vector, that we denote by $N$, is too big, the Kalman estimator is subject to the "curse of dimensionality" [22] and cannot be used in practice. In fact, at each time step, the method requires the computation of the covariance matrix $P$, which is a full $N \times N$ matrix. As already mentioned, this drawback is particularly important when considering two-dimensional and three-dimensional problems. As our application is a one-dimensional problem, Kalman filtering is a suitable strategy.

In the case of a large state dimension, however, alternative solutions that do not require computing the matrix $P$ should be investigated. In the following, we propose two of these solutions. A variety of other methods called *reduced-order filtering methods* [67, 158, 119] have also been designed to avoid the computation of the full matrix $P$.

### Back-and-forth nudging

We recall that, in the Kalman method, we need to compute the operator $P$ because it defines the optimal Kalman gain, namely $K = PC^*W^{-1}$. In order to circumvent the difficulty of computing $P$, a first approach is based on building a sequential estimator associated to a different gain. The aim is to find a simple gain that guarantees a reduction in the estimation error, but does not require high computational cost. In particular, the gain proposed is not associated with the minimisation of a variational criterion. This approach is known as *nudging* in the data assimilation community, but also as *Luenberger filter design* since it was first introduced by Luenberger in [113, 114]. Let us consider $u$ the solution of the following system

$$\begin{cases} \dot{u}(t) &= A(u(t),t), \qquad t \in [0,\tau], \\ u(0) &= u_\diamond + \xi. \end{cases}$$

and the observations

$$z(t) = C(u(t), t) + \chi(t), \qquad t \geq 0.$$

We recall that a sequential estimator is given by the solution of a dynamical system as follows

$$\left\{ \begin{array}{lcl} \dot{u}_{\mathrm{L}}(t) & = & A(u_{\mathrm{L}}(t), t) + G\big(z - C(u_{\mathrm{L}}(t), t)\big), \qquad t \in [0, \tau], \\ u_{\mathrm{L}}(0) & = & u_\diamond. \end{array} \right.$$

Luenberger then proposed to control the behaviour of $u_{\mathrm{L}}$, by finding the simplest gain $G$ which stabilises the dynamics of the observer error

$$\widetilde{u} = u_{\mathrm{L}} - u$$

to zero. In our case, the model operator and the observation operator are linear and time-independent. Therefore, the observer error $\widetilde{u}$ is the solution of the following system

$$\left\{ \begin{array}{lcl} \dot{\widetilde{u}}(t) & = & \big(A - GC\big)\widetilde{u} + G\chi(t), \qquad t \in [0, \tau], \\ u(0) & = & \xi. \end{array} \right.$$

A choice of $G$ such that $A - GC$ is dissipative, guarantees a reduction in the observer error over time. More precisely, as time goes by, $\widetilde{u}$ goes exponentially fast to zero, regardless of the value of $\xi$.

For instance, in [83] the author proves that, for conservative problems, the gain to $G = C^*$ gives the exponential stabilisation of the observer error. In our application the model operator is dissipative, but only because of the loss of information at the boundary $x = 0$. Therefore, we believe that the gain $G = C^*$ may still be a good choice. In the following we easily show that it is at least stabilising. Let us consider the energy of the observer error, it solves the following differential equation

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\widetilde{u}\|^2 = \int_0^\ell \widetilde{u}\frac{\partial b\widetilde{u}}{\partial x}\mathrm{d}x - \int_0^\ell \widetilde{u}x^n \int_0^\ell y^n \widetilde{u}\mathrm{d}y\mathrm{d}x.$$

In the case of a constant transport velocity $b$, for instance, the error has decreasing energy since

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\widetilde{u}\|^2 = -b\widetilde{u}^2(0) - \Big(\int_0^\ell x^n \widetilde{u}\mathrm{d}x\Big)^2.$$

Luenberger estimators are usually designed as asymptotic estimators of the target trajectory. However, in our case, the interest of such a property is limited as, over time, the state $u$ exits the space domain $[0, \ell]$ from the boundary $x = 0$, at a constant rate. Therefore, in finite time we have $u = 0$. One can thus object that the constantly null estimator is an equally good asymptotic estimator. The interest of the Luenberger approach lies in the fact that it can also be used to estimate the initial condition $u_0$.

The initial condition estimation can be performed with a "back and forth" approach [147, 153, 9, 83]. The idea is to define an estimator using the Luenberger observer forward and backward in time. Specifically, we compute the Luenberger observer over a certain time window. When the estimator $u_{\mathrm{L}}$ exits the domain, we store its value at the boundary. We then use these values to compute another Luenberger observer backward in time, over the same time domain. The key idea in our case is to reinject the information, that has been lost at the boundary $x = 0$ during the forward propagation, in the backward propagation.

We denote by $u_{\mathrm{L}}$ and $u_{\mathrm{L}b}$ the forward and backward observers, respectively. The iteration of the following back-and-forth cycles

$$\begin{cases} \dfrac{\partial u_{\mathrm{L}}^k}{\partial t} - b\dfrac{\partial u_{\mathrm{L}}^k}{\partial x} = x^n\Big(z - \int_0^\ell x^n u_{\mathrm{L}}^k \mathrm{d}x\Big), & \forall (x,t) \in [0,\ell]\times[0,\tau] \\ u_{\mathrm{L}}^k(\ell,t) = 0, & t \in [0,\tau], \\ u_{\mathrm{L}}^k(x,0) = {u_{\mathrm{L}}}_b^{k-1}(x,0), & x \in [0,\ell] \end{cases}$$

and

$$\begin{cases} \dfrac{\partial {u_{\mathrm{L}}}_b^k}{\partial t} - b\dfrac{\partial {u_{\mathrm{L}}}_b^k}{\partial x} = -x^n\Big(z - \int_0^\ell x^n {u_{\mathrm{L}}}_b^k \mathrm{d}x\Big), & \forall (x,t) \in [0,\ell]\times[0,\tau] \\ {u_{\mathrm{L}}}_b^k(0,t) = u_{\mathrm{L}}^k(0,\tau - t), & s \in [0,\ell] \\ {u_{\mathrm{L}}}_b^k(x,\tau) = u_{\mathrm{L}}^k(x,\tau), & x \in [0,\ell] \end{cases}$$

gives better and better estimations and

$$\lim_{k\to+\infty} \|u_{\mathrm{L}}^k(0) - u_0\| \to 0.$$

We show now an example of initial condition estimation obtained by back-and-forth nudging for a backward transport model with transport velocity $b = 0.2$ through the observation of the moment of order 0 over the time domain $[0,5]$. In Figure 6.1, we present the target initial condition $u_0 = e^{-\frac{(x-0.3)^2}{10^{-2}}}$ and a selection of nine functions $u_{\mathrm{L}}^k(0)$ for several cycle iterations $k$. In particular, the first subfigure corresponds to the first cycle and the last one to the 100-th cycle.

This approach provides good estimations with a very simple gain and, importantly, it does not require computing the matrix $P$. The possible drawbacks of the nudging approach are the difficulty of designing the gain $G$ and the potentially high number of cycle iterations needed to reach convergence when the stability improvement obtained by using the gain $G$ is limited.

## Moments

The second approach aims at defining a reduced model which describes the same physical system as in System (6.2). The advantage of a reduced state dimension comes at the cost of some additional assumptions on the state function $u$. In the following, we present the preliminary considerations of an on-going study.

Inspired by the fact that our observations consist of some moments of the state function, we consider the differential system solved by the moments of the state function $u$ to be a good candidate. Let us introduce the mapping

$$M: \quad w \in \mathcal{U} \quad \longmapsto \quad (\mu)_i = \Big(\int_{\mathbb{R}_+} x^i w dx\Big)_i \in \mathbb{R}^m$$

which associates any function $w$ to the vector of its first $m$ moments.

Let us call $u$ the solution of System (6.2), the differential equations solved by the components of $\mu = M(u)$ are

$$\dot{\mu}_i = \int_{\mathbb{R}_+} x^i \frac{\partial}{\partial t} u \,\mathrm{d}x = \int_{\mathbb{R}_+} x^i \frac{\partial}{\partial x}(bu)\,\mathrm{d}x = x^i b u_{|\mathbb{R}_+} - \int_{\mathbb{R}_+} i x^{i-1} b u \,\mathrm{d}x.$$

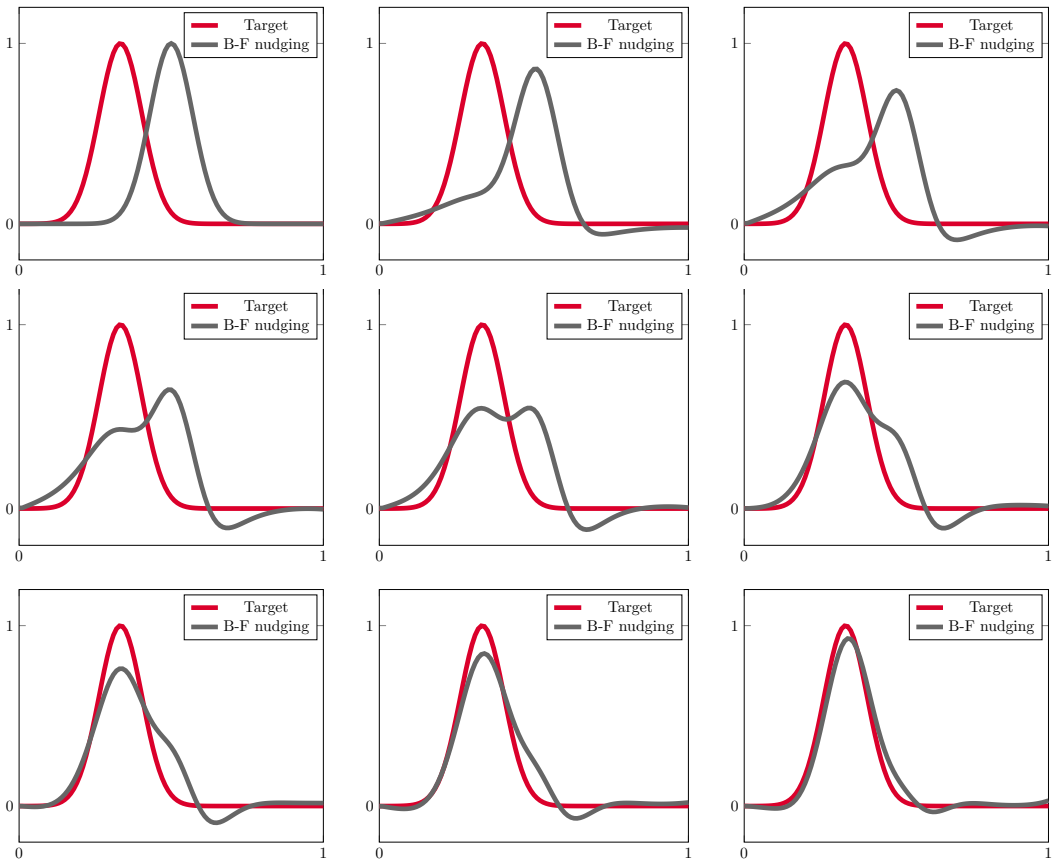



FIGURE 6.1 – Luenberger estimation of the initial condition.

Considering a constant transport velocity $b$, we have

$$\begin{cases} \dot{\mu}_i &= -ib\mu_{i-1}, \\ \dot{\mu}_0 &= -bu(0,t). \end{cases} \tag{6.4}$$

Once we have fully investigated this approach in the case of a constantly depolymerising system we can, without further difficulty, consider the complete Lifshitz-Slyozov system

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) + \dfrac{\partial}{\partial x}\Big(\big(a(x)v(t) - b(x)\big)u(x,t)\Big) = 0, & x \geq 0,\ t \geq 0, \\ v(t) + \int_0^\infty xu(x,t)\mathrm{d}x = \rho > 0, & t \geq 0 \\ v(0) = v_0, \\ u(x,0) = u_0(x), \end{cases} \tag{6.5}$$

with constant polymerisation rate $a$. In fact, in this case, the moment system is given by

$$\begin{cases} \dot{\mu}_i(t) &= -i(av(t) - b)b\mu_{i-1} = i(a(\rho - \mu_1(t)) - b)\mu_{i-1}, \\ \dot{\mu}_0(t) &= (av(t) - b)u(0,t) = (a(\rho - \mu_1(t)) - b)u(0,t). \end{cases} \tag{6.6}$$

In the case of a variable rate, a possible approach would be to consider a polynomial approximation of $b(x)$ and $a(x)$. With this strategy, we obtain more complex dependency between the moments in the definition of $\dot{\mu}_i$, and the same expression for $\dot{\mu}_0$.

We notice that, in this approach, as pointed out in the presentation of the nudging approach, the values of the function $u$ at the boundary $x = 0$ play an important role. Writing $u(0,t)$ with respect to the moments $\mu_i$ is the main difficulty of this approach and it would provide a closure condition for the system. The closure of the moment system is no trivial problem, we refer, for instance, to [160, 121, 151] in which the authors deal with the problem of finding a closure condition for the moment of order $m$ depending on the moment of order $m + 1$.

Let us consider now the constant depolymerisation case. Given the observation of a moment of order $n \leq m$ of $u$, we aim at estimating the initial condition $u_0$ in two steps :

1. We consider the model in System (6.4) and the selection of the $n$-th component of $\mu$ as observation operator. We solve the inverse problem on the reduced moment system, thereby estimating the initial condition of the moments $\mu(0)$.
2. The estimation of $u_0$ is given by the unique function having $\mu(0)$ as first $m$ moments.

We notice that this strategy can only apply when there is an injective correspondence from the moments to the functions. We thus make the assumption that $u_0$ belongs to a family of functions $\mathcal{F}$ whose elements are completely determined by knowing their moments. We can subsequently extend this approach to a general case in which the function $u_0$ can be approximated by a linear combination of the functions in $\mathcal{F}$. The approximation can then be computed, for instance, by minimisation of a least square criterion.

A first difficulty is to define a convenient family $\mathcal{F}$. Recalling that, in our application, $u$ represents the protein aggregate distribution, we consider $\mathcal{F}$ as a family of continuous distributions. An advantage of this choice is that we can refer to an extensive literature on distributions and their moments [80, 92, 34, 15]. We can thus find a sufficient condition

characterising the distributions that are determined by their moments. For instance, we know that log-normal distributions are not a good candidate for $\mathcal{F}$, as shown in an example in [90].

We look for a family of parametric functions defined on the semi-interval $[0, +\infty)$. Let us define

$$\Theta : \quad \theta \in \mathbb{R}^p \quad \longmapsto \quad f_\theta \in \mathcal{F}.$$

We consider the family of truncated Gaussian distributions (TGD) of mean $\nu \in \mathbb{R}$, variance $\sigma^2 \in \mathbb{R}_+$ and area $\omega$ defined as follows

$$u(x) = \begin{cases} \dfrac{\omega}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\nu)^2}{2\sigma^2}} & \text{if } x \geq 0 \\ \\ 0 & \text{if } x < 0. \end{cases}$$

The mapping $\Theta$ thus reads

$$\Theta : \quad \theta \in \mathbb{R}^3 \quad \longmapsto \quad \frac{\theta_1}{\sqrt{2\pi}\theta_3} e^{-\frac{(x-\theta_2)^2}{2\theta_3^2}} \mathbb{1}_{x\geq 0}, \tag{6.7}$$

where $\mathbb{1}_{x\geq 0}$ is the characteristic function of the semi-interval $[0, +\infty)$. The advantage of this choice is twofold : on the one hand we can represent the TGD by only three parameters, and on the other hand, given an initial condition $u_0 \in \mathcal{F}$, we have $u(\cdot, t) \in \mathcal{F}$ for all $t$. We recall that it is this second property that allows us to close the moment system.

The moments of a TGD of parameters $\theta = (\nu, \sigma, \omega)$ can be written as follows

$$\begin{aligned} \mu_{\theta i} = \int_{-\infty}^{+\infty} x^i u(x) dx = \int_0^{+\infty} x^i \frac{\omega}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\nu)^2}{2\sigma^2}} dx = \\ \int_{\frac{-\nu}{\sigma}}^{+\infty} (\nu + \sigma y)^i \frac{\omega}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \omega \sum_{r=0}^{i} \binom{i}{r} \nu^{i-r} \sigma^r I_r \left( \frac{-\nu}{\sigma} \right), \end{aligned} \tag{6.8}$$

where $I_r\left(\frac{-\nu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{\frac{-\nu}{\sigma}}^{\infty} x^r e^{-\frac{x^2}{2}} dx$.

Let us call

$$F : \quad \theta \in \mathbb{R}^3 \quad \longmapsto \quad \mu_\theta \in \mathbb{R}^m$$

and

$$L : \quad \mu_\theta \in \mathbb{R}^m \quad \longmapsto \quad \Theta \circ F^{-1}(\mu_\theta) \in \mathcal{F}. \tag{6.9}$$

The operator $L$ provides a way to close the differential systems on the moments. Because of the non-linearity of $L$, System (6.4) is non-linear. To perform the first step of our approach we can consider the EKF method. Once we have estimated the initial condition of the moments $\mu(0)$, we can use $L$ to obtain $u_0$, the state initial condition. This ends our state estimation approach.

As mentioned before, this approach can then be extended to the case of a generic $u \in \mathcal{U}$. In this case, we cannot solve $u = L(M(u))$ exactly, therefore we consider the approximation $\bar{u} \in \mathcal{F}$ of $u$, obtained by minimising the least square criterion as $\bar{u} = \arg\min_{u\in\mathcal{F}} \|u - L(M(u))\|^2$.

**Numerical simulations** To test our strategy, we need to discretise the moment model. Let us consider a uniform time grid $0 = t_0 < \ldots < t_N = \tau$. We denote by $\mu_i^k$ the approximation of the $i$-th moment at time $t_k$. We consider a $\theta$-scheme of parameter $\alpha$ to obtain the following discrete model

$$\begin{cases} \dfrac{\mu_i^{k+1} - \mu_i^k}{\delta t} &= -ib(\alpha\mu_{i-1}^{k+1} + (1-\alpha)\mu_{i-1}^k), \qquad 0 < i < m \\ \dfrac{\mu_0^{k+1} - \mu_0^k}{\delta t} &= -bL(\alpha\mu^{k+1} + (1-\alpha)\mu^k)(0). \end{cases} \tag{6.10}$$

Let $\mu^{k+1} \in \mathbb{R}^m$ be the vector of the approximations of the first $m$ moments at time $t_k$. We thus compute $\mu^{k+1}$ as the solution of $g(\mu, \mu^k) = 0$ where $g = (g_0, \ldots, g_{m-1})$ and

$$\begin{cases} g_i(\mu, \mu^k) = \dfrac{\mu_i - \mu_i^k}{\delta t} + ib(\alpha\mu_{i-1} + (1-\alpha)\mu_{i-1}^k) \quad \text{for } 0 < i < m \\ g_0(\mu, \mu^k) = \dfrac{\mu_0 - \mu_0^k}{\delta t} + bL(\alpha\mu + (1-\alpha)\mu^k)(0). \end{cases} \tag{6.11}$$

This system can be solved by applying Newton's method. To this end, we need to compute the derivative of $g$ with respect to $\mu$. Furthermore, this computation is useful when solving the inverse problem because it will be required for computing the model tangent when applying EKF method, see Section 1.3.5.

We can easily see that the equations for the moments of order $i > 0$ are linear. Calling $g_i = (g_1, \ldots, g_{m-1})$, we find that its derivative with respect to $\mu$ is

$$d_\mu g_i(\mu, \mu^k) = \begin{pmatrix} b\alpha & \delta t^{-1} & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & (m-1)b\alpha & \delta t^{-1} \end{pmatrix}.$$

The derivative of $g_0$ with respect to $\mu$ is given by

$$d_\mu g_0(\mu, \mu^k) = \frac{\delta_{i,0}}{\delta t} + b\alpha d_\mu L(\alpha\mu + (1-\alpha)\mu^k)(0).$$

where $\delta_{i,0}$ is a row vector with the first component equal to 1 and 0 elsewhere. The computational difficulty of solving System (6.11) lies in the term $d_\mu L(\alpha\mu + (1-\alpha)\mu^k)$. From the definition of $L$ (6.9), we have $d_\mu L(\mu) = d_\theta\Theta(F^{-1}(\mu))d_\mu F^{-1}(\mu)$, where $d_\theta\Theta$ is the derivative of $\Theta$ with respect to the parameters $\theta$.

Let us first deal with the derivative of $\Theta$. From the definition of $\Theta$ (6.7), we compute its partial derivatives

$$\begin{aligned} \frac{\partial\Theta}{\partial\theta_1} &= \frac{1}{\sqrt{2\pi}\theta_3} e^{-\frac{(x-\theta_2)^2}{2\theta_3^2}} \mathbb{1}_{x\geq 0}, \\ \frac{\partial\Theta}{\partial\theta_2} &= \frac{\theta_1}{\sqrt{2\pi}\theta_3} e^{-\frac{(x-\theta_2)^2}{2\theta_3^2}} \frac{x-\theta_2}{\theta_3^2} \mathbb{1}_{x\geq 0}, \\ \frac{\partial\Theta}{\partial\theta_3} &= \frac{\theta_1}{\sqrt{2\pi}} e^{-\frac{(x-\theta_2)^2}{2\theta_3^2}} \left(\frac{(x-\theta_2)^2}{\theta_3^4} - \frac{1}{\theta_3^2}\right) \mathbb{1}_{x\geq 0}. \end{aligned}$$

Let us now focus of the operator $F$. To discretise this operator, the first difficulty we need to handle is the computation of the integral in the Equation (6.8) and more precisely in the quantities $I_r\left(\frac{-\nu}{\sigma}\right)$. Let us call for simplicity $h = \frac{-\nu}{\sigma}$, in the following we deduce a recursive formula to compute the integrals $I_r(h)$. The first terms are :

$$I_0(h) = \frac{1}{\sqrt{2\pi}} \int_h^\infty e^{-\frac{x^2}{2}} dx = (1 - \Phi(h)),$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$ is the cumulative distribution function of the standard Gaussian distribution, and

$$I_1(h) = \frac{1}{\sqrt{2\pi}} \int_h^\infty x e^{-\frac{x^2}{2}} dx = -\left[\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\right]_h^\infty = \frac{e^{-\frac{h^2}{2}}}{\sqrt{2\pi}}.$$

For all $r \geq 2$, we have

$$I_r(h) = \frac{1}{\sqrt{2\pi}} \int_h^\infty x^r e^{-\frac{x^2}{2}} dx \frac{h^{r-1}}{\sqrt{2\pi}} e^{-\frac{h^2}{2}} + (r-1) I_{r-2}(h) = h^{r-1} I_1(h) + (r-1) I_{r-2}(h).$$

We can thus define the operator $F$ as

$$F : \quad \theta \in \mathbb{R}^3 \quad \longmapsto \quad (\mu)_i = \left(\theta_1 \sum_{r=0}^i \binom{i}{r} \theta_2^{i-r} \theta_3^r I_r\left(-\frac{\theta_2}{\theta_3}\right)\right)_i \in \mathbb{R}^m.$$

The tangent of $F$ is given in the following lines.

$$\begin{aligned}
\frac{\partial F_i}{\partial \theta_1} =& \sum_{r=0}^i \binom{i}{r} \theta_2^{i-r} \theta_3^r I_r\left(-\frac{\theta_2}{\theta_3}\right), \\
\frac{\partial F_i}{\partial \theta_2} =& \theta_1 \sum_{r=0}^i \binom{i}{r} \theta_3^r \left((i-r)\theta_2^{i-r-1} I_r\left(-\frac{\theta_2}{\theta_3}\right) + \theta_2^{i-r} dI_r\left(-\frac{\theta_2}{\theta_3}\right) \frac{-1}{\theta_3}\right) \\
& \theta_1 \sum_{r=0}^i \binom{i}{r} \left((i-r)\theta_2^{i-r-1} \theta_3^r I_r\left(-\frac{\theta_2}{\theta_3}\right) + (-1)^r \frac{\theta_2^i}{\theta_3} \frac{e^{-\frac{\theta_2^2}{2\theta_3^2}}}{\sqrt{2\pi}}\right) \\
\frac{\partial F_i}{\partial \theta_3} =& \theta_1 \sum_{r=0}^i \binom{i}{r} \theta_2^{i-r} \left(r\theta_3^{r-1} I_r\left(-\frac{\theta_2}{\theta_3}\right) + \theta_3^r dI_r\left(-\frac{\theta_2}{\theta_3}\right) \frac{\theta_2}{\theta_3^2}\right) \\
& \theta_1 \sum_{r=0}^i \binom{i}{r} \left(r\theta_2^{i-r} \theta_3^{r-1} I_r\left(-\frac{\theta_2}{\theta_3}\right) - (-1)^r \frac{\theta_2^{i+1}}{\theta_3^2} \frac{e^{-\frac{\theta_2^2}{2\theta_3^2}}}{\sqrt{2\pi}}\right).
\end{aligned}$$

in which we used the fact that $dI_r(h) = d\left(\frac{1}{\sqrt{2\pi}} \int_h^\infty x^r e^{-\frac{x^2}{2}} dx\right) = -\frac{h^r}{\sqrt{2\pi}} e^{-\frac{h^2}{2}}$.

To have a rough idea of the nonlinear relation linking $\theta$ to $F(\theta)$, let us consider a simple example. For opportune choices of mean $\theta_2$ and covariance $\theta_3$, the tgd can be approximated by the Gaussian distribution associated to the same parameters $\theta$. In this case, considering the first three moments we can explicitly write $F$ and its tangent as follows

$$F(\theta) = \begin{pmatrix} \theta_1 I_0 \\ \theta_1(\theta_2 I_0 + \theta_3 I_1) \\ \theta_1\left((\theta_2^2 + \theta_3^2) I_0 + \theta_2 \theta_3 I_1\right) \end{pmatrix},$$

$$d_\theta F(\theta) = \begin{pmatrix} I_0 & \frac{\theta_1}{\theta_3} I_1 & -\frac{\theta_1\theta_2}{\theta_3^2} I_1 \\ \theta_2 I_0 + \theta_3 I_1 & \theta_1 I_0 & \theta_1 I_1 \\ (\theta_2^2 + \theta_3^2) I_0 + \theta_2\theta_3 I_1 & 2\theta_1(\theta_2 I_0 + \theta_3 I_1) & 2\theta_1\theta_3 I_0 \end{pmatrix}.$$

Now that we have a precise definition of the operator $F$, we want to compute its inverse operator $F^{-1}$. We recall that knowing $F^{-1}$ and its derivative would allow us to compute the derivative of the operator $L$ and consequently solve the discrete model (6.10).

Since we know $F$ and its derivative, we perform this last step using Newton's method to approximate the roots of $G(\theta) = 0$, where $G(\theta) = F(\theta) - \mu$. The algorithm returns the approximation of $\theta$ together with the value of $d_\theta F(\theta)$. Therefore, we can conclude by noticing that $d_\mu F^{-1}(\mu) = (d_\theta F(\theta))^{-1}$.

We point out that one should pay particular attention when using Newton's method to compute this last result. As is well-known, the approximation provided by Newton's method strongly depends on the choice of the initial guess. When applying the method in our case to compute the moment vector $\mu^{k+1}$, a natural choice for this initial guess would be the vector of parameters relative to the function at time $t_k$, namely $\theta^k = F^{-1}(\mu^k)$. It is important to choose a family of functions $\mathcal{F}$ such that the parameters associated with the transported function over time are controlled. This represents a key aspect that is worth investigating further. For instance, it is possible to verify that the family of Gamma functions is unsuitable for this strategy.

In the following, we illustrate the moment method with an example. Let us consider the initial condition $u_0 = \frac{1}{0.3\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2*(0.3)^2}} \mathbb{1}_{x\geq 0}$. Let $u$ be the solution of the backward transport model, with initial condition $u_0$ and transport velocity $b = 0.5$. We discretise the transport model with an upwind scheme. We consider a uniform discretisation of the size domain with step $\delta x = 0.5 \cdot 10^{-3}$ and a uniform discretisation of the time domain with step $\delta x = 10^{-3}$.

In Figure 6.2, we compare the first three moments of the state function, which solves the discrete transport model described above, and the solution of System (6.10) with $m = 2$ and initial condition $M(u_0)$. The distances between the moments obtained with the two methods are provided in the following table

| mom | 0th | 1st | 2nd |
|---|---|---|---|
| $\|\|\|_\infty$ | 0.0166 | $6.7 \cdot 10^{-4}$ | $7.2 \cdot 10^{-4}$ |
| $\|\|\|_2$ | 0.05 | 0.02 | 0.03 |

.

We solve now the inverse problem of estimating the initial condition $u_0$. We take into account the synthetic observations of the second moment of the solution of the discrete transport model. We consider as initial *a priori* the tgd $u_\diamond = \frac{1}{0.3\sqrt{2\pi}} e^{-\frac{(x-1.8)^2}{2*(0.3)^2}}$. We are thus making an error in the estimation of the peak position. In Figure 6.3 we show the target initial condition and our *a priori* estimation.

We estimate the initial condition through the solution of the discrete moment system. Specifically, we take into account the first three moments, namely $m = 2$. We denote by $\breve{\mu}_0$ the target initial condition and $\mu_\diamond$ our *a priori*. In the following table we report their numerical values.

| mom | 0th | 1st | 2nd |
|---|---|---|---|
| $\breve{\mu}_0$ | 1 | 2 | 4.09 |
| $\mu_\diamond$ | 1 | 1.8 | 3.33 |

.

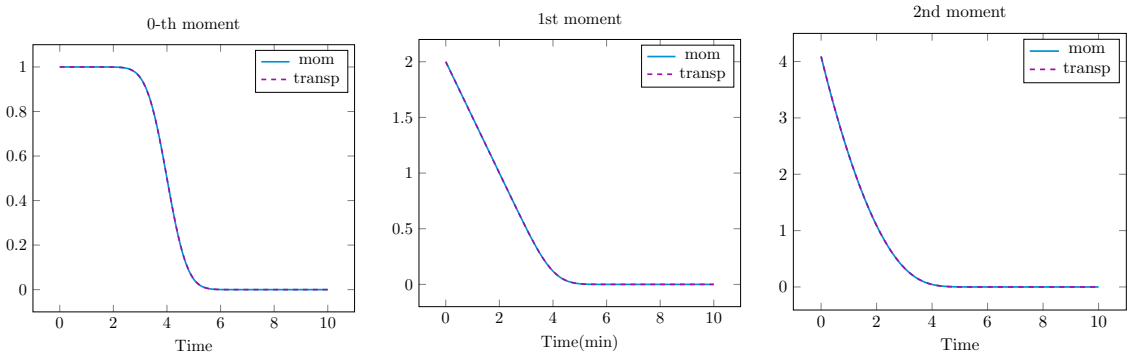



FIGURE 6.2 – Comparison between the moments computed by solving the moment system (blue solid line) and the moments of the solution of the transport model (red dashed line), for the initial condition $u_0 = \frac{1}{0.3\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2*(0.3)^2}}$ and the transport velocity $b = 0.5$. From left to right we compare the 0th, 1st and 2nd moments over the time domain $[0, 10]$.

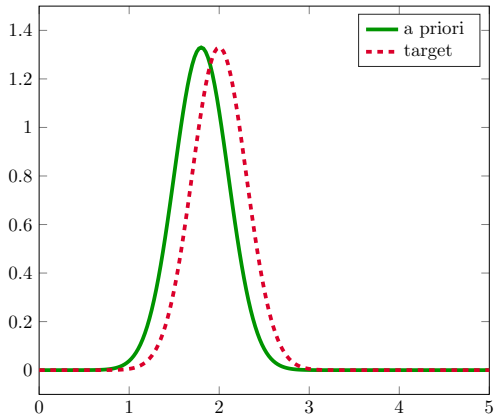



FIGURE 6.3 – Target initial condition (red dashed line) and Kalman estimator *a priori* (green line).

We set the initial covariance operator $P_0$ to the diagonal matrix with diagonal elements $(10^{-10}, 10^{-2}, 10^{-1})$. In Figure 6.4, we present the initial condition estimations when considering increasing levels of noise on the observations of the second moment. More precisely we consider the noise levels of 0%, 0.05%, 0.5%. The measurement covariances are set to $10^{-8}$, $10^{-6}$ and $10^{-4}$, respectively to the three increasing levels of noise. In Figure 6.4 we show the resulting estimations and in Figure 6.5 the associated second moment observations.

In the following table we report the moment estimations in the three cases and the function parameters associated.

| noise | $\hat{\mu}_0$ | | | $\hat{\theta}$ | | |
|---|---|---|---|---|---|---|
| 0% | 1 | 1.9998 | 4.09 | 1 | 1.9998 | 0.3013 |
| 0.05% | 0.9997 | 2.001 | 4.0907 | 0.9997 | 2.0010 | 2.9447 |
| 0.5% | 1 | 2.0045 | 4.0930 | 1 | 2.0045 | 2.738 |

.

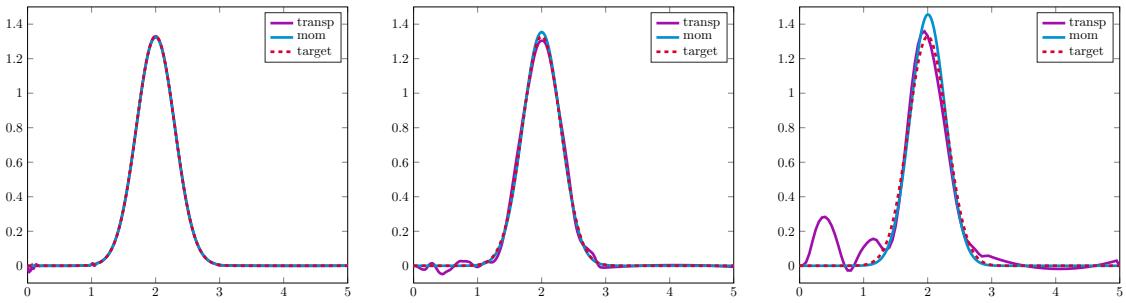



FIGURE 6.4 – Estimations of $u_0$ (red dashed line) obtained by EKF from the discrete moment model (blue line) and the discrete transport model (purple line) and the observations in Figure 6.5.

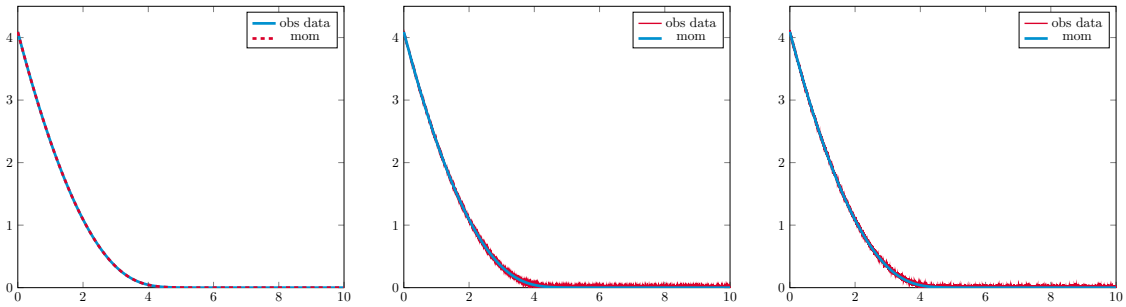



FIGURE 6.5 – Second moment observations associated to the target initial condition $u_0$ (red line) and second moment observations associated to the initial condition estimation in Figure 6.4 (blue line).

We conclude by noticing that the estimations obtained with this strategy seem to be more accurate than the estimations obtained by considering the discrete transport model, see Figure 6.4. These preliminary results open up new mathematical questions such as the analysis of the estimation error and the observability of this reduced problem. In particular, this last point corresponds to analysing observability conditions in the case of finite-dimensional nonlinear problems [89]. Furthermore, computing the distance between two functions through the

distance of their moments, rather than using an $\mathbb{L}^2$ norm, may be a more suitable strategy when dealing with transport problems. In [69] the authors investigate the use of alternative norms to the $\mathbb{L}^2$ norm in these cases.

### Lifshitz-Slyozov model

The final aim of our work is to solve the initial condition estimation problem for the complete Lifshitz-Slyozov model (6.5). The difficulty of this model lies in its non-linearity. Let us briefly discuss some of the possible strategies to handle this problem. Let us consider the transport equation in the Lifshitz-Slyozov model (6.5). By replacing the monomer concentration by the expression in the second line of the system we obtain

$$\frac{\partial u}{\partial t}(x,t) + \frac{\partial}{\partial x}\Big(\big(a(x)\rho - b(x) - a(x)\int_0^\infty xu(x,t)\mathrm{d}x\big)u(x,t)\Big) = 0.$$

In practice, when we consider this equation in the data assimilation framework, we find two major problems. The first is due to the discretisation error due to the approximation of the integral. Accurate discretisation schemes are then required. Adopting these schemes, however, may constitute a problem when we want to solve an inverse problem since they demand a high computational cost. The second problem is the accuracy of the inverse problem solution, that in a non-linear setting strongly depends on a precise estimation of the amount of error in the data and the initial condition *a priori*.

A possible strategy to treat the complete Lifshitz-Slyozov model, but circumvent the difficulty of nonlinearity, is to consider the monomer concentration as known. In this case, the model would be linear in the state. Measuring the monomer concentration experimentally is indeed possible. We can measure it directly by SEC, as shown in the first part of this thesis, or indirectly by Thioflavin T fluorescence which records the polymerised mass $\int_0^\infty xu(x,t)$. However, these data are affected by noise. Let us assume the case of an additive noise $\varepsilon$. By replacing the $v$ by its empirical measurement $v+\varepsilon$, we are thus introducing some noise into the model that results in

$$\frac{\partial u}{\partial t}(x,t) + \frac{\partial}{\partial x}\Big(\big(a(x)v(t) - b(x)\big)u(x,t)\Big) + \varepsilon\frac{\partial}{\partial x}\big(a(x)u(x,t)\big) = 0. \tag{6.12}$$

Let us define the operators

$$A(t): \left|\begin{array}{ccc} \mathcal{D}(A) \subset \mathbb{L}^2([0,\ell]) & \longrightarrow & \mathbb{L}^2([0,\ell]) \\ f & \longmapsto & \dfrac{\mathrm{d}}{\mathrm{d}x}\big((av(t)-b)f\big) \end{array}\right.$$

with domain $\mathcal{D}(A) = \{f \in \mathbb{H}^1([0,\ell]) \mid f(\ell) = 0\}$ and

$$B: \left|\begin{array}{ccc} \mathcal{D}(B) \subset \mathbb{L}^2([0,\ell]) & \longrightarrow & \mathbb{L}^2([0,\ell]) \\ f & \longmapsto & \dfrac{\mathrm{d}}{\mathrm{d}x}\big(af\big) \end{array}\right.$$

with domain $\mathcal{D}(B) = \mathcal{D}(A)$. Equation (6.12) can thus be written in the state-space formalism as follows

$$\dot{u} = A(t)u + \varepsilon Bu.$$

The difference with the framework presented in this thesis is that, this time, the noise in the model is not additive. Designing inverse problem strategies in this new setting is a new mathematical challenge.

## Biological short-term perspectives

An interesting immediate consequence of this work would be the design of a new set of experiments to physically separate the two oligomer species. Such experiments could give rise to new lines of investigation. In association with a mathematical analysis, it could lead to the detection of new oligomer species and, ultimately, a full characterisation of ovine prion oligomers. In this more ample framework, understanding the species evolutions and their interactions will constitute the first challenge.

Furthermore, it is possible to apply our data analysis methodology to SLS and SEC data collected on other kinds of proteins. As observed in the course of our work, this data analysis represents a valuable tool in itself, as it enables us to detect the reliable data and ensure their correct interpretation.

As an example, we can consider the study currently being carried out on the mutant protein H190A [40]. In Figure 6.6, we show the SLS measurements on the mutant oligomer system, at several concentrations between $0.6\mu M$ and $8\mu M$. The oligomers made up by H190A monomers are supposed to undergo the same processes as ovPrP. By means of these first experimental observations, we notice faster experimental times, likely associated to higher kinetic rates. Furthermore, we can identify a multi-phasic behaviour of SLS data.

Another short-term perspective, would be using our model as a starting point in the study of other prion or prion-like proteins. A straightforward application of our investigation strategy, presented on ovPrP, to other types of proteins represents a simple way to test the validity of our ODE model in these studies. For instance, we could test the suitability of our model to describe the SLS data in Figure 6.6 for concentrations $\rho = 6,\ 8\mu M$ up to 70min. More generally, we recommend applying our approach to test the validity of any model, as presented in Appendix A.

## Biological medium-term perspectives

One of the interesting points highlighted and not fully investigated in the course of this work is the analysis of the noise in the SLS data, which is characterised by spikes. Our preliminary conclusions were that the frequency and/or the intensity of these spikes seem to be higher for small concentrations. A more in depth understanding of this noise can reveal new mechanisms governing polymer aggregation. We notice that, after establishing the nature of the noise, it is possible to estimate the kinetic parameters and the initial condition by means of a modified version of the extended Kalman filter that takes into consideration different kinds of noise.

A preliminary set of experiments on human prion fibrils displays an oscillatory behaviour of the SLS data. Our model may serve as a starting point in the study of this phenomenon. In fact, a first mathematical analysis on a toy model suggests that it is possible to explain the oscillations by considering the interactions between several species. This study will represent one of the subjects explored by Mathieu Mezache in his PhD work.

## Biological long-term perspectives

Our study reveals the presence of two oligomer species. The characterisation of their thermodynamical stability, provided by our work, may be used to propose new therapeutic

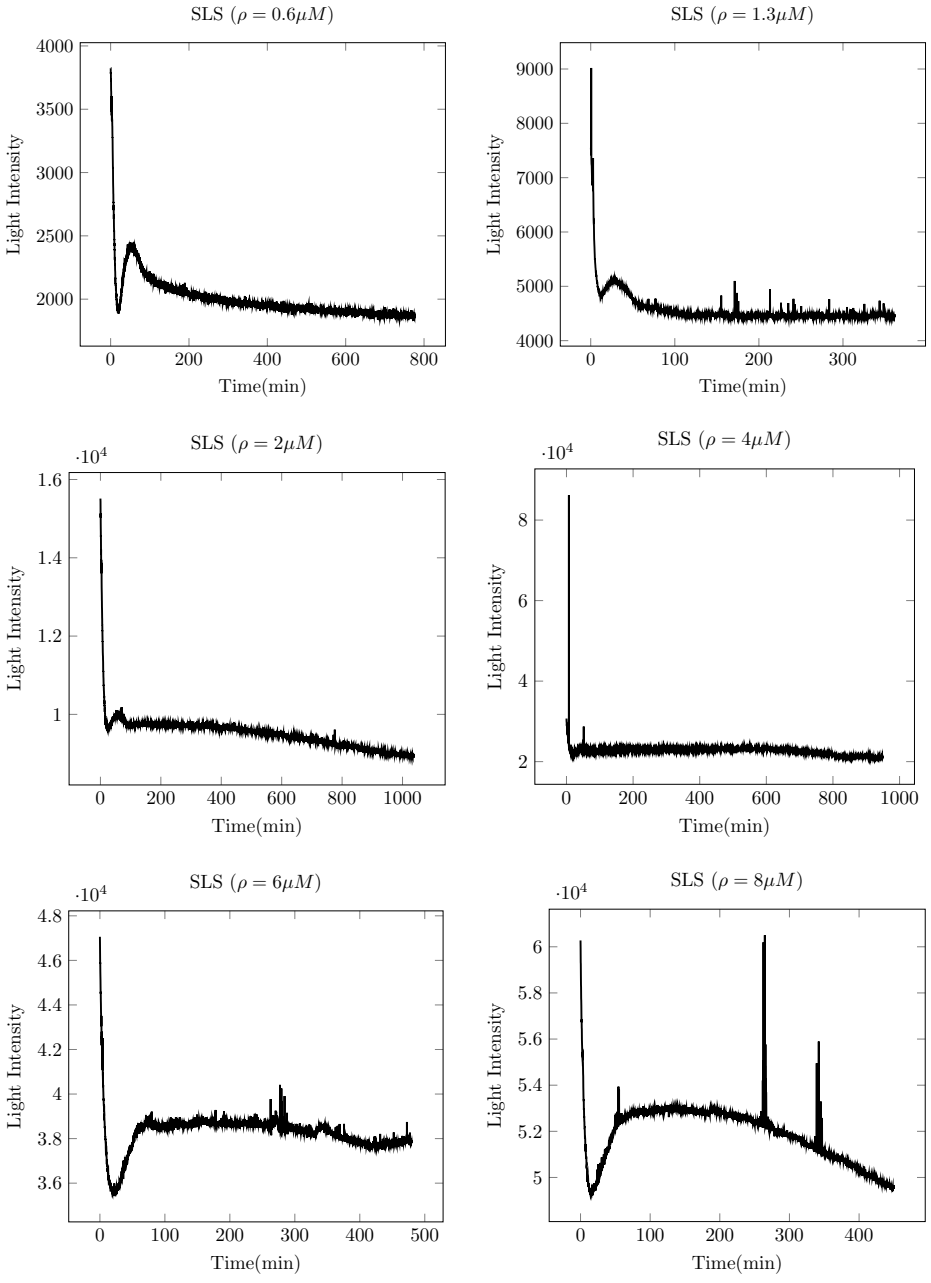



FIGURE 6.6 – SLS of H190A oligomers.

strategies. In the context of multidisciplinary collaborations, the efficiency of these strategies can be, first of all, tested with numerical simulations and *in vitro* experiments. Designing *in vivo* experiments and finding their representation through mathematical models represent the last challenges to solve in order to finally find a cure for amyloid diseases.

# Bibliographie


[1] G. K. Ackers. Molecular sieve methods of analysis. *The proteins*, 1 :1–94, 1975.

[2] A. Aguzzi and A. K. Lakkaraju. Cell biology of prions and prionoids : a status report. *Trends in cell biology*, 26(1) :40–51, 2016.

[3] M. Aizenman and T. A. Bak. Convergence to equilibrium in a system of reacting polymers. *Communications in Mathematical Physics*, 65(3) :203–230, 1979.

[4] T. Alper and W. A. Cramp. Does the agent of scrapie replicate without nucleic acid ? *Nature*, 214 :764766, May 1967.

[5] B. D. Anderson and J. B. Moore. Optimal filtering. *Eaglewood Cliffs, NJ : Prentice-Hall*, 1979.

[6] A. Armiento, M. Doumic, P. Moireau, and H. Rezaei. Estimation from moments measurements for amyloid depolymerisation. *Journal of theoretical biology*, 397 :68–88, 2016.

[7] S. Arrhenius. *Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte*. Wilhelm Engelmann, 1889.

[8] S. Arrhenius. Über die reaktionsgeschwindigkeit bei der inversion von rohrzucker durch säuren. *Zeitschrift für physikalische Chemie*, 4 :226–248, 1889.

[9] D. Auroux and J. Blum. A nudging-based data assimilation method : the back and forth nudging (bfn) algorithm. *Nonlinear Processes in Geophysics*, 15(2) :305–319, 2008.

[10] J. M. Ball and J. Carr. Asymptotic behaviour of solutions to the becker-döring equations for arbitrary initial data. *Proceedings of the Royal Society of Edinburgh : Section A Mathematics*, 108(1-2) :109–116, 1988.

[11] J. M. Ball, J. Carr, and O. Penrose. The Becker-Döring cluster equations : basic properties and asymptotic behaviour of solutions. *Comm. Math. Phys.*, 104(24) :657–692, 1986.

[12] J. Ballabrera-Poy, A. J. Busalacchi, and R. Murtugudde. Application of a reduced-order kalman filter to initialize a coupled atmosphere-ocean model : Impact on the prediction of el nino. *Journal of climate*, 14(8) :1720–1737, 2001.

[13] H. T. Banks, M. Doumic-Jauffret, and C. Kruse. Efficient numerical schemes for nucleation-aggregation models : Early steps. Mar 2014.

[14] H. T. Banks, K. L. Sutton, W. C. Thompson, G. Bocharov, D. Roose, T. Schenkel, and A. Meyerhans. Estimation of cell proliferation dynamics using cfse data. *Bulletin of mathematical biology*, 73(1) :116–150, 2011.

[15] D. R. Barr and E. T. Sherrill. Mean and variance of truncated normal distributions. *The American Statistician*, 53(4) :357–361, 1999.

[16] I. V. Baskakov, G. Legname, M. A. Baldwin, S. B. Prusiner, and F. E. Cohen. Pathway complexity of prion protein assembly into amyloid. *Journal of Biological Chemistry*, 277(24) :21140–21148, 2002.

[17] R. W. Bass, V. D. Norum, and L. Schwartz. Optimal multichannel nonlinear filtering. *Journal of Mathematical Analysis and Applications*, 16(1) :152–164, 1966.

[18] K. M. Batzli and B. J. Love. Agitation of amyloid proteins to speed aggregation measured by ThT fluorescence : A call for standardization. *Materials Science and Engineering : C*, 48(0) :359 – 364, 2015.

[19] R. Becker and W. Döring. Kinetische Behandlung der Keimbildung in bersttigten Dmpfen. *Ann. Phys*, 24 :719–752, 1935.

[20] B. M. Bell and F. W. Cathey. The iterated kalman filter update as a gauss-newton method. *IEEE Transactions on Automatic Control*, 38(2) :294–297, 1993.

[21] J. F. Bellantoni and K. W. Dodge. A square root formulation of the kalman-schmidt filter. *AIAA journal*, 5(7) :1309–1314, 1967.

[22] R. Bellman. Adaptive control processes : a guided tour princeton university press. *Princeton, New Jersey, USA*, 1961.

[23] A. Bensoussan. *Filtrage Optimal des Systèmes Linéaires*. Méthodes Mathématiques de l'informatique. Dunod, 1971.

[24] C. Bertoglio, P. Moireau, and J.-F. Gerbeau. Sequential parameter estimation for fluid–structure problems : Application to hemodynamics. *International Journal for Numerical Methods in Biomedical Engineering*, 28(4) :434–455, 2012.

[25] K. Binder. Theory for the dynamics o "clusters." ii. critical diffusion in binary systems and the kinetics of phase separation. *Physical Review B*, 15(9) :4425, 1977.

[26] M. F. Bishop and F. A. Ferrone. Kinetics of nucleation-controlled polymerization. a perturbation treatment for use with a secondary pathway. *Biophys J.*, 46(5) :631644, November 1984.

[27] J. Bisschops. Gelation of concentrated polyacrylonitrile solutions. *Journal of Polymer Science*, XVII :89–98, 1955.

[28] M. L. Bolognesi and G. Legname. Approaches for discovering anti-prion compounds : lessons learned and challenges ahead. *Expert opinion on drug discovery*, 10(4) :389–397, 2015.

[29] D. C. Bolton, M. P. McKinley, and S. B. Prusiner. Identification of a protein that purifies with the scrapie prion. *Science*, 218(4579) :1309–1311, 1982.

[30] A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis : the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, 1997.

[31] H. Brézis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext Series. Springer, 2010.

[32] R. G. Brown and P. Y. C. Hwang. Introduction to random signal analysis and kalman filtering with matlab exercises and solutions, 1996.

[33] A. E. Bryson and M. Frazier. Smoothing for linear and nonlinear dynamic systems. In *Proceedings of the optimum system synthesis conference*, pages 353–364, 1963.

[34] J. Burkardt. The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*, 2014.

[35] A. Caiazzo, F. Caforio, G. Montecinos, L. O. Müller, P. J. Blanco, and E. F. Toro. Assessment of reduced order kalman filter for parameter identification in one-dimensional blood flow models using experimental data. submitted.

[36] V. Calvez, N. Lenuzza, M. Doumic, J.-P. Deslys, F. Mouthon, and B. Perthame. Prion dynamics with size dependency–strain phenomena. *Journal of Biological Dynamics*, 4(1) :28–42, 2010.

[37] V. Calvez, N. Lenuzza, D. Oelz, J.-P. Deslys, P. Laurent, F. Mouthon, and B. Perthame. Size distribution dependence of prion aggregates infectivity. *Mathematical biosciences*, 217(1) :88–99, 2009.

[38] J. A. Canizo and B. Lods. Exponential convergence to equilibrium for subcritical solutions of the becker–döring equations. *Journal of Differential Equations*, 255(5) :905–950, 2013.

[39] P. Cavaliere, J. Torrent, S. Prigent, V. Granata, K. Pauwels, A. Pastore, H. Rezaei, and A. Zagari. Binding of methylene blue to a surface cleft inhibits the oligomerization and fibrillization of prion protein. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1832(1) :20–28, 2013.

[40] N. Chakroun, S. Prigent, C. A. Dreiss, S. Noinville, C. Chapuis, F. Fraternali, and H. Rezaei. The oligomerization properties of prion protein are restricted to the H2H3 domain. *The FASEB Journal*, 24(9) :3222–3231, sep 2010.

[41] D. Chapelle, M. Fragu, V. Mallet, and P. Moireau. Fundamental principles of data assimilation underlying the verdandi library : applications to biophysical model personalization within euheart. *Medical & biological engineering & computing*, 51(11) :1221–1233, 2013.

[42] G. Chavent. *Nonlinear least squares for inverse problems : theoretical foundations and step-by-step guide for applications*. Springer Science & Business Media, 2010.

[43] F. Chen and M. Dunnigan. Comparative study of a sliding-mode observer and kalman filters for full state estimation in an induction machine. *IEE Proceedings-Electric Power Applications*, 149(1) :53–64, 2002.

[44] Z. Chen, R. Rodrigo, V. Parsa, and J. Samarabandu. Using ultrasonic and vision sensors within extended kalman filter for robot navigation. *Canadian Acoustics*, 33(3) :28–29, 2005.

[45] N. Cindea, A. Imperiale, and P. Moireau. Data assimilation of time under-sampled measurements using observers, the wave-like equation example. *ESAIM : Control, Optimisation and Calculus of Variations*, 21(3) :635–669, 2015.

[46] J. Collet and T. Goudon. On solutions of the Lifshitz-Slyozov model. *Nonlinearity*, 13 :1239– 1262, 2000.

[47] J. Collet, T. Goudon, F. Poupaud, and A. Vasseur. The BekerDöring system and its Lifshitz-Slyozov limit. *SIAM J. APPL. MATH.*, 62(5) :14881500, 2002.

[48] J.-M. Coron. *Control and nonlinearity*. Number 136. American Mathematical Soc., 2007.

[49] M. Costanzo and C. Zurzolo. The cell biology of prion-like spread of protein aggregates : mechanisms and implication in neurodegeneration. *Biochemical Journal*, 452(1) :1–17, 2013.

[50] F. H. Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[51] R. Curtain and A. J. Pritchard. The infinite-dimensional riccati equation for systems defined by evolution operators. *SIAM Journal on Control and Optimization*, 14(5) :951–983, 1976.

[52] R. F. Curtain. Infinite-dimensional filtering. *SIAM Journal on Control*, 13(1) :89–104, 1975.

[53] R. F. Curtain and H. Zwart. *An introduction to infinite-dimensional linear systems theory*, volume 21. Springer Science & Business Media, 2012.

[54] R. Daley. *Atmospheric data analysis*. Number 2. Cambridge university press, 1993.

[55] P. Debye. Light scattering in solutions. *Journal of Applied Physics*, 15 :338–342, apr 1944.

[56] L. Denis-Vidal and G. Joly-Blanchard. Equivalence and identifiability analysis of uncontrolled nonlinear dynamical systems. *Automatica*, 40(2) :287–292, 2004.

[57] A. Dickinson and G. Outram. The scrapie replication-site hypothesis and its implications for pathogenesis. *Slow transmissible diseases of the nervous system*, 2 :13–31, 1979.

[58] M. Doumic, T. Goudon, and T. Lepoutre. Scaling limit of a discrete prion dynamics model. *Communications in Mathematical Sciences*, 7(4) :839–865, 2009.

[59] M. Doumic, B. Perthame, and J. P. Zubelli. Numerical solution of an inverse problem in size-structured population dynamics. *Inverse Problems*, 25(4) :045008, 2009.

[60] R. Drake. *Topics in Current Aerosol Research : International Reviews in Aerosol Physics and Chemistry*, volume 2. Oxford Pergamon Press, 1972.

[61] K. R. E. A new approach to linear filtering and prediction problems. *Journal of Bsic Engeneering*, 82 :3545, 1960.

[62] F. Eghiaian, T. Daubenfeld, Y. Quenet, M. van Audenhaege, A. P. Bouin, G. van der Rest, J. Grosclaude, and H. Rezaei. Diversity in prion protein oligomerization pathways results from domain expansion as revealed by hydrogen/deuterium exchange and disulfide linkage. *Proc Natl Acad Sci U S A.*, 104(18) :7414–7419, may 2007.

[63] H. Eibern and H. Schmidt. A four-dimensional variational chemistry data assimilation scheme for eulerian chemistry transport modeling. *Journal of Geophysical Research : Atmospheres*, 104(D15) :18583–18598, 1999.

[64] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, 1996.

[65] H. Engler, J. Prüss, and G. F. Webb. Analysis of a model for the dynamics of prions ii. *Journal of mathematical analysis and applications*, 324(1) :98–117, 2006.

[66] P. L. Falb. Infinite-dimensional filtering : The kalman-bucy filter in hilbert space. *Information and Control*, 11(1) :102–137, 1967.

[67] B. F. Farrell and P. J. Ioannou. State estimation using a reduced-order kalman filter. *Journal of the Atmospheric Sciences*, 58(23) :3666–3680, 2001.

[68] F. A. Ferrone. Assembly of a$\beta$ proceeds via monomeric nuclei. *Journal of molecular biology*, 427(2) :287–290, 2015.

[69] N. Feyeux, M. Nodet, and A. Vidard. Optimal Transport for Data Assimilation. working paper or preprint, July 2016.

[70] M. Fliess and S. T. Glad. An algebraic approach to linear and nonlinear control. In *Essays on Control*, pages 223–267. Springer, 1993.

[71] D. Fraser and J. Potter. The optimum linear smoother as a combination of two optimum linear filters. *IEEE Transactions on Automatic Control*, 14(4) :387–390, 1969.

[72] S. K. Friedlander. On the particle size spectrum of a condensing vapour. *Physics of Fluids (1958-1988)*, 3(5) :693–696, 1960.

[73] J.-P. Gauthier and I. Kupka. *Deterministic observation theory and applications*. Cambridge university press, 2001.

[74] A. Gelb. *Applied optimal estimation*. MIT press, 1974.

[75] M. Ghil and P. Malanotte-Rizzoli. Data assimilation in meteorology and oceanography. *Advances in geophysics*, 33 :141–266, 1991.

[76] J. R. Glover and S. Lindquist. Hsp104, hsp70, and hsp40 : a novel chaperone system that rescues previously aggregated proteins. *Cell*, 94(1) :73–82, 1998.

[77] J. Graham and D. Oppenheimer. Orthostatic hypotension and nicotine sensitivity in a case of multiple system atrophy. *Journal of Neurology, Neurosurgery & Psychiatry*, 32(1) :28–34, 1969.

[78] M. L. Greer, L. Pujo-Menjouet, and G. F. Webb. A mathematical analysis of the dynamics of prion proliferation. *Journal of theoretical biology*, 242(3) :598–606, 2006.

[79] J. S. Griffith. Self-replication and scrapie. *Nature*, 215(5105) :1043, 1967.

[80] G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford university press, 2001.

[81] C. M. Guldberg and P. Waage. Studies concerning affinity. *C. M. Forhandlinger : Videnskabs-Selskabet i Christiana*, 35, 1864.

[82] S. Haïk, G. Marcon, A. Mallet, M. Tettamanti, A. Welaratne, G. Giaccone, S. Azimi, V. Pietrini, J.-R. Fabreguettes, D. Imperiale, et al. Doxycycline in creutzfeldt-jakob disease : a phase 2, randomised, double-blind, placebo-controlled trial. *The Lancet Neurology*, 13(2) :150–158, 2014.

[83] A. Haraux. Une remarque sur la stabilisation de certains systèmes du deuxième ordre en temps. *Portugal. Math.*, 46(3) :245–258, 1989.

[84] S. E. Harding. *Protein Hydrodynamics. "Protein : A Comprehensive Treatise"*, volume 2. JAI Press Inc., Stamford CT, 1999.

[85] S. Hariz and J. F. Collet. A modified version of the lifshitz-slyozov model. *Applied mathematics letters*, 12(1) :81–85, 1999.

[86] S. Hariz. *Une version modifiée du modele de Lifshitz-Slyozov : existence et unicité de la solution, simulation numérique.* PhD thesis, 1999.

[87] S. S. Haykin et al. *Kalman filtering and neural networks.* Wiley Online Library, 2001.

[88] E. M. Hendriks, M. H. Ernst, and R. M. Ziff. Coagulation equations with gelation. *Journal of Statistical Physics*, 31(3) :519–563, 1983.

[89] R. Hermann and A. J. Krener. Nonlinear controllability and observability. *IEEE Transactions on automatic control*, 22(5) :728–740, 1977.

[90] C. C. Heyde. On a property of the lognormal distribution. *J. Royal Stat. Soc.*, (29) :392–393, 1963.

[91] O. Hijab. Asymptotic bayesian estimation of a first order equation with small diffusion. *The Annals of Probability*, pages 890–902, 1984.

[92] W. C. Horrace. Moments of the truncated normal distribution. *Journal of Productivity Analysis*, 43(2) :133–138, 2015.

[93] P. Huang, F. Lian, Y. Wen, C. Guo, and D. Lin. Prion protein oligomer and its neurotoxicity. *Acta biochimica et biophysica Sinica*, page gmt037, 2013.

[94] P.-E. Jabin and B. Niethammer. On the rate of convergence to equilibrium in the becker–döring equations. *Journal of Differential Equations*, 191(2) :518–543, 2003.

[95] J. T. Jarrett and P. T. Lansbury. Seeding "one-dimensional crystallization" of amyloid : a pathogenic mechanism in alzheimer's disease and scrapie ? *Cell*, 73(6) :1055–1058, 1993.

[96] A. M. Jazwinski. *Stochastic processes and filtering theory.* Academic Press, New York, 1970.

[97] T. Kailath. Lectures on wiener and kalman filtering. In *Lectures on Wiener and Kalman Filtering*, pages 1–143. Springer, 1981.

[98] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of basic engineering*, 83(1) :95–108, 1961.

[99] E. Kalnay. *Atmospheric modeling, data assimilation and predictability.* Cambridge university press, 2003.

[100] Y. E. Karapetyan, G. F. Sferrazza, M. Zhou, G. Ottenberg, T. Spicer, P. Chase, M. Fallahi, P. Hodder, C. Weissmann, and C. I. Lasmézas. Unique drug screening approach for prion diseases identifies tacrolimus and astemizole as antiprion agents. *Proceedings of the National Academy of Sciences*, 110(17) :7044–7049, 2013.

[101] D. K. V. Kumar, S. H. Choi, K. J. Washicosky, W. A. Eimer, S. Tucker, J. Ghofrani, A. Lefkowitz, G. McColl, L. E. Goldstein, R. E. Tanzi, and R. D. Moir. Amyloid-$\beta$ peptide protects against microbial infection in mouse and worm models of alzheimers disease. *Science translational medicine*, 8(340) :340ra72–340ra72, 2016.

[102] G. H. Lathe and C. R. Ruthven. The separation of substances on the basis of their molecular weights, using columns of starch and water. *The Biochemical Journal*, 60(4) :xxxiv, 1955.

[103] P. Laurençot. Weak solutions to the lifshitz-slyozov-wagner equation. *Indiana University mathematics journal*, 50(3) :1319–1346, 2001.

[104] P. Laurençot and S. Mischler. From the discrete to the continuous coagulation–fragmentation equations. *Proceedings of the Royal Society of Edinburgh : Section A Mathematics*, 132(05) :1219–1248, 2002.

[105] F.-X. Le Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations : theoretical aspects. *Tellus A*, 38(2) :97–110, 1986.

[106] G. Legname, I. V. Baskakov, H.-O. B. Nguyen, D. Riesner, F. E. Cohen, S. J. Dearmond, and S. B. Prusiner. Synthetic mammalian prions. *Science*, 305(5684) :673–676, 2004.

[107] N. Lenuzza. *Modélisation de la réplications des Prions : Implication de la dépendance en taille des agrégats de PrP et de l'hétérogénéité des populations cellulaires.* PhD thesis, Ecole Centrale Paris, 2009.

[108] R. J. LeVeque. *Finite-Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[109] J. M. Lewis, S. Lakshmivarahan, and S. Dhall. *Dynamic data assimilation : a least squares approach*, volume 13. Cambridge University Press, 2006.

[110] F. Leyvraz and H. Tschudi. Singularities in the kinetics of coagulation processes. *Journal of Physics A : Mathematical and General*, 14(12) :3389, 1981.

[111] I. Lifshitz and V. V. Slyozov. The kinetics of precipitation from supersaturated solid solutions. *Journal of physics and chemistry of solids*, 19 :35–50, 1961.

[112] L. Ljung and T. Glad. On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2) :265–276, 1994.

[113] D. G. Luenberger. *Determining the state of a linear system with observers of low dynamic order*. Department of Electrical Engineering, Stanford University., 1963.

[114] D. G. Luenberger. An introduction to observers. *IEEE Transactions on automatic control*, 16(6) :596–602, 1971.

[115] L. Manuelidis, Z.-X. Yu, N. Barquero, and B. Mullins. Cells infected with scrapie and creutzfeldt–jakob disease agents produce intracellular 25-nm virus-like particles. *Proceedings of the National Academy of Sciences*, 104(6) :1965–1970, 2007.

[116] J. Masel, V. A. A. Jansen, and M. A. Nowak. Quantifying the kinetic parameters of prion replication. *Biophysical chemistry*, 77(2) :139–152, 1999.

[117] P. S. Maybeck. *Stochastic models, estimation, and control*, volume 3. Academic press, 1982.

[118] J. Mendham. *Analyse chimique quantitative de Vogel*. De Boeck Supérieur, 2005.

[119] P. Moireau and D. Chapelle. Reduced-order unscented kalman filtering with application to parameter identification in large-dimensional systems. *ESAIM : Control, Optimisation and Calculus of Variations*, 17(2) :380–405, 2011.

[120] A. M. Morris, M. A. Watzky, and R. G. Finke. Protein aggregation kinetics, mechanism, and curve-fitting : a review of the literature. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1794(3) :375–397, 2009.

[121] I. Nåsell. An extension of the moment closure method. *Theoretical population biology*, 64(2) :233–239, 2003.

[122] I. M. Navon. Data assimilation for numerical weather prediction : a review. In *Data assimilation for atmospheric, oceanic and hydrologic applications*, pages 21–65. Springer, 2009.

[123] B. Niethammer and R. L. Pego. On the initial-value problem in the lifshitz–slyozov–wagner theory of ostwald ripening. *SIAM Journal on Mathematical Analysis*, 31(3) :467–485, 2000.

[124] B. Niethammer. On the evolution of large clusters in the becker-döring model. *Journal of Nonlinear Science*, 13(1) :115–122, 2003.

[125] B. Niethammer and R. L. Pego. Non-self-similar behavior in the lsw theory of ostwald ripening. *Journal of statistical physics*, 95(5-6) :867–902, 1999.

[126] B. Niethammer and R. L. Pego. The lsw model for domain coarsening : Asymptotic behavior for conserved total mass. *Journal of Statistical Physics*, 104(5-6) :1113–1144, 2001.

[127] M. A. Nowak, D. C. Krakauer, A. Klug, and R. M. May. Prion infection dynamics. *Integrative Biology Issues News and Reviews*, 1(1) :3–15, 1998.

[128] N. Ohsawa, C.-H. Song, A. Suzuki, H. Furuoka, R. Hasebe, and M. Horiuchi. Therapeutic effect of peripheral administration of an anti-prion protein antibody on mice infected with prions. *Microbiology and immunology*, 57(4) :288–297, 2013.

[129] F. Oosawa and S. Asakura. *Thermodynamics of the Polymerization of Protein*. Academic Press, 1975.

[130] S. Pant, B. Fabrèges, J.-F. Gerbeau, and I. E. Vignon-Clementel. A methodological paradigm for patient-specific multi-scale cfd simulations : from clinical measurements to parameter estimates for individual analysis. *International journal for numerical methods in biomedical engineering*, 30(12) :1614–1648, 2014.

[131] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3) :1065–1076, 1962.

[132] O. Penrose. The becker-döring equations at large times and their connection with the lsw theory of coarsening. *Journal of statistical physics*, 89(1-2) :305–320, 1997.

[133] O. Penrose. The becker-döring equations for the kinetics of phase transitions. In *Math. Proc. Camb. Phil. Soc*, volume 96, 2001.

[134] O. Penrose and J. Lebowitz. *Towards a rigorous theory of metastability.*, volume VII of *Studies in statistical mechanics*. Amsterdam : North-Holland, 1979.

[135] B. Perthame and J. P. Zubelli. On the inverse problem for a size-structured population model. *Inverse Problems*, 23(3) :1037, 2007.

[136] D. T. Pham, J. Verron, and M. C. Roubaud. A singular evolutive extended kalman filter for data assimilation in oceanography. *Journal of Marine systems*, 16(3) :323–340, 1998.

[137] H. Pohjanpalo. System identifiability based on the power series expansion of the solution. *Mathematical biosciences*, 41(1) :21–33, 1978.

[138] E. T. Powers and D. L. Powers. The kinetics of nucleated polymerizations at high concentrations : amyloid fibril formation near and above the "supercritical concentration". *Biophysical journal*, 91(1) :122–132, 2006.

[139] S. Prigent, A. Ballesta, F. Charles, N. Lenuzza, P. Gabriel, L. M. Tine, H. Rezaei, and M. Doumic. An Efficient Kinetic Model for Assemblies of Amyloid Fibrils and Its Application to Polyglutamine Aggregation. *PLoS ONE*, 7(11), 11 2012.

[140] S. Prigent and H. Rezaei. Prp assemblies. *Prion*, 5(2) :69–75, 2011. PMID : 21788728.

[141] S. B. Prusiner. Novel proteinaceous infectious particles cause scrapie. *Science*, 216(4542) :136–144, 1982.

[142] S. B. Prusiner. Molecular biology of prion diseases. *Science*, 252(5012) :1515–1522, 1991.

[143] S. B. Prusiner. Prions. *Proceedings of the National Academy of Sciences*, 95(23) :13363–13383, 1998.

[144] S. B. Prusiner, D. Groth, A. Serban, R. Koehler, D. Foster, M. Torchia, D. Burton, S.-L. Yang, and S. J. Dearmond. Ablation of the prion protein (prp) gene in mice prevents scrapie and facilitates production of anti-prp antibodies. *Proceedings of the National Academy of Sciences*, 90(22) :10608–10612, 1993.

[145] S. B. Prusiner, A. L. Woerman, D. A. Mordes, J. C. Watts, R. Rampersaud, D. B. Berry, S. Patel, A. Oehler, J. K. Lowe, S. N. Kravitz, D. H. Geschwindf, D. V. Gliddenh, G. M. Hallidayi, L. T. Middletonj, S. M. Gentlemank, L. T. Grinbergb, and K. Gilesa. Evidence for $\alpha$-synuclein prions causing multiple system atrophy in humans with parkinsonism. *Proceedings of the National Academy of Sciences*, 112(38) :E5308–E5317, 2015.

[146] J. Prüss, L. Pujo-Menjouet, G. Webb, and R. Zacher. Analysis of a model for the dynamics of prions. *Discrete Contin. Dyn. Syst. Ser. B*, 6(1) :225–235, 2006.

[147] K. Ramdani, M. Tucsnak, and G. Weiss. Recovering the initial state of an infinite-dimensional system using observers. *Automatica*, 46(10) :1616–1625, 2010.

[148] H. E. Rauch, C. T. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8) :1445–1450, 1965.

[149] H. E. Rauch, C. T. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8) :1445–1450, 1965.

[150] H. Rezaei, D. Marc, Y. Choiset, M. Takahashi, G. Hui Bon Hoa, T. Haertlé, J. Grosclaude, and P. Debey. High yield purification and physico-chemical properties of full-length recombinant allelic variants of sheep prion protein linked to scrapie susceptibility. *European Journal of Biochemistry*, 267(10) :2833–2839, 2000.

[151] J. Ruess, A. Milias-Argeitis, S. Summers, and J. Lygeros. Moment estimation for chemically reacting systems by extended kalman filtering. *The Journal of chemical physics*, 135(16) :165102, 2011.

[152] M. J. Sadowski, A. Verma, and T. Wisniewski. Infectious disease of the nervous system : prion diseases. *Bradley, WG. ; Daroff, RB. ; Fenichel, GM*, pages 1567–1581, 2008.

[153] H. Shim, A. Tanwani, and Z. Ping. Back-and-forth operation of state observers and norm estimation of estimation error. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 3221–3226. IEEE, 2012.

[154] S. K. Shoffner and S. Schnell. Estimation of the lag time in a subsequent monomer addition model for fibril elongation. *Physical Chemistry Chemical Physics*, 2016.

[155] J. R. Silveira, G. J. Raymond, A. G. Hughson, R. E. Race, V. L. Sim, S. F. Hayes, and B. Caughey. The most infectious prion protein particles. *Nature*, 437(7056) :257–261, 2005.

[156] R. Simha. Kinetics of degradation and size distribution of long chain polymers. *Journal of Applied Physics*, 12(7) :569–578, 1941.

[157] D. Simon. *Optimal State Estimation : Kalman, H Infinity, And Nonlinear Approaches.* Springer, 2006.

[158] D. Simon. Reduced order kalman filtering without model reduction. *Control and Intelligent Systems*, 35(2) :169, 2007.

[159] S. Simoneau, H. Rezaei, N. Salès, G. Kaiser-Schulz, M. Lefebvre-Roque, C. Vidal, J. Fournier, J. Comte, F. Wopfner, J. Grosclaude, H. Schätzl, and C. Lasmézas. In vitro and in vivo neurotoxicity of prion protein oligomers. *PLoS Pathog*, 3(8), Aug 2007.

[160] A. Singh and J. P. Hespanha. Approximate moment dynamics for chemically reacting systems. *IEEE Transactions on Automatic Control*, 56(2) :414–418, 2011.

[161] S. V. Slezov, V.V. Diffusive decomposition of solid solutions. *Soviet Physics Uspekhi*, 30(1) :23, 1987.

[162] M. Smoluchowski. Drei vortrage uber diffusion. brownsche bewegung und koagulation von kolloidteilchen. *Z. Phys.*, 17 :557–585, 1916.

[163] M. Smoluchowski. Versuch einer mathematischen theorie der koagulationskinetik kolloider lösungen. *Zeitschrift fuer physikalische Chemie*, pages 129–168, 1917.

[164] M. Smoluchowski and R. Criterion. Rayleigh scattering explained.

[165] D. Some. Light-scattering-based analysis of biomolecular interactions. *Biophysical Reviews*, 5(2) :147–158, 2013.

[166] Y. Song and J. W. Grizzle. The extended kalman filter as a local asymptotic observer for nonlinear discrete-time systems. In *American Control Conference, 1992*, pages 3365–3369. IEEE, 1992.

[167] J. Spouge. An existence theorem for the discrete coagulation-fragmentation equations. *MATH. Proc.Camb. Phil. Soc.*, 96 :351357, 1984.

[168] J. W. Strutt. Lviii. on the scattering of light by small particles. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(275) :447–454, 1871.

[169] J. W. Strutt. Xv. on the light from the sky, its polarization and colour. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(271) :107–120, 1871.

[170] W. Swietnicki, M. Morillas, S. G. Chen, P. Gambetti, and W. K. Surewicz. Aggregation and fibrillization of the recombinant human prion protein huprp90-231. *Biochemistry*, 39(2) :424–431, 2000.

[171] L. M. S. Tine. *Analyse Mathématique et Numérique de Modèles de Coagulation-Fragmentation.* PhD thesis, University of Lille and University Gaston Berger, 2011.

[172] L. M. Tine, T. Goudon, and F. Lagoutiere. The lifschitz-slyozov equation with space-diffusion of monomers. 2010.

[173] P. Tixador, L. Herzog, F. Reine, E. Jaumain, J. Chapuis, A. Le Dur, H. Laude, and V. Béringue. The physical relationship between infectivity and prion protein aggregates is strain-dependent. *PLoS Pathog*, 6(4) :1–12, 04 2010.

[174] M. F. Tuite. Genetics-psi no more for yeast prions. *Nature*, 370(6488) :327–328, 1994.

[175] R. Van Der Merwe. *Sigma-point Kalman filters for probabilistic inference in dynamic state-space models.* PhD thesis, Oregon Health & Science University, 2004.

[176] J. A. D. Wattis and J. R. King. Asymptotic solutions of the becker-döring equations. *Journal of Physics A : Mathematical and General*, 31(34) :7169, 1998.

[177] G. Welch and G. Bishop. An introduction to the kalman filter. department of computer science, university of north carolina, 2003.

[178] J. C. Willems. Deterministic least squares filtering. *Journal of econometrics*, 118(1) :341–373, 2004.

[179] W.-F. Xue, S. W. Homans, and S. E. Radford. Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly. *Proceedings of the National Academy of Sciences*, 105(26) :8926–8931, 2008.

[180] H. Zetterberg and K. Blennow. Biomarker evidence for uncoupling of amyloid build-up and toxicity in alzheimer's disease. *Alzheimer's & Dementia*, 9(4) :459–462, 2013.

[181] J. Zhang and M. Muthukumar. Simulations of nucleation and elongation of amyloid fibrils. *The Journal of chemical physics*, 130(3) :035102, 2009.