Information Content in Data Sets for a Nucleated-Polymerization Model

H.T. Banks¹, M. Doumic²,³, C. Kruse²,³, S. Prigent²,³, H.Rezaei⁴

¹ Center for Research in Scientific Computation
North Carolina State University, Raleigh, NC 27695-8212

²Institut National de Recherche en Informatique et Automatique
Paris-Rocquencourt, France

³Pierre et Marie Curie University
Paris, France

⁴Institut National de Recherche Agronomique
Jouy-en-Josas, France

April 28, 2015

Abstract
We illustrate the use of statistical tools (asymptotic theories of standard error quantification using appropriate statistical models, bootstrapping, model comparison techniques) in addition to sensitivity analysis that may be employed to determine the information content in data sets. We do this in the context of recent models [25] for nucleated polymerization in proteins, about which very little is known regarding the underlying mechanisms; thus the methodology we develop here may be of great help to experimentalists. We conclude that the investigated data sets will support with reasonable levels of uncertainty only the estimation of the parameters related to the early steps of the aggregation process.

Key Words: Inverse problems, polyglutamine and aggregation modeling, information content, sensitivity, uncertainty quantification

Mathematics Subject Classification: 65M32,62P10,64B10,49Q12
1 Introduction

As mathematical models become more complex with multiple states, increasing numbers of parameters need to be estimated using experimental data. Thus there is a need for critical analysis in model validation related to the reliability of parameter estimates obtained in model fitting. A recent concrete example involves previous HIV models [1, 6] with 15 or more parameters to be estimated. In [4], using recently developed parameter selectivity tools [5] based on parameter sensitivity based scores, it was shown that a number of these parameters could not be estimated with any degree of reliability. Moreover, we found that quantifiable uncertainty varies among patients depending upon the number of treatment interruptions (perturbations of therapy). This leads to a fundamental question: how much information with respect to model validation can be expected in a given data set or collection of data sets?

Here we illustrate the use of other tools: asymptotic theories (as the data sample size increases without bound) of standard error quantification using appropriate statistical models, bootstrapping (repeated sampling of synthetic data similar to the original data), and statistical (ANOVA type) model comparison techniques—these are discussed in some detail in the recent monograph [9] as well as in numerous statistical texts). Such techniques are employed in addition to sensitivity theory in order to determine the information content in data sets. We pursue this in the context of recent models [25] for nucleated polymerization in proteins.

After briefly outlining the biological context of amyloid formation, we describe the model in Section 2. In Section 3, we investigate the statistical model to be used with our noisy data. This is a necessary step in order to use the correct error model in our generalized least squares (GLS) minimization. This also reveals information on our experimental observation process. Once we have found parameters which allow a reasonable fit, we determine the confidence we may have in our estimation procedures. We do this in Section 4, using both the condition number of the covariance matrix and a sensitivity analysis. This reveals a smaller number of parameters (than those estimated in [25]) to be reasonably sensitive to the data sets, whereas others do not really affect the quality of the fits to our data. To further support our sensitivity findings, we then apply a bootstrapping analysis in Section 5. We are lead to four main parameters and compare their resulting errors with the asymptotic confidence intervals of Section 4. Finally, in Section 6, we carry
out model comparison tests [8, 9, 11] as used in [3], and these lead us to select three parameters out of the nine original ones estimated in [25] that can be reliably estimated.

1.1 Protein Polymerization

It is now known that several neuro-degenerative disorders, including Alzheimer's disease, Huntington's disease and Prion diseases e.g., mad cow, are related to aggregations of proteins presenting an abnormal folding. These protein aggregates are called amyloids and have become a focus of modeling efforts in recent years [12, 25, 29, 30, 31]. One of the main challenges in this field is to understand the key aggregation mechanisms, both qualitatively and quantitatively. In order to test our methodology on a relatively simple case, we focus here on polyglutamine (PolyQ) containing proteins. This was also the case study chosen to illustrate the fairly general ODE-PDE model proposed in [25]; the reason for our choice is that, as shown in [25], the polymerization mechanisms prove to be simpler for PolyQ aggregation than for other types of proteins, e.g. PrP [21]. To understand data sets from experiments carried out by Human Rezaei and his team at INRA (Virologie et Immunologie Moleculaires), see [25], we adapt the general model to this context. The data sets (DS1-DS4) of interest to us here are depicted in Figure 1 below and record the evolution of normalized total polymerized mass in time.

In [25] and a subsequent effort in [7], the authors sought to investigate several questions including (i) understanding the key polymerization mechanisms, (ii) how to numerically approximate the model, and (iii) how to select parameters and calibrate the model. Here we briefly summarize results related to (ii) and focus primarily on (iii).

2 The Model

2.1 Original ODE Model

We briefly outline the model that is the same as that of [25]. Let \((V, V^*, c_i)\) be the main variables of interest consisting respectively of concentrations of the normal proteins that we will call monomers (basic subunits that are repeated in a chainlike fashion), of the monomeric proteins presenting an abnormal configuration that we will call conformers, and of the \(i\)-polymers made of
Figure 1: The data sets of interest from [25, 7]. The total polymerized mass is measured by Thioflavine T (ThT), which is one of the most common experimental tools for \textit{in vitro} protein polymerization (see [25, 29]).

\( c_0 = 200 \mu M \)

\( \text{time (hour)} \)

\( \% \text{ of the total polymerized mass} \)

\( 0 \)

\( 1 \)

\( 0.1 \)

\( 0.2 \)

\( 0.3 \)

\( 0.4 \)

\( 0.5 \)

\( 0.6 \)

\( 0.7 \)

\( 0.8 \)

\( 0.9 \)

\( 1 \)

\( 0 \)

\( 1 \)

\( 2 \)

\( 3 \)

\( 4 \)

\( 5 \)

\( 6 \)

\( 7 \)

\( 8 \)

The following comprise the fundamental dynamics modeled in [25] which the interested reader may consult for further details:

- Monomer-conformer exchange: \( V \xrightarrow{k_j^+} V^* \). This represents the spontaneous formation of an active form of the monomer, denoted here \( V^* \), out of the initially present inert form denoted \( V \). The inert form cannot react and form fibrils, whereas the active conformer may.

- Nucleation: \( V^* + V^* + ... + V^* \xrightarrow{k_{on}^N} k_{off}^N c_{i_0} \). This is the spontaneous formation of the smallest stable polymer, formed by the addition of a certain number \( i_0 \) of active conformers. This smallest stable polymer is called the \textit{nucleus}.

- Polymerization by conformer addition: \( c_i + V^* \xrightarrow{k_{on}^c} c_{i+1} \). Once a nu-
cleus is formed, its size grows progressively by addition of an active conformer.

Other reactions like fragmentation and coalescence are negligible for the case of polyglutamine containing proteins (see [25] for experimental justification).

The law of mass action in the deterministic framework (see [11, 27] and the numerous references therein), translates \( A + \frac{k_i^+}{k_i^-} B \rightleftharpoons A' + B' \) into the ordinary differential equation \( \frac{d[A]}{dt} = -k^+[A][B] + k^-[A'][B]' \). Using these basic ideas we obtain the infinite system of ordinary differential equations (ODEs) studied in [25]

\[
\begin{align*}
\frac{dV}{dt} &= -k_I^+ V + k_I^- V^*, \\
\frac{dV^*}{dt} &= k_I^+ V - k_I^- V^* + i_0 k_{off} N c_{i_0} - V^* \sum_{i \geq i_0} k_{on}^i c_i, \\
\frac{dc_{i_0}}{dt} &= k_{on}^N (V^*)^{i_0} - k_{off} N c_{i_0} - i_0 k_{on}^i c_{i_0} V^*, \\
\frac{dc_i}{dt} &= V^* (k_{on}^{i-1} c_{i-1} - k_{on}^i c_i), \quad i = i_0 + 1, \ldots \ \\
\end{align*}
\]

with initial conditions

\[
V(0) = c_0, \quad V^*(0) = 0, \quad c_{i_0}(0) = c_i(0) = 0
\]

and the mass balance equation

\[
\frac{d}{dt} \left( V + V^* + \sum_{i = i_0}^{\infty} ic_i \right) = 0.
\]

The experiments of interest to us measure the a-dimensional total polymerized mass, or \( Madim \) which is given by

\[
M(t) = \sum_{i \geq i_0} ic_i(t).
\]
2.2 An Approximate PDE System and the Associated Forward Problem

Amyloid formations are characterized by very long polymers (a fibril may contain up to $10^6$ monomer units). A PDE version of the standard model, where a continuous variable $x$ approximates the discrete sizes $i$, is thus a reasonable approximation for large amyloid polymers. However, for small polymer sizes this resulting continuum approximation does not work very well. Thus we take a “hybrid approach” of retaining the ODE for smaller sizes and using the PDE for larger ones, see [7].

We define a small parameter $\varepsilon = \frac{1}{i_M}$, and let $x_i = i\varepsilon$ with $i_M \gg 1$ being the average polymer size defined by

$$i_M = \frac{\sum_{i \geq i_0} i c_i}{\sum c_i}.$$ 

Then after definition of dimensionless quantities

$$c^\varepsilon(t, x) = \sum c_i 1_{[x_i, x_{i+1}]}$$ 

we may obtain a partial differential equation (PDE) to replace the infinite ODE system. Rigorous derivations of such continuous integro-PDE models may be found in [22] for coagulation-fragmentation equations, in [15] for the limit of the so-called Becker-Dring system (polymerization and depolymerization reactions are the only reactions considered) toward its continuous limit called the Lifshitz-Slyozov model, and in [19] for the growth-fragmentation “Prion Mode”. A formal derivation for a full model, also including nucleation, is carried out in [25].

Let $N_0 \in \mathbb{N}$. We then use the approximation

$$\frac{dV}{dt} = -k^+_IV + k^-_IV^*,$$

$$\frac{dV^*}{dt} = k^+_IV - k^-_IV^* + i_0k_{off}^Nc_{i_0} - V^* \sum_{i \geq i_0} k_{on}^ic_i,$$ 

(5)

$$\frac{dc_{i_0}}{dt} = k_{on}^N(V^*)^i - k_{off}^Nc_{i_0} - k_{on}^i c_{i_0}V^*,$$ 

(6)

$$\frac{dc_i}{dt} = V^*(k_{on}^{i-1}c_{i-1} - k_{on}^i c_i), \quad i \leq N_0,$$ 

(7)

$$\partial_x c^\varepsilon(x, t) = -V^*\partial_x (k_{on}(x)c^\varepsilon(x, t)), \quad x \geq N_0,$$ 

(8)
with initial conditions
\[ V(0) = c_0, \quad V^*(0) = 0, \quad c_{i_0}(0) = c_i(0) = 0, \quad c^\varepsilon(x,0) = 0, \]
and the boundary condition
\[ c^\varepsilon(x = N_0, t) = c_{N_0}(t). \]
Here we have passed to the continuous representations for chain lengths larger than \( i = N_0 \).

Then an assumed mass balance equation becomes
\[
\frac{d}{dt} \left( V + V^* + \sum_{i=i_0}^{N_0} i c_i + \int_{N_0}^\infty x c^\varepsilon(x) \, dx \right) = 0.
\]

In [7] we considered requirements for a good discretization scheme including (i) it should conserve the a-dimensional total polymerized mass (\( \text{Madim} \)), (ii) it should be fast and most importantly, (iii) it should be accurate.

To ensure the conservation of mass, we replace the ODE for \( V^* \) by the mass conservation equation and obtain
\[
\frac{dV}{dt} = -k_i^+ V + k_i^- V^*,
\]
\[
V^* = c_0 - V - \sum_{i=i_0}^{N_0} i c_i - \int_{N_0}^\infty x c^\varepsilon(x) \, dx,
\]
\[
\frac{dc_{i_0}}{dt} = k_{i_0}^N (V^*)^{i_0} - k_{i_0}^{i_1} c_{i_0} - k_{i_1}^{i_0} c_{i_0} V^*,
\]
\[
\frac{dc_i}{dt} = V^*(k_{i-1}^{i_1} c_{i-1} - k_{i_0}^i c_i), \quad i \leq N_0,
\]
\[
\partial_x c^\varepsilon(x, t) = -V^* \partial_x (k_{on}(x)c^\varepsilon(x,t)), \quad x \geq N_0,
\]
with initial and boundary conditions as before.

We developed methodology for forward solutions in [7]. We first point out that the desired spatial computational domain is very large as determined by the maximum size of observed polymers, with range up to \( 10^6 \). The peak in the distribution is at the left side of the domain of interest; for larger polymer sizes, the distribution is almost linearly decreasing.
Based on these and other considerations discussed in [7], the PDE was approximated by the Finite Volume Method (see [23] for discussions of Upwind, Lax-Wendroff and flux limiter methods) with an adaptive mesh, refined toward the smaller polymer sizes. Furthermore, we kept the ratio between the step size and the corresponding mesh element constant, i.e., we used $\frac{\Delta x_i}{x_i} = q < 1$ so that $x_i = \frac{1}{1-q}x_{i-1}$. This mesh is quasi-linear in the sense of $\frac{\Delta x_{i+1}}{\Delta x_i} = 1 + O(q)$. The resulting Upwind and Lax-Wendroff schemes are then consistent on the progressive mesh (see [23]). For further details on these schemes including examples demonstrating convergence properties, the interested reader may consult [7].

3 The Inverse Problem

A major question in formulating the model for use in inverse problem scenarios consists of how to best parametrically represent the polymerization parameters $k_i^{on}$ of (9) and the function $k_{on}$ of (10) for our application. We do this in a combined piecewise continuous formulation.

For this continuous polymerization function $k_{on}(x)$, we use the piecewise linear representation (see Figure 2)

$$k_{on}(x) = \begin{cases} k_{on}^{min} + \frac{x_{i_{max}}^2 - k_{on}^{min}}{x_{i_{max}}^2 - x_{i_{min}}}, & x \leq x_{i_{max}} \\ k_{on}^{max}, & x_{i_{max}} < x \leq x_{2i_{max}} \\ k_{on}^{max} - \frac{x_{i_{max}}^2}{x_{i_{max}}^2(1-x_{2})}, & x_{2i_{max}} < x \leq i_{max} \\ 0, & x \geq i_{max} \end{cases}$$

In the numerical approximations, we chose $i_0 = 2, N_0 = 500$. The discrete polymerization parameters $k_{on}^i, i = i_0, ..., N_0$, are then obtained as

$$k_{on}^i := k_{on}(x = i).$$

Following [25], we chose to approximate $k_{on}$ by a function as depicted in Figure 2. (According to our discussions with S. Prigent, H. Rezaei and J. Torrent, other choices like a Gaussian bell curve are also possible, but as we will subsequently conclude, the presently available data will not support estimation of parameters in these representations). Thus with this parametrization we have 5 more parameters $k_{on}^{min}, k_{on}^{max}, x, x_{2}, i_{max}$ in addition to the 4 basic parameters $k_{I}^+, k_{I}^-, k_{on}, k_{off}^N$ to be estimated using our data sets.
Thus we seek to estimate (with acceptable quantification of uncertainties) the nine parameters $k_I^+, k_I^-, k_{on}, k_{off}, k_{on}^N,$ and $k_{on}$ (represented in parametrical form depicted above with the 5 additional unknowns $k_{on}^{min}, k_{on}^{max}, x_1, x_2, i_{max}$) that fit the data best! To do this we need an efficient discretization method as discussed above for the forward problem as well as a correct assumption on the measurement errors in the inverse problem.

### 3.1 Estimation of Parameters

We make some standard statistical assumptions (see [9, 11, 17, 28]) underlying our inverse problem formulations.

- Assume that there exists a true or nominal set of parameter $\theta_0 = (k_I^-, ..., i_{max})$
- Let $\mathcal{E}_i$ be i.i.d with $\mathbb{E}(\mathcal{E}_i) = 0$ and $\text{cov}(\mathcal{E}_i, \mathcal{E}_i) = \sigma_0^2$, where $i = 1, ..., n$ and $n$ is the number of observations or data points in the given data set.

Denote as $\hat{\theta}$ the estimated parameter for $\theta_0$. The inverse problem is based on statistical assumptions on the observation error in the data.

If we assume an absolute error data model, then data points are taken with equal importance. This is represented by observations (here $\epsilon_i$ is a
realization of the error process \( E_i \)
\[
y_i = M(t_i, \theta_0) + \epsilon_i. \tag{11}
\]

On the other hand, if one assumes some type of relative error data model, then the error is proportional in some sense to the measured polymerized mass. This can be represented by observations of the form
\[
y_i = M(t_i, \theta_0) + M(t_i, \theta_0) \gamma \epsilon_i, \quad \gamma \in (0, 1]. \tag{12}
\]

Absolute model error formulations dictate we use Ordinary Least Squares (OLS) inverse problem \([9, 11]\) given by
\[
\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^{n} (y_i - M(t_i, \theta))^2 \tag{13}
\]
while for relative error models one should use inverse problem formulations with Generalized Least Squares (GLS) cost functional
\[
\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^{n} \left( \frac{y_i - M(t_i, \theta)}{M(t_i, \theta) \gamma} \right)^2, \quad \gamma \in (0, 1]. \tag{14}
\]

3.1.1 The Residual Plots

To obtain a best statistical model, we used residual plots (see \([9, 11]\) for more details) with residuals given by
\[
r_i = \frac{y_i - M(t_i, \hat{\theta})}{M(t_i, \theta) \gamma}, \quad \gamma \in [0, 1].
\]

To illustrate what we are seeking for our data sets, we first used simulated relative error data (simulated data for \( \gamma = 1 \)), then carried out the inverse problems for both a relative error cost functional (i.e., \( \gamma = 1 \)) and an ordinary least squares cost functional (i.e., \( \gamma = 0 \)). We then plotted the corresponding residuals vs. time and also residuals vs. the model values. The first plots are related to the correctness of our assumption of independency and identical distributions \textit{i.i.d.} for the data whereas the second plots contain information as to the correctness of the form of our proposed statistical model.
Figure 3: Plots with simulated data: (a) Correct cost function vs. time ($\gamma = 1$); (b) Incorrect cost function vs. time ($\gamma = 0$)

Figure 4: Plots with simulated data: (a) Correct cost function vs. model ($\gamma = 1$); (b) Incorrect cost function vs. model ($\gamma = 0$)
3.2 Statistical Models of Noise

We next carried out similar inverse problems with data set (DS) 4 of our experimental data collection. We first used DS 4 on the interval \( t \in [0, 8] \). Based on some earlier calculations we also chose the nucleation index \( i_0 = 2 \) for all our subsequent calculations. The residual plots given below in Figures 5 and 6 suggest strongly that neither of the first attempts of assumed statistical models and corresponding cost functionals (absolute error and OLS or relative error with \( \gamma = 1 \) and simple GLS) are correct.

Figure 5: (a) \( M(t_k) \) (Madim) with OLS; (b) Residuals vs. Model: OLS

Figure 6: (a) \( M(t_k) \) with GLS, \( \gamma = 1 \); (b) Residuals vs. Model: GLS

Based on these initial results and the speculation that early periods of the polymerization process may be somewhat stochastic in nature, we chose to
subsequently use all the data sets on the intervals \([t_0, 8]\) where \(t_0\) is the first time when \(M(t_0) > 0.12\) (thus 12% of the a-dimensional total polymerized mass). Moreover, we decided to use other values of \(\gamma\) between 0 and 1 to test data set 4. Setting \(i_0 = 2\), we focused on the question of the most appropriate values of \(\gamma\) to use in a generalized least squares approach (again see [9] for further motivation and details). We then obtained the results with data set 4 depicted in Figure 7. Analysis of these residuals suggest that either \(\gamma = 0.6\) or \(\gamma = 0.7\) might be satisfactory for use in a generalized least squares setting.

Motivated by these results, we next investigated the inverse problems for each of the four experimental data sets with initial concentration \(c_0 = 200\) \(\mu\)mol and \(i_0 = 2\). We carried out the optimization over all data points with \(M(t_k) \geq 0.12\) and used the generalized least squares method with \(\gamma = 0.6\). The resulting graphics depicted in Figure 8 again suggest that \(\gamma = 0.6\) is a reasonable value to use in our subsequent analysis of the polyglutamine data with regard to its information content for inverse problem estimation and parameter uncertainty quantification.
Figure 8: Residuals for the 4 experimental data sets using $\gamma = 0.6$. 
4 Standard Errors and Asymptotic Analysis

4.1 Standard Errors for Parameters Using GLS

We employed first the asymptotic theory (as $n \to \infty$) for parameter uncertainty summarized in [9, 11, 17] and the references therein. In the case of generalized least squares, the associated standard errors for the estimated parameters $\hat{\theta} = (k_1^\top, \ldots, i_{\max})$ (vector length $\kappa_\theta = 9$) are given by the following construction (for details see Chap. 3.2.5 and 3.2.6 of [9]).

We may define the standard errors by the formula

$$SE_k = \sqrt{\Sigma_{kk}(\hat{\theta})}, \quad k = 1, \ldots, 9,$$

where the covariance matrix $\Sigma$ is given by

$$\Sigma(\hat{\theta}) = \hat{\sigma}^2(\chi^T(\hat{\theta})W(\hat{\theta})\chi(\hat{\theta}))^{-1}.$$

Here

$$\chi = \frac{\partial M}{\partial \theta} = \left(\frac{\partial M(t_1; \hat{\theta})}{\partial \theta}, \ldots, \frac{\partial M(t_n; \hat{\theta})}{\partial \theta}\right)$$

is the sensitivity matrix of size $n \times \kappa_\theta = n \times 9$ ($n$ being the number of data points and $\kappa_\theta = 9$ being the number of estimated parameters) and $W$ is defined by

$$W^{-1}(\hat{\theta}) = \text{diag}(M(t_1; \hat{\theta})^{2\gamma}, \ldots, M(t_n; \hat{\theta})^{2\gamma}).$$

We use the approximation of the variance

$$\sigma^2_0 \approx \hat{\sigma}(\hat{\theta})^2 = \frac{1}{n - \kappa_\theta} \sum_{i=1}^{n} \frac{1}{M(t_i; \hat{\theta})^{2\gamma}} (M(t_i, \hat{\theta}) - y_i)^2.$$

To obtain a finite standard error using asymptotic theory, the $\kappa_\phi \times \kappa_\theta = 9 \times 9$ matrix $F = \chi^T(\hat{\theta})W(\hat{\theta})\chi(\hat{\theta})$ thus must be invertible. In the above problem we do indeed obtain a good fit of the curve and good residuals (for the sake of brevity, not depicted here!). However, we also found that the condition number of the matrix

$$F = \chi^T(\hat{\theta})W(\hat{\theta})\chi(\hat{\theta})$$

is $\kappa = 10^{24}$. Looking more closely at the matrix $F$ reveals a near linear dependence between certain rows, hence the large condition number. We thus quickly reach the following conclusions:
1. We obtain a set of parameters for which the model fits well, but we cannot have any reasonable confidence in them using the asymptotic theories from statistics, e.g., see the references given above.

2. We suspect that it may not be possible to obtain sufficient information from our data set curves to estimate all 9 parameters with a high degree of confidence! This is based on our calculations with the corresponding Fisher matrices as well our prior knowledge in that the graphs depicted in Figure 1 are very similar to Logistic or Gompertz curves. These curves can be quite well fit with parameterized models with only 2 or 3 carefully chosen parameters!

To assist in initial understanding of these issues, we consider components of the associated sensitivity matrices \( \chi = \frac{\partial M}{\partial \theta} \).

### 4.2 Sensitivity Analysis

For the sensitivity analysis, we follow [9, 11] and carry out all computations using the differential system of sensitivity equations as detailed in those references. Hereafter all our analysis will be carried out using data set 4 and the best estimate \( \hat{\theta} \) obtained for the latter. We find that the model is sensitive mainly to four parameters: \( k^+_I, k^-_I, k^+_N, k^N_{off} \). The sensitivities for the remaining parameters are on an order of magnitude of \( 10^{-6} \) or less. It also shows some sensitivity with respect to \( x_1 \). However, the parameter \( x_1 \) appears in the model only as the factor \( x_1i_{max} \). The sensitivities depicted below use \( \hat{\theta} \) for the nine best fit GLS parameters, i.e., \( \hat{\theta} \) for \( \kappa_\theta = 9 \). We note that since we use the a-dimensional quantity \( M \) in the cost functionals (OLS or GLS), it is the sensitivity of this quantity with respect to the parameters \( \theta \), rather than any relative sensitivities, that will determine changes in the cost functionals to be minimized with respect to changes in the parameters.
Figure 9: (a) Sensitivity w.r.t. $k_I^-$; (b) Sensitivity w.r.t. $k_I^+$

Figure 10: (a) Sensitivity w.r.t. $k_{on}^N$; (b) Sensitivity w.r.t. $k_{off}^N$
Figure 11: (a) Sensitivity w.r.t. $k_{on}^\text{min}$; (b) Sensitivity w.r.t. $k_{off}^\text{max}$

Figure 12: (a) Sensitivity w.r.t. $x_1$; (b) Sensitivity w.r.t. $x_2$
5 Sensitivity Motivated Inverse Problems

Based on the sensitivity findings depicted above, we investigated a series of inverse problems in which we attempted to estimate an increasing number of parameters beginning first with the fundamental parameters $k^+_I$ and $k^-_I$. In each of these inverse problems we attempted to ascertain uncertainty bounds for the estimated parameters using both the asymptotic theory described above and a generalized least squares version of bootstrapping [13, 14, 16, 18, 20].

A quick outline of the appropriate bootstrapping algorithm is given next.

5.1 Bootstrapping Algorithm: Nonconstant Variance Data

We suppose now that we are given experimental data $(t_1, y_1), \ldots, (t_n, y_n)$ from the underlying observation process

$$Y_i = M(t_i; \theta_0) + M(t_i; \theta_0)\gamma \tilde{E}_i,$$

where $i = 1, \ldots, n$ and the $\tilde{E}_i$ are i.i.d. with mean zero and constant variance $\sigma^2_0$. Then we see that $E(Y_i) = M(t_i; \theta_0)$ and $Var(Y_i) = \sigma^2_0 M^2\gamma(t_i, \theta_0)$, with associated corresponding realizations of $Y_i$ given by

$$y_i = M(t_i; \theta_0) + M(t_i; \theta_0)\gamma \tilde{e}_i.$$
A standard algorithm can be used to compute the corresponding bootstrapping estimate \( \hat{\theta}_{boot} \) of \( \theta_0 \) and its empirical distribution. We treat the general case for nonlinear dependence of the model output on the parameters \( \theta \). The algorithm is given as follows.

1. First obtain the estimate \( \hat{\theta}^0 \) from the entire sample \( \{y_i\} \) using the GLS given in (14) with \( \gamma = 1 \). An estimate \( \hat{\theta}_{boot} \) can be solved for iteratively as follows.

2. Define the nonconstant variance standardized residuals

\[
\bar{s}_i = \frac{y_i - M(t_i; \hat{\theta}^0)}{M(t_i; \theta^0)\gamma}, \quad i = 1, 2, \ldots, n.
\]

Set \( m = 0 \).

3. Create a bootstrapping sample of size \( n \) using random sampling with replacement from the data (realizations) \( \{\bar{s}_1, \ldots, \bar{s}_n\} \) to form a bootstrapping sample \( \{s^m_1, \ldots, s^m_n\} \).

4. Create bootstrapping sample points

\[
y^m_i = M(t_i; \hat{\theta}^0) + M(t_i; \hat{\theta}^0)\gamma \bar{s}_i^m, \quad \text{where } i = 1, \ldots, n.
\]

5. Obtain a new estimate \( \hat{\theta}^{m+1} \) from the bootstrapping sample \( \{y^m_i\} \) using GLS.

6. Set \( m = m + 1 \) and repeat steps 3–5 until \( m \geq M \) where \( M \) is large (e.g., \( M = 1000 \)).

We then calculate the mean, standard error (SE), and confidence intervals using the formulae

\[
\hat{\theta}_{boot} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^m,
\]

\[
\text{Var}(\hat{\theta}_{boot}) = \frac{1}{M - 1} \sum_{m=1}^{M} (\hat{\theta}^m - \hat{\theta}_{boot})^T (\hat{\theta}^m - \hat{\theta}_{boot}),
\]

\[
\text{SE}_k(\hat{\theta}_{boot}) = \sqrt{\text{Var}(\theta_{boot})_{kk}},
\]

where \( \theta_{boot} \) denotes the bootstrapping estimator.
5.2 Estimation of two parameters

We first carried out estimation for the 2 parameters $k_I^+$ and $k_I^-$. We use the GLS formulation with $\gamma = 0.6$. We fix globally (based on previous estimations with DS 4) the parameter values $k_N^{\text{on}}$ of $k_N^{\text{off}}$ on $k_{\text{on}}^{\text{max}}$ on $x_1$ $x_2$ $i_{\text{max}}$

<table>
<thead>
<tr>
<th>$k_{\text{on}}^{\text{N}}$</th>
<th>$k_{\text{off}}^{\text{N}}$</th>
<th>$k_{\text{on}}^{\text{min}}$</th>
<th>$k_{\text{on}}^{\text{max}}$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$i_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4616.962</td>
<td>93.332</td>
<td>1684.381</td>
<td>1.5152 · 10^9</td>
<td>0.0626</td>
<td>0.859</td>
<td>3.542 · 10^6</td>
</tr>
</tbody>
</table>

and used the initial guesses for the parameters given by

<table>
<thead>
<tr>
<th>q0</th>
<th>$k_I^+$</th>
<th>$k_I^-$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.1600</td>
<td>10.9270</td>
</tr>
</tbody>
</table>

We then used the bootstrapping algorithm and obtained the following means and standard errors for $M = 1000$ which, as reported below, compare quite well with the asymptotic theory estimates. The corresponding distributions are shown in Figures 14 and 15.

<table>
<thead>
<tr>
<th>$k_I^+(\text{boot})(\text{GLS})$</th>
<th>$k_I^-(\text{boot})(\text{GLS})$</th>
<th>$k_I^+(\text{asymp})(\text{GLS})$</th>
<th>$k_I^-(\text{asymp})(\text{GLS})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>2.158</td>
<td>10.911</td>
<td>2.157</td>
</tr>
<tr>
<td>SE</td>
<td>0.0044</td>
<td>0.0247</td>
<td>0.00396</td>
</tr>
</tbody>
</table>

5.3 GLS Estimation of 3 Parameters

We tried next to estimate 3 parameters using the GLS formulation with $\gamma = 0.6$. Once again we fixed all the parameters describing the domain and the polymerization function $k_{\text{on}}$ and we also fixed either $k_{\text{off}}^{\text{N}}$ or $k_{\text{on}}^{\text{N}}$ in the corresponding inverse problems.

5.4 GLS Estimation for $k_I^+$, $k_I^-$ and $k_{\text{on}}^{\text{N}}$

We fixed values as follows:

<table>
<thead>
<tr>
<th>$k_{\text{off}}^{\text{N}}$</th>
<th>$k_{\text{on}}^{\text{min}}$</th>
<th>$k_{\text{on}}^{\text{max}}$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$i_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>93.33</td>
<td>1684.38</td>
<td>1.5 · 10^9</td>
<td>0.062</td>
<td>0.859</td>
<td>3.5 · 10^6</td>
</tr>
</tbody>
</table>

We used as initial parameter values:

<table>
<thead>
<tr>
<th>q0</th>
<th>$k_I^+$</th>
<th>$k_I^-$</th>
<th>$k_{\text{on}}^{\text{N}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.1600</td>
<td>10.9270</td>
<td>4616.962</td>
</tr>
</tbody>
</table>
Figure 14: Two parameters estimation ($k_I^+, k_I^-$). Bootstrapping distribution for $k_I^+$. We use GLS and $M=1000$ runs.

Figure 15: Two parameters estimation ($k_I^+, k_I^-$). Bootstrapping distribution for $k_I^-$. We use GLS and $M=1000$ runs.
We obtained the estimated parameters together with the corresponding standard errors, variances and the condition numbers $\kappa$ of the corresponding sensitivity matrices for the four data sets as reported below. The 95% confidence results based on the asymptotic theory are also depicted for DS 4 in Figure 16.

<table>
<thead>
<tr>
<th></th>
<th>$k_I^+$</th>
<th>$k_I^-$</th>
<th>$k_{on}^N$</th>
<th>$SE$</th>
<th>$\sigma^2$</th>
<th>$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>2.26</td>
<td>13.49</td>
<td>4616.96</td>
<td>(.012, .099, 53.925)</td>
<td>8.52 · 10^{-6}</td>
<td>8.89 · 10^{10}</td>
</tr>
<tr>
<td>DS2</td>
<td>2.99</td>
<td>16.20</td>
<td>4616.96</td>
<td>(.021, .151, 56.691)</td>
<td>9.67 · 10^{-6}</td>
<td>4.37 · 10^{10}</td>
</tr>
<tr>
<td>DS3</td>
<td>2.18</td>
<td>15.76</td>
<td>9840.31</td>
<td>(.011, .103, 90.466)</td>
<td>6.45 · 10^{-6}</td>
<td>3.94 · 10^{11}</td>
</tr>
<tr>
<td>DS4</td>
<td>2.16</td>
<td>10.91</td>
<td>4616.96</td>
<td>(0.0089, 0.0649, 45.262)</td>
<td>6.36 · 10^{-6}</td>
<td>7.14 · 10^{10}</td>
</tr>
</tbody>
</table>

Figure 16: Confidence Intervals

To compare these asymptotic results with bootstrapping, we carried out bootstrapping with Data Set (DS) 4 for the estimation of $k_I^+$, $k_I^-$ and $k_{on}^N$ with the same initial values as above. We then obtained the following means and standard errors (SE) for a run with $M = 1000$, in comparison to the asymptotic theory.

<table>
<thead>
<tr>
<th></th>
<th>$k_I^+$ (boot)</th>
<th>$k_I^-$ (boot)</th>
<th>$k_{on}^N$ (boot)</th>
<th>$k_I^+$ (asymp)</th>
<th>$k_I^-$ (asymp)</th>
<th>$k_{on}^N$ (asymp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>2.153</td>
<td>10.887</td>
<td>4616.962</td>
<td>2.157</td>
<td>10.910</td>
<td>4616.962</td>
</tr>
<tr>
<td>SE</td>
<td>0.0039</td>
<td>0.0219</td>
<td>0.00003</td>
<td>0.0089</td>
<td>0.0649</td>
<td>45.262</td>
</tr>
</tbody>
</table>
Of particular interest are the values obtained for $k^N_{on}$ and the bootstrapping standard errors for $k^N_{on}$ which are extremely small. It should be noted that the sensitivity of the model output on $k^N_{on}$ is also very small. Thus one might conjecture that the iterations in the bootstrapping algorithm do not change the values of $k^N_{on}$ very much and hence one observes the extremely small SE that are produced for the bootstrapping estimates.

Figure 17: Estimation for $k^+_I$, $k^-_I$ and $k^N_{on}$: Bootstrapping distribution for $k^-_I$ for GLS and $M = 1000$ runs.
Figure 18: Estimation for $k^+_I$, $k^-_I$ and $k^N_{on}$. Bootstrapping distribution for $k^+_I$ for GLS and $M = 1000$ runs.

Figure 19: Estimation for $k^+_I$, $k^-_I$ and $k^N_{on}$. Bootstrapping distribution for $k^N_{on}$ for GLS and $M = 1000$ runs.
5.5 GLS estimation for $k_I^+, k_I^-$ and $k_{off}^N$

In another test, we fixed $k_{on}^N$ and instead estimate $k_{off}^N$ (along with $k_I^+$ and $k_I^-$). We use the fixed values:

$$
\begin{array}{|c|c|c|c|c|}
\hline
k_{on}^N & k_{on}^{min} & k_{on}^{max} & x_1 & x_2 \\
\hline
4616.962 & 1684.381 & 1.5152 \cdot 10^9 & 0.0626 & 0.859 \cdot 3.542 \cdot 10^9 \\
\hline
\end{array}
$$

and the initial guesses for the parameters to be estimated given by:

$$
\begin{array}{|c|c|c|c|}
\hline
q_0 & k_I^+ & k_I^- & k_{off}^N \\
\hline
2.1600 & 10.9270 & 108.256 \\
\hline
\end{array}
$$

We obtained the estimated parameters and corresponding SE.

$$
\begin{array}{|c|c|c|c|c|c|c|}
\hline
 & k_I^+ & k_I^- & k_{off}^N & SE & \sigma^2 & \kappa \\
\hline
DS1 & 2.203 & 12.997 & 99.861 & (.011, .091, 1.208) & 8.165 \cdot 10^{-6} & 4.912 \cdot 10^9 \\
DS2 & 2.893 & 15.474 & 100.019 & (.019, .137, 1.279) & 9.323 \cdot 10^{-6} & 2.486 \cdot 10^9 \\
DS3 & 2.168 & 15.631 & 41.935 & (.011, .102, 0.424) & 6.435 \cdot 10^{-6} & 9.125 \cdot 10^9 \\
DS4 & 2.181 & 11.090 & 90.536 & (.009, .066, 0.936) & 6.289 \cdot 10^{-6} & 3.043 \cdot 10^9 \\
\hline
\end{array}
$$

Also in this case, we carried out bootstrapping for DS 4. The bootstrapping distributions for $k_I^+$, $k_I^-$ and $k_{off}^N$ are found in Figures 20-22. We then obtained the following means and standard errors for a run with $M = 1000$ in comparison to the asymptotic theory.

$$
\begin{array}{|c|c|c|c|c|c|c|}
\hline
 & k_I^+(boot) & k_I^-(boot) & k_{off}^N(boot) & k_I^+(asymp) & k_I^-(asymp) & k_{off}^N(asymp) \\
\hline
mean & 2.169 & 11.013 & 91.254 & 2.181 & 11.090 & 90.536 \\
SE & 0.0094 & 0.0699 & 1.0392 & 0.009 & 0.066 & 0.936 \\
\hline
\end{array}
$$

26
Figure 20: Three parameters estimation ($k^+_I$, $k^-_I$ and $k^N_{off}$): Bootstrapping distribution for $k^+_I$. We used GLS and $M = 1000$ runs.

Figure 21: Three parameters estimation ($k^+_I$, $k^-_I$ and $k^N_{off}$): Bootstrapping distribution for $k^-_I$. We used GLS and $M = 1000$ runs.
5.6 Estimation of 4 main parameters

Following the sensitivity analysis detailed above, we tried to estimate a combination of the parameters $k_I^+, k_I^-, k_{on}^N, k_{off}^N$ for the parameter set with $\kappa_\theta = 4$. Parameters were fixed as follows from the original 9 parameter fit:

<table>
<thead>
<tr>
<th>$k_{on}^\text{min}$</th>
<th>$k_{on}^\text{max}$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$i_{max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1684</td>
<td>$1.5 \cdot 10^9$</td>
<td>0.062</td>
<td>0.859</td>
<td>$3.5 \cdot 10^9$</td>
</tr>
</tbody>
</table>

We obtained the following result for the estimation of the four parameters using the data sets 1 to 4. In all of them, the condition number of the Fischer’s information matrix $\kappa$ is too large to invert. This along with the sensitivity results above strongly suggests that the data sets do not contain sufficient information to estimate 4 or more parameters with any degree of certainty attached to the estimates.

<table>
<thead>
<tr>
<th></th>
<th>$k_I^+$</th>
<th>$k_I^-$</th>
<th>$k_{on}^N$</th>
<th>$k_{off}^N$</th>
<th>$\sigma^2$</th>
<th>$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DS1$</td>
<td>2.1431</td>
<td>12.4751</td>
<td>4616.962</td>
<td>108.259</td>
<td>$8.7219 \cdot 10^{-6}$</td>
<td>$6.1226 \cdot 10^{19}$</td>
</tr>
<tr>
<td>$DS2$</td>
<td>2.7995</td>
<td>14.7630</td>
<td>4616.957</td>
<td>108.4308</td>
<td>$9.8694 \cdot 10^{-6}$</td>
<td>$1.4442 \cdot 10^{19}$</td>
</tr>
<tr>
<td>$DS3$</td>
<td>2.180</td>
<td>15.757</td>
<td>4618.599</td>
<td>41.369</td>
<td>$6.4622 \cdot 10^{-6}$</td>
<td>$1.881 \cdot 10^{17}$</td>
</tr>
<tr>
<td>$DS4$</td>
<td>2.161</td>
<td>10.9278</td>
<td>4617.3316</td>
<td>93.3265</td>
<td>$6.374 \cdot 10^{-6}$</td>
<td>$2.144 \cdot 10^{18}$</td>
</tr>
</tbody>
</table>
6 Model Comparison Tests

A type of Residuals Sum of Squares (RSS) based model selection criterion [8, 9, 11] can be used as a tool for model comparison for certain classes of models. In particular this is true for models such as those given in [3] in which potentially extraneous mechanisms can be eliminated from the model by a simple restriction on the underlying parameter space while the form of the mathematical model remains unchanged. In other words, this methodology can be used to compare two nested mathematical models where the parameter set $Ω_H^θ$ (this notation will be defined explicitly in Section 6.1 below) for the restricted model can be identified as a linearly restricted subset of the admissible parameter set $Ω_θ$ of the unrestricted model. Indeed, the RSS based model selection criterion is a useful tool to determine whether or not certain terms in the mathematical models are important in describing the given experimental data.

6.1 Ordinary Least Squares

We now turn to the statistical model (11), where the measurement errors are assumed to be independent and identically distributed with zero mean and constant variance $σ_0^2$. In addition, we assume that there exists $θ_0$ such that the statistical model

$$Y_j = M(t_j; θ_0) + E_j, \quad j = 1, 2, \ldots, n.$$  \hspace{1cm} (17)

correctly describes the observation process. In other words, (17) is the true model, and $θ_0$ is the true value of the mathematical model parameter $θ$.

With our assumption on measurement errors, the mathematical model parameter $θ$ can be estimated by using the ordinary least squares method; that is, the ordinary least squares estimator of $θ$ is obtained by solving

$$θ^n = \arg \min_{θ ∈ Ω_θ} J^n(θ; Y).$$

Here $Y = (Y_1, Y_2, \ldots, Y_n)^T$, and the cost function $J^n$ is defined as

$$J^n(θ; Y) = \frac{1}{n} \sum_{k=1}^{n} (Y_k - M(t_k; θ))^2.$$
The corresponding realization $\hat{\theta}^n$ of $\theta^n$ is obtained by solving 
$$
\hat{\theta}^n = \arg \min_{\theta \in \Omega} J^n(\theta; y),
$$
where $y$ is a realization of $Y$ (that is, $y = (y_1, y_2, \ldots, y_n)^T$).

As alluded to in the introduction, we might also consider a restricted version of the mathematical model in which the unknown true parameter is assumed to lie in a subset $\Omega^H_{\theta} \subset \Omega_{\theta}$ of the admissible parameter space.

We assume this restriction can be written as a linear constraint, $\mathcal{H}\theta_0 = h$, where $\mathcal{H} \in \mathbb{R}^{\kappa_r \times \kappa_q}$ is a matrix having rank $\kappa_r$ (that is, $\kappa_r$ is the number of constraints imposed), and $h$ is a known vector. Thus the restricted parameter space is
$$
\Omega^H_{\theta} = \{ \theta \in \Omega_{\theta} : \mathcal{H}\theta = h \}.
$$
Then the null and alternative hypotheses are
$$
H_0 : \quad \theta_0 \in \Omega^H_{\theta} \quad \text{and} \quad H_A : \quad \theta_0 \not\in \Omega^H_{\theta}.
$$
We may define the restricted parameter estimator as
$$
\theta^{n,H} = \arg \min_{\theta \in \Omega^H_{\theta}} J^n(\theta; Y),
$$
and the corresponding realization is denoted by $\hat{\theta}^{n,H}$. Since $\Omega^H_{\theta} \subset \Omega_{\theta}$, it is clear that
$$
J^n(\hat{\theta}^n; y) \leq J^n(\hat{\theta}^{n,H}; y).
$$
This fact forms the basis for a model selection criterion based upon the residual sum of squares. Using the standard assumptions (given in detail in [9]), one can establish asymptotic convergence result for the test statistics (which is a function of observations and is used to determine whether or not the null hypothesis is rejected)
$$
U^n = \frac{n \left( J^n(\theta^{n,H}; Y) - J^n(\theta^n; Y) \right)}{J^n(\theta^n; Y)},
$$
where the corresponding realization $\hat{U}_n$ is defined as
$$
\hat{U}^n = \frac{n \left( J^n(\hat{\theta}^{n,H}; y) - J^n(\hat{\theta}^n; y) \right)}{J^n(\hat{\theta}^n; y)}.
$$
This asymptotic convergence result is summarized in the following theorem.
Theorem 6.1. Under assumptions detailed in [9, 11] and assuming the null hypothesis $H_0$ is true, then $U^n$ converges in distribution (as $n \to \infty$) to a random variable $U$ having a chi-square distribution with $\kappa_r$ degrees of freedom.

The above theorem suggests that if the sample size $n$ is sufficiently large, then $U^n$ is approximately chi-square distributed with $\kappa_r$ degrees of freedom. We use this fact to determine whether or not the null hypothesis $H_0$ is rejected. To do that, we choose a significance level $\alpha$ (usually chosen to be 0.05) and use $\chi^2$ tables to obtain the corresponding threshold value $\tau$ so that $\text{Prob}(U > \tau) = \alpha$. We next compute $U^n$ and compare it to $\tau$. If $U^n > \tau$, then we reject the null hypothesis $H_0$ with confidence level $(1 - \alpha)100\%$; otherwise, we do not reject. We emphasize that care should be taken in stating conclusions: we either reject or do not reject $H_0$ at the specified level of confidence. The table below illustrates the threshold values for $\chi^2(1)$ with the given significance level.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\tau$</th>
<th>confidence level</th>
</tr>
</thead>
<tbody>
<tr>
<td>.25</td>
<td>1.32</td>
<td>75%</td>
</tr>
<tr>
<td>.1</td>
<td>2.71</td>
<td>90%</td>
</tr>
<tr>
<td>.05</td>
<td>3.84</td>
<td>95%</td>
</tr>
<tr>
<td>.01</td>
<td>6.63</td>
<td>99%</td>
</tr>
<tr>
<td>.001</td>
<td>10.83</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

Similar tables can be found in any elementary statistics text or online or calculated by some software package such as Matlab, and is given here for illustrative purposes and also for use in the examples demonstrated below.

6.2 Generalized Least Squares

The model comparison results outlined can be extended to deal with generalized least squares problems in which measurement errors are independent with $E(\varepsilon_k) = 0$ and $Var(\varepsilon_k) = \sigma^2_0 w^2(t_k, \hat{\theta})$, $k = 1, 2, \ldots, n$, where $w$ is some known real-valued function with $w(t, \hat{\theta}) \neq 0$ for any $t$. This is achieved through rescaling the observations in accordance with their variance (as discussed in [9]) so that the resulting (transformed) observations are identically distributed as well as independent.
6.3 Results for PolyQ Aggregation Models

We then carried out a series of model comparison tests (we again used DS 4) for nested models to determine if an added parameter yields a statistically significantly improved model fit. Our null hypothesis in each case was: $H_0$: The restricted model is adequate (i.e., the fit-to-data is not significantly improved with the model containing the additional parameter as a parameter to be estimated). We obtained the following results.

1. Model with estimation of $\{k_i^+, k_i^-\}$ vs. the model with estimation of $\{k_i^+, k_i^-, k_i^N\}$: We find with $n=699$, $J_n(\hat{\theta}_n^0; Y) = 0.0044192109$, $J_n(\hat{\theta}_n^0; Y) = 0.0043709501$ and $\hat{U}_n = 7.7178$. Thus we reject $H_0$ at a 99% confidence level.

2. Model with estimation of $\{k_i^+, k_i^-\}$ vs. the model with estimation of $\{k_i^+, k_i^-, k_i^N, k_i^N\}$: We find $J_n(\hat{\theta}_n^0; Y) = 0.0044192108$ with $\hat{U}_n = 7.49 \times 10^{-06}$. Thus we don’t reject $H_0$ at a 99% confidence level.

3. Model with estimation of $\{k_i^+, k_i^-, k_i^N\}$ vs. the model with estimation of $\{k_i^+, k_i^-, k_i^N, k_i^N\}$: To the order of computation we find no difference in the cost functions in this case and therefore we do not reject $H_0$ at a confidence level of 99%.

4. Model with estimation of $\{k_i^+, k_i^-, k_i^N\}$ vs. the model with estimation of $\{k_i^+, k_i^-, k_i^N, k_i^N\}$: We find $J_n(\hat{\theta}_n^0; Y) = 0.0043709780$ with $\hat{U}_n = 7.7133$ and hence we reject $H_0$ with a confidence level of 99%.

From these and the preceding results we conclude the information content of the typical data set for the dynamics considered here will support at most 3 parameters estimated with reasonable confidence levels and these are the parameters $\{k_i^+, k_i^-, k_i^N\}$.

7 Conclusions and Suggested Further Efforts

For the efforts reported on above we make several conclusions.

For the majority of data sets, the GLS residual plots with $\gamma = 0.6$ are random when fitted for data points $M(t_k) \geq 0.12$. As conjectured earlier, this may be because the early formation of aggregates is somewhat stochastic in nature which is not well described by either the mathematical and/or
statistical models. It appears that one needs special consideration of smaller polymer sizes. Indeed we suspect from additional discussions with our colleagues that perhaps the nucleation step might be dominated by a stochastic rather than deterministic process in the early stages (i.e., for small polymer sizes). This is a possible direction of further investigation.

Based on several different mathematical/statistical methodologies (sensitivities, asymptotic analysis, bootstrapping, model comparison tests), the data sets we considered do not contain sufficient information for the reliable estimation of all 9 parameters of interest. Indeed our findings suggest that at most 3 parameters can be reliably estimated with the data sets typical of those presented here, and that these parameters are \( \{k_I^+, k_I^-, k_{\text{off}}^N\} \). Recently related efforts [2] suggest that perhaps there are experimental design questions that could be addressed to collect data that might support the more sophisticated models derived in [25], especially in order to investigate information coming from different initial concentrations. Indeed, we have considered here data sets related to experiments carried out with the same initial concentration. Adapting the previously used techniques to simultaneously or successively use all the information content in data sets carried out for different initial concentration is a challenging problem (see [24] for a discussion of the effect of initial concentration on nucleated polymerization).

Here we concluded that at most 3 parameters \( \{k_I^+, k_I^-, k_{\text{off}}^N\} \) can be reliably estimated with the data sets investigated. The two first parameters determine the balance between the normal and abnormal protein concentrations and the third represents the stability of the nucleus against the degradation into monomeric entities. These three parameters are related to the early steps of the aggregation process, and thus we conclude that the model applied to these data sets does not provide any insight into the polymerization of larger polymers. Since this is the case, there is little motivation to modify the polymerization function \( k_{\text{on}} \) depicted in Figure 2 until further data collection procedures are pursued. It is difficult to say whether the initiation parameters are biologically more or less interesting than other parameters in the later aggregation process. (Everything interests the biological community since very little is known with certainty!) What we can more prudently conclude is rather a negative conclusion: this type of experiments seems to be more sensitive to the ignition of the reaction than to secondary pathways, which reveals a limit, and argues for the measurement not only of the total polymerized mass but also of size distributions of fibrils if one wants to really select the proper model for self-acceleration, as for instance in [26, 30, 31].
The methodology we developed here may be carried out for other types of proteins as well as related experiments, and this is of peculiar interest since the ThT measurements as studied here are the most standard method for protein aggregation. For instance, in the seminal paper [29], models are proposed and fit to data; our type of analysis would complement their findings and allow one to assess the quality of these fits and how much confidence the authors can have in their conclusions. These results in turn could be combined with optimal design techniques [2, 9, 10] to design more informative experiments.

Acknowledgements

This research was supported in part (MD, CK) by the ERC Starting Grant SKIPPERAD, in part (HTB) by Grant Number NIAID R01AI071915-10 from the National Institute of Allergy and Infectious Diseases, and in part (HTB) by the Air Force Office of Scientific Research under grant number AFOSR FA9550-12-1-0188. The authors are most grateful to a referee whose careful reading of an earlier version of this manuscript led to substantial improvements in the final version.

References


