Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Stochastic modeling of fluctuations in the NF-kB activity of neoplastic cells

Relatore:
Dott. Enrico Giampieri

Correlatore:
Dott.ssa Mattia Lauriola

Presentata da:
Davide Giosué Lippolis

Anno Accademico 2018/2019

Dedicated to Grandpa

Sommario

In the last decade there has been a thriving development of quantitative methods in the field of biomedical physics, both in the modeling of biological complex systems and in the statistical analysis of experimental results. In this thesis we examine the effect of a variety of stimuli on the NF-kB activity's oscillation in metastatic colorectal cancer cells using probabilistic models such as the Chemical Master Equation and the Bayesian statistics. After a brief introduction (Chapter 1) we explain the theoretical foundations of the Master Equation approach (Chapter 2) and of the Bayesian inference, together with a brief dip in the subject of Hamiltonian Monte Carlo methods (Chapter 3). In the continuation we create a solvable model using the Master Equation approach (Chapter 4) and we gain insights concerning its eigenvalues distribution in the complex plane. We therefore build a Bayesian regression model (Chapter 5) and we use it to analyze the oscillating autocorrelation function of the previous stochastic model in order to test the capabilities of the statistical model. Finally (Chapter 6) we analyze the biological data using the previously created and tested statistical model, exploring and commenting the results (Capitolo 7) and outlining further research directions.

Stochastic processes Bayesian statistic Biological oscillations Citokines Colorectal cancer Chemical master equation

Sommario

Nell'ultimo decennio è avvenuto un fiorente sviluppo di metodi quantitativi applicati al campo della fisica biomedica, sia per quanto riguarda la modellizzazione di sistemi complessi di carattere biologico sia nell'analisi statistica dei risultati sperimentali. In questa tesi si esamina l'effetto di vari stimoli sull'oscillazione dell'attività di NF-kB nelle cellule tumorali metastatiche del cancro al colon-retto usando modelli probabilistici quali la Chemical Master Equation e la statistica Bayesiana. A seguito dell'esposizione delle premesse (Capitolo 1) si sviluppano i fondamenti teorici della Master Equation (Capitolo 2) e dell'inferenza Bayesiana, insieme a un'introduzione ai metodi Monte Carlo Hamiltoniani (Capitolo 3). Successivamente si sviluppa un modello risolubile con l'approccio della Master Equation (Capitolo 4) ottenendo risultati per quanto riguarda la distribuzione degli autovalori nel piano complesso. Conseguentemente si costruisce un modello Bayesiano di regressione (Capitolo 5) e lo si utilizza per analizzare le oscillazioni della funzione di autocorrelazione del precedentemente introdotto processo stocastico al fine di testare le capacità di tale modello statistico. Infine (Capitolo 6) si analizzano i dati biologici usando il modello statistico precedentemente creato e testato, esplorando e commentando i risultati (Capitolo 7) delineando successive direzioni di ricerca.

Contents

Chapter 1

Introduction

In this thesis we analyze the fluctuations of the NF-kB activity in colorectal cancer cells using the fluorescence data derived by the experiments performed in the AlmaIdea Junior 2017 research activity. In the first part of the thesis we introduce the theoretical framework we use, namely the Master Equation methods and the Bayesian inference methods. In the latter part of the thesis we show how we apply these methods in order to get insights of the fluctuations of NF-kB activity in colorectal cancer cells, when treated with different stimuli as anticancer drugs and interleukin. In this context we show and comment the result of our analysis.

1.1 Quantitative modelling of biological complex systems

During the last decade there has been an increasing interest concerning the application of stochastic methods to biochemical systems and biological phenomena. It is observed that almost every biological process possesses intrinsical noise[1], hence their description by means of deterministic methods such as the modeling and resolution of ordinary differential equations (ODE) usually leads to different solutions than those obtained from stochastic methods. Again, for nonlinear systems the stochastic solution is in general different from the noisy linear one.

Moreover, the majority of biological phenomena like cellular growth, gene expression, complex signaling systems and more generally biochemical pathways can't be described by continuous models, therefore even continuous stochastic models like the Fokker-Planck equation approach or the analogous Langevin equation approach don't fit well the analysis of those biochemical systems. The combination of these two properties (discreteness

[1] Actually the intrinsical noise of these systems can be also be described as a consequence of their deterministic chaos given by the presence of multiple variables with nonlinear interactions.

and noisiness) is leading to an increasing exploitation of the Master Equation formalism in order to model biological systems, due to its capability of describing processes that are both discrete and stochastic.

Quantitative modelling of chemical systems has been based on the Law of Mass Action. The applications of this theory to biochemical kinetics leads to the standard kinetic modelling of biochemical reactions for enzymatic reaction systems. However the theory based on the Law of Mass Action doesn't consider fluctuations; namely, it is a deterministic theory. Anyway it is a fully dynamical theory, hence it can be used for systems far outside the equilibrium.

In order to extend the Law of Mass Action to fluctuating systems outside the equilibrium state it is introduced the Chemical Master Equation (CME) approach. It is in fact true that the chemical master equation is the stochastic counterpart of the chemical kinetic equation based on the Law of Mass Action [1], [5].

1.2 Probabilistic analysis of biological data

Statistics is the theoretical framework for rigorous data analysis. It provides all the mathematical tools required to set up probabilistic models and to explain how data are produced by experiments, taking into account the intrinsic uncertainty associated with these processes.

While there isn't an universal statistical framework, one interpretation that has gained strength in the last thirty years is the Bayesian statistical approach, mostly because of the growth of the number of computational resources available. Bayesian statistics offers some capabilities that enable it to solve a variety of complex problems. Moreover, where some conditions are met the bayesian methods' results coincide with those of the more standard Maximum Likelihood methods. However the main appeal of the bayesian methods is their capability to model a variety of complex systems by means of multilevel (hierarchical) models, following a set of simple principles. Moreover, the bayesian analysis leads to more intelligible results with a more intuitive meaning.

In the bayesian framework the model parameters are random variables too, hence we are interested in inference of the overall unknown parameter's distribution, instead of a point estimate or an interval estimate typical of the standard methods. This is useful in order to get insights and predict new data given the model.

Due to its power in the analysis of complex models and due to the fact that it makes the introduction of prior domain information easier, Bayesian approaches are widely used in a variety of scientific applications, including biological research.

Chapter 2

The Master Equation approach

2.1 Markov Processes

The Master Equation is defined in the framework of Markov processes. A stochastic process is defined as Markovian if it satisfies the Markov property as follow:

Definition 1. *Given a set of n ordered times* $\{t_i\}$ *a process satisfies the Markov property if*

$$P(x_n, t_n|x_{n-1}, t_{n-1}, \ldots, x_1, t_1) = P(x_n, t_n|x_{n-1}, t_{n-1}) \tag{2.1}$$

Stated differently, the conditional probability density at t_n given the value at t_{n-1} isn't affected by any information about earlier times. Thus a Markov process is defined uniquely by the two following density functions

$$P(x_1, t_1) \tag{2.2}$$

$$P(x_2, t_2|x_1, t_1) \tag{2.3}$$

Usually the (2.3) is called transition probability. Using those two functions a generic joint probability density function can be constructed as

$$P(x_1, t_1, \ldots, x_n, t_n) = P(x_1, t_1) \prod_{i=n}^{i=2} P(x_i, t_i|x_{i-1}, t_{i-1}) \tag{2.4}$$

Using this property for $n = 3$ and therefore integrating over the middle variable x_2 one obtains

$$P(x_3, t_3, x_1, t_1) = P(x_1, t_1) \int dx_2 P(x_3, t_3|x_2, t_2) P(x_2, t_2|x_1, t_1) \tag{2.5}$$

Finally dividing both sides by $P(x_1, t_1)$ one finds the famous Chapman-Kolmogorov equation for the transition probabilities

$$P(x_3, t_3|x_1, t_1) = \int dx_2 P(x_3, t_3|x_2, t_2) P(x_2, t_2|x_1, t_1) \tag{2.6}$$

For a stationary process the transition probability only depends on the lag time $\tau = t_i - t_{i-1}$ [1], therefore adopting the standard notation

$$P(x_i, t_i|x_{i-1}, t_{i-1}) = T_\tau(x_i|x_{i-1}) \tag{2.7}$$

Inserting this definition in the (2.6) one finds

$$T_{\tau_1+\tau_2}(x_3|x_1) = \int dx_2 T_{\tau_2}(x_3|x_2) T_{\tau_1}(x_2|x_1) \tag{2.8}$$

Using this notation the Chapman-Kolmogorov is nothing else than a product of integral kernels (or matrices).

2.2 Master Equation approach to Markov processes

The Master Equation follows from the (2.6) in the small lag times limit. Although the Chapman-Kolmogorov Equation is a functional equation in the transition probability $T_\tau(x_2|x_1)$ the Master Equation is a differential equation, at least ideally easier to tackle.

Let's begin defining the transition probability per unit time W as

$$T_\tau(x_2|x_1) \approx \tau W(x_2|x_1) \qquad x_2 \neq x_1 \tag{2.9}$$

and using this definition one can define the probability that no transition occurs during a lag τ as

$$Q(x_1) = 1 - \tau \int dx_2 W(x_2|x_1) \tag{2.10}$$

Putting these equations together one finds

$$T_\tau(x_2|x_1) = \tau W(x_2|x_1) + Q(x_1)\delta_(x_2 - x_1) \tag{2.11}$$

Moreover putting it in the (2.6), dividing by τ and letting $\tau \to 0$ one finds

$$\frac{\partial}{\partial \tau} T_{\tau'}(x_3|x_1) = \int dx_2 \Big\{ W(x_3|x_2) T_{\tau'}(x_2|x_1) - W(x_2|x_3) T_{\tau'}(x_3|x_1) \Big\} \tag{2.12}$$

Finally recalling the definition (2.3) and considering $\{x_i\}$ as a set over discrete states one obtains the most common notation for the Master Equation as a balance equation for probabilities of different states

$$\frac{dP_n(t)}{dt} = \sum_m \Big\{ W_{nm} P_m(t) - W_{mn} P_n(t) \Big\} \tag{2.13}$$

[1] This is basically the definition of a stochastic stationary process.

2.3 Expansion in eigenfunctions of the Master Equation solution

The Master Equation (2.13) can be written in a more compact way defining the $\mathbb{W}$ matrix as

$$\mathbb{W}_{nm} = W_{nm} - \delta_{nm} \sum_l W_{nl} \tag{2.14}$$

Thus the (2.13) can be written as

$$\dot{\vec{P}}(t) = \mathbb{W}\vec{P}(t) \tag{2.15}$$

This kind of differential equation has a formal solution

$$\vec{P}(t) = e^{t\mathbb{W}}\vec{P}(0) \tag{2.16}$$

The explicit solution for $\vec{P}$ is usually found expanding it as a sum of eigenvectors. If one allows $\mathbb{W}$ to be only real valued due to the physical meaning of the matrix elements, three cases arise

- $\mathbb{W}$ is symmetric, therefore diagonalizable
- $\mathbb{W}$ isn't symmetric, but still diagonalizable
- $\mathbb{W}$ isn't symmetric and not diagonalizable (it can still be put in Jordan form)

Now the case of an isolated system and diagonalizable $\mathbb{W}$ will be considered. Assuming that $\mathbb{W}$ is indecomposable (it can't be divided in two non-interacting subsystems) then the eigenvalue problem is stated as usual

$$\mathbb{W}\vec{\Phi}_\lambda = \lambda\vec{\Phi}_\lambda \tag{2.17}$$

Therefore the solution of (2.13) is merely

$$\vec{P}(t) = \sum_\lambda C_\lambda \vec{\Phi}_\lambda e^{\lambda t} \tag{2.18}$$

with the constants C_λ defined by the initial probability state $\vec{P}(0)$.
Studying the long time limit one can demonstrate [1] that

$$\lambda_0 = 0 \tag{2.19}$$

$$\mathrm{Re}(\lambda_i) < 0 \qquad for \quad i > 0 \tag{2.20}$$

Finally we want to use the (2.18) to find the autocorrelation of some observable O (in our case it will be the number of particles for each chemical species). For simplicity we consider the $\langle O(t)\rangle_{\mathrm{eq}} = 0$ case. We can rephrase the (2.18) as

$$P_n(t) = \sum_k \Phi_n^{(l)} e^{\lambda_l t} \Big[\sum_m \Phi_m^{(l)} \frac{P_m(0)}{P_m^{\text{eq}}} \Big] \tag{2.21}$$

where n and m span the state space enumeration while l indexes the eigenvalues and the eigenvectors, and $\vec{P}^{\text{eq}}$ is the equilibrium probability distribution. As can be seen the term in the square brackets depends on the initial conditions. Using the 2.21 one can find the joint probability distribution

$$P(n, 0; m, t) = P_n(0) \sum_l e^{\lambda_l t} \Phi_m^{(l)} \frac{\Phi_n^{(l)}}{P_i^{\text{eq}}} \tag{2.22}$$

Moreover, using the definition of equilibrium autocorrelation we find for our observable O

$$k(\tau) = \lim_{t \to \infty} \langle O(t) O(t + \tau) \rangle \tag{2.23}$$

combining it with the previous equations we find

$$\begin{aligned} k(\tau) \quad &= \sum_{n,m} O_n O_m P(n, 0; m, \tau) = \\ &= \sum_{n,m} O_n O_m P_n(0) \sum_l e^{\lambda_l \tau} \Phi_m^{(l)} \frac{\Phi_n^{(l)}}{P_i^{\text{eq}}} = \\ &= \sum_l e^{\lambda_l \tau} \sum_n O_n \Phi_n^{(l)} \sum_m O_m \Phi_m^{(l)} = \\ &\sum_{l=0} e^{\lambda_l \tau} \Big[\sum_n O_n \Phi_n^{(l)} \Big]^2 \end{aligned}$$

where we exploit the stationary of our stochastic process and the fact that at the equilibrium $P_i(0) = P_i^{\text{eq}}$. Finally using the fact that our observable has zero equilibrium mean we find the desired autocorrelation function formula

$$k(\tau) = \sum_{\lambda \neq 0} e^{\lambda \tau} \Big[\vec{n} \cdot \vec{\Phi}_\lambda \Big]^2 \tag{2.24}$$

Notably, as can be seen it doesn't depend on the initial system configuration.

2.4 The Chemical Master Equation

A common approach to modeling biological systems is to find the Master Equation describing the chemical system and then solving the latter. Since the size of the Chemical Master Equation's matrix grows exponentially with the number of species, even if one applies a bound to the total number of particles solving the Master Equation is usually numerically very expensive, therefore this task is often accomplished by mean of Gillespie stochastic simulation algorithms (SSA) or other kinds of approximations [2].

Before choosing the right numerical approach to the current problem one needs to translate the chain of chemical equations in a Master Equation written as a $\mathbb{W}$ matrix. Let's start considering a set of M chemical equations concerning N chemical species

$$a_{1k}S_1 + \cdots + a_{Nk}S_N \xrightarrow{\beta_k} b_{1k}S_1 + \cdots + b_{Nk}S_N \qquad \text{for} \quad k = 1, \ldots, M \tag{2.25}$$

where a and b are the stechiometric coefficients[2] and the β_k are the reaction rates. Therefore one can define the state space in the number of particle basis, as

$$\vec{n} = (n_1, \ldots, n_N) \tag{2.26}$$

Of course it will be a set of N time dependent functions.

To build the $\mathbb{W}$ matrix one need to define the propensity function $\alpha_k(\vec{n}(t))$ of the k reaction so that $\alpha_k(\vec{n}(t))dt$ is the probability of the reaction to occur in the time interval $[t, t + dt]$. Therefore applying the (2.13) one finds

$$\frac{dP(\vec{n}, t)}{dt} = \sum_k \left[\alpha_k(\vec{n}'_k)P(\vec{n}'_k, t) - \alpha_k(\vec{n})P(\vec{n}, t)\right] \tag{2.27}$$

where $\vec{n}'_k = (n_1 + b_{1k} - a_{1k}, \ldots, n_N + b_{Nk} - a_{Nk})$ is the updated state.

The last step needed is to enumerate the state space. Considering a finite state space one can order all the states by an arbitrary index such as $\{\vec{n}_i\}$ is the state space. This procedure leads to the probability function for the state i as $P_i(t) = P(\vec{n}_i, t)$. Recalling the meaning of the $\mathbb{W}$ matrix elements one finds

$$\mathbb{W}_{ij} = \begin{cases} -\sum_k \alpha_k(\vec{n}_j) & \text{for} \quad i = j \\ \alpha_k(\vec{n}_j) & \text{for} \quad \vec{n}_j = \vec{n}'_i \end{cases} \tag{2.28}$$

[2]For reactions with chemical species degradation or creation some of the stechiometric coefficients can be zero.

As a final recap we describe the operative steps to be performed when modeling a complex chemical (biological) system via the Master Equation approach:

- Statement of the physical problem:
  Identification of the physical variables (often chemical species).
- Enumeration of the state space:
  The state spaces can be classified as finite / infinite and discrete / continuous. Often in the framework of the Chemical Master Equation one approximates the state space as finite and discrete for the sake of computational manageability.
- Construction of the $\mathbb{W}$ matrix elements using the (2.28).
- Resolution of the Master Equation via analytical or approximate algorithms.

Chapter 3

A brief introduction to Bayesian inference

3.1 Bayesian workflow

Bayesian methods involve inference using probability models for quantities we observe (data) and quantities we want to know (parameters). In this framework both observable variables and parameters are random variables, therefore the main goal is to find the correct probability distribution for each of the parameters rather than finding a point estimate of them. Every bayesian data analysis process involves three main steps:

- Modeling:
  The first step concerns the detection of a joint probability distribution for all the observable and unobservable variables. It is n this phase that usually both the likelihood and the priors are defined.

- Conditioning:
  The second step is to condition the above model on the observed data. More explicitly, it involves finding the correct posterior density distribution defined as the conditional probability distribution of the unobserved variables (our parameters of interest) given the observed variables (the data).

  The main drawback of using these methods is that for real world problems one can't gain access to the posterior analytically, however it can be done approximately using Markov Chain Monte Carlo methods such Metropolis algorithm, Gibbs sampler or Hamiltonian Monte Carlo methods.

- Testing:
  After fitting the model one needs to evaluate it and see if the conclusions are reasonable. It is done usually using testing techniques and judgement of the results given the experimental context.

We use a bayesian probabilistic approach mainly because it helps an intuitive interpretation of the results. In fact in contrast with a frequentist confidence interval, a bayesian probability interval (usually called Highest Density Interval (HDI) or Highest Posterior Density (HPD)) for an unknown parameter can be directly thinked as having an high probability of containing the actual parameter value. In fact the Highest Density Interval is defined as the interval that covers the most of the distribution (usually 95% or 50%) with the property that the points inside the HDI are more credible than those outside. As a further motivation, the bayesian approach allows great flexibility in building the probabilistic model.

3.2 Building of the probabilistic model

In order to make statements on the unknown parameters one need to provide a joint probability distribution for both the parameters (generally indicated as $\vec{\theta}$) and the data $\vec{y}$. The joint probability density can be divided as

$$p(\vec{\theta}, \vec{y}) = p(\vec{y}|\vec{\theta})p(\vec{\theta}) \tag{3.1}$$

where the first factor is the likelihood and the second one is the prior distribution . Thus using the Bayes rule one finds

$$p(\vec{\theta}|\vec{y}) = \frac{p(\vec{y}, \vec{\theta})p(\vec{\theta})}{p(\vec{y})} \tag{3.2}$$

where the denominator in the above equation is often called prior predictive distribution.

Therefore choosing the prior distribution encoding the prior knowledge for our parameters and the likelihood one can find the posterior distribution simply applying the Bayes rule. As stated before it is mostly done with computational algorithms as various implementations of the MCMC methods.

There are several computational enviroments to build a bayesian inference model. The most famous are Stan, Jags and PyMC3. In the present the latter is used. Aside from differences between them, they all work in two steps: the model definition described above and the sampling.

3.3 Regression in a bayesian framework

In this section we describe a toy model of a simple linear regression model in a bayesian framework in order to lead the way to the actual model described in the next chapters.

Often the data are divided in covariates $\vec{x}$ and variates $\vec{y}$. The covariates are usually known very well, while the variates aren't. The goal is hence figuring out the function

that models our variates as a function of the covariates. We usually assume that our $\vec{x}$ doesn't depend on any parameter. Therefore

$$p(\vec{x}, \vec{y}|\vec{\theta}) = p(\vec{y}|\vec{\theta})p(\vec{x}) \tag{3.3}$$

where the last term $p(\vec{x})$ is just a normalization constant. Now we introduce a deterministic relationship f as

$$p(\vec{x}, \vec{y}|\vec{\theta}) = p(\vec{y}|f(x, \vec{\theta}_1), \vec{\theta}_2)p(\vec{x}) \tag{3.4}$$

so that some subset $\vec{\theta}_1$ of the parameters depends on the covariates, but some other $\vec{\theta}_2$ doesn't. In the case of a simple linear regression model

$$p(x, y|\alpha, \beta, \sigma) = \mathcal{N}(y|\alpha x + \beta, \sigma)p(x) \tag{3.5}$$

Here α and β depend on the covariates but σ doesn't (homoscedastic noise).

Hence if we assume the data as normally distributed with f deterministic (and parametric) function as mean and σ as standard deviation, we can compute the posterior density function using the (3.2) (of course after defining the prior distribution for each parameter) and eventually find the posterior marginal density for the parameters of the function f.

3.4 MCMC algorithms for high dimensional parameter spaces and continuous random variables: the Hamiltonian Monte Carlo Methods

3.4.1 Computation of expectations

The ultimate goal of computational statistics is to evaluate expectation values with respect to some probability distribution. Let's considering $\vec{x}\,\epsilon\,\mathcal{S}$ where $\mathcal{S}$ is the sample space and $\vec{x}$ a vector of real numbers. Therefore knowing the probability density distribution $p(\vec{x})$ one can find the expectation value of some f as an integral

$$\mathbb{E}_p[f] = \int d\vec{x} p(\vec{x}) f(\vec{x}) \tag{3.6}$$

Since almost never it is possible to solve this integral analytically one must approximate it with numerical methods. Since their accuracy is limited we need to find a way to avoid the region of the space with negligible contribution. The simplest answer is to reduce our summation over a set where the integrand is big. Therefore one can choose the so called "typical set" as the region where the product of the probability density function and our function f is big enough. It will be so in the region around the mode of our probability density p.

However this argument fails where the dimension of the space grows enough. The reason is that each actual infinitesimal contribution of the integral is

$$d\vec{x}p(\vec{x})f(\vec{x}) \tag{3.7}$$

therefore one need to take into account the region of the space $d\vec{x}$ over that contribution is computed. In fact it can easily proven that for high dimensional spaces the volume inside any neighbourhood is far lesser than the volume around it, hence the typical set won't be centered anymore near the mode, but it will be spreaded along the tails of the probability density function.

3.4.2 Markov Chain Monte Carlo methods

One of the most generic and useful methods to find the typical set is the Markov Chain Monte Carlo method. It exploits the properties of Markov chains in order to explore the typical set [7]. Given the condition that the Markov transition probability preserves the target distribution p, then at any point in the space the Markov transition will be greater towards the typical set. Therefore if we sample a sufficient large amount of points of the Markov chain, they will become a quantification of the typical set.

Given the sampled set $\{x_1, \ldots, x_N\}$ one can estimate expectations averaging the function f

$$\hat{f}_N = \frac{1}{N}\sum_{i}^{N} f(x_i) \tag{3.8}$$

with the asymptotic behaviour

$$\lim_{N\to\infty} \hat{f}_N = \mathbb{E}_p[f] \tag{3.9}$$

Of course there isn't any way to exploit this asymptotic behaviour directly, so one need to understand how the Markov Chain acts after a finite exploration time. There are many ways to check how well a Markov Chain Monte Carlo is behaving, such as computing the Markov Chain Monte Carlo Standard Error (MCMCSE) or the Effective Sample Size (ESS) [7].

In the ideal behaviour the Markov chain reaches the typical set after a warmup time and after that it remains there exploring better and better that region. The warmup time is particularly important since it it is needed to remove the effects related to the Markov Chain starting point. However if the typical set isn't compatible with the Markov transition, which happens when the typical set has some geometrical patologies, then the Markov chain fails to explore properly the typical set leading to biased estimations.

Now we briefly discuss the Metropolis algorithm. It is basically a random walk with an acceptance-rejection rule ensuring convergence to the target distribution. The algorithm proceeds with the following steps:

- Initialization:
  Draw a starting point θ_0 with $p(\theta_0|x) > 0$.

- Proposal:
  Sample a proposal θ^* from a proposal distribution $J_t(\theta * |\theta_{t-1})$. For the Metropolis algorithm it must be symmetric in θ, while for the Metropolis-Hastings algorithm this requirement is no longer mandatory.

- Calculation:
  Calculate the ratio r. For the Metropolis algorithm it is defined as

$$r = \frac{p(\theta^*|x)}{p(\theta_{t-1}|x)} \tag{3.10}$$

  while for the Metropolis-Hastings algorithm, in order to correct for the asymmetry of J, it is defined as

$$r = \frac{\frac{p(\theta^*|x)}{J_t(\theta*|\theta_{t-1})}}{\frac{p(\theta_{t-1}|x)}{J_t(\theta_{t-1}|\theta^*)}} \tag{3.11}$$

- Jump:
  Set the new point of the chain to

$$\theta_t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta_{t-1} & \text{otherwise} \end{cases} \tag{3.12}$$

  Note that even if the jump is not accepted it counts as an iteration of the algorithm. After the jump return to the first step with $\theta_0 = \theta_t$ or stop the algorithm.

3.4.3 Hamiltonian Monte Carlo

The usual Random Walk MCMC methods such as the Metropolis-Hastings algorithm don't work very well in high dimensional spaces. In fact tipically in these situations the exploration will be extremely slow since for a lot of directions in the space the proposal will lead the chain far outside the typical set and will hence be rejected. This effect is even stronger where there are correlations in the likelihood's variables, as it often happens in hierarchical models.

In order to solve this issue information about the geometry of the typical set need to be exploited. With a continuous sample space one way to access the information about the geometry is to find a vector field aligned with the typical set. Using this vector field the algorithm can therefore follow the directions and sample the typical set much faster.

To build the vector field the gradient of the target probability can be used, but this leads the chain towards the mode of the distribution and away from the typical set. In

order to constrain the chain in the typical set one need to find a way to twist the vector field. To do so the notion of conservative dynamics comes in handy. The main idea is to expand the sample space in a phase space, introducing a set of conjugate momentum parameters $\vec{v}$.

Hence one defines a canonical distribution

$$p(\vec{x}, \vec{v}) = p(\vec{v}|\vec{x})p(\vec{x}) \tag{3.13}$$

Since the momentum $\vec{v}$ are built to be conjugate to the $\vec{x}$ therefore the dynamics induced by the invariant Hamiltonian

$$H(\vec{x}, \vec{v}) = -\ln p(\vec{x}, \vec{v}) \tag{3.14}$$

will conserve the volumes in the phase space. It means that the even if the gradient of the probability density leads the dynamics towards the mode, the Hamiltonian structure will constrain the motion inside the typical set.

Using the analogy with physical systems

$$H(x_i, v_i) = - \ln p(v_i|x_i) - \ln p(x_i) \tag{3.15}$$

$$= K(v_i, x_i) + V(x_i) \tag{3.16}$$

where the first term is analogous to a kinetic energy and the latter is a potential energy. Writing the Hamilton Equations

$$\begin{cases} \frac{dx_i}{dt} = \frac{\partial H}{\partial v_i} = \frac{\partial K}{\partial v_i} \\ \frac{dv_i}{dt} = -\frac{\partial H}{\partial x_i} = -\frac{\partial K}{\partial x_i} - \frac{\partial V}{\partial x_i} \end{cases} \tag{3.17}$$

one finds exactly the vector field needed to build Markov chains that can explore efficiently the typical set. Of course the trajectories generated by the vector fields need to be projected back to the coordinate space, namely the parameter space.

Chapter 4

The Master Equation toy model

4.1 Statement of the problem

Our purpose is to analyze an analytical solvable model using the Chemical Master Equation approach. Since our final purpose is to analyze the oscillations in fluctuations in the biological system described above, we need to build a model whose $\mathbb{W}$ matrix exhibits complex eigenvalues. We know that a system with only two chemical species will have only real eigenvalues. In fact in this case the related Master Equation satisfies the Detailed Balance Condition, namely

$$P_n^{\text{eq}}\mathbb{W}_{nm} = P_m^{\text{eq}}\mathbb{W}_{mn} \tag{4.1}$$

Moreover if this condition is met it can be proven [1] that the $\mathbb{W}$ matrix can be put in a symmetric form with same eigenvalues spectrum. Since a real symmetric matrix has all real eigenvalues, the $\mathbb{W}$ matrix shall also have only real eigenvalues. Hence the simplest toy model with complex eigenvalues one can think of is a closed and isolated system with three chemical species and one particle as in Figure 4.1a.

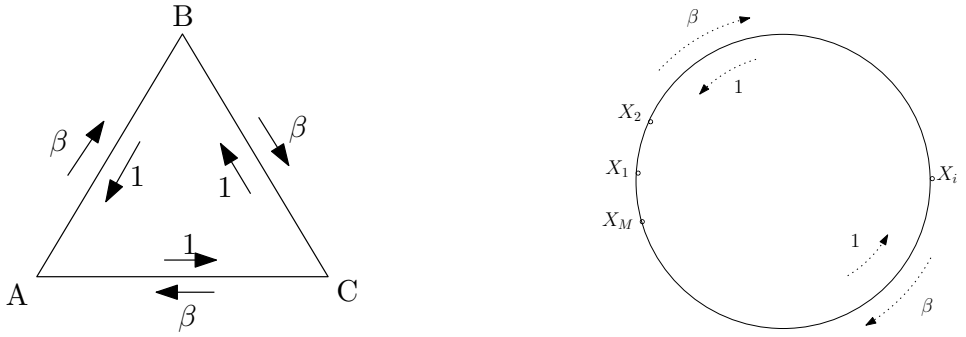
In this model we consider only single particle one step[1] transformations of the kind of $A \leftrightarrow B$. As a further simplification we consider all the clockwise reaction rates (as in the graph of Figure 4.1) as equal to β and all the anti-clockwise reaction rates as equal to 1.

Once defined the structure of the state space this model is extended to a M species and N particles model. Furthermore we expect the intensity of the oscillations to be function of the rate β, therefore we perform the diagonalization of the $\mathbb{W}$ matrix via symbolic algebra, accomplished using the Python package Sympy.

The final task is to simulate the system using the Gillespie SSA algorithm and to use the output data to feed the bayesian model in order to recover the eigenvalue's

[1] We forbid simultaneous transformations of two or more particles. With this approximation our model become a one step process.

distribution and to analyze the matching of the analytical eigenvalues with the inferred ones.



(a) Three species (b) M species

Figure 4.1: Reaction graphs

4.2 Identification and enumeration of the state space

Let's start with the case of three species A, B and C as seen in Figure 4.1a analyzing the number of particle vector written as

$$\vec{n}(t) = \big(n_A(t), n_B(t), n_C(t)\big) \tag{4.2}$$

Since our system is closed then $\vec{n}$ must satisfy the constraint on the total number of particles $\forall t$, explicitly

$$\sum_{i=\{A,B,C\}} n_i(t) = N \qquad \forall t \tag{4.3}$$

In our tridimensional space (it is $\mathbb{N}^3$) this is nothing else than a 2-simplex. Hence from this constrain and the discreteness of state space each point in it is connected at most with six other points. This imply that in the $\mathbb{W}$ matrix, no matter how big are N and M, each row shall have at most 7 non zero components.

Now if we let be $M > 3$ the space become $\mathbb{N}^M$ and our hypersurface where all the accessible states lie become a (M-1)-simplex defined by

$$\sum_{i=1}^{M} n_i(t) = N \qquad \forall t \tag{4.4}$$

The next step concerns the identification of the probability space. As outlined in Chapter 1 if the state space is a finite set of Y elements then the probability space is a

Y-dimensional Hilbert space and $\mathbb{W}$ is a YxY matrix. With a procedure that resembles that used in the Quantum Mechanics to build the Fock space we write

$$(n_1, \ldots, n_M) \longrightarrow |n_1, \ldots, n_M\rangle \tag{4.5}$$
$$|n_1, \ldots, n_M\rangle \longrightarrow (0, \ldots, 1, \ldots, 0) \tag{4.6}$$

with the position of the 1 depends on the enumeration of the state space, where the actual ordering of the enumeration isn't crucial.

Note that the right side of (4.6) is a Y-dimensional vector and resembles the probability of the sure event of being in the state $(n_1, \ldots, n_M)$.

In order to make this step clearer let's set up the procedure for the $N = 2$, $M = 3$ case. We have an $\mathbb{N}^3$ state space with the 2-simplex defined by the constrain

$$n_A + n_B + n_C = 2$$

Thus the total number of elements in the state set is 6. Using the correspondence (4.6) with the probability space one finds

$$\begin{aligned}
P(2,0,0) &\longrightarrow (1,0,0,0,0,0)\\
P(0,2,0) &\longrightarrow (0,1,0,0,0,0)\\
P(0,0,2) &\longrightarrow (0,0,1,0,0,0)\\
P(1,1,0) &\longrightarrow (0,0,0,1,0,0)\\
P(0,1,1) &\longrightarrow (0,0,0,0,1,0)\\
P(1,0,1) &\longrightarrow (0,0,0,0,0,1)
\end{aligned}$$

Keeping this correspondence in mind one can easily build the $\mathbb{W}$ matrix translating the reaction graph 4.1a and recalling the (2.28), obtaining

$$\begin{bmatrix}
-2\beta-2 & 2\beta & 0 & 2 & 0 & 0\\
1 & -2\beta-2 & \beta & \beta & 1 & 0\\
0 & 2 & -2\beta-2 & 0 & 2\beta & 0\\
\beta & 1 & 0 & -2\beta-2 & \beta & 1\\
0 & \beta & 1 & 1 & -2\beta-2 & \beta\\
0 & 0 & 0 & 2\beta & 2 & -2\beta-2
\end{bmatrix}$$

Once obtained the $\mathbb{W}$ matrix elements, the next step is to diagonalize it and to find the eigenvalue spectrum.

4.3 Diagonalization of the $\mathbb{W}$ matrix and its eigenvalues spectrum

As said in the section 5.1 we expect that the intensity of the oscillation modes in (2.18) is dependent on the reaction rate β, therefore we would try to find an analytical form of the solution explicitly dependent to this parameter. In order to accomplish it we need to diagonalize the $\mathbb{W}$ matrix via symbolic algebra.

Since due to the Abel-Ruffini Theorem there isn't a way to solve analytically a general equation of order $N > 4$ we need to find a walkaround for this problem. We first calculate the eigenvectors of the $\mathbb{W}$ for different values of the parameter β via numerical methods. We show in Fig. 4.2 the eigenvector matrices for two different values of β for the case of one particle and three species (although the following consideration has been validated for each value of N and M considered).

At first glance they seem to be different, but they are actually related by a set of row permutations. Since each element in the group of permutations is a similarity transformation, therefore the Eq. 4.7 is invariant under such a transformation. Thus as expected, the eigenvector's set doesn't depend on the parameter β. Hence we can compute it numerically and then using the basis change matrix built with those eigenvectors we can diagonalize the $\mathbb{W}$ matrix directly using the formula

$$D = O^{-1}\mathbb{W}O \tag{4.7}$$

where the columns of O are the eigenvector's basis elements and D is the diagonal matrix with the eigenvalues of $\mathbb{W}$ on the main diagonal.

Finally using the (2.24) we find the analytical form of the autocorrelation function.

Eigenvector matrices for one particle and three species

$$\begin{bmatrix} 0.5773 & 0.2887 - 0.5i & 0.2887 + 0.5i \\ 0.5773 & 0.2887 + 0.5i & 0.2887 - 0.5i \\ 0.5773 & -0.5773 & -0.5773 \end{bmatrix}$$

(a) Eigenvector matrix for $\beta = 10$

$$\begin{bmatrix} 0.5773 & -0.5773 & -0.5773 \\ 0.5773 & 0.2887 - 0.5i & 0.2887 + 0.5i \\ 0.5773 & 0.2887 + 0.5i & 0.2887 - 0.5i \end{bmatrix}$$

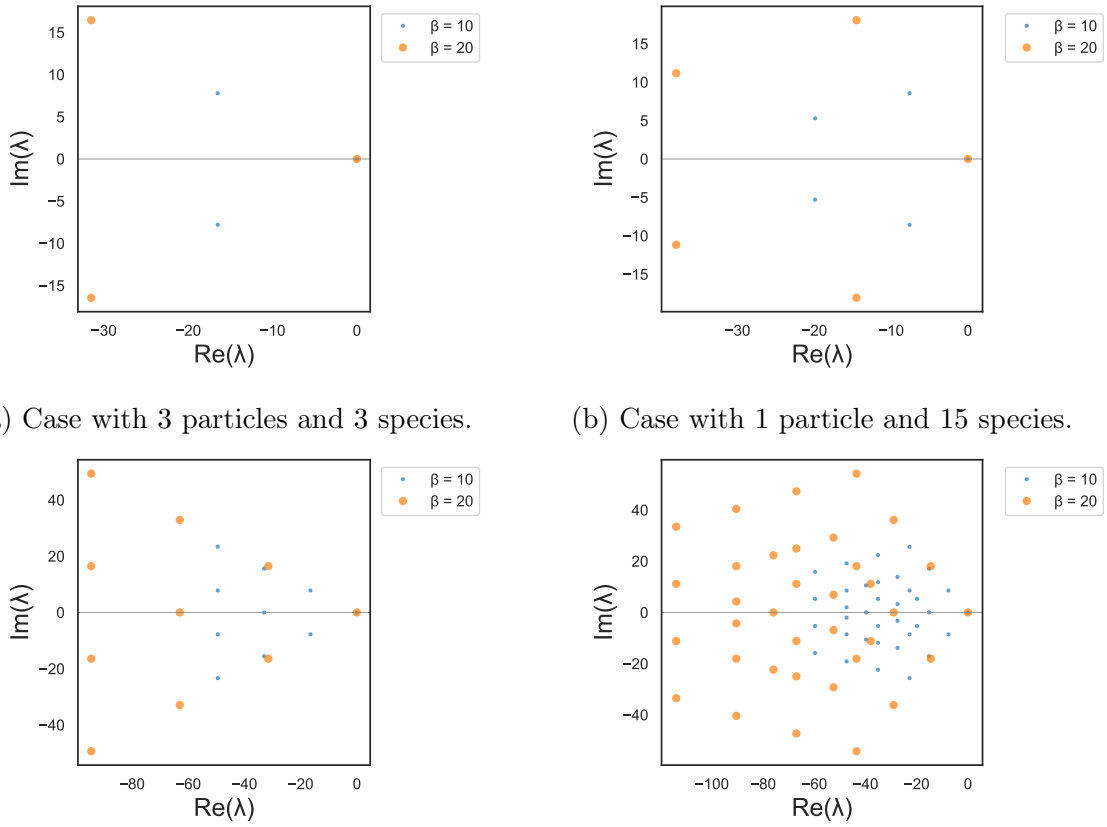
(b) Eigenvector matrix for $\beta = 25$

Figure 4.2: Eigenvector matrices for different values of β. They differ for a set of rows' permutations.

4.4 Analysis of the analytical results

4.4.1 Eigenvalue spectrum

The distribution of the eigenvalues over the complex plane exhibits some interesting properties:



(a) Case with 3 particles and 3 species.

(b) Case with 1 particle and 15 species.

(c) Case with 3 particles and 3 species.

(d) Case with 1 particle and 15 species.

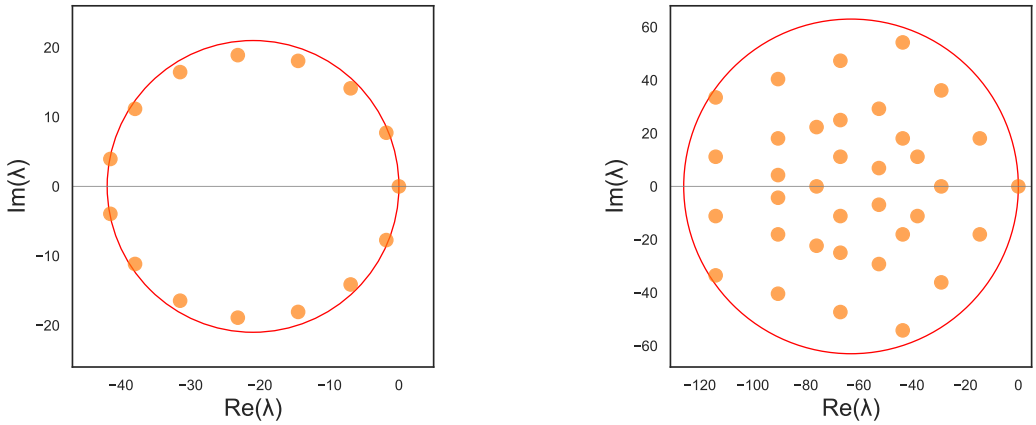
Figure 4.3: Distribution of the eigenvalue spectrum over the complex plane for different configurations of the system.

- Stability of the geometric configuration under change of β:
  In the 1 particle case the eigenvalues distribute over the vertices of a polygon with M edges (Figure 4.3a and 4.3b). Increasing the value of β the modulus of the eigenvalues increases yet leaving unaltered the polygonal structure.

- Stability of the geometric configuration under change of the number of particles:
  Adding particles multiply the range of the eigenvalue spectrum by N, still leaving the same geometric configuration, but increasing the number of particles the eigenvalues start filling the area inside the polygon (Figure 4.3c and 4.3d).

In the $N \gg 1$ and $M >> 1$ limits we expect then to find all the eigenvalues distributed in an approximate circular disc. This congecture first leads to the notion of Geršgorin circle.

Eigenvalue spectrum inside the Geršgorin circle.



(a) Case with 1 particle and 15 species.

(b) Case with 3 particle and 5 states.

Figure 4.4: Distribution of the eigenvalue spectrum inside the Geršgorin circle.

Definition 2. *The curve in the complex plane defined by*

$$\left\{z \,\epsilon\, \mathbb{C} : |z - \mathbb{W}_{ii}| \leq r_i(\mathbb{W})\right\} \tag{4.8}$$

$$r_i(\mathbb{W}) = \sum_j \mathbb{W}_{ij} \qquad for \quad i \neq j \tag{4.9}$$

is called Geršgorin circle, where i is the row index of the matrix and r_i is called i^{th} Geršgorin radius.

By the Geršgorin Theorems [3] one knows that all the eigenvalue spectrum lies inside the union of all the Geršgorin circles.

The properties on the rows of the $\mathbb{W}$ matrix make all the circles collapse onto one with radius $r = -\mathbb{W}_{00}$ and center $c = (\mathbb{W}_{00}, 0)$ on the real line, as in Figure 4.4. Note that even in the case of multiple particles (Fig. 4.4b) the eigenvalues with greater imaginary part lie near the Geršgorin circle.

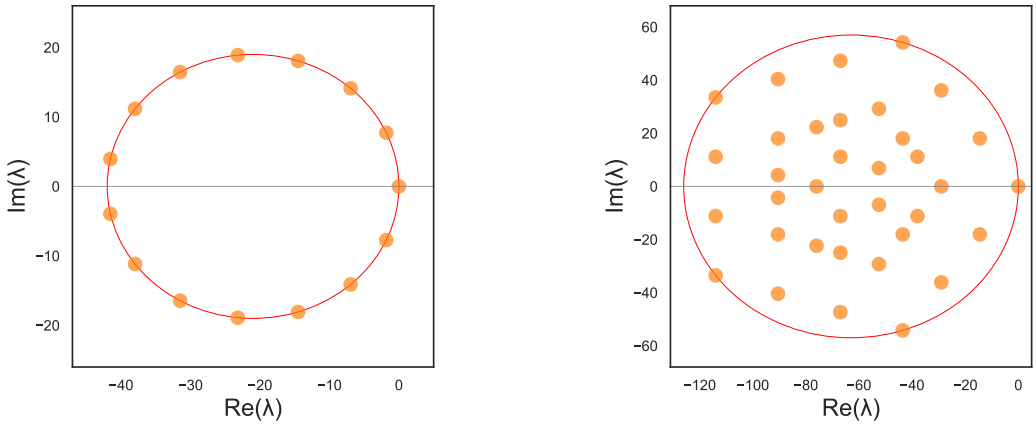
Studying the one particle configuration and dividing the Geršgorin circle in M sections, therefore approximating the actual eigenvalues, we find that the relative differences between the real parts of the analytical and approximate eigenvalues are zeros, while the relative differences between the imaginary parts are all equal (and dependent on β). This is an hint of a distribution of the eigenvalues on an ellipse rather than on a circle.

Finally we find that the correct ellipse equation is

$$E = \left\{z \,\epsilon\, \mathbb{C},\ z = x + iy \middle| \frac{(x - \mathbb{W}_{00})^2}{\mathbb{W}_{00}^2} + \frac{y^2}{\tau^2} = 1\right\} \tag{4.10}$$

$$\tau = \mathbb{W}_{01}\mathbb{W}_{10} - N \tag{4.11}$$

Eigenvalue spectrum onto the elliptical curve.



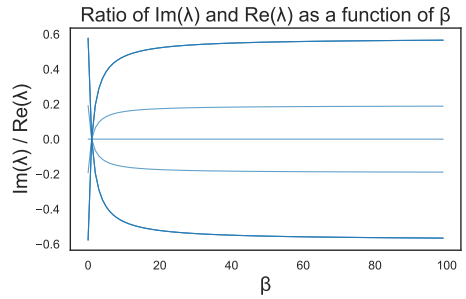
(a) Case with 1 particle and 15 species.

(b) Case with 3 particle and 5 states.

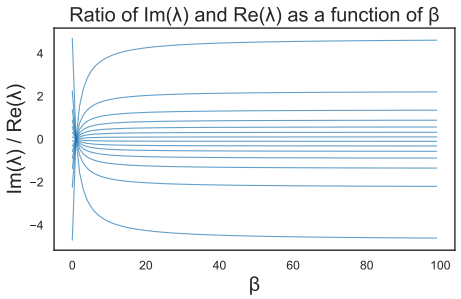
Figure 4.5: Distribution of the eigenvalue spectrum inside the ellipse.

where N is as usual the total number of particles and τ is related to the autocorrelation coefficient in the Girko Elliptic Law [4]. Since this theorem holds for an ensemble of random matrices, we find this experimental elliptical law considering our matrix as element of an ensemble of identical $\mathbb{W}$ matrices. In this manner the correlation needed to find the ellipse $\langle \mathbb{W}_{01}\mathbb{W}_{10} \rangle$ becomes trivially the product of the elements $\mathbb{W}_{01}$ and $\mathbb{W}_{10}$.

Using this ellipse equation we find (Fig. 4.5) that the eigenvalue distribution matches the ellipse for each configuration of number of species M for $N = 1$ (Fig. 4.5a). As stated above in the $N > 1$ case it is still true that the eigenvalues on the vertices of the bigger polygon follow the same elliptical equation (Fig. 4.5b).



(a) Case with 3 particles and 3 species.



(b) Case with 1 particle and 15 species.

Figure 4.6: Ratio of imaginary and real part of each eigenvalue as a function of beta.

4.4.2 Autocorrelation oscillations

We know that since each term of the Master Equation solution (2.18) and therefore of the autocorrelation is of the form

$$e^{\mathrm{Re}(\lambda_i)t}\cos(\mathrm{Im}(\lambda_i)t) \tag{4.12}$$

we need the imaginary part of the i^{th} eigenvalue (the angular velocity) to be greater[2] of the real part of it (the decay rate) for the oscillations to be visible before getting suppressed by the exponential term. In Figure 4.6 we show the ratio between the imaginary part and the real part of eigenvalues as a function of the parameter β.

In the first case of 3 particles and 3 species (Figure 4.6a) the ratio is always lesser than one even if growing with β, while in the case of 1 particle and 15 species most of the eigenvalues stay below the ratio of one, but there are few of them that go well above the ratio of one.

Both of the situations graphed in Figure 4.6 share the fact that the ratio is saturated around the value of $\beta = 20$, thus we shall use that value in order to obtain the most visible oscillations in the autocorrelation. Furthermore, analyzing cases with same number of species and different number of particles we find that the maximal ratios are common to all these cases (they don't depend on the number of particles).

Using the (2.24) with the eigenvalues and eigenvectors found solving the Master Equation we build the autocorrelation function as a parametric function of β. Exploiting the eigenvalue distribution we expect the eigenvalues on the ellipse to be responsible for the major oscillating modes, hence we can restrict ourself to the one particle case.

As seen in Figure 4.7 and stated above the autocorrelation function is oscillating when β is large enough, therefore we use a value of $\beta = 20$ in the SSA simulations.

[2]Since the real part of the eigenvalues is negative and the imaginary part will come in conjugate pairs due to the reality of the $\mathbb{W}$ matrix both terms shall be taken in absolute value.

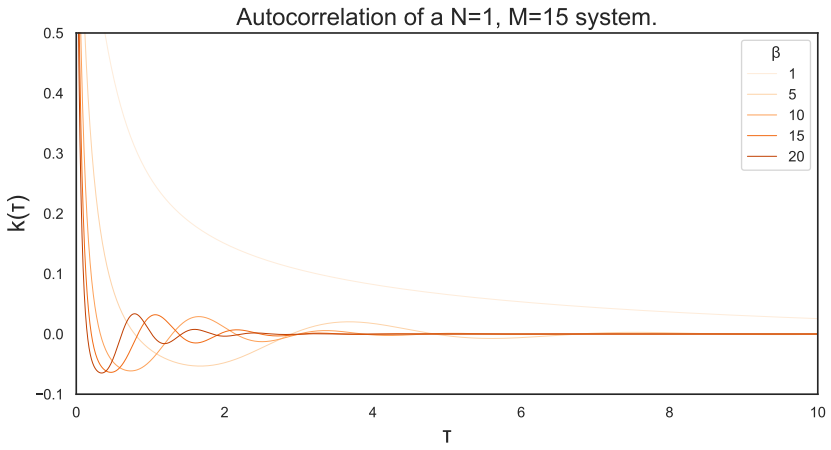


Figure 4.7: Autocorrelation of the analytical solution of the system with 1 particle and 15 species.

4.5 Simulation of the system

We simulate our system using the simplest SSA algorithm: the Gillespie algorithm [6]. The Gillespie algorithm, as each SSA, doesn't solve numerically the Chemical Master Equation, since this would mean finding the conditional probability $P(\vec{X}|\vec{X}_0)$. Conversely, it generates directly single trajectories $\vec{X}(t)$ (they are realizations of the stochastic process defined by $\vec{X}$).

To accomplish that one defines a new probability function $p(\tau, j|\vec{x}, t)$ as the probability (given $\vec{x} = \vec{X}(t)$) that a reaction occurs in the time interval $\left[t+\tau, t+\tau+d\tau\right]$ and that it is the j^{th} reaction.

Given the definition of propensities as for the CME (2.27) one finds that

$$p(\tau, j|\vec{x}, t) = \alpha_j(\vec{x})e^{-\alpha_0(\vec{x})\tau} \tag{4.13}$$

$$\alpha_0(\vec{x}) = \sum_i \alpha_i(\vec{x}) \tag{4.14}$$

where the summation is over all the propensities. This implies that τ is a random variable exponentially distributed, while j is an integer independent random variable with probability $\frac{\alpha_j(\vec{x})}{\alpha_0(\vec{x})}$.

In the present we use the so called direct method; explicitly:

$$\tau = \frac{1}{\alpha_0(\vec{x})} \ln\left(\frac{1}{r_1}\right) \tag{4.15}$$

$$j = \min\left\{j' | \sum_{i=1}^{j'} \alpha_i(\vec{x}) > r_2\alpha_0(\vec{x})\right\} \tag{4.16}$$

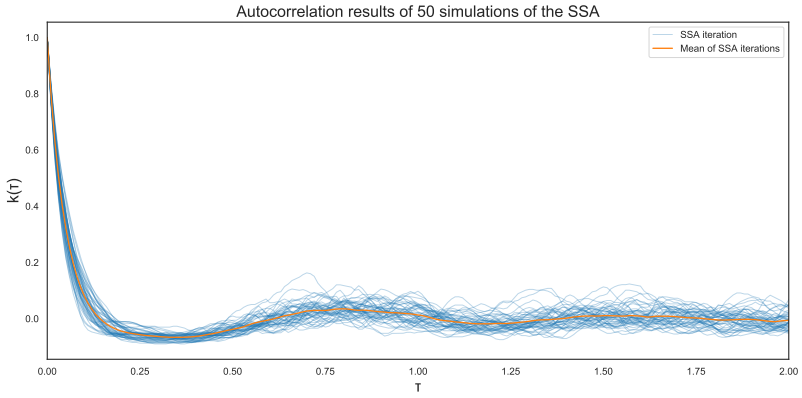


Figure 4.8: Autocorrelation of the SSA iterations of the system with 1 particle and 15 species. The 50 blue solid lines are the autocorrelation functions for each SSA simulation. The orange solid line represents their mean.

with r_1 and r_2 drawed from an uniform distribution.

The SSA algorithm steps are thus:

- Initialization
  Set $t = t_0$ and $\vec{x} = \vec{x}_0$.
- Evaluation
  Evaluate all the possible $\alpha_j(\vec{x})$.
- Generation
  Generate r_1 and r_2 and therefore τ and j.
- Reaction
  Simulate the occurred reaction updating t and $\vec{x}$.
- Storage
  Record $(\vec{x}, t)$ and return to the Initialization step (or end the algorithm with the desired condition[3])

We end the present chapter showing the results of the Gillespie algorithm used to find the right regression Bayesian model for time series' autocorrelations of stochastic processes.

In Figure 4.8 there are the autocorrelations from each iteration of the SSA, that are the input data of the bayesian model. Note that the mean of those autocorrelation resembles the analytical solution. Since there isn't any way to perform the Gillespie algorithm without choosing a value for the parameter β we put it to $\beta = 20$.

[3]Usually a time limit or some other conditions on the reaching of an absorbing state.

Chapter 5

Analysis of the oscillations in autocorrelation with a Bayesian regression model

We first start analyzing our regression problem. We have prior information on the form of the regression function here called f; explicitly, we know that it should have the form of a finite superposition of complex exponentials:

$$f_i = w_i e^{(\mathrm{Re}(\lambda_i)+i\,\mathrm{Im}(\lambda_i))t}$$

W can therefore exploit the information about the reality of the $\mathbb{W}$. Since every real valued non symmetric matrix has its eigenvalues in complex conjugate pairs, we can rewrite the complex exponential functions as

$$f_i(t) = w_i e^{\mathrm{Re}(\lambda_i)t}\cos(\mathrm{Im}(\lambda_i)t)$$

Furthermore we assume that our data is distributed as a Normal distribution

$$y(t) \sim \mathcal{N}\big(\mu = \sum_i f_i(t), \sigma = s\big) \tag{5.1}$$

Now we need to define prior distribution over all the parameter in the model. Since we know that both s and w_i are positive and possibly below the value of 1 we set for them the following prior distributions:

$$w_i \sim \mathrm{InverseGamma}(\mu = 0.5, \sigma = 3) \tag{5.2}$$
$$s \sim \mathrm{InverseGamma}(\mu = 0.5, \sigma = 3) \tag{5.3}$$

where we assume homoscedastic noise. We choose in the present an Inverse Gamma distribution (Fig. 5.2) following the considerations in [8] because we don't want the weights and the noise s to tend toward zero.

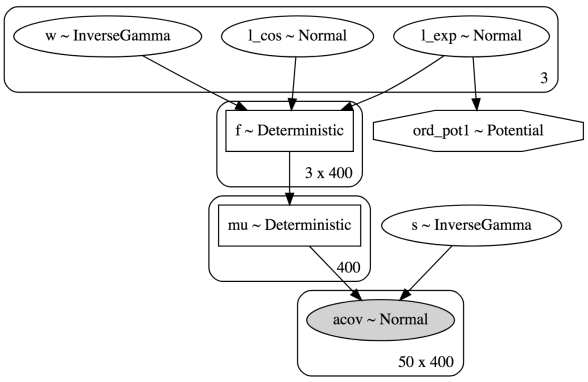


Figure 5.1: Graph of the model used to fit the autocorrelation of the toy model with 1 particle and 15 species.

Moreover, watching Fig. 4.8 we expect the decay rates in the range $\big[0, 50\big]$ and the oscillation angular velocities to be in the range $\big[0, 20\big]$. Hence we put as prior distributions

$$\mathrm{Re}(\lambda_i) \equiv \mathrm{l\_exp}_i \sim \mathcal{N}(\mu = 20, \sigma = 10) \tag{5.4}$$

$$\mathrm{Im}(\lambda_i) \equiv \mathrm{l\_cos}_i \sim \mathcal{N}(\mu = 10, \sigma = 5) \tag{5.5}$$

In order to complete our model we need to get rid of the multimodality induced by the fact that all the components f_i are modeled in the same way. In fact our posterior

Prior distribution for the weights w_i and for s

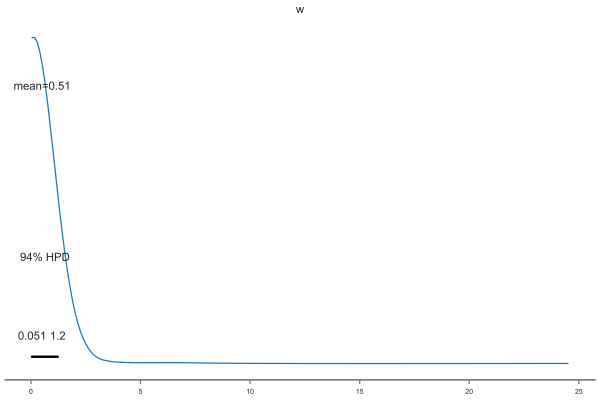


Figure 5.2: Inverse gamma distribution with $\mu = 0.5$ and $\sigma = 3$. It is shown the mean value and the 94% Highest Density Interval.

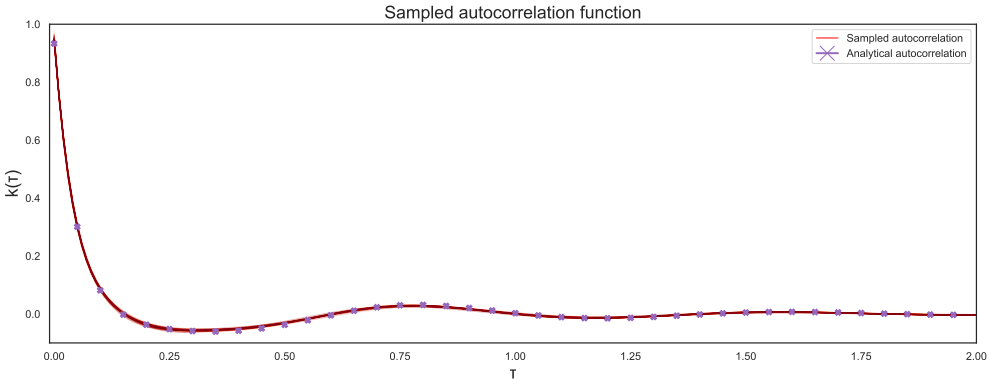


Figure 5.3: Plot of the autocorrelation samples drawn from the posterior distribution in the $N = 1$, $M = 15$, $\beta = 20$ configuration.

shall be the same under the exchange of a pair $(\text{l\_cos}_i,\ \text{l\_exp}_i) \leftrightarrow (\text{l\_cos}_j, \text{l\_exp}_j)$. A solution to this problem is using an ordering technique to break this symmetry. The quicker way to do it in PyMC3 is using a Potential, that allows to add an arbitrary term to the log probability function. For our purposes we build it in a way that penalizes any kind of different order of the parameters from the proper one. Let's note that in order to avoid bad initializations of the Markov Chain we need to inform our model that it should begin the chain with some ordered initial values.

The last information the model needs is the number of components f_i. Unfortunately there isn't any easy way to deal with this kind of problem, even if a set of different solutions has been proposed in [10], [11], [12] . Anyway, in our model something remarkable happens when the number of components isn't right. When the algorithm runs if there are too many components two different things can happen. In a first scenario, two components pick the same values for the parameters, with half weight, but we dealt with this possibility through the ordering potential. Another thing that can happen is that the weight of a component tends towards zero. In this case the variance of the distribution of such a weight tends to zero too, causing a bad fit by the NUTS algorithm, signaled by a *max_treedepth* warning by PyMC3. Therefore in the present we run the Markov chain with a different number of components f_i, starting with ten and hence reducing the components until the right model is found.

In Figure 5.4 there are the results of a NUTS sampling with four different chains that converge quite fast. In Table 5.1 there are the results of the fit for the relevant parameters, with the analytical values attached for comparison. It can be seen that our model fits quite good the smallest angular velocities (the largest periods) and the related decay rates, that is exactly what we are looking for in order to fit the biological data. The distribution of the largest pair of parameters (l_exp, l_cos) is more spreaded

	mean	analytical	sd	hpd_3%	hpd_97%	r_hat
l_exp[0]	1.9	1.82	0.17	1.57	2.23	1.0
l_exp[1]	8.14	6.95	1.72	5.12	11.42	1.0
l_exp[2]	30.7	31.5	3.41	24.98	37.2	1.0
l_cos[0]	7.83	7.73	0.1	7.65	8.01	1.0
l_cos[1]	14.33	14.12	0.84	12.74	15.91	1.0
l_cos[2]	14.61	16.45	2.1	10.34	18.13	1.0

Table 5.1: Results of the NUTS run with 3 exponential components for the relevant variables. It is shown the posterior distribution mean and the standard deviation, together with 94% Highest Posterior Density intervals. As can be seen the real (analytical) values are included in those intervals. The $\hat{r}$ statistics [9] is near 1 when the chains converged.

because it tries to incorporate different components at small lag time of the analytical autocorrelation. To probe the flexibility of our model it has been checked for a variety of configurations ($M = (10, 15, 20)$ and $\beta = (10, 20)$).

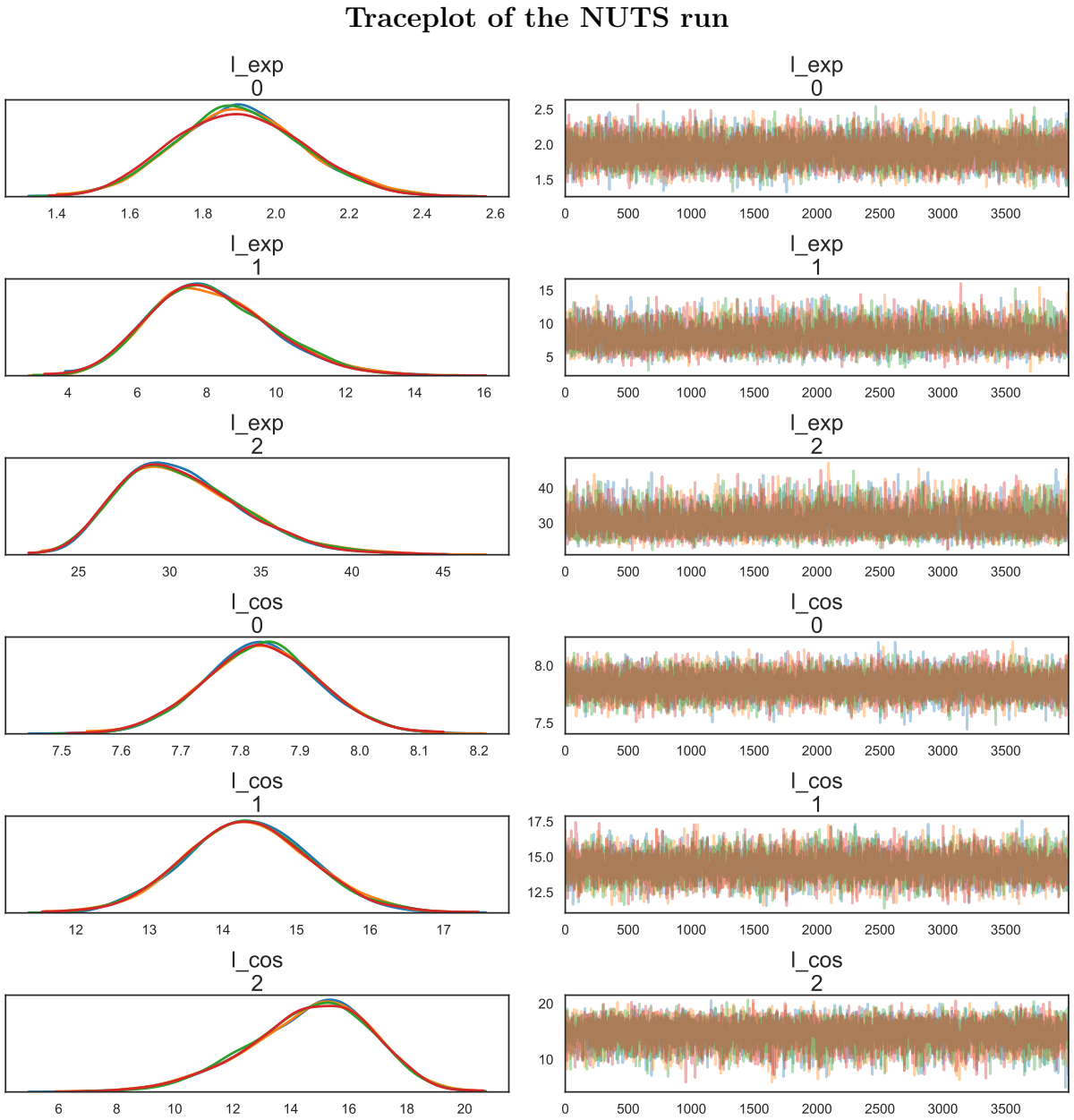


Figure 5.4: Traceplot of the relevant parameters from a NUTS run with 4 independent chains and three exponential components.

Chapter 6

Analysis of the fluctuations in NF-kB activity in colorectal cancer cells

6.1 Introduction

6.1.1 NF-kB signaling system

Aim of the project is the analysis of the oscillations of NF-kB activity in metastatic colorectal cancer cells. The NF-kB protein complex is a genes regulator that controls cell proliferation, cell survival and protection from apoptosis. Therefore this protein complex can be considered as a pleiotropic mediator of gene expression control. The NF-kB is composed at a molecular level by two subunits. The former (p50) is responsible for DNA binding, while the latter (p65) lacks DNA binding activity and affects the susceptibility of the protein complex to his inhibitory proteins, namely IkBα and IkBβ.

In absence of stimuli the NF-kB is retained in the cytoplasm by IkB inhibitor proteins that bound the complex and inactivate his transcription factor. As can be seen in Fig. 6.1 it is a cytokine activated protein kinase complex, namely IKK, that is responsible for the dissociation of NF-kB and IkB by phosphorylation of the latter, thus leading to degradation of the inhibitor protein via the proteasome by an ubiquination process. Once the inhibitor is detatched from the NF-kB it can subsequently enter the nucleus where it can change the expression of specific genes, thus leading to some physiological response.

There are two NF-kB signaling pathways regulated by two multiprotein IKK complexes. In the so called canonical NF-kB signaling pathway IKKβ alone is sufficient for the phosphorylation of the protein inhibitor IkBα. This pathway is mainly associated with inflammatory response. Conversely, in the non canonical pathway it is the IKKα complex that activates the NF-kB. An important factor in the analysis of the NF-kB regulatory system is that one of the earliest NF-kB responsive promoters is its own inhibitor protein, namely the IkBα. This induces a negative feedback producing a propensity for

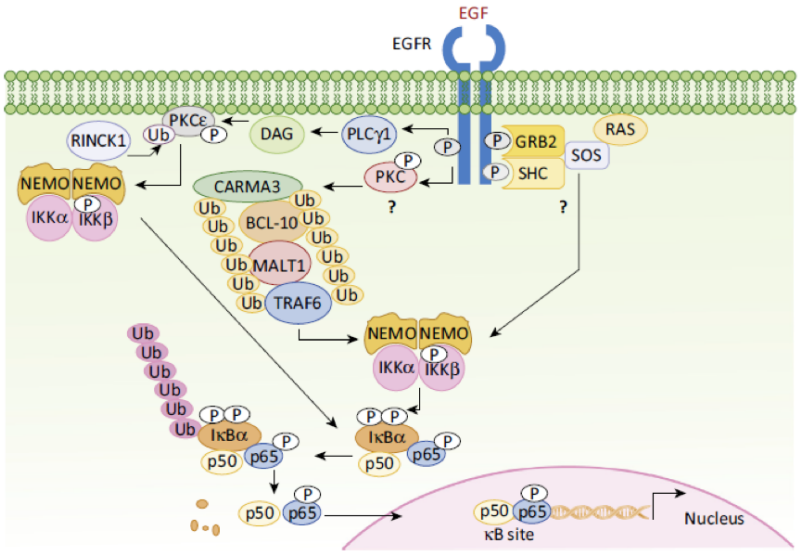


Figure 6.1: Molecular pathway of activation of NF-kB by EGF. (Shostak et al. "EGFR and NF-kB : partners in cancer", 2015)

oscillations in the NF-kB activity [13].

NF-kB activity enhances angiogenesis, cell proliferation and promotes metastasis and cell invasion. Moreover, there is evidence that shows correlation between EGFR signaling and NF-kB activity in solid tumours like colorectal cancer. Indeed, the Epidermal Growth Factor protein binds to his receptor EGFR (Fig. 6.1) whose signal leads to NF-kB activation through the proteasome mediated degradation of the IkB inhibitor.

6.1.2 Cell cultures configuration and luciferase assay

There have been studied two colorectal cancer cell lines, one sensitive and one resistant to the Cetuximab treatment. Notably, the resistant cells were derived from the sensitive ones, by continuous exposure to an anticancer drug (Cetuximab). For the experimental design, cell were treated with a combination of stimuli, including Cetuximab, IL1A, IL1B alone and in combination. Thus, these two model systems share the same genetic background. The IL1A and IL1B interleukins are important mediators of the inflammatory response and are involved in a variety of cell activities. Moreover, they are the most important pro-inflammatory cytokines in tumoral cells' environment, thus they have a role in the enhancement of the NF-kB activity.

In order to quantify the oscillations in the NF-kB activity the luciferase assay has been used. Every 20 minutes fluorescent signal values of luciferase have been recorded for each cell colony, for a total of 216 measurements for well, in a time interval of 72 hours. Moreover the luciferase values have been preprocessed by subtracting the median

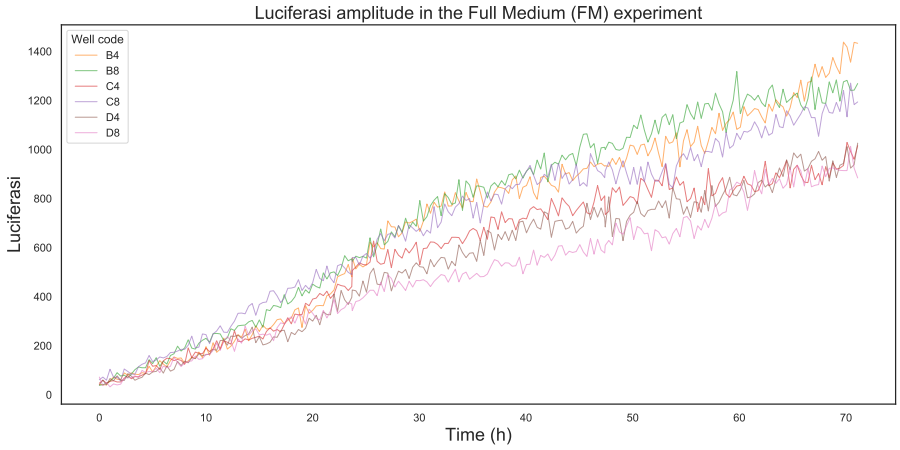


Figure 6.2: Luciferasi amplitude in the Full Medium experiment for each of the 6 wells (cell colonies)

value of the background signal for each luciferasi detection cycle.

There are six different treatments for each cell line (Parental cells and Cetuximab resistant cells (CXR)):

- Full medium with no treatment (FM)
- Treatment with IL1α (FM + IL1A)
- Treatment with IL1β (FM + IL1B)
- Treatment with Cetuximab (FM + CTX)
- Treatment with Cetuximab and IL1α (FM + CTX + IL1A)
- Treatment with Cetuximab and IL1β (FM + CTX + IL1B)

For each treatment there are 6 wells, which are completely pooled in the analysis, performed separately for each of the 12 different treatments.

6.2 Preprocessing operations on the raw data

During the experiment time the luciferasi amplitude for each well increases constantly due to cell proliferation, as shown in Fig. 6.2. Since for this analysis we are only interested in the fluctuations of NF-kB activity, the first step is a detrending operation in order to extract the fluctuating component of the signal. The technique chosen to

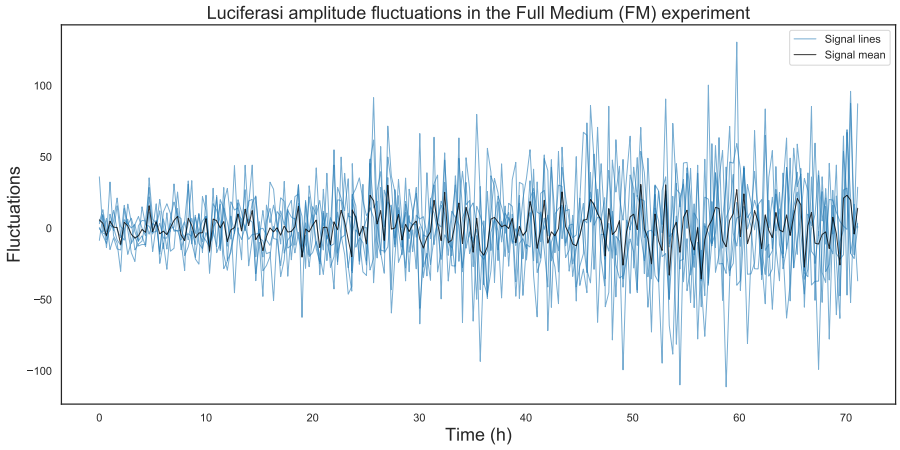


Figure 6.3: Luciferasi amplitude after the subtraction of the LOWESS smoothing curve for the Full Medium experiment for each of the 6 cell colonies

accomplish this is the LOWESS, a kind of nonlinear regression technique particularly useful for our purposes.

In parametric nonlinear regression models one defines

$$y = f(\vec{x}, \vec{\theta}) + \epsilon \tag{6.1}$$

where f is a parametric function of the parameters $\vec{\theta}$ specified a priori and ϵ is usually specified as distributed according to a Normal distribution

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{6.2}$$

The general nonparametric regression model in defined in the same way, the only difference being that the function f isn't dependent on any parameter and it is left unspecified[1].

Let's consider now the LOcally WEighted Scatterplot Smoothing (LOWESS) for the simple regression case (one predictor)

$$y = f(x) + \epsilon \tag{6.3}$$

Be x_0 some x value; let's first start performing a k^{th} order polynomial regression

$$y = a_0 + a_1(x - x_0) + \cdots + b_k(x - x_0)^k + e \tag{6.4}$$

[1]In most of the nonparametric regression models f is assumed to be a smooth function.

Moreover a weight function W is introduced in order to weight the cases according to the distance $x - x_0$. A commonly used W function is the tricubic weight function

$$W(z) = \begin{cases} \left(1 - |z|^3\right)^3 & \text{for} \quad |z| < 1 \\ 0 & \text{for} \quad |z| \geq 1 \end{cases} \tag{6.5}$$

with $z = \frac{x - x_0}{h}$, where h is the window width called smoothing parameter. Since the predictor is conveniently centered in x_0 the predicted value of y is nothing more than $\hat{y}_0 = a_0$. Repeating the procedure for all the x values one finally finds the desired smoothing regression curve. As a final remark on the robustness of this technique, one disadvantage of this procedure is that it is quite prone to the effect of outliers. To cut out the relevance of the outliers a robust iterative version [14] of the algorithm has been used.

For our purposes the window width is set to $h = 72/10$, which means using 21 points for each local weighted regression, and the polynomial order is set to $k = 1$ (linear regression). Finally, subtracting the result of the LOWESS regression from each of the luciferasi amplitudes we find (Example in Fig. 6.3) the oscillating component of the luciferasi amplitude.

The last step before implementing the bayesian regression model concerns finding the signals' autocorrelations. The force brute computation method defines the autocorrelation for the discrete time stationary signal as

$$k_x(\tau) = \sum_t x_t x^*_{t-\tau} \tag{6.6}$$

Since this brute force algorithm has computational cost of order N^2 where N is the number of data points, we exploit the Wiener-Khinchin theorem to perform the calculation. Let's define for a continuous time stochastic process $X(t)$ the power spectral density as

$$S_X(\omega) = \left| \mathcal{F}_\omega \big[X(t) \big] \right|^2 \tag{6.7}$$

and the autocorrelation function for a stationary process as

$$K_X(\tau) = \mathbb{E}\big[X(0) X^*(\tau) \big] - \mu(0)\mu^*(\tau) \tag{6.8}$$

In its simplest form, the Wiener-Khinchin theorem states that if both the power spectral density of the signal and the autocorrelation function are continuous and absolutely integrable in the Lebesgue sense, hence

$$S_X(\omega) = \int_{-\infty}^{\infty} d\tau K_X(\tau) e^{-i\omega\tau} \tag{6.9}$$

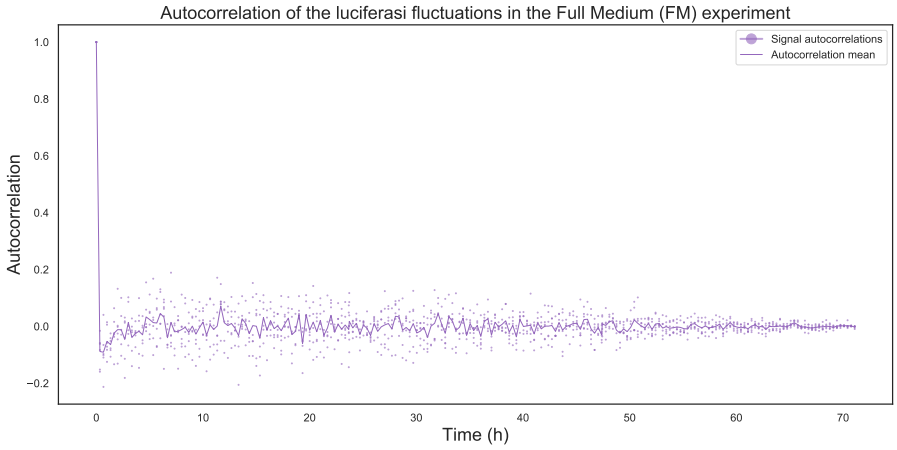


Figure 6.4: Luciferasi fluctuations' autocorrelations for the Full Medium experiment for each of the 6 cell colonies.

Using this property of the power spectral density and the autocorrelation of being Fourier-transform pairs one can therefore compute the autocorrelation in a more efficient manner (computational cost of order $N \ln N$) using two Fourier transforms[2]

$$F_K(\omega) = \mathcal{F}_\omega\big[X(t)\big] \tag{6.10}$$
$$S(\omega) = F_K(\omega)F_K^*(\omega) \tag{6.11}$$
$$K(\tau) = \mathcal{F}_\tau^{-1}\big[S(\omega)\big] \tag{6.12}$$

Using this algorithm we compute the autocorrelation for each set of different treatments (Example in Fig. 6.4) that are the input data of our regression model.

6.3 Bayesian regression model for the luciferase autocorrelations' oscillations

While at first glance (Fig. 6.4) there aren't visible oscillations in the signal's autocorrelation, we expect them to be present underneath the noise anyway, due to our information on the biological system to which the data refer.

Therefore we build the bayesian regression model as in Chapter 5. First we run a control version of the model without any constrain on the parameters in order to obtain an overview of the distribution of our parameters. As opposed to the toy model analyzed in Chapter 5, here (Fig. 6.5) there isn't any order for the exponential rates, but rather

[2]We use the fast Fourier transform approximation in the actual computation.

it happens to be on the oscillation angular velocities. Given this preliminary analysis we modify our model to allow the decay rates to be the same and ordering the angular velocities, moreover constraining them on an interval $\left[0, 110\right]$ to avoid faster and faster oscillating modes.

With these modifications we build the probabilistic model, recalling the usual considerations on the number of exponential components. Since we are more interested in the decay time and the oscillation period, the decay rate and the oscillation angular velocity are inverted with a deterministic transformation in the model, in order to gain access to the posteriors of the biological relevant variables directly.

6.4 Analysis of the results

The bayesian model has been used separately for each experiment to infer the parameters' distributions. In particular, we are interested in the oscillation periods and decay times. Moreover, in order to obtain a good quality fit of the autocorrelation data we needed to equip our model with both large and small oscillation periods, but we are mainly interested in the bigger ones, therefore in the following Tables 6.1 and 6.2 we show the largest oscillation period mode for each experiment. However the interested reader can find the complete information (Posterior plots and some other informative plots) in the Appendix.

As can be seen the oscillation periods' range is roughly of six hours at most. Conversely, the decay time interval is not as well constrained as for the oscillation periods. This isn't surprising, since we already knew our model worked better with the periods than the decay parameters. About the weight parameters, it can be seen in Fig. 6.6 that for the largest oscillation period components they don't differ sensibly, while the weights of the components with smaller oscillation periods depend strongly on the total number of oscillation components, where a lower number of components leads to higher weight values. With these considerations, we assume that the weight parameters don't contain quantifiables biological information in this setup.

From Fig. 6.6 it results that the oscillation periods can be divided in three classes with different values for the oscillation periods. To the first class belong the cell colonies with the largest oscillation period roughly between five and six hours. Furthermore for these experiments the decay times are well defined around the value of five hours. For the second class the largest oscillation is roughly between the values of three and four hours. In this class the decay times for some components have distributions with longer tails, but still the highest probability is around five hours. Finally, in the third class not only the largest period is located between two and three hours, but the other oscillation periods are also far closer. Moreover, for this class the decay rates are greater than those of the other two classes mentioned before.

Traceplot of the NUTS run for the FM cell colonies

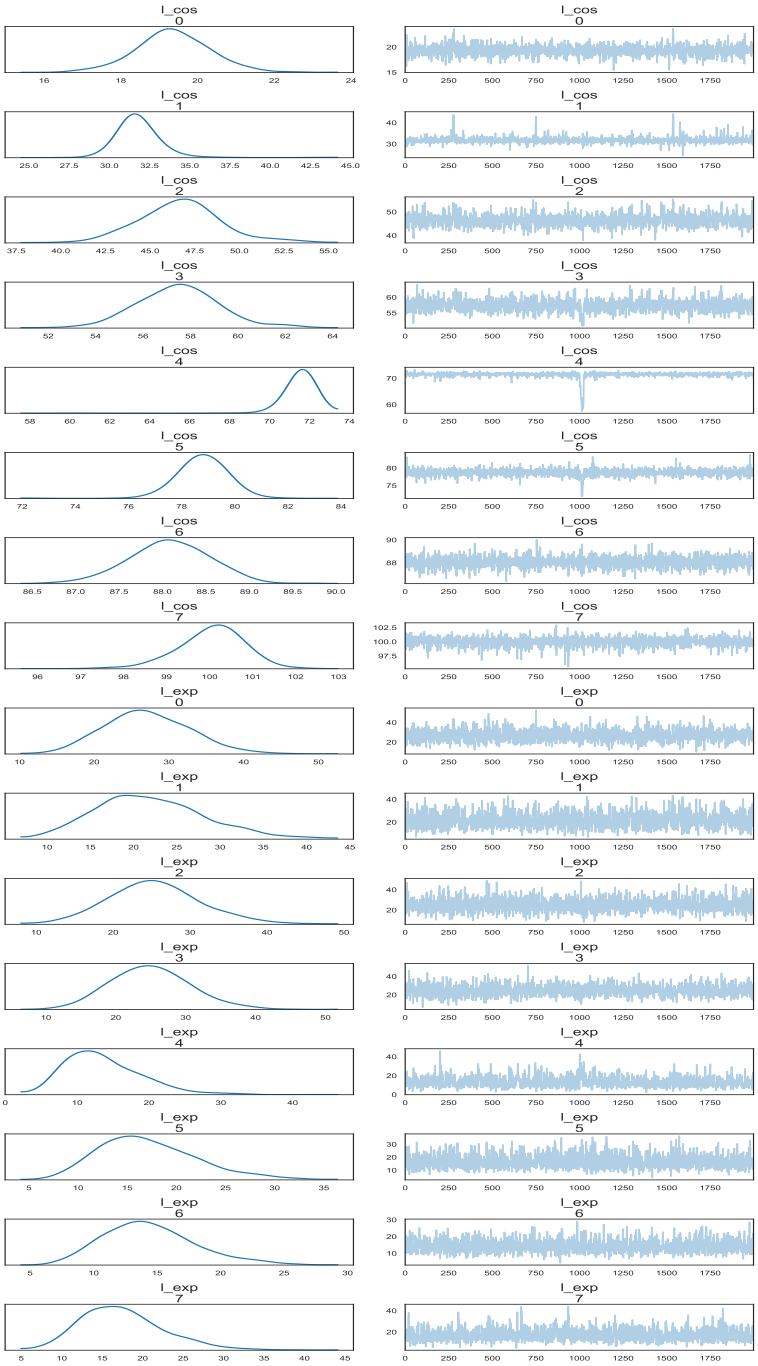


Figure 6.5: Traceplot results for the decay rates and the oscillation angular velocities of the analysis of Full Medium cell colonies.

Graphical results of the statistical analysis for the three longest oscillation periods' components

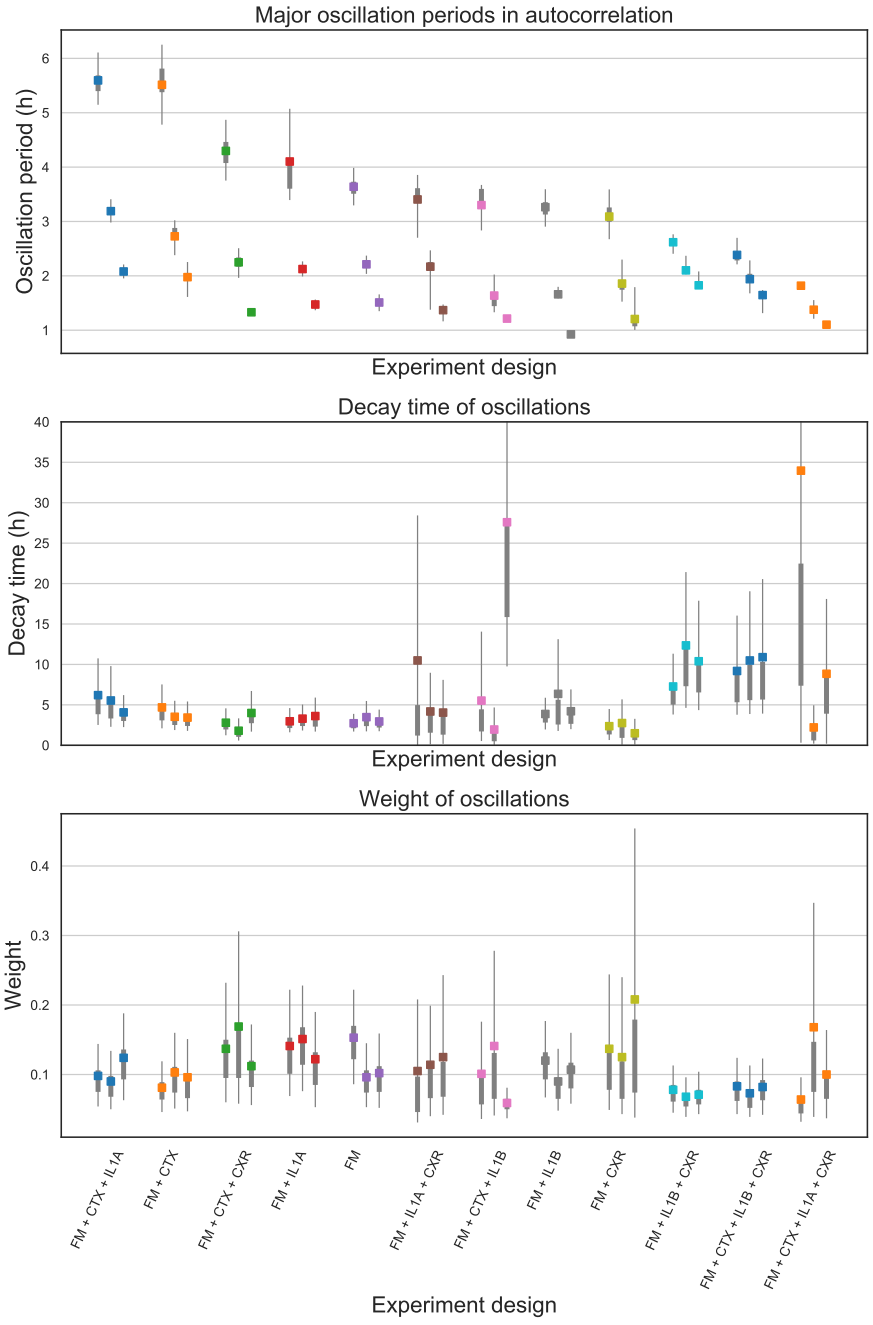


Figure 6.6: Boxplots for the oscillation periods, the decay times and the components' weights of the three modes with the largest oscillation periods.

Analysis results for the major oscillation mode in Parental cell colonies

		mean	sd	hpd_3%	hpd_97%	hpd_25%	hpd_75%
FM	cos_t[0]	3.639	0.201	3.296	3.986	3.51	3.731
	exp_t[0]	2.724	0.633	1.698	3.885	2.108	2.838
	w[0]	0.153	0.037	0.086	0.222	0.122	0.17
FM + CTX	cos_t[0]	5.514	0.444	4.781	6.252	5.378	5.812
	exp_t[0]	4.685	1.618	2.096	7.526	3.081	4.897
	w[0]	0.081	0.02	0.046	0.119	0.064	0.089
FM + IL1A	cos_t[0]	4.103	0.484	3.394	5.073	3.603	4.133
	exp_t[0]	2.967	0.954	1.606	4.603	2.132	3.034
	w[0]	0.141	0.042	0.069	0.222	0.101	0.153
FM + CTX + IL1A	cos_t[0]	5.595	0.262	5.147	6.108	5.398	5.687
	exp_t[0]	6.201	2.593	2.515	10.75	3.842	6.429
	w[0]	0.098	0.025	0.054	0.144	0.075	0.106
FM + IL1B	cos_t[0]	3.263	0.183	2.905	3.593	3.128	3.354
	exp_t[0]	3.872	1.161	1.953	5.878	2.81	4.068
	w[0]	0.12	0.03	0.067	0.177	0.093	0.132
FM + CTX + IL1B	cos_t[0]	3.301	0.274	2.835	3.672	3.281	3.599
	exp_t[0]	5.527	4.72	0.526	14.057	1.71	4.447
	w[0]	0.101	0.045	0.036	0.176	0.057	0.102

Table 6.1: Results of the bayesian analysis for the major oscillation period component for each experiment of cell colonies not resistant to Cetuximab treatment.

Analysis results for the major oscillation mode in Cetuximab resistant cell colonies (CXR)

		mean	sd	hpd_3%	hpd_97%	hpd_25%	hpd_75%
FM	cos_t[0]	3.09	1.234	2.674	3.59	2.995	3.258
	exp_t[0]	2.363	1.131	0.65	4.488	1.341	2.533
	w[0]	0.137	0.074	0.049	0.244	0.078	0.137
FM + CTX	cos_t[0]	4.298	0.308	3.752	4.87	4.075	4.463
	exp_t[0]	2.791	0.954	1.236	4.569	1.955	3.026
	w[0]	0.137	0.053	0.06	0.232	0.095	0.15
FM + IL1A	cos_t[0]	3.405	0.508	2.701	3.856	3.427	3.612
	exp_t[0]	10.494	24.032	0.034	28.429	1.2	4.971
	w[0]	0.105	0.064	0.031	0.208	0.046	0.097
FM + CTX + IL1A	cos_t[0]	1.819	0.031	1.789	1.842	1.807	1.823
	exp_t[0]	33.968	43.859	0.316	104.57	7.367	22.471
	w[0]	0.064	0.027	0.032	0.096	0.044	0.065
FM + IL1B	cos_t[0]	2.618	0.192	2.406	2.764	2.58	2.69
	exp_t[0]	7.27	2.204	3.805	11.327	5.026	7.358
	w[0]	0.078	0.018	0.045	0.113	0.061	0.085
FM + CTX + IL1B	cos_t[0]	2.385	0.259	2.211	2.699	2.283	2.349
	exp_t[0]	9.19	5.612	3.779	16.04	5.323	8.825
	w[0]	0.083	0.023	0.043	0.124	0.062	0.09

Table 6.2: Results of the bayesian analysis for the major oscillation period component for each experiment of cell colonies resistant to Cetuximab treatment (CXR).

Graphical results of the statistical analysis for the three longest oscillation periods

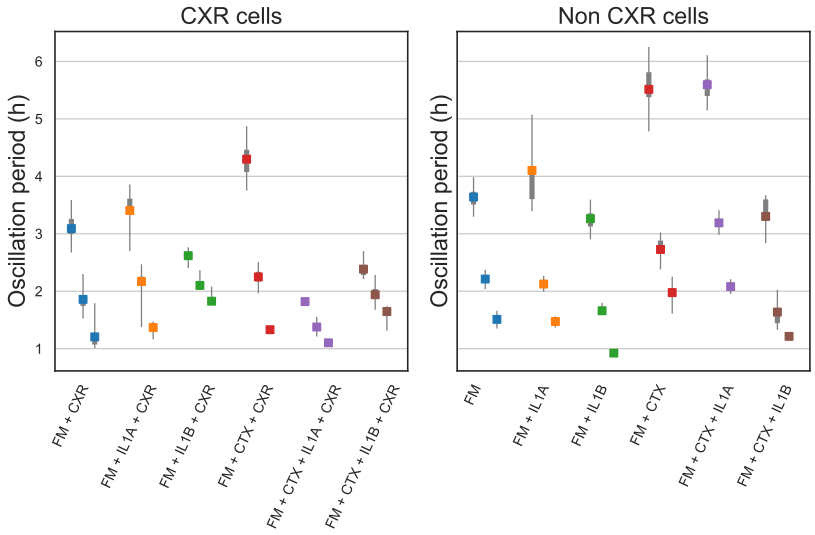
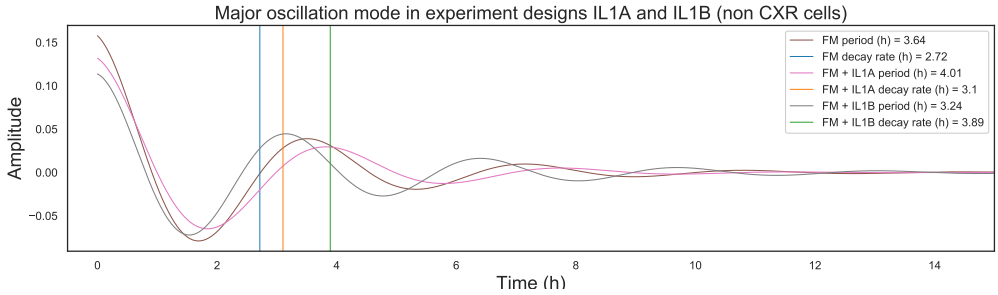


Figure 6.7: Boxplots for the three largest oscillation periods of each experiment. In the left plot there are the ones with Parental cell colonies, while in the right plot there are the Cetuximab resistant cell colonies (CXR).

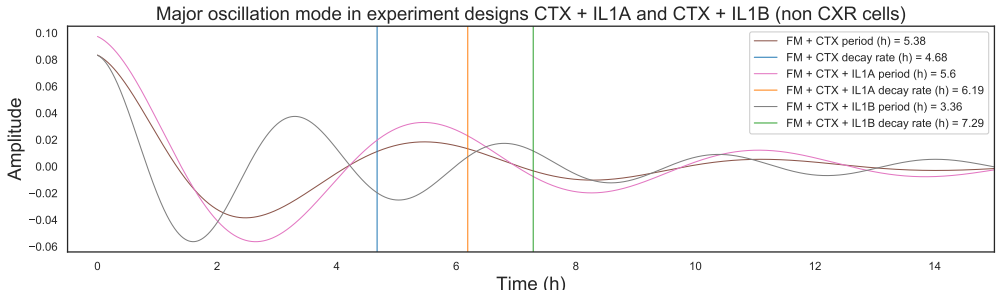
We want now to further refine our analysis of the results by focusing on the oscillation periods. We divide the results in the Cetuximab resistant group and in the Parental group. Looking at Fig. 6.7 we notice at first that the first class (highest periods) belongs to the Parental group, while the third class (lowest periods) belongs to the Cetuximab resistant group. Moreover, the second class appears to be distributed between these two groups.

First of all, the Cetuximab resistant group has lower oscillation periods for each experiment respect to the Parental one. Moreover, the presence of IL1B largely reduces the oscillation periods, both for the Parental and the Cetuximab resistant cells (green and brown boxplots), compared to the respective without that kind of interleukin. Conversely, the insertion of IL1A alone (orange boxplots) doesn't have such an evident effect, either in the Parental group and in the CXR one.

Let's look now at the experiment with the addition of Cetuximab. Focusing on the Parental cells, It is rather clear that with this treatment the periods increase, except for the case with IL1B (right plot, brown boxplot), where the effect of the Cetuximab treatment seems to be inhibited by the Interleukin. Shifting our attention to the Cetuximab resistant group, we notice that the effect of the anticancer drug is partially suppressed, especially when combined with both kinds of Interleukine. Remarkably, the only experiment with a visible response to the IL1A is the Full Medium of Cetuximab



(a) Oscillation for the experiments without Cetuximab treatment.



(b) Oscillation for the experiments with Cetuximab treatment.

Figure 6.8: Major oscillation mode for the Parental cells autocorrelation with different treatments.

resistant cells with Cetuximab treatment and IL1A (left plot, violet boxplot). In this case the oscillation periods are the lowest among all the experiments.

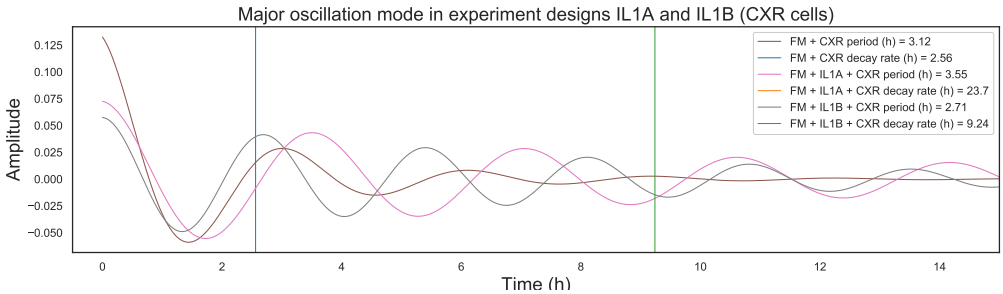
We can further consolidate our considerations looking at Figure 6.8 and 6.9.

Figure 6.8a shows the autocorrelation component with largest oscillation period for Parental cells without Cetuximab treatment. As it can be seen adding IL1A and IL1B interleukin lead to a moderate increase in decay time, while as said before in the IL1B treated colony the oscillation period decreases, while in the IL1A treated colony it increases by a moderate amount.

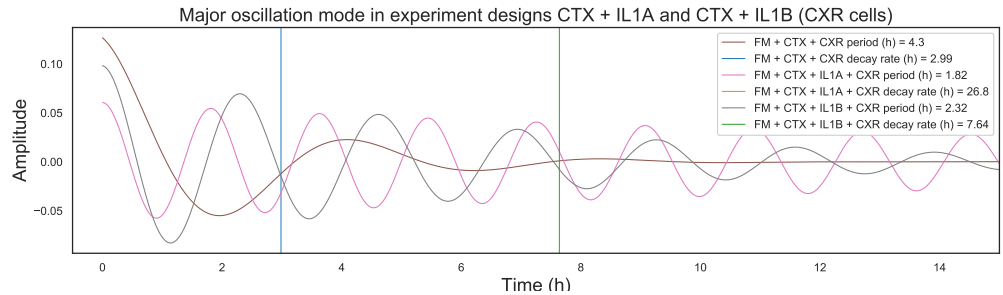
Moreover, in Figure 6.8b there is the autocorrelation component with largest oscillation period for Parental cells with Cetuximab treatment. In this case while the FM cell colony and the IL1A treated colony share a similar oscillation period, treating the cell colony with IL1B leads to a significantly smaller period.

Looking at Figure 6.9 we can notice that the decay times for experiments with IL1A and IL1B possess decay times far greater than in the Parental cell colonies. Moreover, for the Cetuximab resistant cells without treatment with the anticancer drug (Fig. 6.9a) the same considerations as for the analogous Parental cells can be done. Finally, although we see an increasing oscillation period for cell colonies treated with Cetuximab, the combination of the latter with IL1A or IL1B produces the opposite effect, namely a

decrease in the oscillation period.



(a) Oscillation for the experiments without Cetuximab treatment.



(b) Oscillation for the experiments with Cetuximab treatment.

Figure 6.9: Major oscillation mode for the Cetuximab resistant cells autocorrelation with different treatments.

Chapter 7

Conclusions

In the first part of this thesis we presented a singular property of the eigenvalue distribution of our toy model, namely that all the eigenvalues for the one particle case (Fig. 4.5a) distribute on the ellipse defined in Equation 4.11. For the N particles' case it is conjectured that all the eigenvalues lie on ellipses with same center as the one defined above with major and minor axes as some fractions of the ones in Eq. 4.11.

The results of the analysis performed on the biological data show that different stimuli lead to different kinds of oscillations in the NF-kB activity's autocorrelation. Moreover, performing some other analysis of the cell colonies growth like the colony-forming assay (Fig. **??**) one can merge the insights in order to find the stimulus reaction that inhibits

Colony-forming assay assay results for Parental and Cetuximab cell colonies

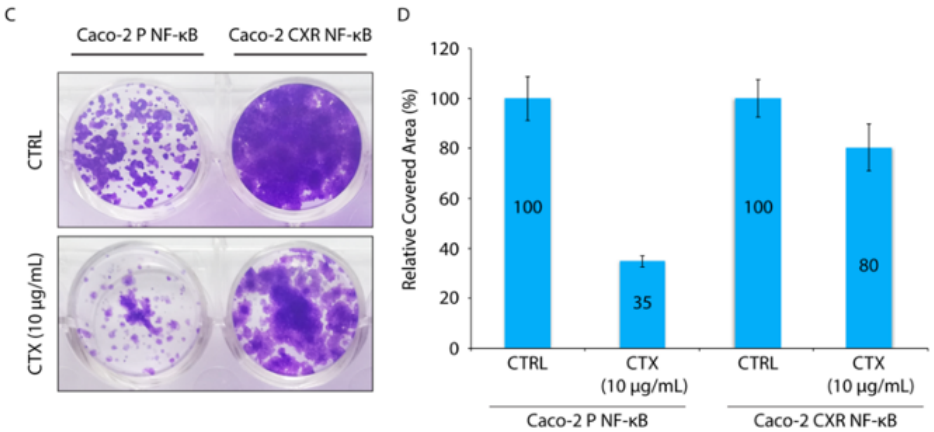


Figure 7.1: Colony-forming assay of the Parental cells (left) and Cetuximab resistant cells (right) following Cetuximab treatment. Cell were grown with or without Cetuximab for eight days before being fixed, stained with Crystal Violet and photographed. Quantification of the covered areas by ImageJ is provided.

the anomalous growth. Knowing which types of oscillations of the NF-kB activity are correlated to slower growth we can alter the biochemical pathway by other means, namely modifying the gene expression of the IkB (the NF-kB inhibitors) proteins, hence obtaining an oscillating cycle with the right properties.

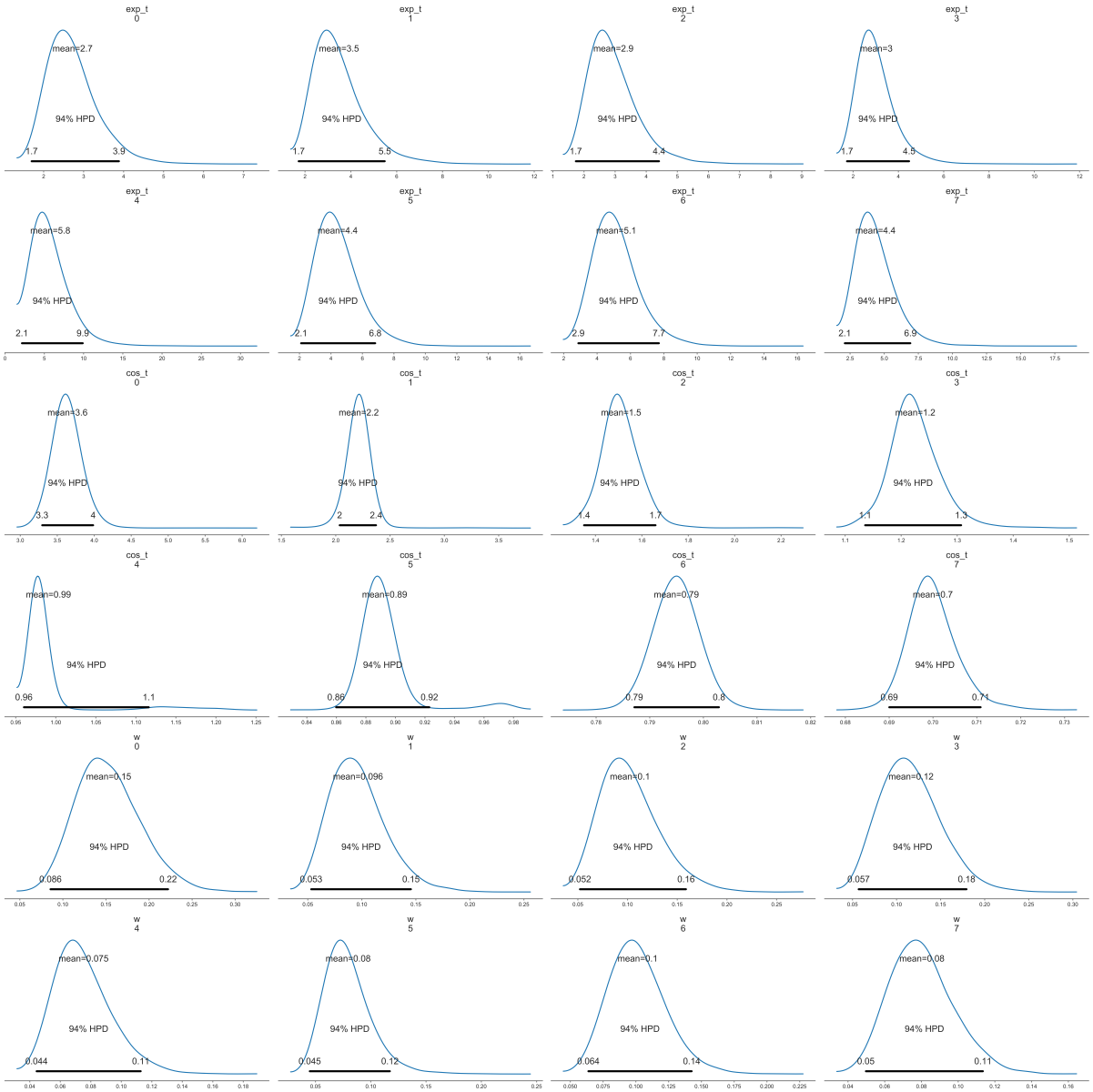
Since it is obviously too expensive to run actual experiments for each different biological configuration of the system, some simulation techniques need to be employed to optimize the selection of the experiments, in order to choose those with biologically relevant outcomes. In fact, using the information of the biochemical pathway, we can build and solve a Chemical Master Equation of the actual system. Following the same route of the one in Chapters 4 and 5 we can possibly extract the oscillating modes to be compared to those found with the present analysis. This could lead to new ways to exploit the knowledge of the biological pathway in order to control the population growth of colorectal cancer cells.

Appendices

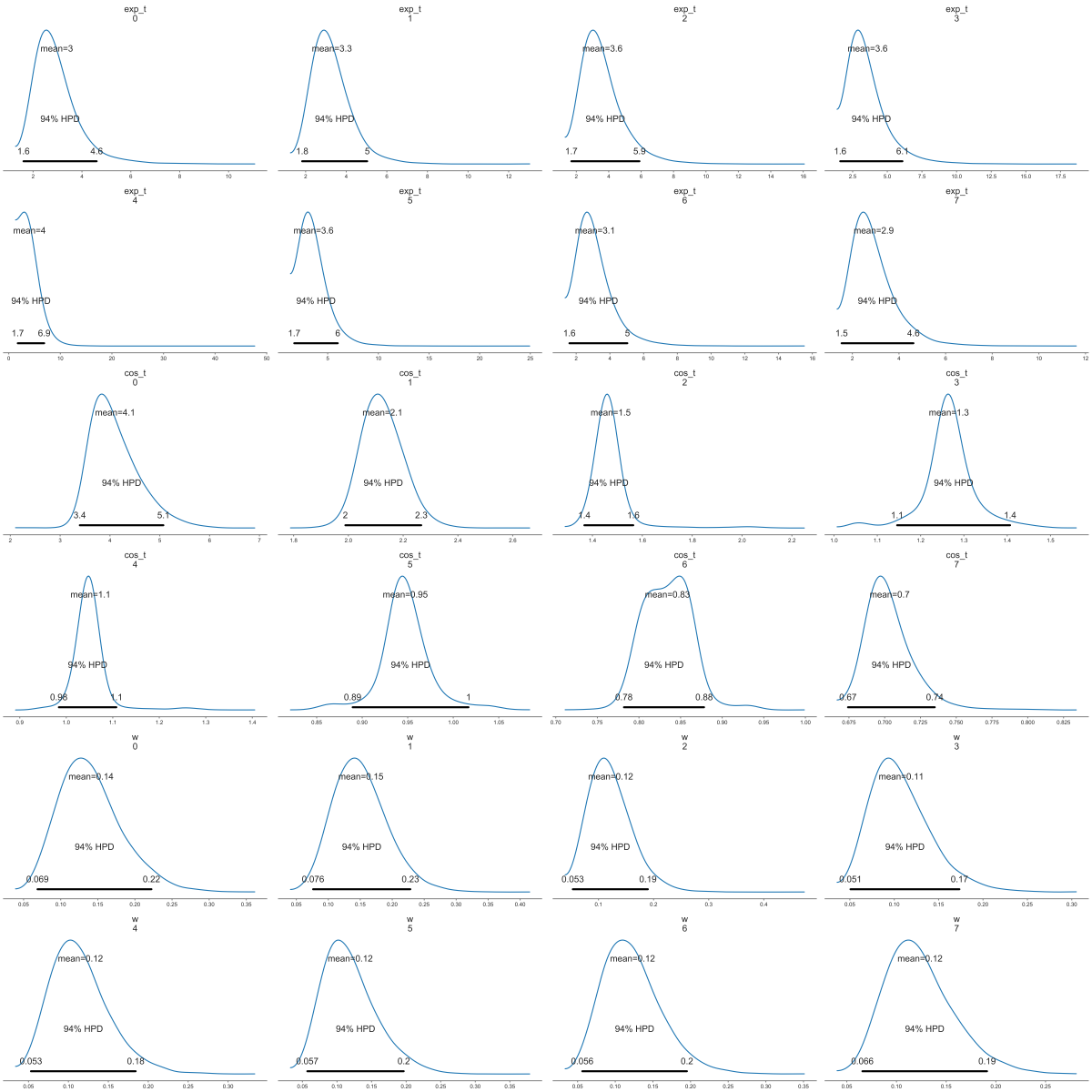
Appendix A

Posterior distribution plots

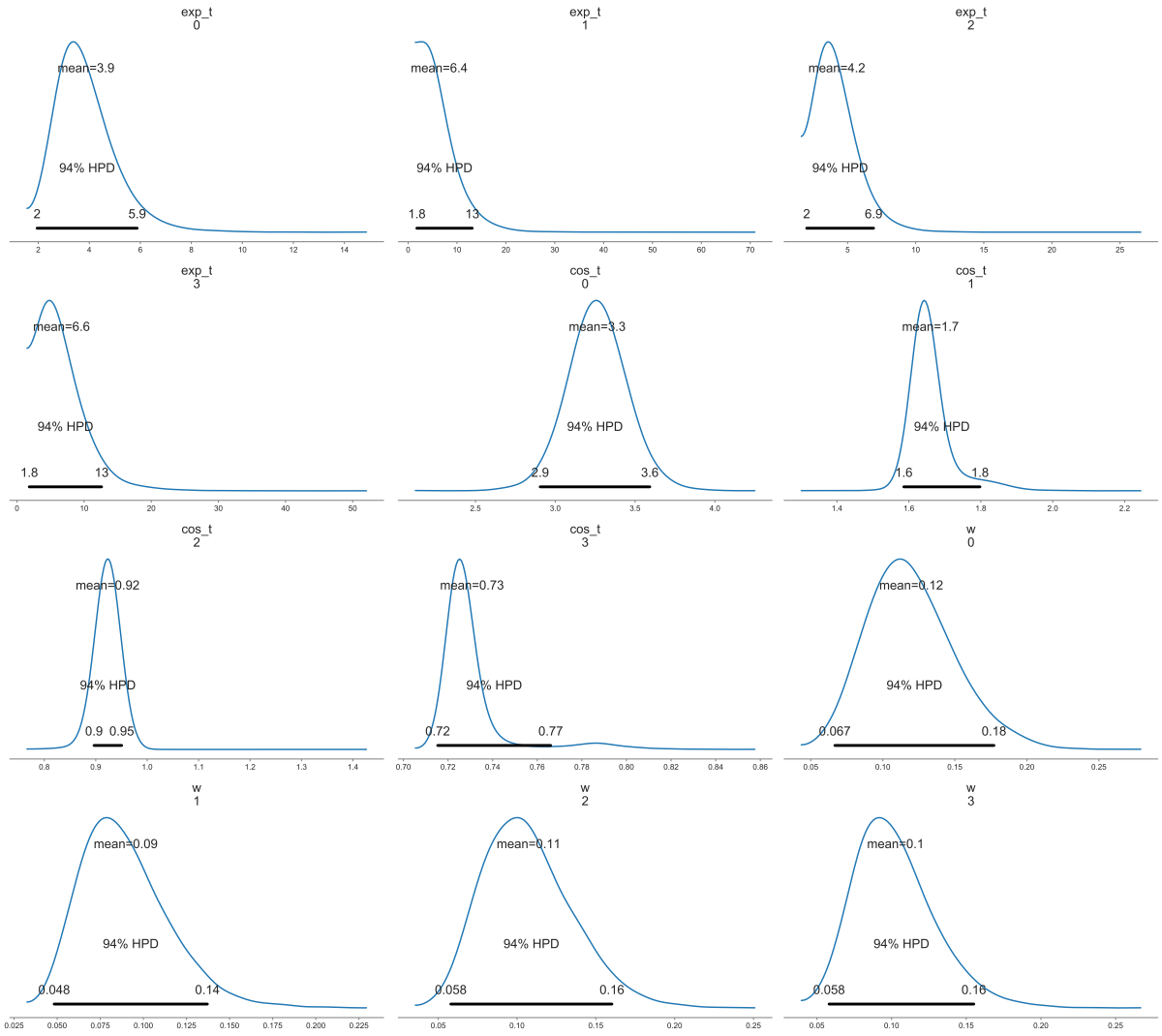
A.1 Full Medium experiment



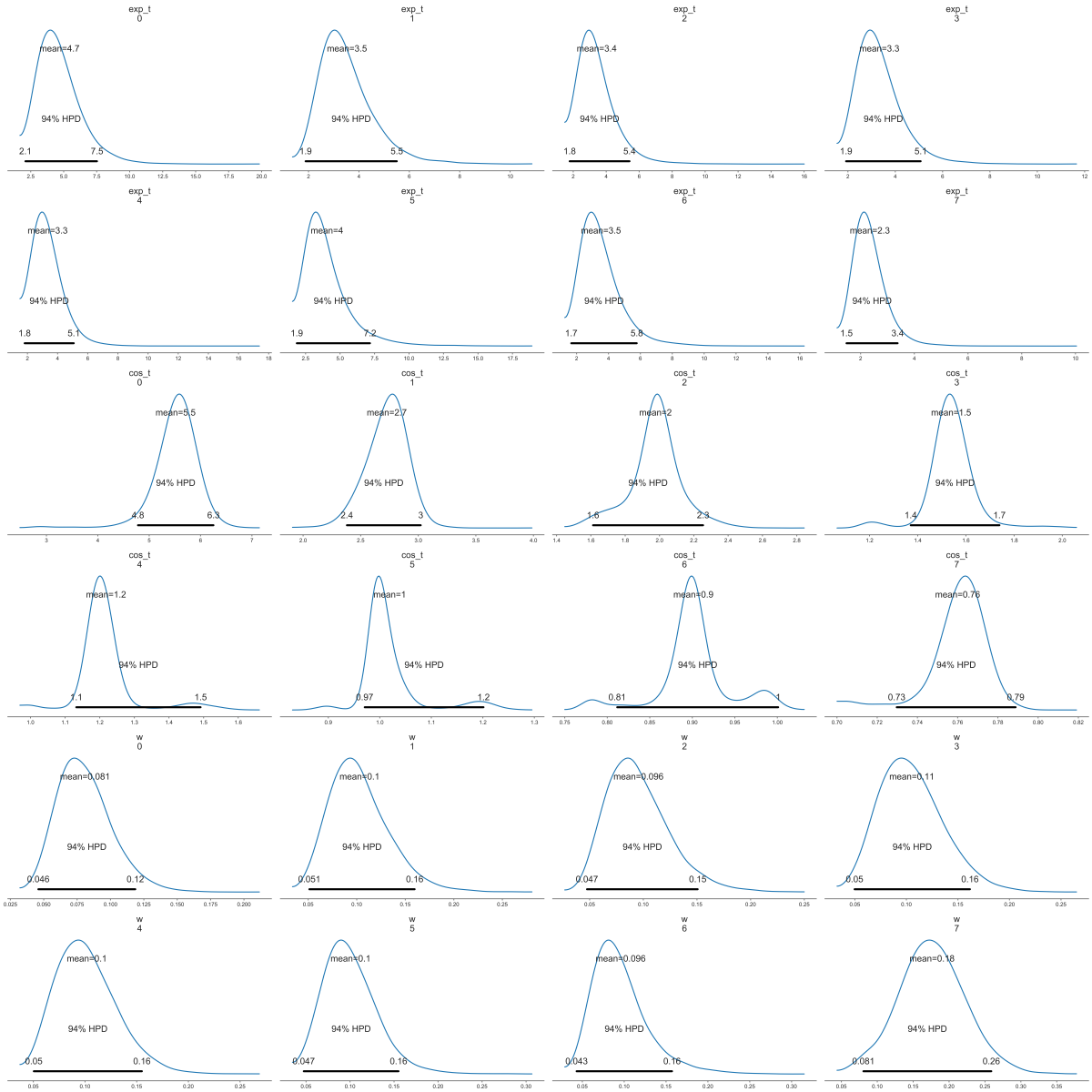
A.2 Full Medium + IL1A experiment



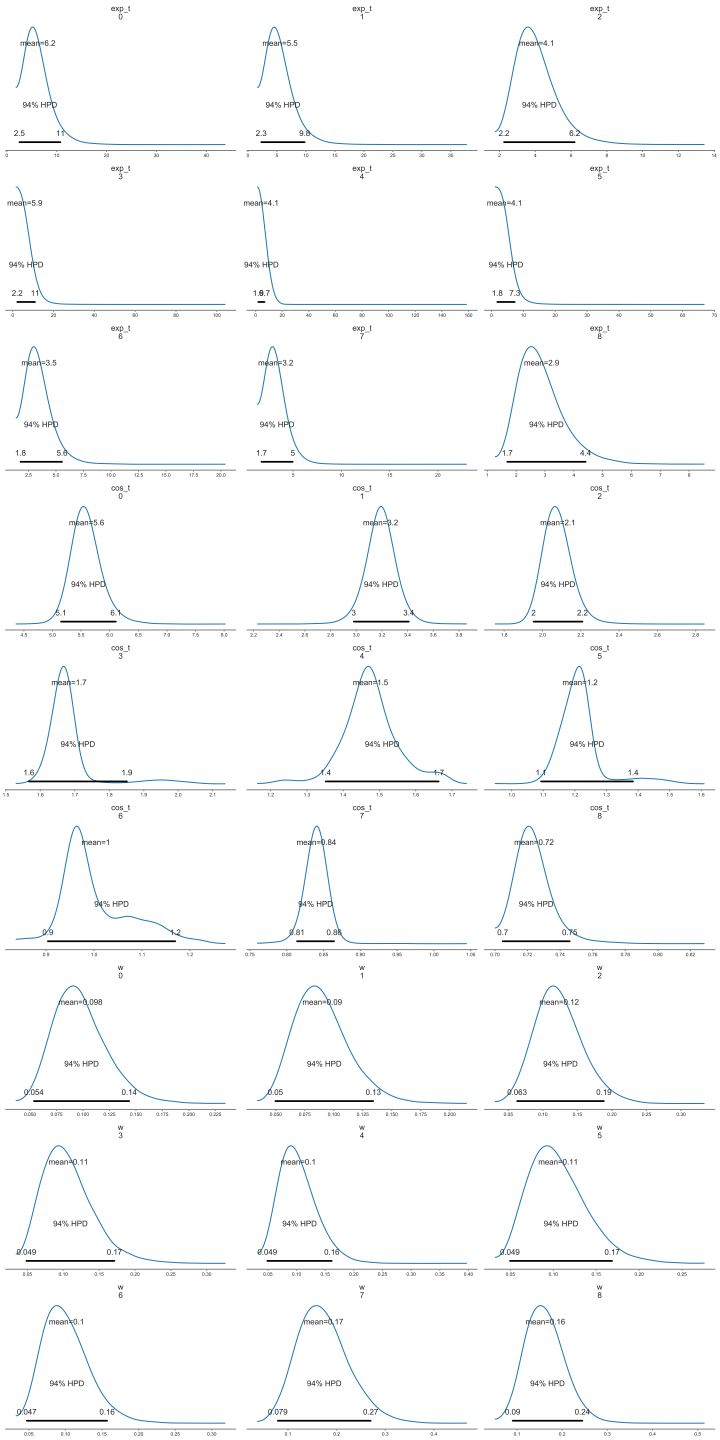
A.3 Full Medium + IL1B experiment



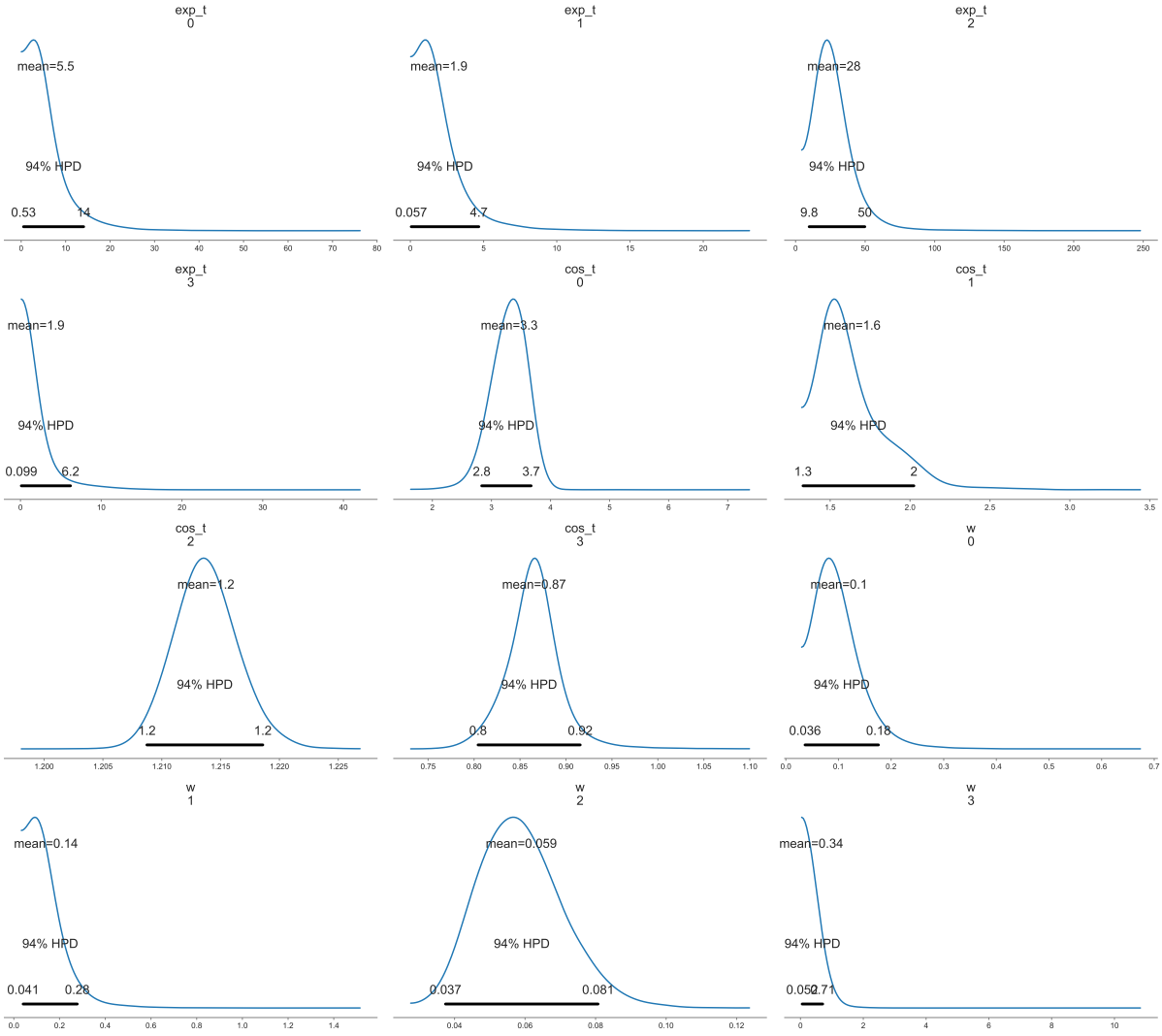
A.4 Full Medium + Cetuximab experiment



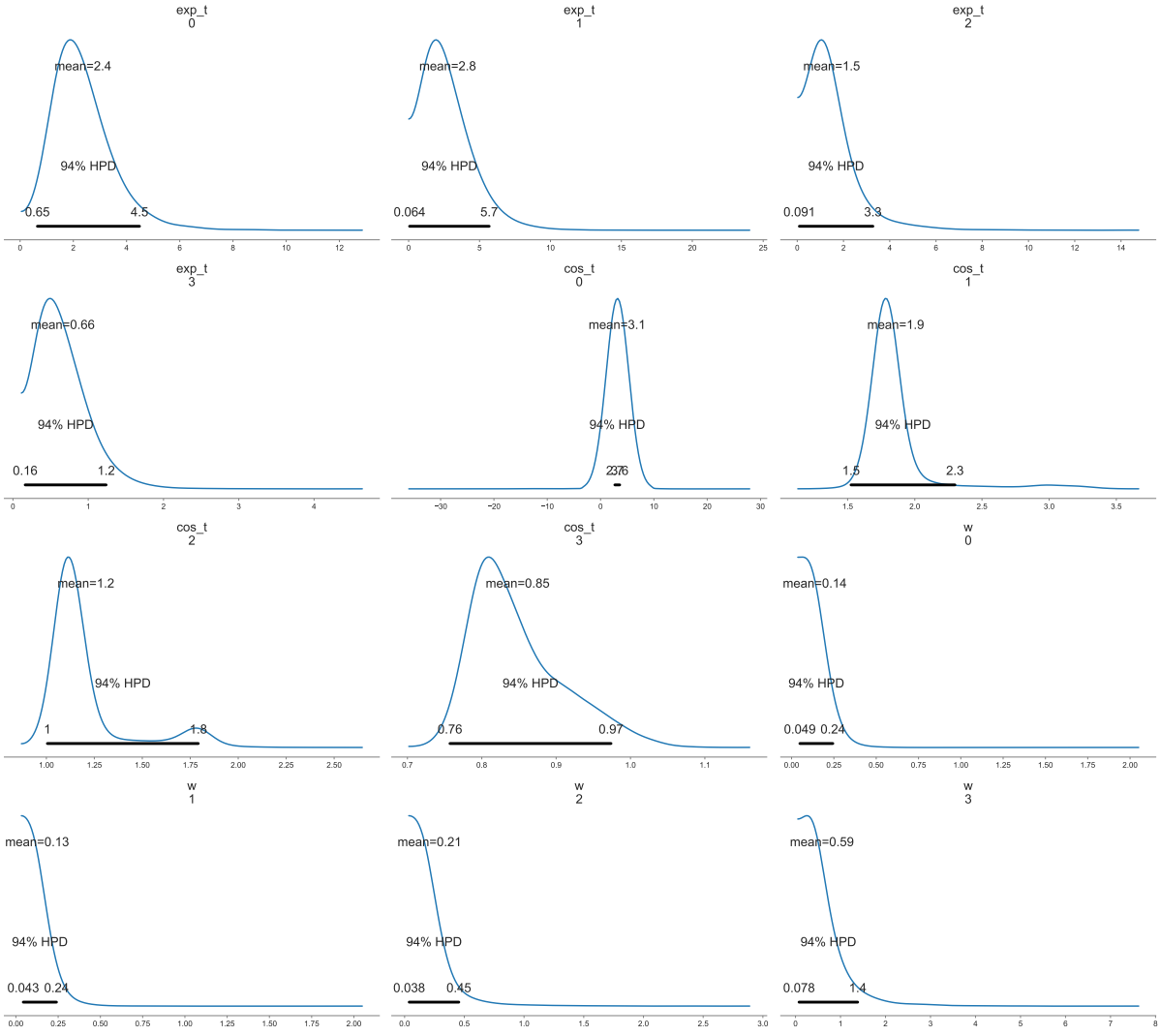
A.5 Full Medium + Cetuximab + IL1A experiment



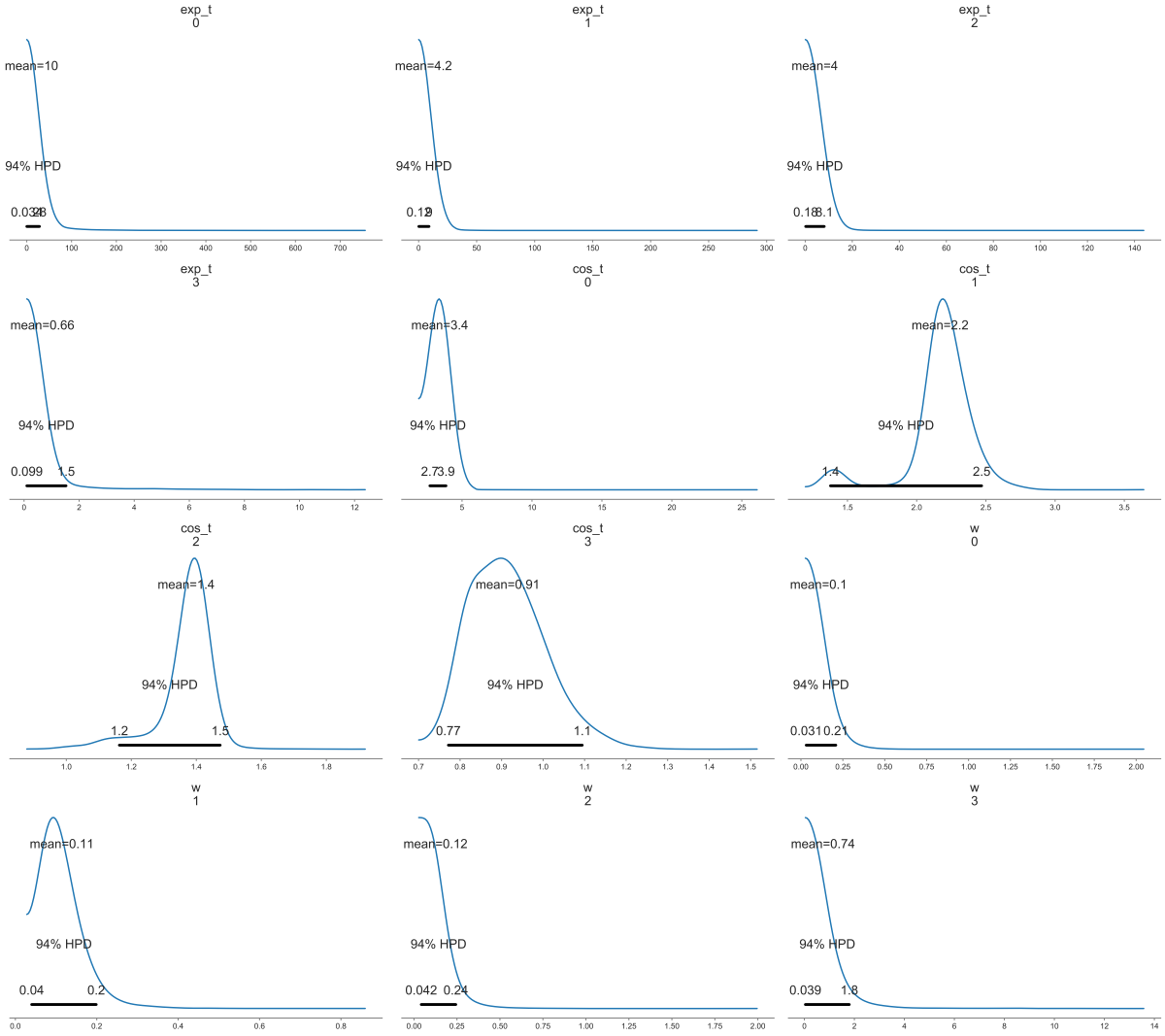
A.6 Full Medium + Cetuximab + IL1B experiment



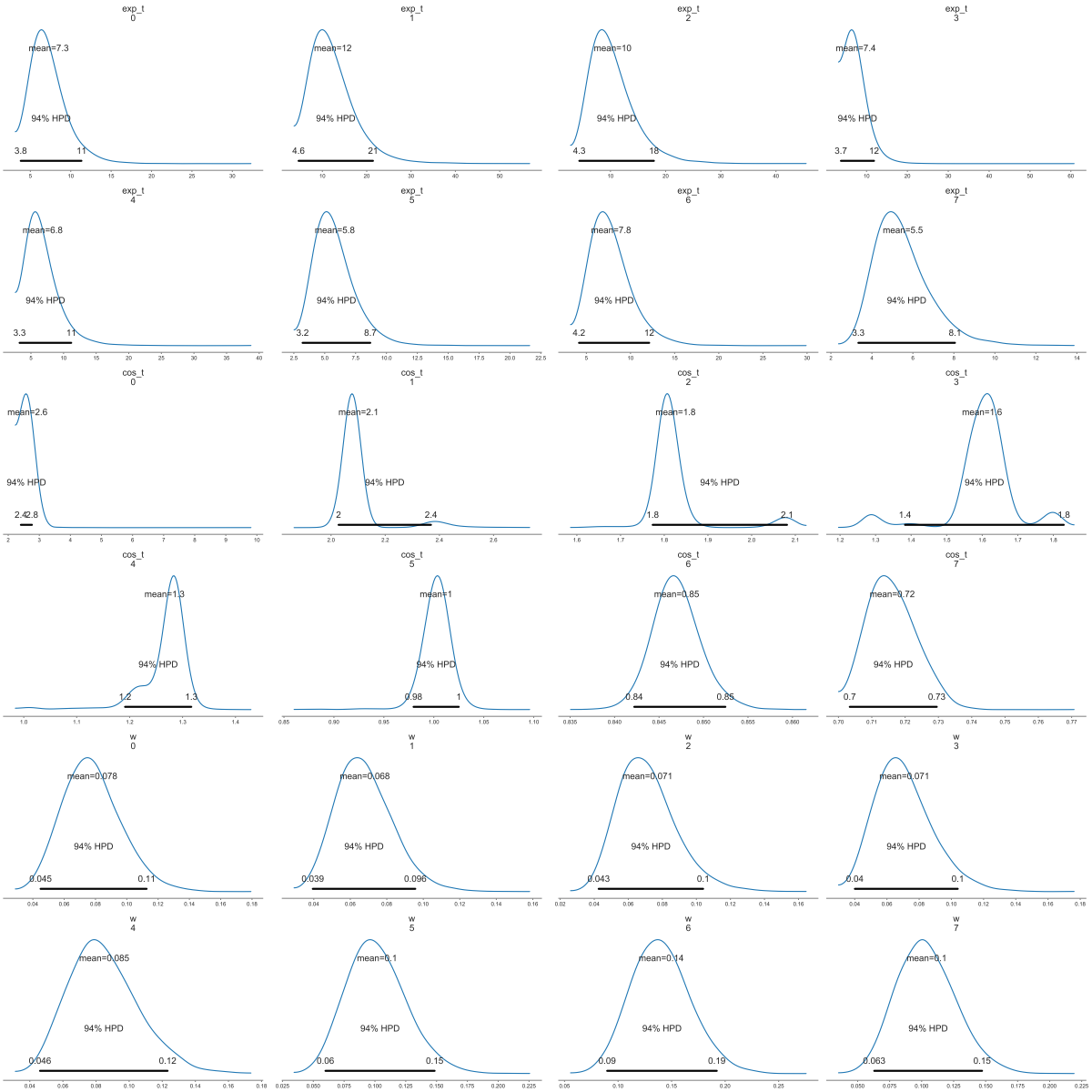
A.7 Cetuximab resistant Full Medium experiment



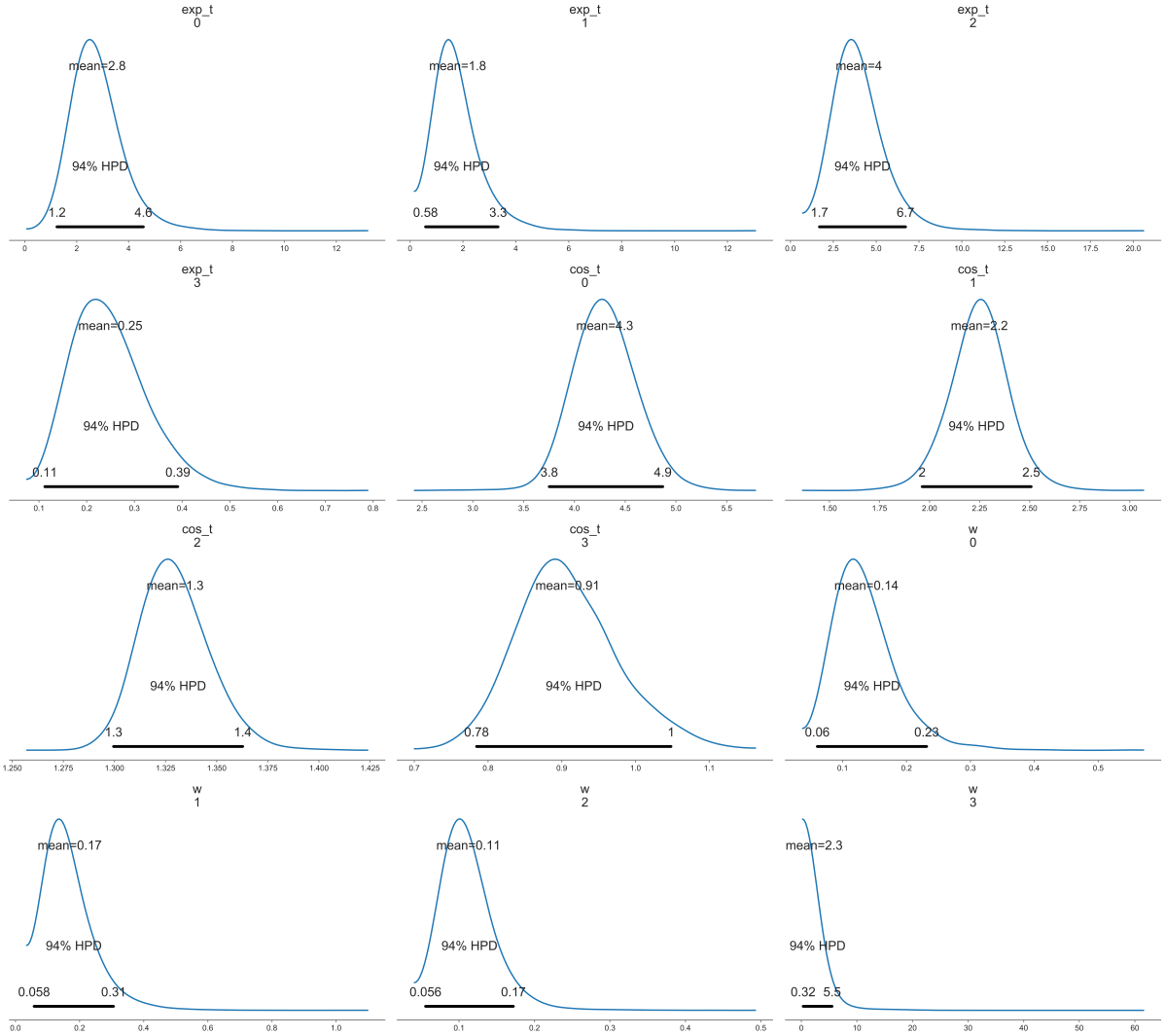
A.8 Cetuximab resistant Full Medium + IL1A experiment



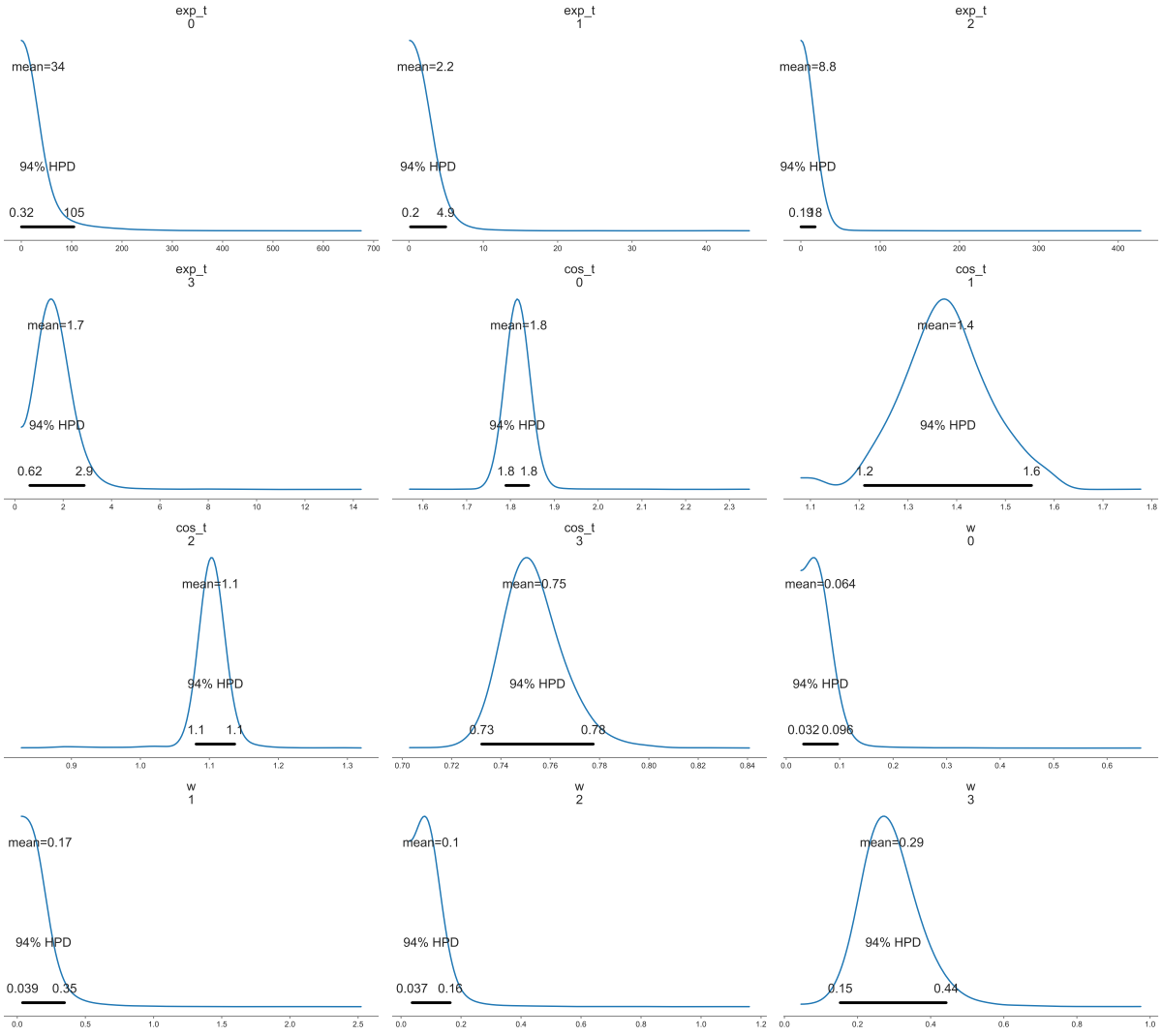
A.9 Cetuximab resistant Full Medium + IL1B experiment



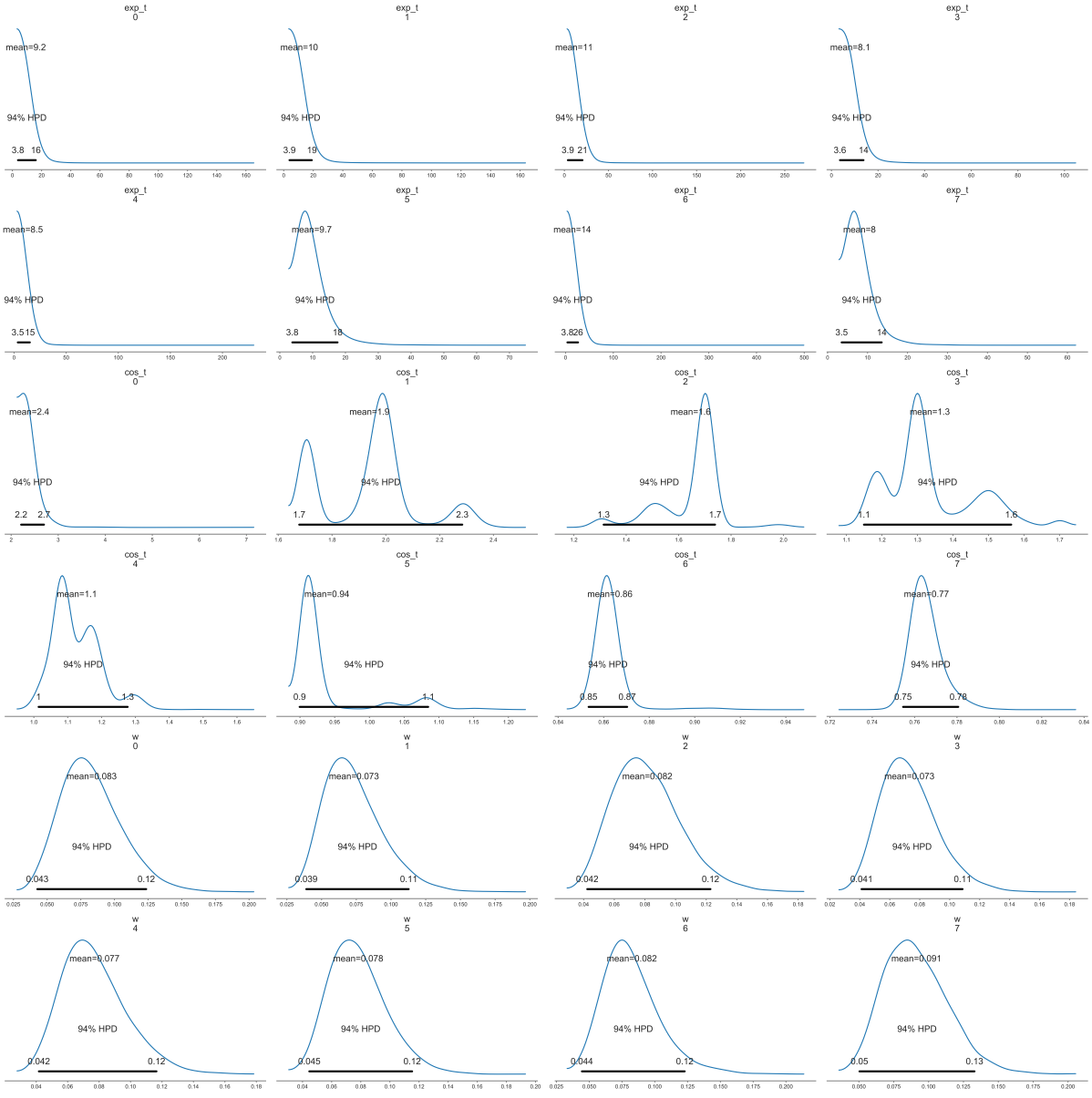
A.10 Cetuximab resistant Full Medium + Cetuximab experiment



A.11 Cetuximab resistant Full Medium + Cetuximab + IL1A



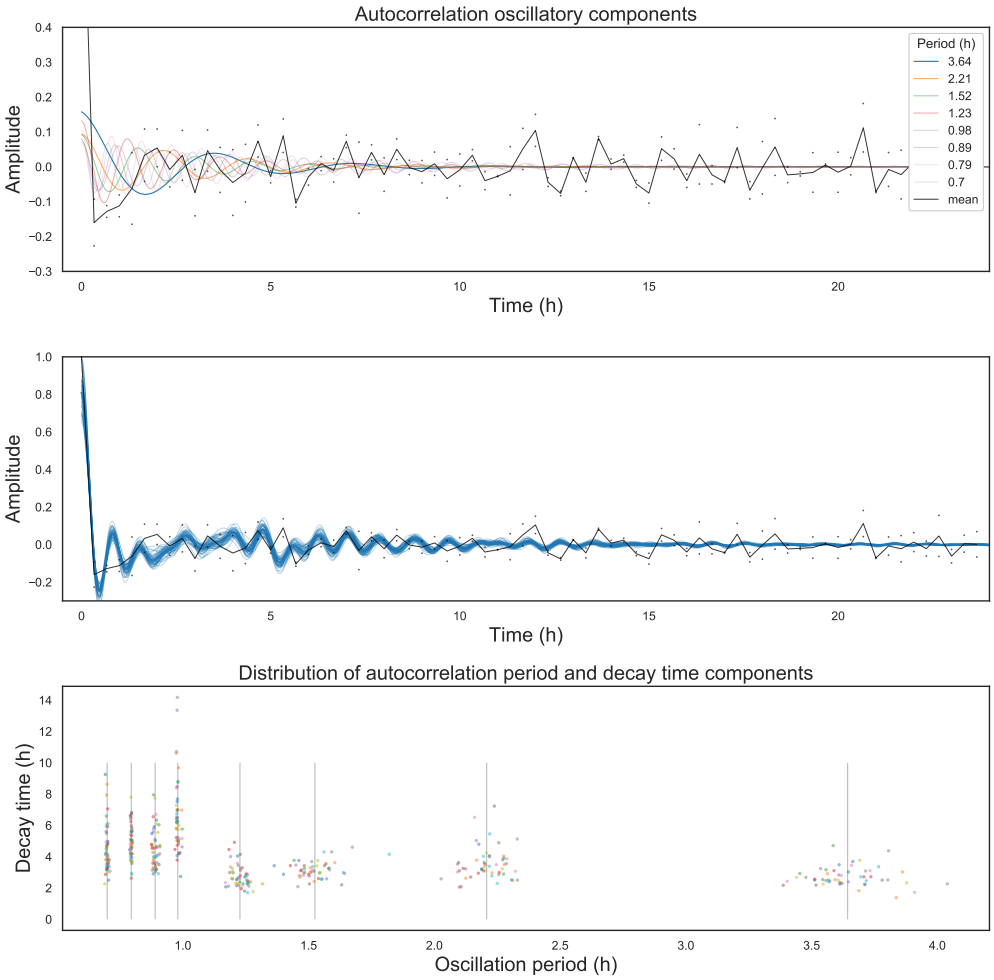
A.12 Cetuximab resistant Full Medium + Cetuximab + IL1B



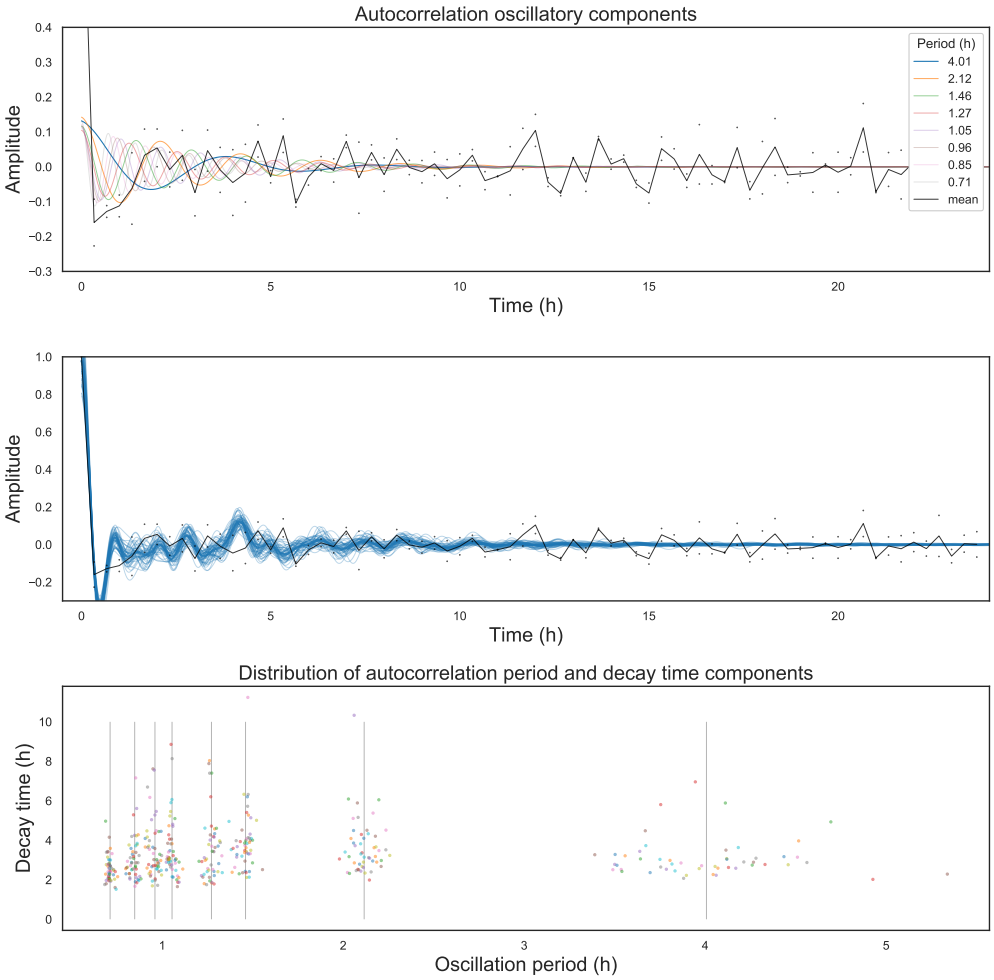
Appendix B

Samples of the parameters from the posterior distributions

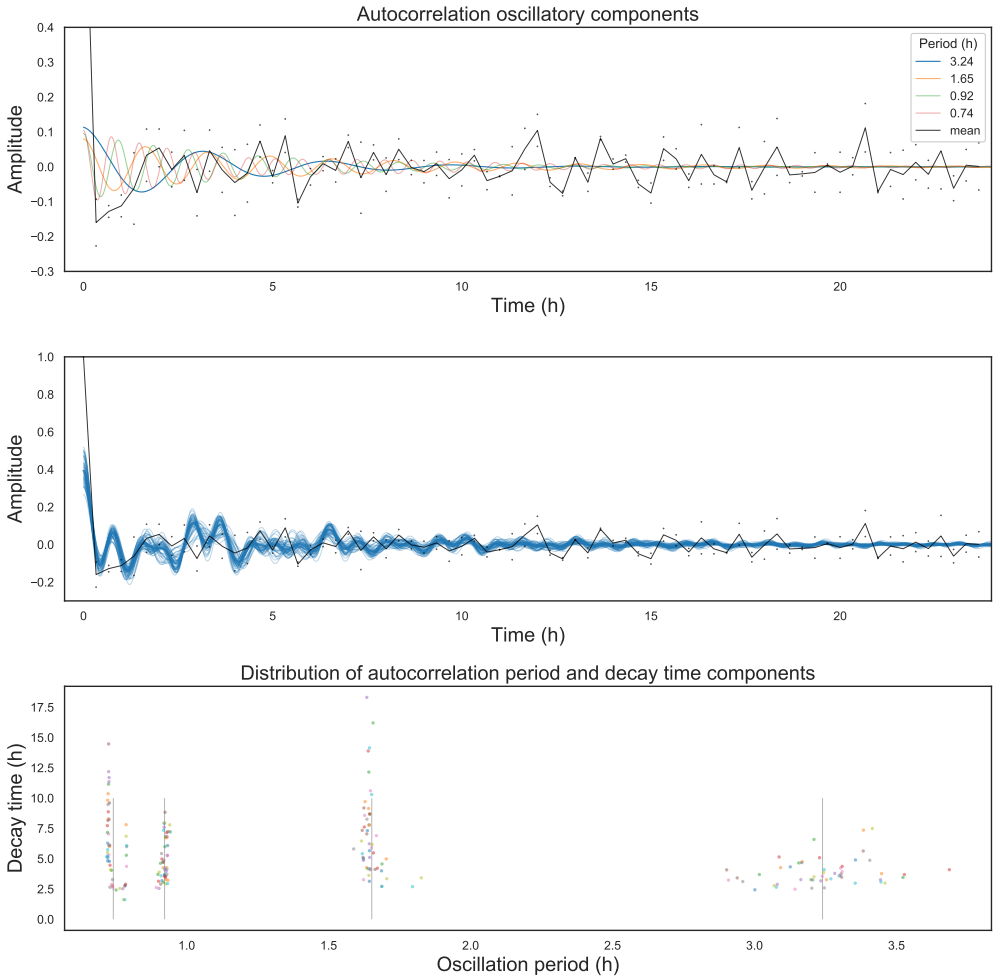
B.1 Full Medium experiment



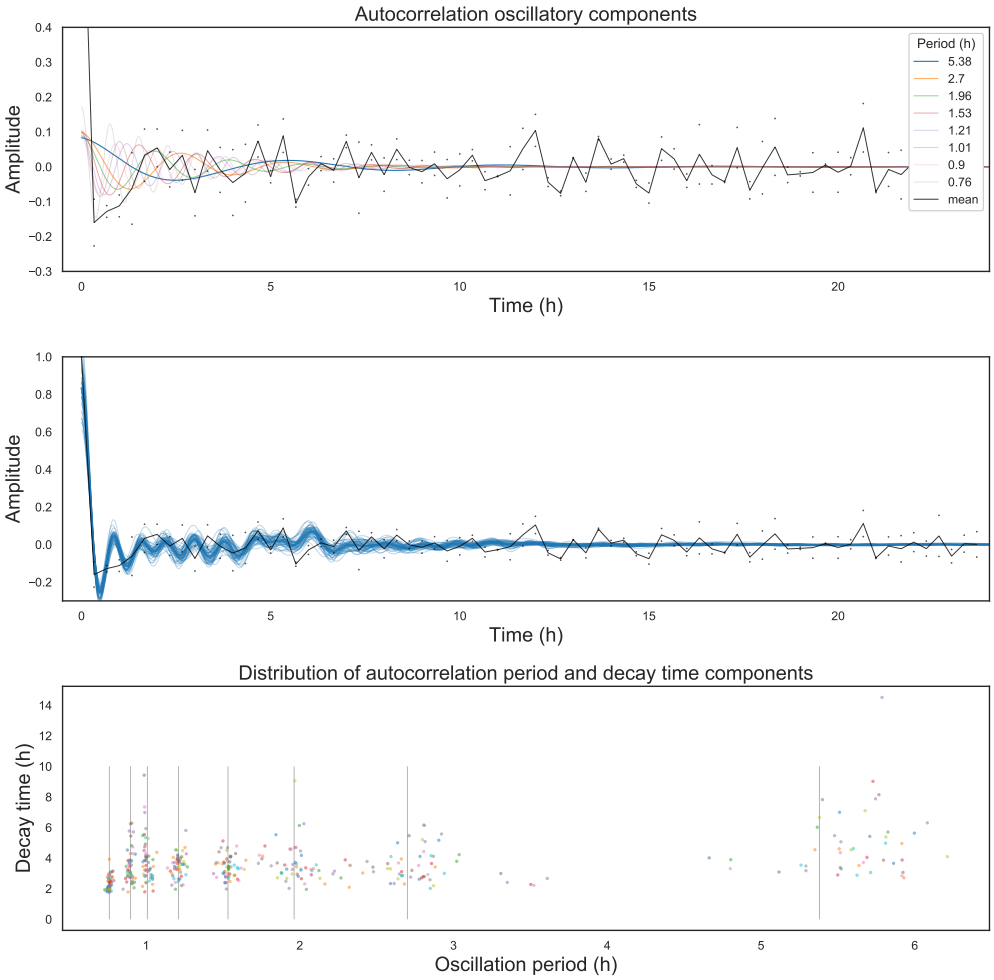
B.2 Full Medium + IL1A experiment



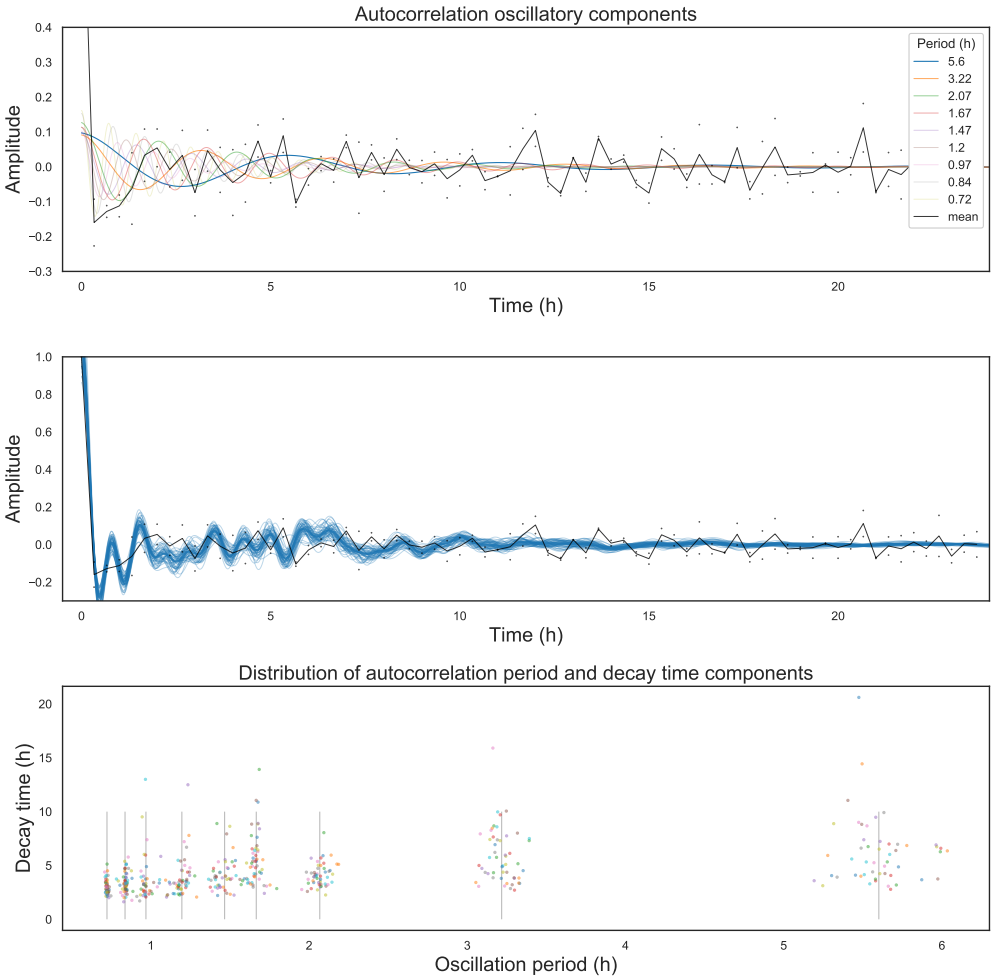
B.3 Full Medium + IL1B experiment



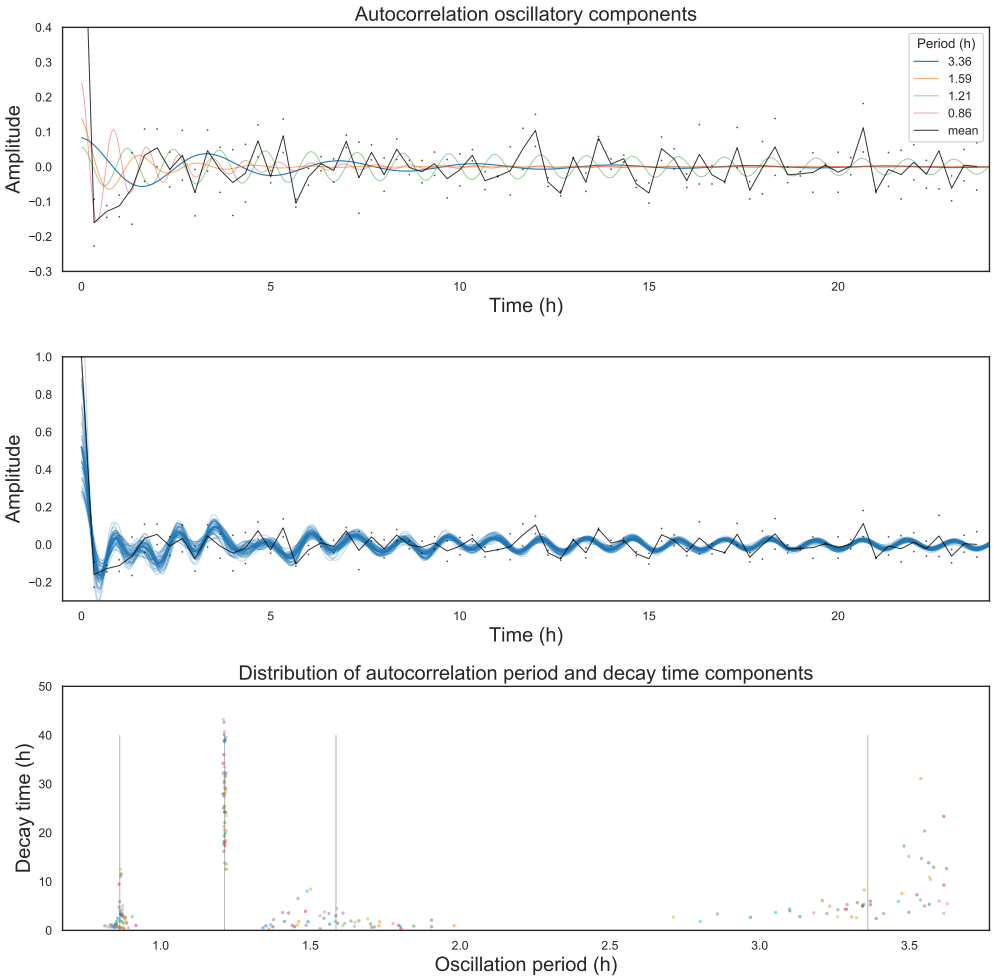
B.4 Full Medium + Cetuximab experiment



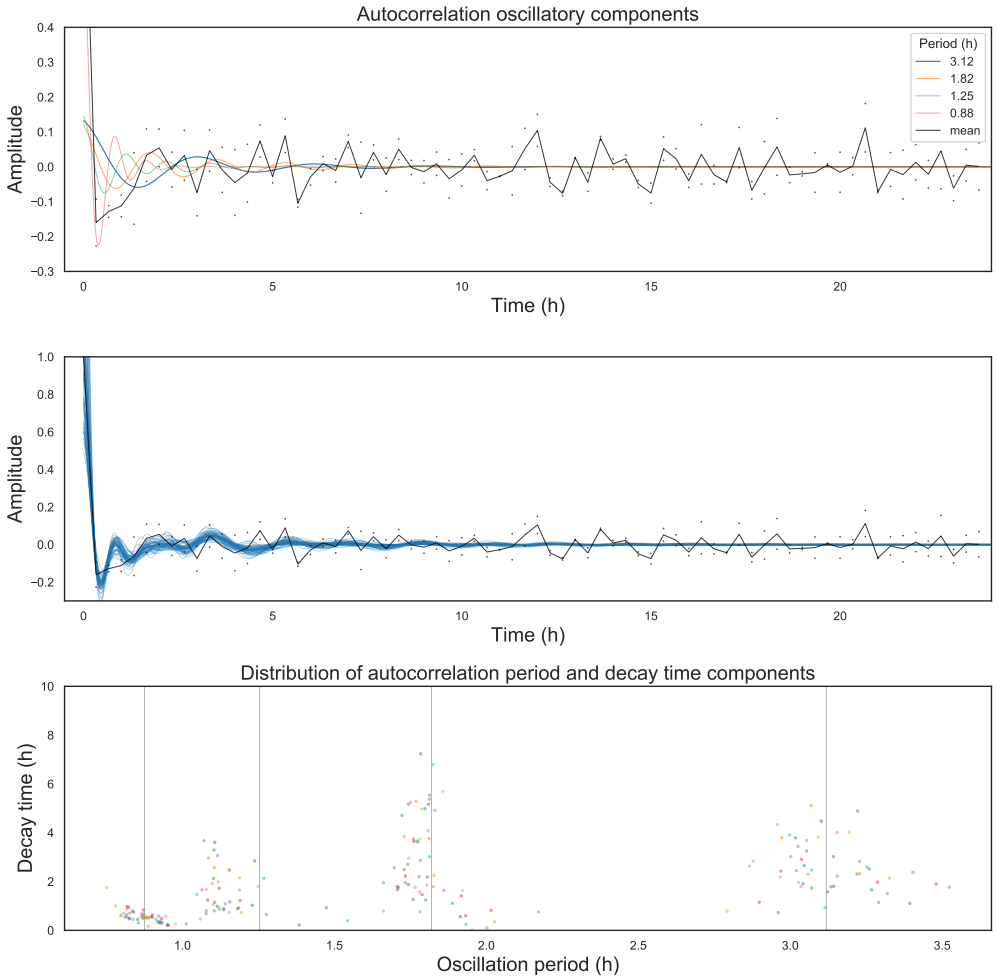
B.5 Full Medium + Cetuximab + IL1A experiment



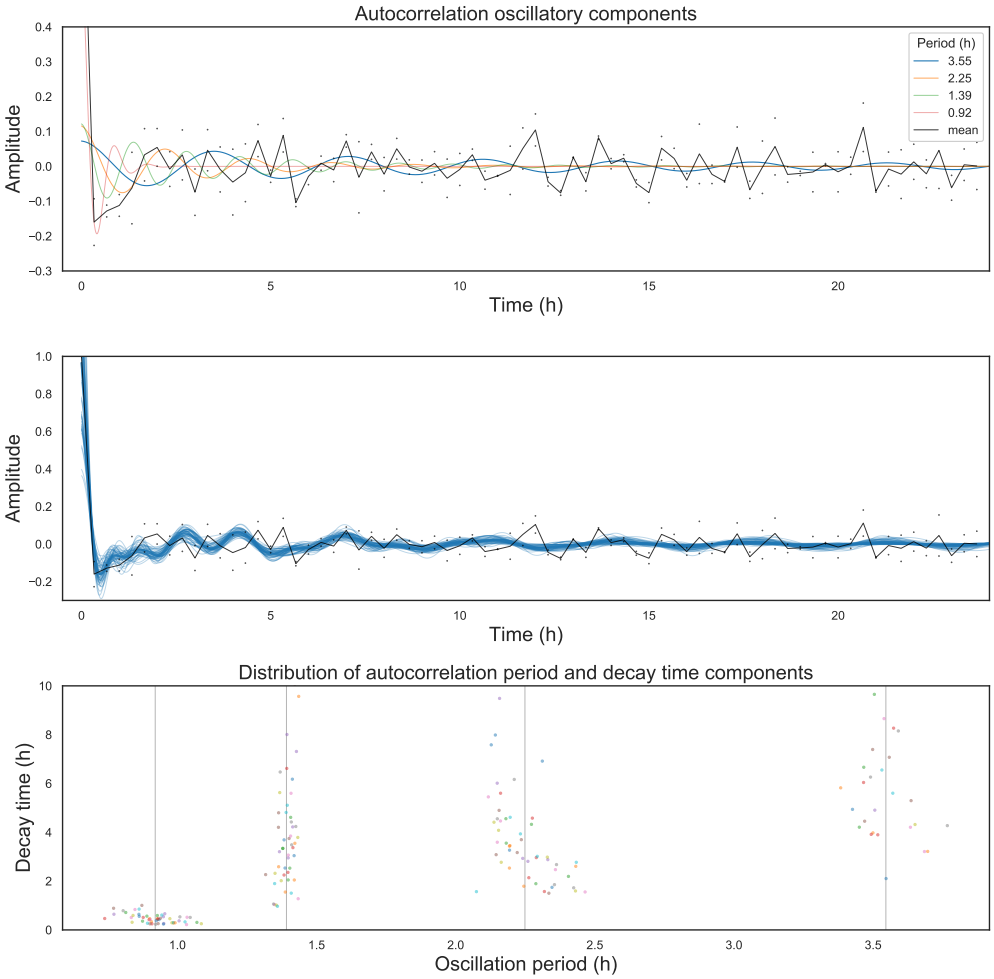
B.6 Full Medium + Cetuximab + IL1B experiment



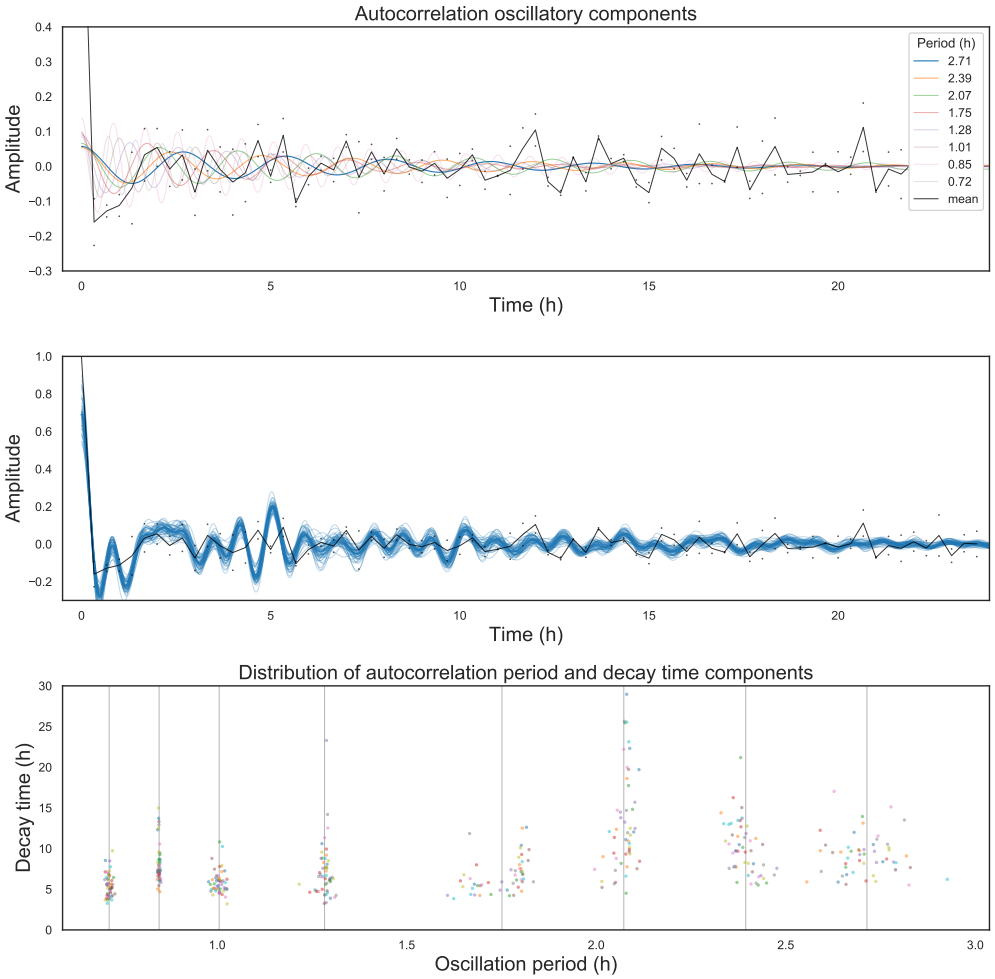
B.7 Cetuximab resistant Full Medium experiment



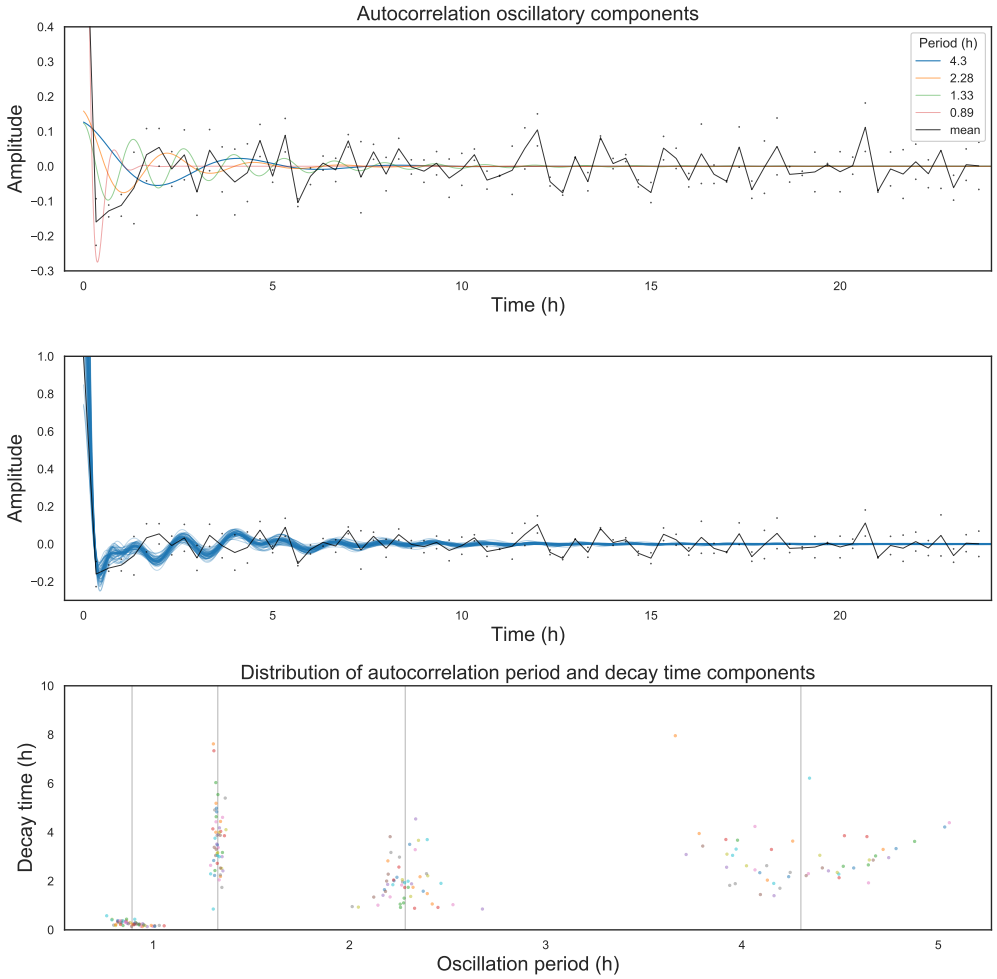
B.8 Cetuximab resistant Full Medium + IL1A experiment



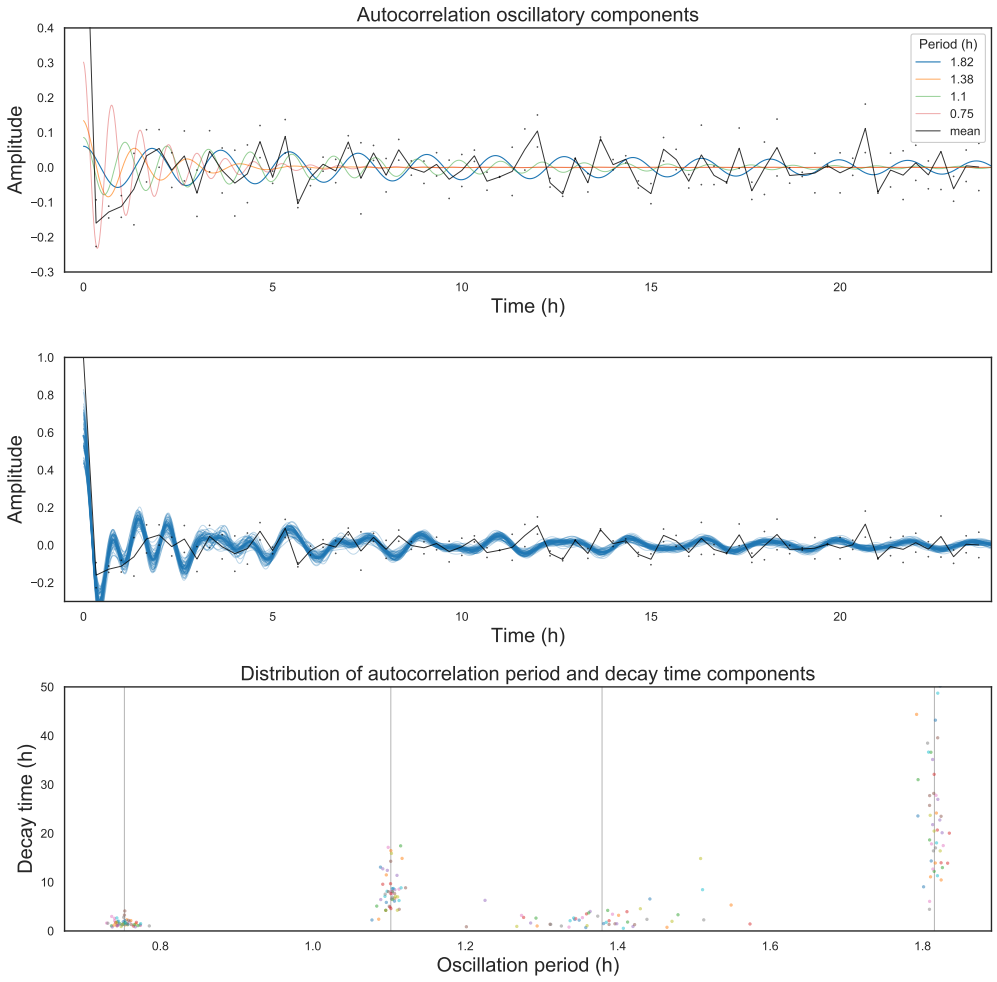
B.9 Cetuximab resistant Full Medium + IL1B experiment



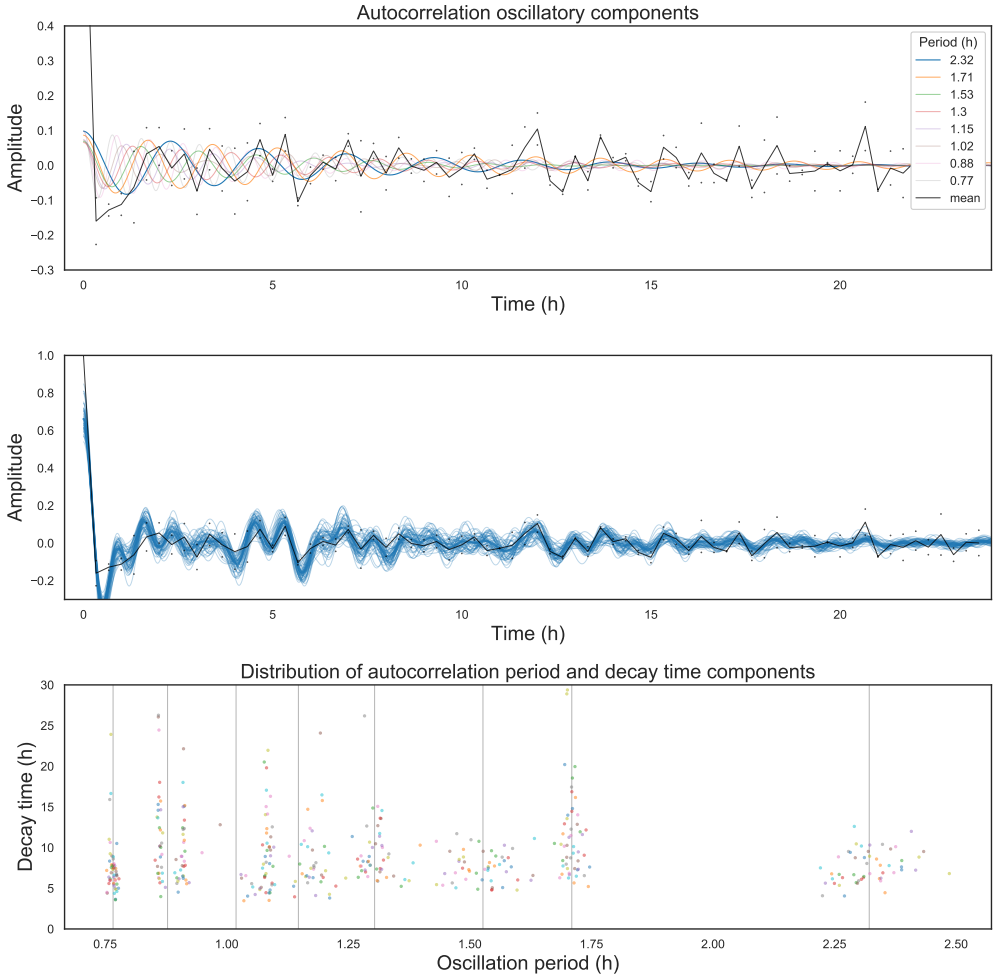
B.10 Cetuximab resistant Full Medium + Cetuximab experiment



B.11 Cetuximab resistant Full Medium + Cetuximab + IL1A



B.12 Cetuximab resistant Full Medium + Cetuximab + IL1B



Bibliography

[1] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, Elsevier B.V (1992)

[2] K. N. Dinh and R. B. Sidje. Understanding the finite state projection and related methods for solving the chemical master equation, *Phys. Biol.* 13 035003, 2016

[3] S. Geršgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk* 6 (1931)

[4] V. L. Girko. The elliptic law. *Theory Probab. Appl.* 30 (1986)

[5] D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A 188* 404-425 (1992)

[6] D. T. Gillespie. Stochastic Simulation of Chemical Kinetics. *Annu. Rev. Phys. Chem.* 2007.58:35-55

[7] M. Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. arXiv:1701.02434v2 *stat.ME*, 2018

[8] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1, Number 3, pp. 515–533 (2006)

[9] S. P. Brooks, A. Gelman. General Methods for Monitoring Convergence of Iterative Simulations. *American Statistical Association, Institute of Mathematical Statistics, and Interjiace Foundation of North America Journal of Computational and Graphical Statistics*, Volume 7, Number 4, Pages 434455 (1998)

[10] R.J. Steele, A. Raftery. Performance of Bayesian model selection criteria for Gaussian mixture models. *Frontiers of Statistical Decision Making and Bayesian Analysis.* 2010

[11] G. Celeux, F. Forbes, C.P. Robert, D.M. Titterington. Deviance information criteria for missing data models. *Bayesian analysis.* 2006

[12] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika.* 1995

[13] A. Hoffmann, A. Levchenko, M. L. Scott, D. Baltimore. The IkB-NF-kB signaling module: Temporal control and selective gene activation. *Science (80-.).* (2002) doi:10.1126/science.1071914.

[14] W. S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* (1979) 74 (368): 829-836 1979

Acknowledgements

First of all I would like to thank my thesis supervisor, teacher and friend Enrico Giampieri. He showed me in many ways where the dedication can lead.

I am deeply grateful to my parents and my whole family, for their continuous support throughout my journey.

I would like to thank my Sophia, she has always been there for me.

To my Apulian friends and to my Via del Porto's friends, thanks for everything they gave to me.

I would also thank my new housemates, who welcomed me in my new house, and my martial arts friends, who sometimes helped me to relieve my pain.

Last but not least, my special thanks go to Buba, who enlightened me with his wisdom.