Stochastic modelling of transcriptional regulation with applications to circadian genes

by

Silvia Calderazzo

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics, University of Warwick

September 2016

Contents

Acknowledgments

I first wish to thank my supervisor Prof. Bärbel Finkenstädt for advice on the research work and motivation during the last years, and for her contribution to the improvement of the present manuscript through helpful feedback and suggestions.

I am also grateful to I. Carré and M. Hastings and his group, for kindly providing the experimental data motivating this work, for explaining the experimental procedures, and for outlining the relevant biological processes and questions.

Moreover, I wish to thank to the participants to the weekly group meetings, for valuable insights and discussions.

This work has been funded by the Department of Statistics at the University of Warwick, and by the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/K503204/1.

Finally, special thanks go to my family and friends. To Panayiota, Fiona, Kirsty, Pier and Simone, 'travel-mates' of this experience, for solidarity in the bad days, and shared happiness in the good ones.

To Maria, for constant encouragement and trust in my capabilities, and to my parents, for their infinitely patient and loving support.

To Licia, Daniela, Illary, Elisabetta and Carlotta, who, despite the physical distance and the busy schedules, have always found a spare moment to share the experiences of our 'new' lives.

Declarations

I hereby declare that this thesis contains my own original work, unless otherwise acknowledged and referenced.

I also declare that this thesis is submitted for consideration for the award of Doctor of Philosophy at the University of Warwick, and it has not been submitted for any other degree.

Part of the analyses included in this work are performed on experimental data kindly provided by the Carré lab. at the University of Warwick, and by the Hastings lab. at the MRC in Cambridge.

List of Abbreviations

ABRE	Abscisic acid regulated element
bZIP	Basic leucine zipper proteins
Ca^{++}	Calcium
CBS	Circadian clock associated-1 binding site
CLE	Chemical Langevin equation
CME	Chemical master equation
CRE	Calcium/cAMP-responsive element
CT	Circadian time
DNA	Deoxyribonucleic acid
E-Box	Enhancer box motif
EE	Evening element, which can be denoted as EE-1A, EE-2A, EE-3A or EE-4A depending on the number of As at the beginning of the sequence
EKBF	Extended Kalman-Bucy filter
EKF	Extended Kalman filter
GABA	γ-aminobutyric acid
HEX	Hexamer
HPDI	Highest posterior density interval
LHY	Late elongated hypocotyl
LNA	Linear noise approximation
luc	Luciferase
MCMC	Markov chain Monte Carlo
MEME	Multiple EM for motif elicitation
MRE	Macroscopic rate equation
mRNA	Messanger RNA
ODE	Ordinary differential equation
QSSA	Quasi steady-state assumption
RNA	Ribonucleic acid
RNAP	RNA polymerase

RNAPc	RNA polymerase complex (basal transcriptional complex)
SCN	Suprachiasmatic nucleus
SD	Standard deviation
SDE	Stochastic differential equation
SNF	Second order nonlinear filter
SSA	Stochastic simulation algorithm
TF	Transcription factor
TTFL	Transcriptional and translational feedback loop
VIP	Vasoactive intestinal peptide

Abstract

Circadian rhythms, i.e. rhythms exhibiting a cyclic behaviour with a period of approximately 24 hours, are present in the metabolism of most living organisms. The transcriptional processes, i.e. the processes associated with mRNA synthesis, critically contribute to their origination, and are responsible for most of the mechanisms which regulate gene expression levels in cells. Inhibition or activation of a putative transcriptionally regulated 'child' gene can be achieved via binding of proteins called transcription factors (TFs) to the gene promoter, a region of the DNA containing protein-specific binding sites.

In this work, we investigate modelling and inference approaches for different scenarios of circadian transcriptional regulation. We focus on a system which comprises two transcription factors and a regulated child gene. We first perform parameter inference in the context of state-space models on simulated data from a mechanistic stochastic model describing this scenario. Additionally, we investigate the effect of data aggregation across different cells, and derive the smoothing equations for a destructive sampling scenario.

In the second part of this work, we consider a situation in which an important regulator of a child gene has not been observed. We apply our model to mRNA expression levels of a subset of circadian genes of the *Arabidopsis Thaliana* model plant. Inference is in this case aimed at estimating both the model parameters and the unobserved transcription factor profile. We compare *a posteriori* the inferred transcription factor profiles with available time-series data for one important circadian regulator in the *Arabidopsis Thaliana*, namely late elongated hypocotyl (LHY), and identify similarities for a several genes known to belong to the central clock.

Finally, we focus on a scenario of transcriptional regulation which includes an auto-regulatory negative feedback loop. This modelling framework is motivated by the availability of spatio-temporal imaging data of genes belonging to the mammalian central clock in mice suprachiasmatic nucleus (SCN), and in particular here we focus on *Cry*1. We introduce a distributed delay to account for nuclear export, translation, protein complex formation, and nuclear import, of the molecular species involved. To perform inference, we develop a novel filtering algorithm that can be applied to any system with distributed delays. We finally apply the methodology to *Cry-luc* spatio-temporal data, and find that parameter estimates are spatially distributed, with a marked difference between central and peripheral SCN regions.

Outline

Since gene expression profiles have become available with the recent advancement of sequencing and imaging technologies, one of the main challenges that biologists and biostatisticians have had to face is to understand the unobserved complex network of interactions that causes the observed patterns of expression.

It is currently believed that genes are mainly regulated during transcription (Latchman, 2007, Chapter 4), by proteins resulting from transcription and translation of specific genes. We will refer to the regulated genes as 'child' genes while the regulatory proteins are called transcription factors (TFs). While experiments now allow determining whether a particular TF binds the promoter of a potentially regulated gene, understanding the actual *activity* of the TF is still an open issue, that is of crucial importance to understanding the behaviour of the full network.

This work focuses in particular on circadian dynamics, i.e. daily oscillatory patterns of expression arising in genes belonging to, or being regulated by, cellular circadian clocks. The circadian clock is a robust and self-sustained mechanism common to most living organisms, which has the function to optimise the metabolism in accordance with daily light and temperature changes (McClung, 2006; Harmer, 2009).

This thesis is divided into three parts. In the first part, comprising Chapters 1 and 2, we investigate modelling and inference for a general framework of stochastic transcriptional regulation of a child gene by two TFs. We recognise that real interactions may involve several genes and TFs. However, even the simplest setting, where only two TFs and one child gene are present, already implies several possible scenarios of transcriptional dynamics, which we will investigate.

In Chapter 1 we introduce the modelling approach, where we start from a description of the reaction network at a stochastic single-cell level, and then gradually move to approximate modelling approaches, available when different time-scales can be assumed for the network reactions, or a large number of molecules is involved. In this first part, we also investigate the assumptions required for data aggregation across different cells, a situation often encountered in real data scenarios. We conclude the first chapter by defining our model in the broader context of state-space models.

In Chapter 2 we address the issue of parameter estimation, and in particular we introduce the concepts of filtering and smoothing in state-space models, with a special focus on Kalman filtering methodologies. We review current approaches for the computation of the likelihood when a new unit from the same process is observed at each observation time-point, in a sampling methodology known as 'destructive sampling'. Such a sampling process is not uncommon in experiments, and also characterises the available *Arabidopsis Thaliana* data, the analysis of which is the focus of the second part of this work. Estimation results on simulated data are presented, and in particular we show that in a low measurement error and high frequency sampling scenario, it is possible to infer parameters describing the increase in transcriptional activity due to effect of the TF, as well as parameters related to the binding of the TFs to the promoter and their binding cooperativity.

In the second part, which includes Chapters 3 and 4, we focus on modelling transcriptional regulation of a subset of genes of the *Arabidopsis Thaliana* model plant, whose mRNA levels are available from a Nanostring experiment from the Carré lab. at Warwick.

In Chapter 3 we present the biological concepts related to gene regulation, its relevance to a particular class of genes (i.e. the core genes of the circadian clock in plants), and the data available from the Nanostring experiment. Moreover, we describe a set of additional experiments and results, which provide a deeper insight into the activity and binding properties of an important known regulator in the *Arabidopsis Thaliana*, namely late elongated hypocotyl (LHY). Finally, we propose three models of transcriptional regulation for the available data.

In Chapter 4 we first deal with the validation of inference on simulated data for the three modelling approaches introduced in Chapter 3, and then perform the real data analysis on the genes with rhythmic mRNA profiles in the Nanostring experiment. We apply a model which assumes one unobserved transcription factor as regulator of the child genes. We finally compare the inferred transcription factors of each child gene to the LHY protein time-series, to assess the degree of similarity, and we cluster the inferred unobserved child gene mRNA profiles, to identify a possible relationship between cluster of expression and selected characteristics of the genes related to LHY activity. We find that the inferred unobserved TF profile has a strong correlation with LHY protein time-series for a number genes known to belong to the *Arabidopsis Thaliana* central clock, namely ELF3, PRR9, CAB1, CCA1, TOC1, ELF4, and LUX. We also observe a possible correlation between cluster of expression and presence of binding sites in the promoter regions of the analysed genes.

In the third and last part of the thesis, comprising Chapters 5 and 6, we extend our approach to be applied to genes belonging to the central circadian clock in mammals, in particular in cells located in the suprachiasmatic nucleus (SCN) of mice. Regulatory dynamics include in this case a negative feedback loop, modelled by means of a distributed delay.

In Chapter 5, we provide a brief overview of the biological knowledge, and current modelling approaches, of the mammalian circadian clock in the SCN. We then introduce the proposed model for a sub-set of central clock genes, and propose a novel methodology which allows to perform filtering in stochastic dynamical systems comprising distributed delays.

In Chapter 6, we approach inference for both simulated data and *Cry*1-*luc* spatio-temporal observations across the SCN. We merge the inferential results from three independent experimental replicates, by means of a hierarchical Bayesian meta-analytic model, and, finally, investigate patterns of spatial variation of the parameters across the SCN. We observe, among other results, a decreasing mean trend in both intrinsic noise and standard deviation of the distributed delay as we move from central towards more peripheral locations of the SCN. The spatial distribution of the parameters related to intrinsic noise and to the variability of the delay distribution, which may have a role in the underlying signalling dynamics, highlights the relevance of stochastic modelling of transcriptional dynamics in this scenario, and motivates the adoption of a distributed delay.

Finally, in the conclusions, we summarise our main findings, discuss possible limitations and outline directions for future developments of our work.

Part I

Modelling stochastic transcriptional regulation

Chapter 1

Modelling transcriptional regulation by two transcription factors

Stochasticity is believed to be a relevant characteristic of gene expression (Elowitz et al., 2002). Differences in the expression levels of a single gene can be both cell-specific, i.e. due to the so-called extrinsic noise, and due to the random interactions of particles, i.e. due to what is known as intrinsic noise (Elowitz et al., 2002). Measurement error is a further relevant source of noise.

Depending on the level of detail required, different modelling and simulation techniques can be employed in order to reproduce the system under study, and the relevant source of noise. The possible size of resolution at which the system is studied can be broadly divided into three classes: the microscopic, mesoscopic and macroscopic level (Lachowicz, 2011).

Data at a microscopic level usually involve very few molecules, approximately less than 10, and therefore exhibit a high level of intrinsic stochasticity. In this context an exact stochastic description of the system is usually required. Since the present state of the system conveys all the necessary information to describe probabilistically its future evolution, the microscopic dynamics can be rigorously modelled with a Markov jump process (Anderson and Kurtz, 2011). Despite the fact that its transition probabilities are usually intractable, it is possible perform exact simulation with the stochastic simulation algorithm (SSA) (Gillespie, 1977; Doob, 1945).

The mesoscopic level involves a higher number of molecules, such that intrinsic stochasticity still plays a role in the system, but a discrete model of number of molecules and reactions is no longer required. Essentially, we move from a Markov process in continuous time - discrete state-space, to a continuous time - continuous state-space one. In this context, two widely applied approximate simulation and modelling approaches are the diffusion approximation, or chemical Langevin equation, (CLE) (Gillespie, 2000; Golightly and Wilkinson, 2005; Wilkinson, 2012, Chapter 8) and the linear noise approximation (LNA) (Kurtz, 1972; Van Kampen, 1992, Chapter 10; Komorowski et al., 2009; Stathopoulos and Girolami, 2013; Fearnhead et al., 2014; Anderson and Kurtz, 2011; Ferm et al., 2008). Moreover, when a subset of reactions is taking place at a fast timescale with respect to the overall dynamics, other approximations are possible, for example the quasi-steady state assumption (QSSA) (Rao and Arkin, 2003). The QSSA assumes that a subset of species has reached its deterministic equilibrium, and only models the stochasticity arising from the reactions happening at a low rate, or involving very few molecules.

Finally, at the macroscopic level intrinsic noise is assumed to be negligible. In fact, as the number of molecules and the volume of the container increase, it is proven (Kurtz, 1972; Anderson and Kurtz, 2011) that the system reaches its deterministic equilibrium. The residual stochasticity is mostly due to measurement error, and ordinary differential equations (ODEs) can be assumed to be an appropriate modelling and simulation approach.

It has to be noted that the available analytical modelling approaches assume, at *any level*, a well stirred and thermally equilibrated environment. This is not the case when data are aggregated from different cells, as each cell represents a different ‘container’. We study this scenario and write the approximate aggregate hazards, stating the condition under which aggregation may be a sensible approximation. Finally, we check the accuracy of the approximated aggregate hazards for our set of reactions.

Here we focus on a scenario of transcriptional regulation comprising two transcription factors and one corresponding transcriptionally regulated ‘child’ gene, and first develop the model and the simulation at a microscopic level, where each reaction is modelled separately. We then move to a mesoscopic level, and, finally, consider the deterministic limit of the model linking the result to existing macroscopic models, i.e. the Hill function and the thermodynamic approach. We conclude this chapter by providing the state-space representation of the proposed model for the child gene mRNA at a mesoscopic scale, assuming that the two transcription factors inputs are fully observed. Extensions to scenarios comprising one unknown transcription factor or a negative feedback loop, are provided in Chapters 3 and 5, respectively.

1.1 Reaction networks

It is useful to first introduce the notation and the concepts underlying a reaction network, as we extensively refer to it in the following sections. Define a system with p chemical species, whose molecule numbers at time t define the vector of random variables $X(t) = (X_1(t), ..., X_p(t))^T$. A realisation of $X(t)$ is denoted with the lower case letter $x(t)$. The species participate in a set of r reactions, $R_1, ..., R_r$. Define the stoichiometry matrix S as a $p \times r$ matrix with elements $s_{i,k}$, each given by the difference between the number of molecules of the i-th species produced, and the number of molecules of the i-th species consumed by the k-th reaction (Wilkinson, 2012, Chapter 6).

As noted in Gillespie (1992), the knowledge of the number of molecules of each species at a specific time t, is in itself not sufficient to determine the future evolution of the system in a deterministic way, due to the fact that the positions and the momenta of the single particles remain unknown. However, Gillespie (1992) shows that, under the assumption of a *thermally equilibrated* and *well stirred* environment, it is possible to rigorously derive the transition probabilities of the system. This set of p differential equations takes the name of chemical master equation (CME).

In order to write the general form of the CME, it is useful to define the hazard of the k-th reaction, $h_k(X, c_k)$. Conditional on the system being in state x at time t, $h_k(x, c_k)dt$ gives the probability of the k-th reaction occurring in the infinitesimal time-interval $[t, t + dt)$ (Wilkinson, 2012, Chapter 6). The parameter c_k is referred to as the rate constant of the k-th reaction, and it is given by the probability that a random selection of molecules involved in the k-th reaction actually *collides* and *reacts* (Gillespie, 1992). To obtain the hazard $h_k(X, c_k)$, c_k must therefore be multiplied by all the possible combinations of available molecules for the k-th reaction. Formally, we have (Wilkinson, 2012, Chapter 6)

$$h_k(X, c_k) = c_k \prod_{i=1}^{p} \binom{X_i}{p_{k,i}}, \tag{1.1}$$

where $p_{k,i}$ denotes the number of molecules of the i-th species consumed by the k-th reaction.

Anderson and Kurtz (2011) then define the state of the process $X(t)$ as

$$X(t) = X(0) + SY\left(\int_0^t h(X(s), c)ds\right), \tag{1.2}$$

where Y is a vector of independent inhomogeneous Poisson processes counting the occurrence of the r reactions, and $h(X(s), c) = (h_1(X(s), c_1), ..., h_r(X(s), c_r))^T$. The Markov property is satisfied because the probabilities of the next transition - i.e. the reactions hazards - only depend on the present state of the system, and not on the past trajectories. The process $X(t)$ is therefore a Markov jump process in continuous time and discrete state-space, and its Kolmogorov's forward equation has the following form (Anderson and Kurtz, 2011; Wilkinson, 2012, Chapter 6)

$$\frac{d}{dt}p(x, t|x_0, t_0) = \sum_{k=1}^{r} \left[h_k(x - s_{\cdot,k}, c_k)p(x - s_{\cdot,k}, t|x_0, t_0) - h_k(x, c_k)p(x, t|x_0, t_0)\right],$$

where $s_{\cdot,k}$ is the k-th column of the stoichiometry matrix S, and $p(x, t|x_0, t_0) = P\left[X(t) = x(t)|X(0) = x(0)\right]$.

This is indeed the CME of Gillespie (1992) (see Wilkinson, 2012, Chapter 6). Note that the CME can be solved explicitly only in a restricted number of cases, reviewed in McQuarrie (1967).

1.2 Regulation by two transcription factors: the microscopic level

Consider now a system with two transcription factors (TFs), called A and B, each with their own binding site. Their protein levels are here denoted P_A and P_B. In order to start transcription, the binding of the basal transcriptional complex, composed of other TFs and the RNA polymerase (RNAP), is also required: for simplicity, we denote the full complex by RNAPc and assume a molecular number constant in time. We refer to Section 3.1 for a more detailed biological introduction of the transcriptional process. A graphical representation of the system is given in Figure 1.1.

We extend the approach presented for a single TF setting in Tkačik and Walczak (2011), and assume four possible states for the promoter: (0), when empty, (A), when only A is bound, (B), when only B is bound, or (A,B), when both A and B are bound. It is crucial to note that these states are mutually exclusive. Each state of the promoter is treated as a chemical species, which has molecule number equal to 1, when the state is visited, or 0 otherwise.

We define with k_{+i} the rate at which P_i binds the promoter, while k_{-i} denotes the rate of unbinding. Different rates can be specified for $i = A, B$.

It is also possible that one TF is more or less likely to bind or unbind, if the other TF has already bound the promoter. This is denoted by *cooperativity*

[1], and is obtained by multiplying the individual binding and unbinding rates of A and B by the cooperativity coefficients k_{+c} and k_{-c}, respectively. We assume the cooperativity coefficients to be equal for the two TFs.

We also assume that, once the promoter is in an active configuration, the mRNA of the child gene, M_g, is produced at a rate R_{state}. Again, this rate can be differentiated for each state. An inactive configuration has a transcription rate equal to 0, following an approach analogous to the one presented in Ribeiro et al. (2006). Independently of the state of the promoter, the mRNA of the child gene is degraded or translated at a rate μ_{M_g}.

Finally, the set of reactions also incorporate transcription and translation of the two TFs. Time-dependent transcription rates $\nu_A(t)$ and $\nu_B(t)$ are assumed for their mRNA, denoted respectively by M_A and M_B. The mRNA of the two transcription factors is then either degraded or translated into the proteins P_A and P_B.

A summary of all the reactions and of their hazards is presented in Table 1.1. The rates in the table account for the most general setting: different transcription rates, depending on the state of the promoter, different binding and unbinding rates for the two TFs, and cooperativity.

1.2.1 Simulation

The stochastic simulation algorithm (SSA) (Gillespie, 1977; Doob, 1945) can be used in order to simulate exactly from the system defined by the reactions in Table 1.1. The availability of simulated data from a known set of reactions and rates is particularly important when it comes to compare the original model with its approximations, and to perform parameter inference. In the former, we can in fact compare the distribution of the simulated data under the approximate models, and under the original one. In the latter, we can check the accuracy of our estimation algorithm by applying it to the simulated data, and then compare the estimated values of the parameters with the true ones.

Let $h_{tot}(X, c) = \sum_{k=1}^{r} h_k(X, c_k)$ be the cumulative hazard. The SSA has then been summarised in Wilkinson (2012, Chapter 6), and here reported in Algorithm 1.

The SSA can be slightly modified in order to record the species counts at fixed time-intervals of length dt (Wilkinson, 2012, Chapter 6), and this is the approach

[1] Cooperativity may be a slightly misleading term, since two TFs can either attract or repulse each other. However, it denotes here any form of interaction in the binding and unbinding to the promoter.

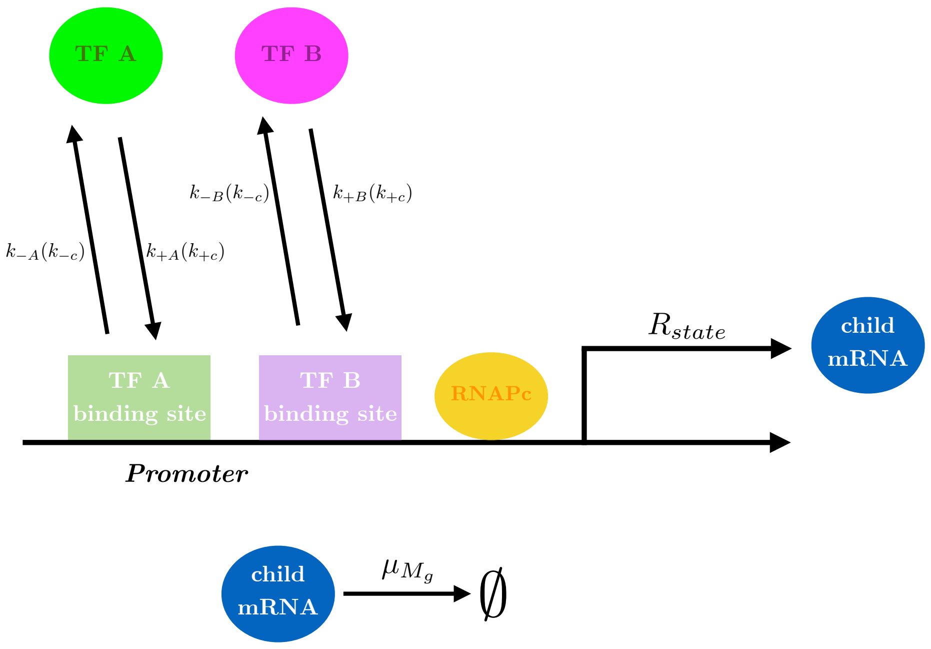


Figure 1.1: System: two TFs, A and B, can bind the promoter at their binding sites. Binding happens at a rate k_{+i} ($k_{+i}k_{+c}$ if the other TF is already bound to the promoter), unbinding at a rate k_{-i} ($k_{-i}k_{-c}$ if the other TF is already bound to the promoter) for P_i, $i = A, B$. Depending on the regulatory logic, transcription can take place when one or both the binding sites are occupied, or empty. The RNAPc has to bind the promoter in order to start transcription. M_g represents the child gene mRNA, which is produced at a rate R_{state}, depending on the state of the binding sites (occupied or empty). M_g is then degraded at a rate μ_{M_g}.

Algorithm 1 Stochastic simulation algorithm (SSA)

```
1: Set $t = 0$, $x = (x_1(0), ..., x_p(0))$ and $c = (c_1, ..., c_r)$;
2: Compute $h_k(x, c_k)$ for all $k$, and $h_{tot}(x, c)$;
3: Sample the time $\tau$ to the next reaction from an exponential random variable
   with mean $1/h_{tot}(x)$;
4: Sample the type $k$ of reaction from a discrete random variable with support
   $I = 1, ..., r$, and probabilities equal to $h_k(x, c_k)/h_{tot}(x, c)$;
5: Set $t = t + \tau$ and $x = x + s_{\cdot,k}$;
6: If the current time is less than the maximum simulation time, return to 2.
```

here followed, taking $dt = 0.1\,\mathrm{h}$. Moreover, here we assume destructive sampling in the experimental design, as motivated by our first application of the proposed model

	Reaction	Rate	Hazard
1	$RNAPc + Pro$ $\rightarrow RNAPc + Pro + M_g$	R_0	$R_0 X_{RNAPc} X_{Pro}$
2	$RNAPc + ProP_A$ $\rightarrow RNAPc + ProP_A + M_g$	R_A	$R_A X_{RNAPc} X_{ProP_A}$
3	$RNAPc + ProP_B$ $\rightarrow RNAPc + ProP_B + M_g$	R_B	$R_B X_{RNAPc} X_{ProP_B}$
4	$RNAPc + ProP_AP_B + M_g$ $\rightarrow RNAPc + ProP_AP_B + M_g$	$R_{A,B}$	$R_{A,B} X_{RNAPc} X_{ProP_AP_B}$
5	$Pro + P_A \rightarrow ProP_A$	k_{+A}	$k_{+A} X_{Pro} X_{P_A}$
6	$ProP_A \rightarrow Pro + P_A$	k_{-A}	$k_{-A} X_{ProP_A}$
7	$Pro + P_B \rightarrow ProP_B$	k_{+B}	$k_{+B} X_{Pro} X_{P_B}$
8	$ProP_B \rightarrow Pro + P_B$	k_{-B}	$k_{-B} X_{ProP_B}$
9	$ProP_A + P_B \rightarrow ProP_AP_B$	$k_{+B} k_{+c}$	$k_{+B} k_{+c} X_{ProP_A} X_{P_B}$
10	$ProP_B + P_A \rightarrow ProP_AP_B$	$k_{+A} k_{+c}$	$k_{+A} k_{+c} X_{ProP_B} X_{P_A}$
11	$ProP_AP_B \rightarrow ProP_A + P_B$	$k_{-B} k_{-c}$	$k_{-B} k_{-c} X_{ProP_AP_B}$
12	$ProP_AP_B \rightarrow ProP_B + P_A$	$k_{-A} k_{-c}$	$k_{-A} k_{-c} X_{ProP_AP_B}$
13	$M_g \rightarrow \emptyset$	μ_{M_g}	$\mu_{M_g} X_{M_g}$
14	$\emptyset \rightarrow M_A$	$\nu_A(t)$	$v_A(t)$
15	$M_A \rightarrow \emptyset$	μ_M	$\mu_M X_{M_A}$
16	$M_A \rightarrow M_A + P_A$	α_M	$\alpha_M X_{M_A}$
17	$P_A \rightarrow \emptyset$	μ_P	$\mu_P X_{P_A}$
18	$\emptyset \rightarrow M_B$	$\nu_B(t)$	$v_B(t)$
19	$M_B \rightarrow \emptyset$	μ_M	$\mu_M X_{M_B}$
20	$M_B \rightarrow M_B + P_B$	α_M	$\alpha_M X_{M_B}$
21	$P_B \rightarrow \emptyset$	μ_P	$\mu_P X_{P_B}$

Table 1.1: Reactions for the system under study, rates and hazards. $RNAPc$ is the basal transcriptional complex, Pro is the promoter, M_A and P_A are the mRNA and protein of TF A, M_B and P_B are the mRNA and protein of TF B. $ProP_A$, $ProP_B$ and $ProP_AP_B$ are the complexes formed by TF A, B and both A and B, when bound to the promoter; M_g denotes the mRNA of the child gene. X is the symbol indicating the molecules number. Finally, transcription rates $v_A(t)$ and $v_B(t)$ for the two TFs are time-dependent, and we assume for simplicity the same translation and degradation rates for the two TFs. Reactions are separated according to their role: the first four are related to the transcription of the child gene, reactions 5 to 12 represent the binding and unbinding of the TFs to the promoter, reaction 13 is the mRNA degradation of the child gene, and, finally reactions 14 to 21 account for the TFs transcription and translation.

to the *Arabidopsis Thaliana* available data. A scenario which does not comprise destructive sampling is assumed in our second data application, presented in Chapters 5 and 6. An appropriate simulation technique in the destructive sampling scenario involves simulation of independent paths for each data-point, i.e. starting each time from $t = 0$ (Stathopoulos and Girolami, 2013). Note, also, that the transcription rates of the two TFs are time-dependent, changing at some assumed switch time-points, thus generating the circadian cyclicity. The hazard function in Algorithm 1 is therefore time-dependent, and has more appropriate notation $h_k(x, t, c_k)$. At each time-step dt, we check if any switch-time has been reached, and update the hazards accordingly. However, as noted in Wilkinson (2012, Chapter 8), this approach is only approximately correct, as the time to the next reaction is computed before the possible hazard change, and may therefore overlap two transcriptional regimes. An exact solution is provided again in Wilkinson (2012, Chapter 8). On the other hand, in our case, the transcriptional rates of each TF are piece-wise constant, and change only five times during the whole simulation time; moreover, the waiting times to the next reaction event are generally relatively short: in one sample simulation of the full system for a single cell, the average waiting time is equal to $7.1 \times 10^{-3}\,\mathrm{s}$ and the maximum equal to $9.6 \times 10^{-2}\,\mathrm{s}$ for simulation scenario A of Figure 1.2, and $6.4 \times 10^{-3}\,\mathrm{s}$ and $6.1 \times 10^{-2}\,\mathrm{s}$, respectively, for simulation scenario B. Therefore we believe that the effect of the adopted approximate simulation approach on the simulated paths is minimal.

TF A transcriptional rates and switch time-points are set according to the parameter estimates obtained by running the Switch Tool of Section B.2 in Appendix (in an earlier version working with a log transformation of the time-series and with a time-constant variance) on available mRNA levels of a particular TF, namely late elongated hypocotyl (LHY). LHY has a pivotal role in the analysis of Chapters 3 and 4, and we refer to these later chapters for further details. TF B is slightly delayed with respect to TF A, and incorporates a shoulder in each peak, i.e. an intermediate transcriptional rate increase between the minimum and maximum rate, a feature often present in real data. Translation and degradation rates of TF B are set equal to that of A. In particular, degradation for TF A and B mRNA is set to $0.5\,\mathrm{h}^{-1}$, translation to $1\,\mathrm{h}^{-1}$, and degradation for TF A and B protein is set to $0.34\,\mathrm{h}^{-1}$.

In our formulation we assume that both the TFs and the child gene mRNA exhibit circadian rhythmicity. This is motivated by available data from both plants and mice, whose analysis is the focus of Chapters 3 to 6. However, the model here presented is applicable to any system where the transcription of a child gene is regulated by two transcription factors, i.e. when changes in the mRNA levels of the child gene can be associated to changes in the levels of the TFs, following a functional form which arises from the set of reactions of Table 1.1, and that is explicitly stated in Section 1.3.2.

With respect to the binding and unbinding rates, here we assume the same binding rate, $k_{+A} = k_{+B} = k_{+} = 2\,\text{molecules}^{-1}\text{h}^{-1}$. Previous work on the *diffusion limit* states that k_{+i} has an upper limit, that experiments in bacteria have observed to be often reached, and that does only depend on the linear dimension of the binding site (Tkačik and Walczac, 2011; Bialek and Setayeshgar, 2005). We are then assuming approximately the same linear dimension for the binding sites of the two TFs; it has to be noted, however, that in an inferential framework it is usually possible to accurately estimate only the dissociation coefficients, i.e. the ratios between k_{-i} and k_{+i}. The assumption of an equal binding rate does therefore not affect the inferential results. The unbinding rates are set to $60\,\text{h}^{-1}$ for TF A and $40\,\text{h}^{-1}$ for TF B. The choice of higher unbinding than binding rates is motivated in Forger and Peskin (2005), and references therein. The ratio between the unbinding and binding rates of each TF provides dissociation coefficients close to the mean expression levels of the TFs themselves, which, as we investigate in more detail in Section 1.3.2, means that both TFs are significantly influencing the dynamics of the child gene.

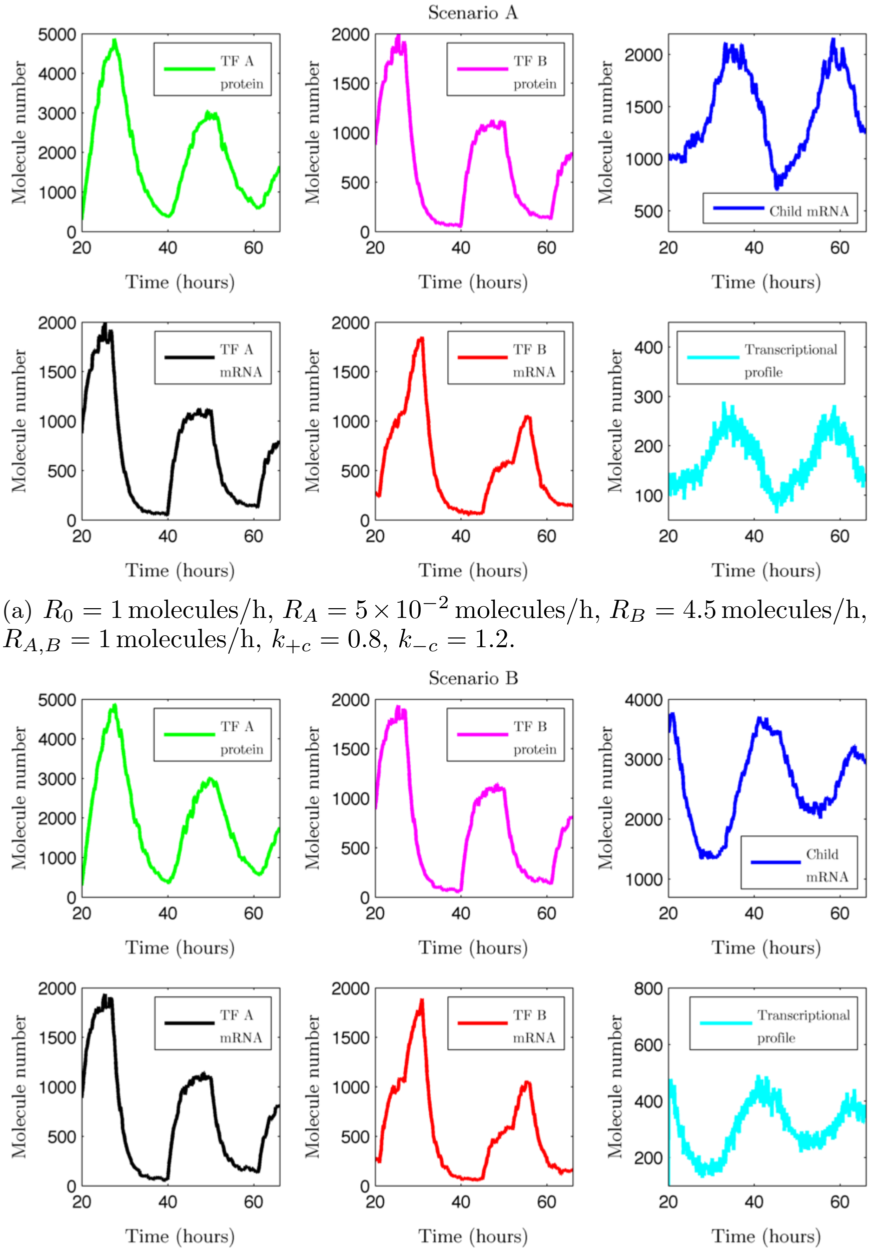
Rates for the transcription of the child gene mRNA, and the cooperativity between the two TFs, are differentiated depending on the regulatory logic implemented. Results and rates of the simulations are shown in Figure 1.2.

The figures show the simulated time-series for the mRNA and protein of TF A and B, the mRNA of the child gene, and the transcriptional profile, i.e. the number of molecules produced in each time step. Simulations have been carried out on a single cell level, and values summed up over 100 cells.

We can see that the behaviour of the child gene mRNA varies according to the TFs time-series and the values fixed for the transcription rates.

In the scenario illustrated in Figure 1.2 (a), for example, TF A represses transcription on its own, while TF B activates it; moreover, when TF A and B are contemporarily bound to the promoter, the resulting transcriptional rate is equal to R_0, i.e. the rate observed when the promoter is empty. We can see in fact that the child gene mRNA level tends to be high when TF B is at its maximum, while low levels can be observed between hours 40 and 50, when TF B is low, and TF A is increasing.

In a second scenario, illustrated in Figure 1.2 (b), we have instead that both TFs are repressors, and even stronger repression is achieved in interaction. We can indeed see that high levels of the child gene mRNA are reached when the two TFs are low, and therefore their repressive effect is released; on the other hand, when the TFs levels are high, transcription is strongly inhibited, resulting in minimal levels



(a) $R_0 = 1\,\text{molecules/h}$, $R_A = 5 \times 10^{-2}\,\text{molecules/h}$, $R_B = 4.5\,\text{molecules/h}$, $R_{A,B} = 1\,\text{molecules/h}$, $k_{+c} = 0.8$, $k_{-c} = 1.2$.

(b) $R_0 = 6\,\text{molecules/h}$, $R_A = 3.5\,\text{molecules/h}$, $R_B = 2.5\,\text{molecules/h}$, $R_{A,B} = 5 \times 10^{-2}\,\text{molecules/h}$, $k_{+c} = 1.2$, $k_{-c} = 0.8$.

Figure 1.2: Simulated data from the SSA algorithm for the reactions in Table 1.1. Common parameters: $\nu_A(t) = [9.6, 0.29, 5.6, 0.72, 4.2]$ (in molecules per hour) with switch times (in hours) $Swt_A = [27, 40, 50, 61]$, $\nu_B(t) = [0.7, 5.7, 9.7, 0.3, 3.0, 5.6, 0.7]$ (in molecules per hour) and switch times (in hours) $Swt_B = [21, 28, 45, 52, 56]$, $\mu_M = 0.5\,\text{h}^{-1}$, $\alpha_M = 1\,\text{h}^{-1}$, $\mu_P = 0.34\,\text{h}^{-1}$, $\mu_{M_g} = 1.2\,\text{h}^{-1}$, $k_+ = 2\,\text{molecules}^{-1}\text{h}^{-1}$, $k_{-A} = 60\,\text{h}^{-1}$, $k_{-B} = 40\,\text{h}^{-1}$. $X_{RNAPc} = 10\,\text{molecules}$. Aggregated over 100 cells.

of the child gene mRNA.

Clearly, other logical forms of regulation by two TFs can be implemented. The key role is played by the binding, unbinding and transcription rates. As noted earlier, the promoter can only be in one of the four possible configurations at any time. An active configuration is characterised by a transcription rate of the child gene greater than 0. However, if the promoter has in the empty state the same transcription rate as the one reached when one TF is bound, we can assume that the TF is independently neither an activator, nor a repressor. Moreover, if the transcription rate of the child gene does not change when both TFs are bound, with respect to the case when only one of them is bound, then we can conclude that there is no interaction effect on transcription and $R_A = R_B = R_{A,B}$. Alternatively, TF A, or B, is dominating the dynamics if $R_{A,B} = R_A$, or $R_{A,B} = R_B$, respectively. We provide a more formal explanation of this point in Section 1.3.2.

Binding and unbinding rates also play a role. In fact, for given levels of the TFs, their role is to define the transitions between the four promoter states. We refer again to Section 1.3.2 for a more detailed explanation of this point.

It is finally worth remarking that binding and unbinding rates will only influence the probability of the promoter being in one of the four states, but different transcription rates have to be specified in order to implement an actual regulatory logic: in the trivial case of equal transcription rates for all states, the production of the child gene mRNA would just be constant, irrespectively of the levels of A and B, or their binding and unbinding rates.

1.3 The mesoscopic level

The specification of the model at a microscopic scale, and for a single cell, represents a useful description of the *exact* dynamics underlying the system under study, and a powerful tool to perform exact simulation. However, in real data settings, the information may be collected at an aggregate level, and its resolution may not be sufficient to obtain sensible estimates of all the reaction rates involved. Some reactions may indeed happen at a fast rate, if compared to the time resolution of the data. When the number of molecules involved or the rate for a subset of reactions is high, approximate approaches are available.

1.3.1 Aggregation

In order to deal with a model for aggregate data, as required by the *Arabidopsis Thaliana* data, we first need to know the corresponding reaction rates. As the number of cells increases, the number of molecules of the reactants, and the volume, increase as well. Suppose that there are n cells in the system, and assume the total volume of all cells multiplied by the Avogadro number to be $n\Omega$. The Avogadro number is equal to 6.02×10^{23}, and represents the number of molecules per unit mole. Let $X_{sum,i}$ be the sum of the number of molecules of the i-th species over all the n cells. It is useful to rescale the hazards with the volume of the container. In particular, following Anderson and Kurtz (2011) Equation 1.1 becomes

$$h_k^{n\Omega}(X_{sum}, c_k) = c_k \frac{\prod_i p_{k,i}!}{(n\Omega)^{|p_{k-1}|}} \prod_i \binom{X_{sum,i}}{p_{k,i}} \tag{1.3}$$

where $|p_k| = \sum_i p_{k,i}$.

We can move one step further and define the hazards in terms of concentrations, obtaining the mass action kinetics form $\tilde{h}_k$. Let $Z_i = X_{sum,i}/n\Omega$. For the reactions of interest in our system, we have

- Zero-th order reactions ($\emptyset \to Product$): $h_k^{n\Omega}(X_{sum}, c_k) = n\Omega\tilde{h}_k(Z, c_k) = n\Omega c_k$;
- First order reactions ($X \to Product$): $h_k^{n\Omega}(X_{sum}, c_k) = n\Omega\tilde{h}_k(Z, c_k) = n\Omega c_k Z$;
- Second order reactions ($X_1 + X_2 \to Product$): $h_k^{n\Omega}(X_{sum}, c_k) = n\Omega\tilde{h}_k(Z, c_k) = n\Omega c_k Z_1 Z_2$;

It is important to note that we have here applied a significant simplification, in that we have assumed the n cells pulled together belong to the same system, in a well stirred and thermally equilibrated environment. This is not true, as the cells represent different 'containers'. However, if we assume independence between the cells, we can exploit the fact that the sum of independent Poisson processes is a new Poisson process, having as mean the sum of the means of the original independent processes. We can then, indeed, obtain the cumulative hazards by just summing up the hazards of the single cells. This leads to the *cumulative hazard* of Equation 1.3 for zero-th and first order reactions, and is consistent with Oates and Mukherjee (2012).

However, that this is not the case for second order reactions, as it is generally not true that the sum of the hazards equals the hazard of the sum (Oates and Mukherjee, 2012). This seems a quite restrictive condition for performing inference on aggregated data, as a model derived from a system-size expansion of the reaction network may not lead in this case to meaningful inferences for the rates involved.

On the other hand, there are conditions under which the system-size expansion can still be 'safely' performed for aggregated data. We show that this is the case if at least one of the reactants involved in a second order reaction has reached its deterministic equilibrium in each cell. Formally, assume for species i in cell j

$$X_{j,i} = \Omega Z_{j,i} \approx \Omega z_i$$

so that

$$X_{sum,i} = \sum_j X_{j,i} = \sum_j \Omega Z_{j,i} \approx n\Omega z_i,$$

where z_i denotes the deterministic equilibrium of species i. Since we assume that in each cell we have the same stochastic process, it must be $z_{j,i} = z_i$, for all $j = 1, ..., n$. Consider now the case of a second order reaction, and assume X_1 to be the species reaching the equilibrium. We have

$$\begin{aligned} \sum_j h_k^{\Omega}(X_{j,\cdot}, c_k) &= \sum_j \frac{1}{\Omega} c_k X_{j,1} X_{j,2} = \sum_j c_k Z_{j,1} X_{j,1} \approx c_k z_1 X_{sum,2} \\ &= \frac{1}{n\Omega} c_k n\Omega z_1 X_{sum,2} \approx \frac{1}{n\Omega} c_k X_{sum,1} X_{sum,2} = h_k^{n\Omega}(X_{sum}, c_k), \end{aligned}$$

Intuitively, under the assumption that $X_{j,1}$ is at equilibrium in each cell, the second order reaction can be approximately considered a first order one. This allows to perform the summation over the different cells.

For the transcriptional regulation framework that we are considering, this assumption is fulfilled if the TFs expression levels are approximately the same in the different cells. We check the accuracy of this approximation through simulation for the two assumed simulation scenarios of Figures 1.2 (a) and (b). In each cell, the peak levels of TF A and B are of about 50 molecules, while their average levels are of about 15-20 molecules. Figure 1.3 shows a comparison between aggregated trajectories simulated from the original model, and trajectories simulated with aggregate hazards. We can see a slight difference in the accuracy of approximation between the two simulation scenarios. In particular, the mean and variance of the simulated trajectories overlap more precisely in scenario A than in scenario B. It is possible that the higher cooperativity coefficient makes second order reactions of binding and unbinding more likely in scenario B, and therefore they have a stronger influence on the dynamics of the child gene mRNA. However, if we simulate the latter scenario by increasing the mRNA and protein TFs molecules numbers, as well as their dissociation coefficient, by a factor of 15, and we maintain the same regulatory logics, the mean and the variability intervals of the child gene mRNA under the original model and the aggregate hazards model, are overlapping more precisely, as we can observe in Figure 1.4

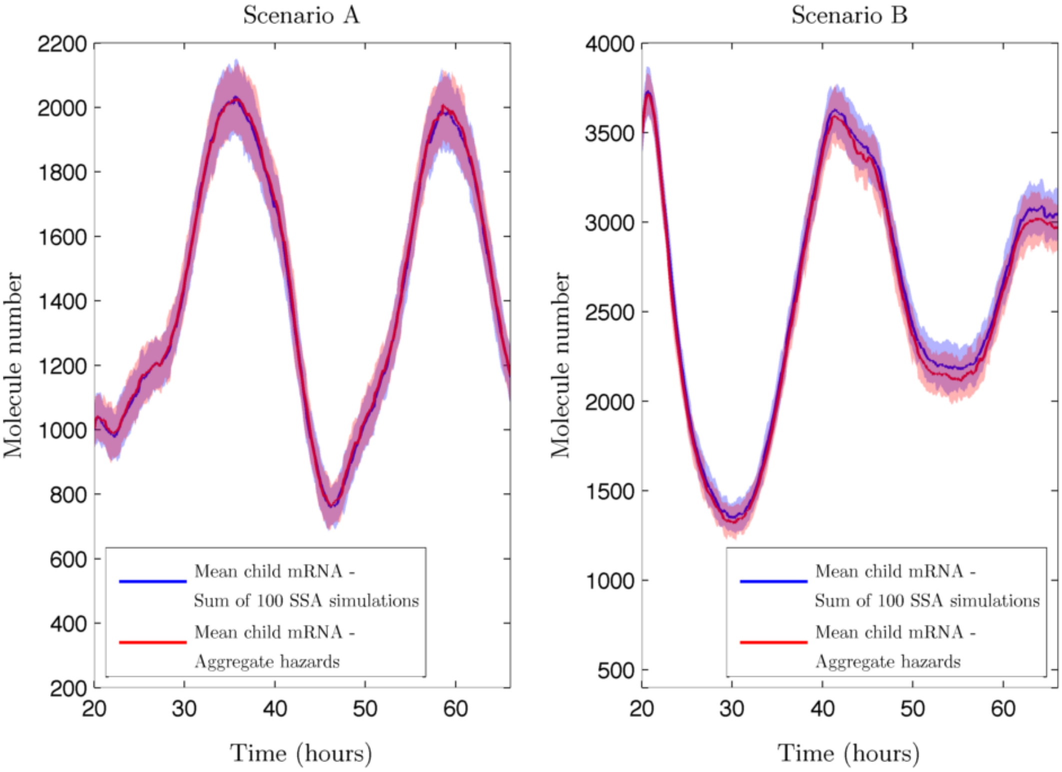


Figure 1.3: Mean and ±2 standard deviation (SD) variability intervals for 50 simulations from the original model (blue), the aggregate hazards model (red). Simulation scenario of Figure 1.2 (a) (left) and (b) (right).

1.3.2 The quasi-steady state assumption

A further approximation of the Markov jump process defined by the set of reactions of Table 1.1, can be applied given that binding and unbinding reactions are believed to happen at a fast timescale, if compared to the production and degradation of the child gene mRNA. This assumption is coherent with previous work on the mammalian circadian clock (we refer again to Forger and Peskin, 2005, and references therein). In this scenario, it is then reasonable to apply the quasi-steady state assumption (QSSA) (Rao and Arkin, 2003).

In our case, the QSSA involves the promoter occupancies, and assumes that the switches between the bound and the unbound states are fast enough if compared to the child gene mRNA production and degradation, that the promoter can be treated as being at equilibrium (Tkačik and Walczak, 2011). In order to derive and solve the deterministic ODEs for the promoter occupancy, we follow again the derivation and notation of Anderson and Kurtz (2011). We refer from now on to the aggregate system, thus simply writing $X_{sum,i}(t)$ as $X_i(t)$.

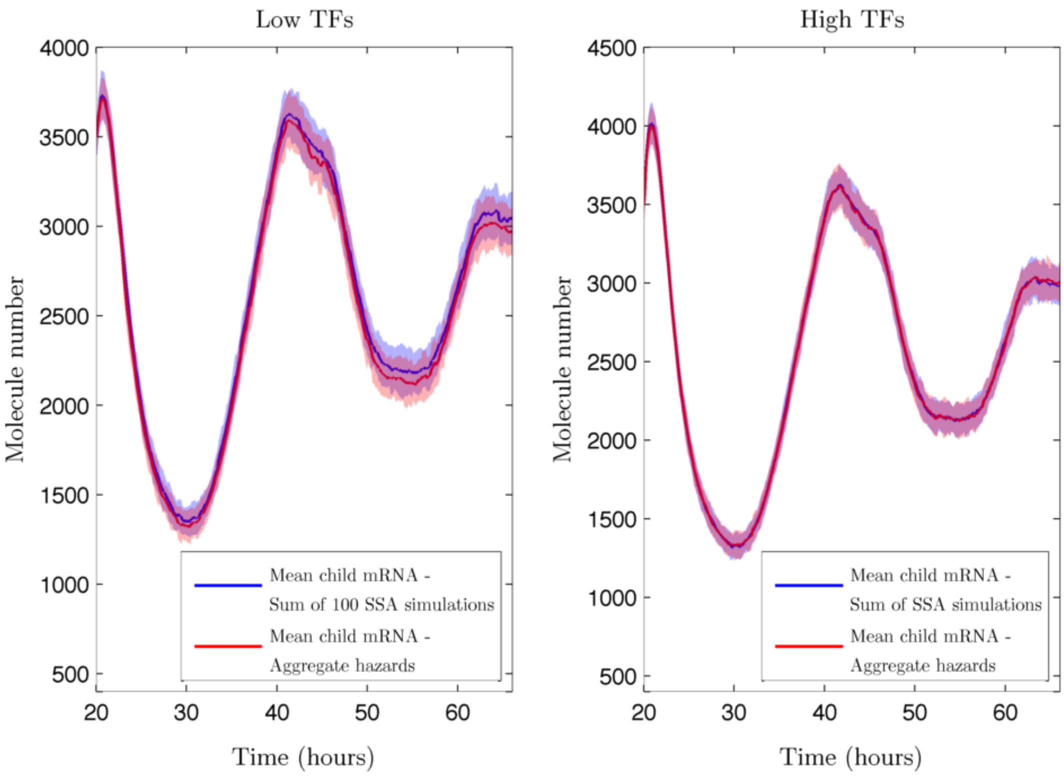


Figure 1.4: Mean and ±2 SD variability intervals for 50 simulations from the original model (blue), the aggregate hazards model (red). Simulation scenario of Figure 1.2 (b) with low molecules counts for the TFs (left) and high molecules counts for the TFs (right).

In order to obtain the deterministic ODE form, the authors apply the so-called classical scaling and rewrite Equation 1.2 in terms of concentrations,

$$Z(t) = Z(0) + \frac{1}{n\Omega} SY \left(\int_0^t h^{n\Omega}(X(s), c) ds \right),$$

from which they obtain

$$\begin{aligned} Z(t) &\approx Z(0) + \frac{1}{n\Omega} SY \left(n\Omega \int_0^t \tilde{h}(Z(s), c) ds \right) \\ &= Z(0) + \frac{1}{n\Omega} S\tilde{Y} \left(n\Omega \int_0^t \tilde{h}(Z(s), c) ds \right) + \int_0^t S\tilde{h}(Z(s), c) ds, \qquad (1.4) \end{aligned}$$

where $\tilde{Y}(u) = Y(u) - u$, i.e. the centred process, which is crucial for the last part of the proof. In fact, from the law of the large numbers for the Poisson process (Anderson and Kurtz, 2011), as $n \to \infty$, $\tilde{Y}(nu)/n \to 0$. In the limit, the process therefore becomes deterministic and has continuous ODE form (Anderson and Kurtz, 2011;

Kurtz, 1972)

$$\frac{dz(t)}{dt} = S\tilde{h}(z(t), c).$$

The last equation is also known as the macroscopic rate equation (MRE).

Moving back to our system, according to the reactions in Table 1.1, we can write the system of ODEs for the binding site occupancies,

$$\begin{aligned}
\frac{dz_0(t)}{dt} &= k_{-B}z_B(t) + k_{-A}z_A(t) - (\tilde{k}_{+A}Z_{P_A}(t) + \tilde{k}_{+B}Z_{P_B}(t))z_0(t) \\
\frac{dz_A(t)}{dt} &= \tilde{k}_{+A}Z_{P_A}(t)z_0(t) + k_{-B}k_{-c}z_{AB}(t) - (\tilde{k}_{+B}k_{+c}Z_{P_B}(t) + k_{-A})z_A(t) \\
\frac{dz_B(t)}{dt} &= \tilde{k}_{+B}Z_{P_B}(t)z_0(t) + k_{-A}k_{-c}z_{AB}(t) - (\tilde{k}_{+A}k_{+c}Z_{P_A}(t) + k_{-B})z_B(t) \\
\frac{dz_{A,B}(t)}{dt} &= \tilde{k}_{A+}k_{+c}Z_{P_A}(t)z_B(t) + \tilde{k}_{B+}k_{+c}Z_{P_B}(t)z_A(t) - k_{-c}(k_{-A} + k_{-B})z_{A,B}(t),
\end{aligned}$$

where we have $z_0 = x_{Pro}/n\Omega$, $z_B = x_{ProP_B}/n\Omega$, $z_A = x_{ProP_A}/n\Omega$ and $z_{A,B} = x_{ProP_AP_B}/n\Omega$. Note also that we have applied the conversion from stochastic to deterministic rates, i.e. $\tilde{k}_{+A} = n\Omega k_{+A}$ and $\tilde{k}_{+B} = n\Omega k_{+B}$. Assume now, for simplicity, $\Omega = 1$. The $z_i(t)$s can be interpreted as the proportion of binding sites in state i in the n cells at time t (this is consistent with Tkačik and Walczak, 2011, for the single TF scenario). Let $\tilde{K}_A = k_{-A}/\tilde{k}_{+A}$, $\tilde{K}_B = k_{-B}/\tilde{k}_{+B}$ and $K_c = k_{-c}/k_{+c}$. By equating the first four ODEs to 0 and imposing the constraint $z_0+z_A+z_B+z_{A,B} = 1$, an explicit solution can be derived, which has the form

$$\begin{aligned}
z_0(t) &= \frac{1}{1 + \frac{Z_{P_A}(t)}{\tilde{K}_A} + \frac{Z_{P_B}(t)}{\tilde{K}_B} + \frac{Z_{P_A}(t)Z_{P_B}(t)}{\tilde{K}_A\tilde{K}_BK_c}} \\
z_A(t) &= \frac{\frac{Z_{P_A}(t)}{\tilde{K}_A}}{1 + \frac{Z_{P_A}(t)}{\tilde{K}_A} + \frac{Z_{P_B}(t)}{\tilde{K}_B} + \frac{Z_{P_A}(t)Z_{P_B}(t)}{K_A\tilde{K}_BK_c}} \\
z_B(t) &= \frac{\frac{Z_{P_B}(t)}{\tilde{K}_B}}{1 + \frac{Z_{P_A}(t)}{\tilde{K}_A} + \frac{Z_{P_B}(t)}{\tilde{K}_B} + \frac{Z_{P_A}(t)Z_{P_B}(t)}{\tilde{K}_A\tilde{K}_BK_c}} \\
z_{A,B}(t) &= \frac{\frac{Z_{P_A}(t)Z_{P_B}(t)}{\tilde{K}_A\tilde{K}_BK_c}}{1 + \frac{Z_{P_A}(t)}{\tilde{K}_A} + \frac{Z_{P_B}(t)}{\tilde{K}_B} + \frac{Z_{P_A}(t)Z_{P_B}(t)}{\tilde{K}_A\tilde{K}_BK_c}}.
\end{aligned}$$

where the denominators are all equal and given by the sum of the four possible numerators.

By plugging-in the promoter equilibrium solution in the child gene mRNA equation, we obtain the QSSA. In particular, start from the stochastic model for

X_{M_g},

$$\begin{aligned} X_{M_g}(t) &= X_{M_g}(0) + Y_1\left(\int_0^t \frac{1}{n\Omega} R_0 X_{RNAPc}(s) X_0(s) ds\right) \\ &+ Y_2\left(\int_0^t \frac{1}{n\Omega} R_A X_{RNAPc}(s) X_A(s) ds\right) \\ &+ Y_3\left(\int_0^t \frac{1}{n\Omega} R_B X_{RNAPc}(s) X_B(s) ds\right) \\ &+ Y_4\left(\int_0^t \frac{1}{n\Omega} R_{A,B} X_{RNAPc}(s) X_{A,B}(s) ds\right) - Y_{13}\left(\int_0^t \mu_{M_g} X_{M_g}(s) ds\right), \end{aligned}$$

where each Poisson process Y refers to a reaction, whose number is given in Table 1.1. We can exploit independence between the Poisson processes describing the transcriptional reactions and write

$$\begin{aligned} X_{M_g}(t) &= X_{M_g}(0) + \\ &+ Y_{tr}\left(\int_0^t \left[R'_0 Z_0(s) + R'_A Z_A(s) + R'_B Z_B(s) + R'_{A,B} Z_{A,B}(s)\right] ds\right) \\ &- Y_{13}\left(\int_0^t \mu_{M_g} X_{M_g}(s) ds\right) \\ &\approx X_{M_g}(0) + \\ &+ Y_{tr}\left(\int_0^t \left[R'_0 z_0(s) + R'_A z_A(s) + R'_B z_B(s) + R'_{A,B} z_{A,B}(s)\right] ds\right) \\ &- Y_{13}\left(\int_0^t \mu_{M_g} X_{M_g}(s) ds\right), \end{aligned}$$

where Y_{tr} is the Poisson process accounting for the overall transcriptional reaction, and we assumed X_{RNAPc} to be approximately constant over time, so that we can drop the dependence on t, and consider it just a multiplying constant for the rates of the different states (Wilkinson, 2012, Chapter 6), therefore $R'_i = X_{RNAPc} R_i$. It is important to note that, with respect to a fully deterministic approach, we have retained a stochastic model formulation for the production and degradation of the child gene mRNA.

It is now interesting to focus on the transcription function. We then have

$$\begin{aligned} \nu(t) = \nu(Z_{P_A}(t), Z_{P_B}(t)) &= (R'_0 z_0 + R'_A z_A + R'_B z_B + R'_{A,B} z_{A,B}) \\ &= \frac{R'_0 + R'_A \frac{Z_{P_A}(t)}{\tilde{K}_A} + R'_B \frac{Z_{P_B}(t)}{\tilde{K}_B} + R'_{AB} \frac{Z_{P_A}(t) Z_{P_B}(t)}{\tilde{K}_A \tilde{K}_B K_c}}{1 + \frac{Z_{P_A}(t)}{\tilde{K}_A} + \frac{Z_{P_B}(t)}{\tilde{K}_B} + \frac{Z_{P_A}(t) Z_{P_B}(t)}{\tilde{K}_A \tilde{K}_B K_c}}. \end{aligned}$$

The overall transcription function is a weighted sum of the transcription rates of each state. The weights are given by the probability of the promoter to be in each of the states, as a function of the TFs concentrations. Note that it is straightforward to substitute concentrations with molecules numbers, and therefore go back to the stochastic rates, i.e.

$$\nu(t) = \frac{R'_0 + R'_A \frac{X_{P_A}(t)}{K_A} + R'_B \frac{X_{P_B}(t)}{K_B} + R'_{AB} \frac{X_{P_A}(t) X_{P_B}(t)}{K_A K_B K_c}}{1 + \frac{X_{P_A}(t)}{K_A} + \frac{X_{P_B}(t)}{K_B} + \frac{X_{P_A}(t) X_{P_B}(t)}{K_A K_B K_c}}, \tag{1.5}$$

where $K_A = k_{-A}/k_{+A}$ and $K_B = k_{-B}/k_{+B}$.

We provide in Appendix A.3 the first partial derivatives with respect to $X_{P_A}(t)$ and $X_{P_B}(t)$, which may help in understanding how repression and activation may be defined in terms of the transcription and binding rates. A positive first derivative means in fact an increase in the transcriptional activity as the corresponding TF increases, thus identifying an activator, while a negative derivative characterises a repressor. The case in which $K_c = 1$ is the most interpretable one. We can see that the first derivatives with respect to $X_{P_A}(t)$ and $X_{P_B}(t)$ are positive respectively for

$$R'_A - R'_0 > (R'_B - R'_{A,B})(X_{P_B}(t)/K_B), \tag{1.6}$$

and

$$R'_B - R'_0 > (R'_A - R'_{A,B})(X_{P_A}(t)/K_A). \tag{1.7}$$

To better understand this relationship, we can use limit arguments. Focussing e.g. on TF A, as TF B tends zero, TF A is an activator if the transcriptional rate of the state in which only A is bound, is higher than R_0, i.e. the rate of the empty promoter. On the other hand, as TF B divided by its dissociation coefficient tends to ∞, the relevant difference becomes the one on the right-hand side of Equations 1.6 and 1.7: if the transcriptional rate when only B is bound is lower than the rate when both A and B are bound, the derivative tends to ∞, while it tends to $-\infty$, if the difference is positive. Finally, if $R'_B = R'_{A,B}$, the sign of the derivative is only defined by the sign of $R'_B - R'_0$.

With respect to the binding and unbinding rates, we noted in Section 1.2.1 that they influence the probability of each promoter state. This point is better understood by resorting again to limit arguments. Indeed, as e.g. $k_{-A}/k_{+A} \to 0$, the probability of TF A being bound to the promoter goes to 1. In the limit, the transcription function of Equation 1.5 reduces to

$$\nu(t) \quad = \quad \frac{R'_A + R'_{A,B}\frac{X_{P_B}(t)}{K_B K_c}}{1 + \frac{X_{P_B}(t)}{K_B K_c}}, \tag{1.8}$$

therefore states (0) and (B) have null probability. Conversely, as $k_{-A}/k_{+A} \to \infty$, the probability of TF A being bound to the promoter goes to 0, and Equation 1.5 approaches

$$\nu(t) \quad = \quad \frac{R'_0 + R'_B\frac{X_{P_B}(t)}{K_B}}{1 + \frac{X_{P_B}(t)}{K_B}}, \tag{1.9}$$

i.e., the two states with null probability are (A) and (A,B). An analogous argument applies to TF B. Note that the two limits are conceptually different, but induce the same model parametrisation. Essentially, TF A has no dynamical influence on the system.

Finally, if we assume cooperativity in the binding or unbinding, i.e. $k_{+c} > (<)1$ or $k_{-c} < (>)1$, it is more (less) likely for the promoter to be in state (A,B). In the limit, as $k_{+c} \to 0$ we have that the transcriptional rate is constant and equal to $R'_{A,B}$. When $k_{+c} \to \infty$, instead, a competitive binding scenario can be assumed: the probability of having both the TFs bound to the promoter goes to zero.

The transcription function in Equation 1.5 has indeed a form that, assuming $K_c = 1$, is available in the literature (see Nachman et al., 2004). Our main contribution is in explicitly deriving its form from a single-cell stochastic model, also introducing cooperativity in the binding. This has allowed to explicitly state all the approximations employed, and to rigorously derive the child gene mRNA intrinsic noise under the assumption of a fast TFs binding and unbinding to the promoter.

It is now of interest to see how the QSSA behaves with respect to the original model and the one with the aggregate hazards. Figures 1.5 and 1.6 show a comparison between the original system, the aggregate hazards one, and the one under the QSSA, for the two simulation scenarios. In particular, we plot mean and variability bandwidths for 50 simulations under each scenario. We can see that, indeed, the QSSA seems a tenable approximation, not leading to any evident mismatch with respect to the aggregate hazards one, while the main source of mismatch can be identified in the aggregate hazards assumption for the simulation scenario of Figure 1.2 (b), as pointed out in the previous section.

It is important to stress that the accuracy of approximation depends on the assumption of fast binding and unbinding reactions; a scenario where lower rates are assumed for these reactions is studied in Chapters 3 and 4.

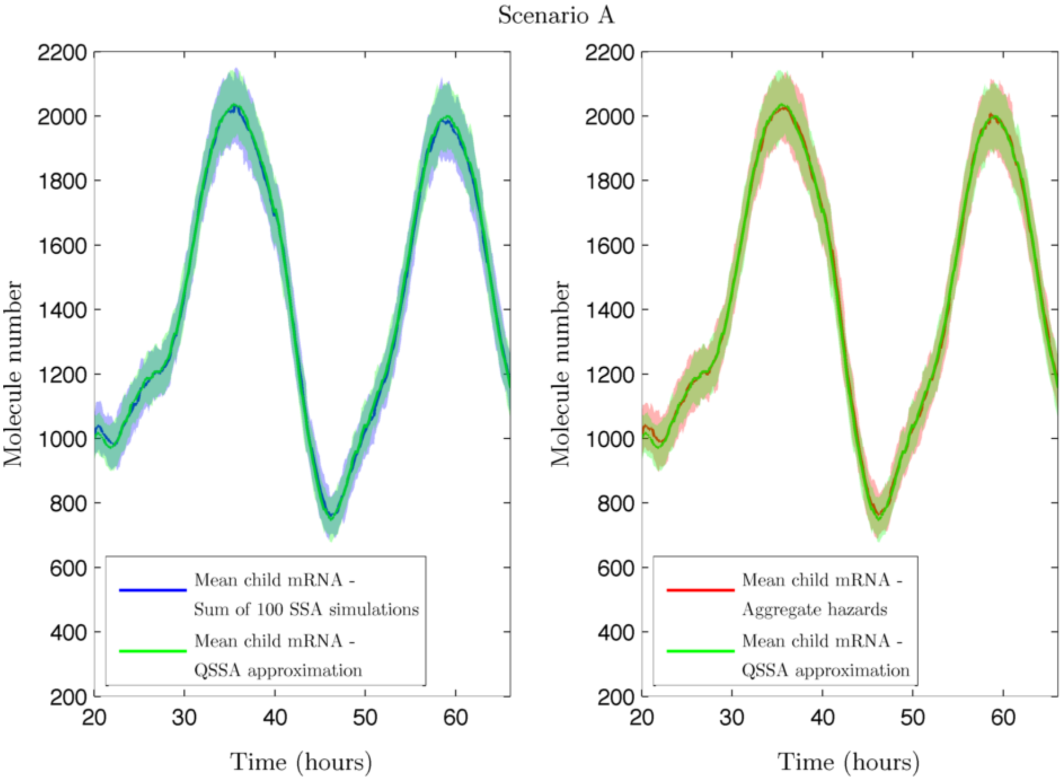


Figure 1.5: Mean and ±2 SD variability intervals for 50 SSA simulations from the the QSSA model (green), the original model (blue), and the aggregated hazards one (red). Simulation scenario of Figure 1.2 (a).

1.3.3 Exact and approximate transition densities

The model obtained from the QSSA is still a Markov jump process in continuous time - discrete state-space, and therefore its chemical master equation/ Kolmogorov's forward equation can be written explicitly. By assuming that the TFs are known, it actually belongs to the few cases in which it can also be solved explicitly, giving as a solution a convolution of a Poisson and a Binomial random variable. The result is indeed more general, and concerns all systems that undergo only monomolecular reactions, with arbitrary initial conditions, and time-varying rates. It is presented in Jahnke and Huisinga (2007), along with an application to an immigration and death process, which straightforwardly applies to our case, and we report here. The

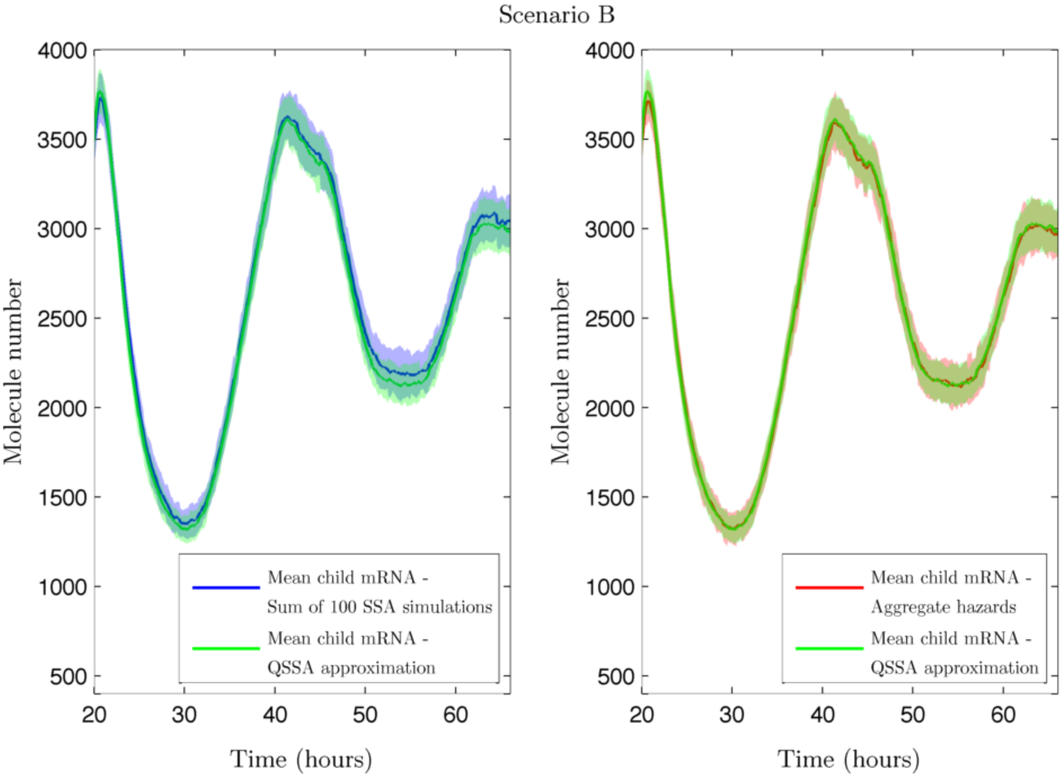


Figure 1.6: Mean and ±2 SD variability intervals for 50 SSA simulations from the the QSSA model (green), the original model (blue), and the aggregated hazards one (red). Simulation scenario of Figure 1.2 (b).

chemical master equation for the child gene mRNA is given by

$$\frac{d}{dt}p(x_{M_g},t) = \nu(t)p(x_{M_g}-1,t) - \nu(t)p(x_{M_g},t) + \mu_{M_g}p(x_{M_g}+1,t) - \mu_{M_g}p(x_{M_g},t), \tag{1.10}$$

where, for ease of notation, the dependence on the initial condition $(x_{M_g,0},t_0)$ has been dropped.

Let $A(t) = -\mu_{M_g}$, $b(t) = \nu(t)$, and solve the system of ODEs of Theorem 1 in Appendix A.1,

$$\begin{aligned}\frac{d\pi(t)}{dt} &= -\mu_{M_g}\pi(t)\\ \frac{d\lambda(t)}{dt} &= -\mu_{M_g}\lambda(t) + \nu(t),\end{aligned}$$

from initial condition $\pi(0) = 1$ and $\lambda(0) = 0$. The solution of Equation 1.10 is then

(Jahnke and Huisinga, 2007)

$$p(x_{M_g}, t) = \sum_{q=0}^{\min(x_{M_g}(0), x_{M_g})} \binom{x_{M_g}(0)}{q} \pi^q(t)(1-\pi(t))^{(x_{M_g}(0)-q)} \frac{\lambda^{(x_{M_g}-q}(t))}{(x_{M_g}-q)!} e^{-\lambda(t)}.$$

It is also of interest to study the mean and variance of the process. In particular we have that, given independence between the Poisson and the Binomial random variable of the convolution (Jahnke and Huisinga, 2007),

$$\begin{aligned} E[X_{M_g}(t)|x_{M_g}(0)] &= x_{M_g}(0)\pi(t)(1-\pi(t)) + \lambda(t) \\ V[X_{M_g}(t)|x_{M_g}(0)] &= x_{M_g}(0)\pi(t) + \lambda(t). \end{aligned}$$

Assume for simplicity $\nu(t) = \nu$, i.e. the transcription rate is constant over time. In our scenario, we can indeed assume the transcription function to be piece-wise constant, for a small enough dt. At the end of the time-interval, we can adopt the current states as the initial conditions of the next interval. We then have

$$\begin{aligned} \pi(t) &= e^{-\delta_{M_g} t} \\ \lambda(t) &= \frac{\nu}{\mu_{M_g}}(1 - e^{-\mu_{M_g} t}), \end{aligned}$$

therefore (Jahnke and Huisinga, 2007)

$$\begin{aligned} E[X_{M_g}(t)|x_{M_g}(0)] &= x_{M_g}(0)e^{-\mu_{M_g} t} + \frac{\nu}{\mu_{M_g}}(1 - e^{-\mu_{M_g} t}) \qquad (1.11) \\ V[X_{M_g}(t)|x_{M_g}(0)] &= x_{M_g}(0)e^{-\mu_{M_g} t}(1 - e^{-\mu_{M_g} t}) + \frac{\nu}{\mu_{M_g}}(1 - e^{-\mu_{M_g} t}). \qquad (1.12) \end{aligned}$$

Although the exact result is in itself interesting, it turns out that it would require computationally time-demanding procedures for inference about the parameters. We hence consider a further approximation, which is appropriate for relatively high molecules numbers which we would expect for aggregate counts.

If, for an infinitesimal time dt, it can be assumed that the hazards remain approximately constant, so that

$$\begin{aligned} X(t + dt) &= X(t) + SY\left(\int_t^{t+dt} h^{n\Omega}(X(s), c)ds\right) \\ &\approx X(t) + SY(h^{n\Omega}(X(t), c)dt), \end{aligned}$$

and the number of occurrences of each reaction is much greater than one, then a multivariate normal random variable can approximate the vector of independent

Poisson random variables Y (Stathopoulos and Girolami, 2013). The **chemical Langevin equation** (CLE), or diffusion approximation, for the process has then the stochastic differential equation (SDE) form (Stathopoulos and Girolami, 2013; Wilkinson, 2012, Chapter 8; Anderson and Kurtz, 2011)

$$dX(t) = Sh^{n\Omega}(X(t), c)dt + S\,\mathrm{diag}\sqrt{\{h^{n\Omega}(X(t), c)\}}dB(t),$$

where $dB(t)$ is an r-dimensional Wiener process. Note that we can equivalently write

$$dX(t) = Sh^{n\Omega}(X(t), c)dt + \mathrm{diag}\sqrt{\{Sh^{n\Omega}(X(t), c)S^T\}}dB(t),$$

where $dB(t)$ is a p-dimensional Wiener process.
In our model, it reads

$$dX_{M_g}(t) = \big(\nu(t) - \mu_{M_g} X_{M_g}(t)\big)\, dt + \sqrt{\nu(t) + \mu_{M_g} X_{M_g}(t)}dB(t). \qquad (1.13)$$

The CLE provides a continuous time - continuous state-space approximation of the process. It turns out that in this simple case the CLE leads also to a transition density with closed form, and is therefore tractable. The see this, assume again $\nu(t) = \nu$, and define the change of variable (Wilkinson, 2012, Chapter 8)

$$Y(t) = X_{M_g}(t) + \frac{\nu}{\mu_{M_g}}.$$

Using Ito's Lemma for the change of variable, we obtain

$$dY(t) = \mu_{M_g}\left(\frac{2\nu}{\mu_{M_g}} - Y(t)\right) dt + \sqrt{\mu_{M_g}}\sqrt{Y(t)}dB(t).$$

This process is indeed the Cox-Ingersoll-Ross process, and has a non-central χ^2 transition density (Wilkinson, 2012, Chapter 5). It also possible to obtain its mean and variance, in particular we have

$$\begin{aligned} E[Y(t)|y(0)] &= y(0)e^{-\mu_{M_g} t} + \frac{2\nu}{\mu_{M_g}}(1 - e^{-\mu_{M_g} t}) \\ V[Y(t)|y(0)] &= y(0)e^{-\mu_{M_g} t}(1 - e^{-\mu_{M_g} t}) + \frac{\nu}{\mu_{M_g}}(1 - e^{-\mu_{M_g} t})^2. \end{aligned}$$

Transforming back to $X_{M_g}(t)$, we obtain the same moments of Equations 1.11 and 1.12. Therefore, the original Binomial and Poisson convolution is approximated by the CLE with a non-central χ^2, which, as expected, matches the correct mean and variance.

The main drawback of the CLE approximation is that the transition density is still non-normal, and therefore not easily tractable for inferential purposes.

A normal transition density can be obtained with the **linear noise approximation** (LNA) (Komorowski et al., 2009; Stathopoulos and Girolami, 2013; Fearnhead et al., 2014; Anderson and Kurtz, 2011). The LNA approximates in fact the exact unknown stochastic process $X(t)$ with a Gaussian stochastic process. This is accomplished by applying the normal approximation to the Poisson process, and by replacing the hazard function with its first order Taylor expansion about the deterministic solution, thus effectively eliminating nonlinearities. Recall that Ω is the volume of the container times the Avogadro number. It can be shown that, assuming $X(0) \sim \mathcal{N}(\Omega z(0), \Omega P(0))$, $X_{LNA}(t)$ satisfies

$$X_{LNA}(t) \sim \mathcal{N}(\Omega z(t), \Omega P(t)),$$

where $z(t)$ and $P(t)$ are the solutions of

$$\begin{aligned}
\frac{dz(t)}{dt} &= S\tilde{h}(z(t), c) \\
\frac{dP(t)}{dt} &= SJ_{\tilde{h}}(z(t))P(t) + P(t)^T J_{\tilde{h}}(z(t))^T S^T + S \operatorname{diag} \tilde{h}(z(t), c) S^T,
\end{aligned}$$

and J denotes the Jacobian. More details about the full derivation are provided in Appendix A.2.

By applying the LNA to our model for the child gene mRNA, we obtain the approximation $X_{M_g}(t) \approx \mathcal{N}(n\Omega z_{M_g}(t), n\Omega P_{M_g}(t))$, where $z_{M_g}(t)$ and $P_{M_g}(t)$ are the solutions of

$$\begin{aligned}
\frac{dz_{M_g}(t)}{dt} &= -\mu_{M_g} z_{M_g}(t) + \nu(t) \\
\frac{dP_{M_g}(t)}{dt} &= -2\mu_{M_g} P_{M_g}(t) + \nu(t) + \mu_{M_g} z_{M_g}(t).
\end{aligned}$$

Assume again, for simplicity that $\nu(t) = \nu$. By solving the mean ODE, and plugging the result into the variance ODE, we obtain the mean and variance at time t, which are again the same as the exact and the CLE solution. This shows that in the case of a system involving only zero-th and first order reactions, the LNA matches exactly the mean and variance of the transition density. However the latter is still approximated by normal random variable, which may be inaccurate for the characteristics related to the higher moments, e.g. symmetry and kurtosis. When reactions of second and higher order are present in the system, the LNA provides just an approximation also of the mean and variance, as they depend on the higher moments. A more formal derivation of this statement is provided in Section 2.3.2 (see also Grima, 2012; Golightly & Gillespie, 2016).

1.4 The macroscopic level

For completeness, we illustrate the deterministic parallel of our model for the child gene mRNA, and how it can be related to the Hill function and the thermodynamic approach. We start from the more general case, i.e. the one allowing for cooperativity in the binding, and then move to the independent binding scenario.

1.4.1 Deterministic model and parallels

It is clear from the previous section, that the deterministic ODE limit for the child gene mRNA is given by

$$\begin{aligned}\frac{dz_{M_g}(t)}{dt} &= \tilde{\nu}(z_{P_A}(t), z_{P_B}(t)) - \mu_{M_g} z_{M_g}(t) \\ &= \frac{1}{n\Omega}\left(R'_0 z_0(t) + R'_A z_A(t) + R'_B z_B(t) + R'_{A,B} z_{A,B}(t)\right) - \mu_{M_g} z_{M_g}(t).\end{aligned}$$

We now show that $\tilde{\nu}(z_{P_A}(t), z_{P_B}(t))$ can be obtained via the thermodynamic approach. The thermodynamic approach deals directly with the steady state, by deriving the probability of each state according to the ratio between the energetic configuration of each state, and the partition sum, i.e. the sum of the energetic configurations of all the possible states. Again, we follow the setup and part of the notation presented in Tkačik and Walczac (2011), and generalise it to the two TFs scenario.

Denote as E_i the energy favouring the binding of TF i, $i = A, B$ to its corresponding binding site. Denote by ξ_i the cost of removing one molecule of TF i from the solution, where $\xi_i(t) = k_B \mathcal{T} \log z_{P_i}(t)$, k_B is the Boltzmann constant and $\mathcal{T}$ the temperature in Kelvin. Also, assume the energy of the empty promoter to be equal to 0, and that η is the amount of additional energy favouring the binding of both TFs at the same time. The partition sum would have the form

$$Z = e^0 + e^{-\beta_c(E_A - \xi_A(t))} + e^{-\beta_c(E_B - \xi_B(t))} + e^{-\beta_c(E_A + E_B + \eta - \xi_A(t) - \xi_B(t))}.$$

Define $\beta_c = 1/k_B\mathcal{T}$. The probabilities of the different states are

$$z_0(t) = \frac{1}{1 + e^{-\beta_c E_A} z_{P_A}(t) + e^{-\beta_c E_B} z_{P_B}(t) + e^{-\beta_c E_A} e^{-\beta_c E_B} e^{-\beta_c \eta} z_{P_A}(t) z_{P_B}(t)}$$

$$\begin{aligned}
z_A(t) &= \frac{e^{-\beta_c E_A} z_{P_A}(t)}{1 + e^{-\beta_c E_A} z_{P_A}(t) + e^{-\beta_c E_B} z_{P_B}(t) + e^{-\beta_c E_A} e^{-\beta_c E_B} e^{-\beta_c \eta} z_{P_A}(t) z_{P_B}(t)} \\
z_B(t) &= \frac{e^{-\beta_c E_B} z_{P_B}(t)}{1 + e^{-\beta_c E_A} z_{P_A}(t) + e^{-\beta_c E_B} z_{P_B}(t) + e^{-\beta_c E_A} e^{-\beta_c E_B} e^{-\beta_c \eta} z_{P_A}(t) z_{P_B}(t)} \\
z_{A,B}(t) &= \frac{e^{-\beta_c E_A} e^{-\beta_c E_B} e^{-\beta_c \eta} z_{P_A}(t) z_{P_B}(t)}{1 + e^{-\beta_c E_A} z_{P_A}(t) + e^{-\beta_c E_B} z_{P_B}(t) + e^{-\beta_c E_A} e^{-\beta_c E_B} e^{-\beta_c \eta} z_{P_A}(t) z_{P_B}(t)}
\end{aligned}$$

where we can obtain the same result as the deterministic approach by noting that $e^{\beta_c E_A} = \tilde{K}_A$, $e^{\beta_c E_B} = \tilde{K}_B$ and $e^{\beta_c \eta} = K_c$.

Finally, in order to have a direct comparison with the Hill function (Goutelle et al. 2008; Alon, 2007, Appendix A), we need to assume that the two TFs bind independently, i.e. $K_c = 1$. The Hill function is usually employed in order to describe input-output relationships in biochemical reactions (Goutelle et al. 2008; Alon, 2007, Appendix A). It is bounded between 0 and 1, and for one activating TF it has the general form

$$f_H(z_P(t)) = \frac{z_P(t)^n}{z_P(t)^n + K^n},$$

where n is called Hill coefficient and is related to the number of binding sites: $n = 1$ for one binding site, and it increases as a function of both the number of binding sites and the cooperativity between molecules of the same TF (Goutelle et al. 2008). K is a coefficient representing the threshold, i.e. the concentration of input required in order to increase the output by 50%.

The value obtained for a specific concentration z can be interpreted as the probability of the promoter being occupied, given a certain concentration of the TF (this is consistent with the approach presented in Bialek and Setayeshgar, 2005). The probability of the promoter *not* being occupied, given a concentration z of the TF, is then just $1 - f_H(z(t); K, n)$, and has the form

$$1 - f_H(z_P(t)) = \frac{K^n}{z_P(t)^n + K^n}.$$

Assume now $n = 1$ (since there is only one binding site for each TF), and assume that the binding of each TF is regulated by a different Hill function. It is possible to derive the equilibrium probabilities for each state of the promoter by multiplying the corresponding Hill functions (we are assuming no cooperativity in the binding)

$$z_0(t) = \frac{\tilde{K}_A \tilde{K}_B}{(z_{P_A}(t) + \tilde{K}_A)(z_{P_B}(t) + \tilde{K}_B)}$$

$$
\begin{aligned}
z_A(t) &= \frac{z_{P_A}(t)\tilde{K}_B}{(z_{P_A}(t)+\tilde{K}_A)(z_{P_B}(t)+\tilde{K}_B)} \\
z_B(t) &= \frac{\tilde{K}_A z_{P_B}}{(z_{P_A}(t)+\tilde{K}_A)(z_{P_B}(t)+\tilde{K}_B)} \\
z_{A,B}(t) &= \frac{z_{P_A}(t) z_{P_B}(t)}{(z_{P_A}(t)+\tilde{K}_A)(z_{P_B}(t)+\tilde{K}_B)}
\end{aligned}
$$

where, again, we obtain the same form of the ODE approach only by rearranging the terms. An analogous result is presented in Nachman et al. (2004).

1.5 State-space representation

In order to finalise our model for mRNA, we have to take into account the presence of measurement error, which also incorporates factors not explicitly modelled.

The Markovian structure of the process describing the evolution of the child gene mRNA levels, and the presence of measurement error, straightforwardly leads to a state-space representation of the model.

A discrete-time state-space model is characterised in the following way. Define as $x_{0:T} = \{x_0, ..., x_T\}$ the unobserved states of a state-space model with observed states $y_{0:T} = \{y_0, ..., y_T\}$. The sets $x_{0:T}$ and $y_{0:T}$ effectively represent time-series, with $x_t \in \mathbb{R}^p$, and $y_t \in \mathbb{R}^q$, $t = 0, ..., T$.

The hidden stochastic process $X_{0:T}$ is assumed to be Markovian. The observed process $Y_{0:T}$ arises as a linear or nonlinear transformation of the hidden process $X_{0:T}$, corrupted with measurement error noise. The observed states are independent between each other, conditionally on the hidden states, i.e. we have for all y_t, $\pi(y_t|x_{0:t}, y_{0:t-1}) = \pi(y_t|x_t)$ (see e.g. Petris et al., 2009, Chapter 2).

We note here that if the experimental design implies destructive sampling, the state-space structure induced is, in some sense, degenerate: we would indeed have $T+1$ independent replications of the unobserved process X, each ending at time t, and having only one corresponding observation y_t, $t = 0, ..., T$. This characteristic has to be taken into account in the inferential process. We discuss this point more extensively in Section 2.4.

Note also that both the unobserved and observed state dynamics may assume a continuous-time form.

1.5.1 Model for two observed transcription factors as regulators

We can now write our model in a state-space form. We assume additive normal noise for the time series describing the dynamic evolution of the mRNA of the child gene. We also introduce a scaling factor, denoted by κ, which accounts for any multiplicative transformation of the child mRNA data (level unobserved).

The state space representation of the model with known TFs inputs, $x_{P_A}(t)$ and $x_{P_B}(t)$ is

$$\begin{aligned} Y_{M_g,t} &= \kappa X_{M_g,t} + \kappa\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0,\sigma_\epsilon^2) \\ dX_{M_g}(t) &= \left(\nu\left(x_{P_A}(t), x_{P_B}(t)\right) - \mu_{M_g} X_{M_g}(t)\right) dt \\ &\quad + \sqrt{\nu\left(x_{P_A}(t), x_{P_B}(t)\right) + \mu_{M_g} X_{M_g}(t)} dB(t). \end{aligned} \tag{1.14}$$

The rates of the reactions, as well as the measurement error variance σ_ϵ^2, the scaling factor κ, and the initial conditions for the mean and variance of the underlying stochastic process $X_{M_g}(t)$ represent a set of parameters, which are unknown in the real data scenario, and we wish to estimate. The state-space formulation provides also a statistical framework to perform inference. This is the main focus of Chapter 2.

Chapter 2

Inference

Parameter estimation is carried out in the context of a nonlinear state-space model. Unobserved dynamics are modelled continuously, while generally only a discrete set of points is observed. Finally, experimental design has to be taken into consideration. In particular, if destructive sampling is assumed, observations come from different sets of cells, and they are therefore independent, conditionally on the parameters of the model (Stathopoulos and Girolami, 2013).

In a Bayesian context, all the information required for the inference of parameters is contained in their posterior distribution. From Bayes' theorem, the posterior distribution is proportional to the product of the prior distribution, representing prior information or beliefs about the parameters, and a likelihood term, which defines the conditional distribution of the data given the parameters. A Markov chain Monte Carlo (MCMC) algorithm can be designed in order to obtain samples from the posterior distribution when, as it is often the case, it is analytically intractable.

In the context of stochastic chemical networks, assuming that all the times and types of reactions happening in the system during the time-interval of interest are known, a *complete data likelihood* is theoretically well defined (see e.g. Wilkinson, 2012, Chapter 10). However, given the timescales of most reactions, and the available technologies, this represents a rather unlikely scenario. Moreover, the available observations generally consist of molecule counts, measured at discrete times and corrupted with a more or less relevant amount of measurement error. As outlined in Chapter 1, their evolution over time is described by a Markov jump process, whose transition densities are nevertheless often intractable. The approximations presented in Chapter 1 provide the framework for the definition of an *approximate observed data likelihood*, which, while not targeting the exact distribution of the data, may still be useful for inference.

We first introduce a general framework for posterior inference in the context of state-space models. We then describe the general methodology known as Kalman filter, available in the context of both discrete and continuous, linear state-space models to perform filtering, and obtain an estimate of the likelihood. We then present one available extension of the filter for nonlinear scenarios, namely the extended Kalman filter. Finally, we provide an application of the latter methodology to our simulation scenario and present inferential simulation results.

2.1 Bayesian inference in state-space models

We now introduce the distributions of interest for inference in state-space models. We here follow Doucet and Johansen (2009) and Wilkinson (2012, Chapter 9). Denote with Ψ the set of all the parameters of a state-space model, defined as in Section 1.5. A representation of a general state-space model dependence structure is provided in Figure 2.1.

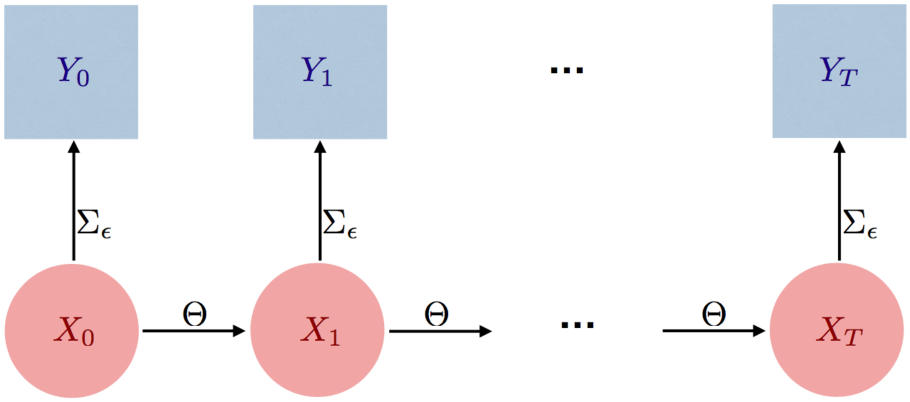


Figure 2.1: General schematic representation of the dependence structure of a state-space model, as defined in Section 1.5. Unobserved process states are indicated by X, observed by Y. Parameters and are in black, arrows indicate dependence, induced by either a linear or nonlinear transformation. Note that $\Psi = \{\Theta, \sigma_\epsilon\}$.

Using Bayes' Theorem, we can write the general posterior distribution for the parameters and the hidden states as

$$\pi(\Psi, x_{0:T}|y_{0:T}) = \frac{\pi(y_{0:T}|x_{0:T}, \Psi)\pi(x_{0:T}|\Psi)\pi(\Psi)}{\pi(y_{0:T})}.$$

State-space models are particularly useful in contexts where information is provided sequentially in time. All the distributions of interest can indeed be rewritten in a sequential form, thus allowing to update the quantities of interest as new observations become available. Thanks to the properties of state-space models outlined in Section 1.5, we can indeed rewrite the posterior distribution for the parameters and the hidden states as

$$\pi(\Psi, x_{0:T}|y_{0:T}) \propto \pi(\Psi)\pi(y_0|x_0, \Psi)\pi(x_0|\Psi)\prod_{t=1}^{T}\pi(y_t|x_t, \Psi)\pi(x_t|x_{t-1}, \Psi).$$

The marginal posterior distribution of the parameters can be obtained by integrating out the hidden states, i.e.

$$\begin{aligned} \pi(\Psi|y_{0:T}) &\propto \int_{X_{0:T}} \pi(\Psi)\pi(y_{0:T}|x_{0:T}, \Psi)\pi(x_{0:T}|\Psi)dx_{0:T} \\ &= \pi(\Psi)\pi(y_{0:T}|\Psi). \end{aligned} \tag{2.1}$$

The marginal likelihood $\pi(y_{0:T}|\Psi)$ can as well be rewritten in a sequential form

$$\pi(y_{0:T}|\Psi) = \pi(y_0|\Psi)\prod_{t=1}^{T}\pi(y_t|y_{0:t-1}, \Psi),$$

where

$$\pi(y_0|\Psi) = \int_{X_0} \pi(y_0|x_0, \Psi)\pi(x_0|\Psi)dx_0,$$

and

$$\pi(y_t|y_{0:t-1}, \Psi) = \int_{X_{t-1:t}} \pi(y_t|x_t, \Psi)\pi(x_t|x_{t-1}, \Psi)\pi(x_{t-1}|y_{0:t-1}, \Psi)dx_{t-1:t}.$$

If the focus of inference lies in both the parameters and the unobserved states, it is of interest to obtain the smoothing density of the model, for reasons which will be explained in more detail later in this chapter.

The smoothing density is defined as

$$\pi(x_{0:T}|y_{0:T}, \Psi) = \pi(x_T|y_{0:T}, \Psi)\prod_{t=0}^{T-1}\pi(x_t|x_{t+1}, y_{0:T}, \Psi),$$

where

$$\pi(x_t|x_{t+1}, y_{0:T}, \Psi) = \pi(x_t|x_{t+1}, y_{0:t}, \Psi) = \frac{\pi(x_{t+1}|x_t, \Psi)\pi(x_t|y_{0:t}, \Psi)}{\pi(x_{t+1}|y_{0:t}, \Psi)}, \tag{2.2}$$

where the equivalences follow again from Bayes' Theorem and the properties of state-space models.

2.2 Discrete-discrete filtering

The posterior distribution is usually analytically intractable, and the likelihood itself is available in a closed form only for a restricted class of state-space models, namely normal and linear state-space models. The Kalman filter recursions (Kalman, 1960) provide an algorithm to sequentially compute the likelihood in this scenario. When non-linearity/non-normality arises, other approaches can be employed. We here focus on the extended Kalman filter (EKF) (see. e.g Kulikov and Kulikova, 2014; Särkkä, 2013; Singer, 2002, and references therein). The main underlying idea is to perform a linearisation of the system by means of a first order Taylor expansion of the nonlinear functions involved. We first focus on the most basic scenario, i.e. we consider that both the unobserved and the observed states have discrete-time dynamics. We also assume, for ease of notation, that the parameters of the model Ψ are set to a known value. In practice, parameters are usually unknown quantities to be estimated. This can be done in the framework of MCMC algorithms discussed later in this chapter.

2.2.1 Kalman filter

The Kalman filter recursions (Kalman, 1960) can be applied in order to obtain the likelihood in the case of discrete, linear and normal state-space models. Specifically, consider a model of the type

$$\begin{aligned} Y_t &= FX_t + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon) \\ X_t &= GX_{t-1} + \upsilon_t \quad \upsilon_t \sim \mathcal{N}(0, \Sigma_\upsilon), \end{aligned}$$

where F is a $q \times p$ matrix, G is a $p \times p$ matrix. Let $\pi(x_0|y_0)$ to be an arbitrary prior distribution with known mean and covariance, i.e. $E[X_0|y_0] = \mu_0$, and $V[X_0|y_0] = M_0$. Then, for $t = 1, ..., T$, the following steps are recursively computed

- *Prediction step* to obtain $\pi(x_t|y_{0:t-1})$.
- *Measurement step* to obtain $\pi(y_t|y_{0:t-1})$.
- *Filtering step* to obtain $\pi(x_t|y_{0:t})$.

These steps are common to all filtering algorithms. However, if we assume $\pi(x_0|y_0)$ to be normal, linearity and normality of the noise imply that all the distributions involved are *exactly* normal. Thus, they are fully characterised by their mean and variance. Following Petris et al. (2009, Chapter 2), we now write their expressions.

- $\pi(x_t|y_{0:t-1}) = \mathcal{N}(\rho_t, P_t)$, where

$$\begin{aligned}
\rho_t &= E[X_t|y_{0:t-1}] = E[GX_{t-1} + \upsilon_t|y_{0:t-1}] = GE[X_{t-1}|y_{0:t-1}] = G\mu_{t-1}, \\
P_t &= V[X_t|y_{0:t-1}] = V[GX_{t-1} + \upsilon_t|y_{0:t-1}] = GV[X_{t-1}|y_{0:t-1}]G^T + \Sigma_\upsilon \\
&= GM_{t-1}G^T + \Sigma_\upsilon.
\end{aligned}$$

- $\pi(y_t|y_{0:t-1}) = \mathcal{N}(\alpha_t, A_t)$, where

$$\begin{aligned}
\alpha_t &= E[Y_t|y_{0:t-1}] = E[FX_t + \epsilon_t|y_{0:t-1}] = FE[X_t|y_{0:t-1}] = F\rho_t, \\
A_t &= V[Y_t|y_{0:t-1}] = V[FX_t + \epsilon_t|y_{0:t-1}] = FV[X_t|y_{0:t-1}]F^T + \Sigma_\epsilon \\
&= FP_tF^T + \Sigma_\epsilon.
\end{aligned}$$

- Finally $\pi(x_t|y_{0:t}) = \mathcal{N}(\mu_t, M_t)$, where

$$\begin{aligned}
\mu_t &= E[X_t|y_{0:t}] = \rho_t + P_tF^TA_t^{-1}(y_t - \alpha_t), \\
M_t &= V[X_t|y_{0:t}] = P_t - P_tF^TA_t^{-1}FP_t.
\end{aligned}$$

The ratio P_tF^T/A_t is often denoted with K_t, and is called the Kalman gain. The last set of moments is derived by applying Bayes' Theorem to $\pi(x_t|y_{0:t})$. We have in fact

$$\pi(x_t|y_{0:t}) = \frac{\pi(y_t|x_t)\pi(x_t|y_{0:t-1})}{\pi(y_t|y_{0:t-1})},$$

where $\pi(y_t|x_t) \sim \mathcal{N}(Fx_t, \Sigma_\epsilon)$, and the other densities involved have the form obtained in the prediction and measurement step.

2.2.2 Extended Kalman filter

When the functions involved are nonlinear, the resulting transition densities are not normal, and the Kalman filter recursions cannot be directly applied. The EKF deals with this problem by approximating the nonlinear functions with their first order Taylor expansion, effectively turning the problem back into a linear problem, where we can rely on the Kalman filter recursions. We deal with a system of the type

$$\begin{aligned}
Y_t &= FX_t + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon) \qquad (2.3) \\
X_t &= g(X_{t-1}) + d(X_{t-1})\upsilon_t \quad \upsilon_t \sim \mathcal{N}(0, I),
\end{aligned}$$

where g and d are arbitrary nonlinear functions. There are more general formulations for a nonlinear/non-normal model, e.g. nonlinearity may arise also in the measurement equation, but for clarity we here concentrate on Model 2.3, as directly linked to our model of mRNA transcription and degradation. We here follow Terejanu (2003) for the derivation.

Suppose that at time $t = 0$ we have an optimal estimate of the mean and covariance of $\pi(x_0|y_0)$, i.e. $\mu_0 = E[X_0|y_0]$ and $M_0 = V[X_0|y_0]$. Our aim is to approximate $E[X_1|y_0]$ and $V[X_1|y_0]$. The EKF does so by Taylor expanding g about the optimal estimate μ_0, i.e.

$$g(X_0) \approx g(\mu_0) + J_g(\mu_0)(X_0 - \mu_0) + \frac{1}{2}(X_0 - \mu_0)H_g(\mu_0)(X_0 - \mu_0)^T, \qquad (2.4)$$

where J_g denotes the Jacobian matrix, and H_g the Hessian matrix. Moreover, define $D(X_0) = d(X_0)d(X_0)^T$, and Taylor expand

$$D(X_0) \approx D(\mu_0) + J_D(\mu_0)(X_0 - \mu_0) + \frac{1}{2}(X_0 - \mu_0)H_D(\mu_0)(X_0 - \mu_0)^T. \qquad (2.5)$$

Truncating the Taylor expansion to the first order, and plugging it into the moments equations, we obtain the approximate mean and covariance of $\pi(x_1|y_0)$,

$$\begin{aligned}
\rho_1 &= E[X_1|y_0] = E[g(X_0) + d(X_0)\upsilon_t|y_0] \\
&\approx E[g(\mu_0) + J_g(\mu_0)(X_0 - \mu_0)|y_0] \\
&= g(\mu_0) + J_g(\mu_0)E[(X_0 - \mu_0)|y_0] \\
&= g(\mu_0), \\
P_1 &= V[X_1|y_0] = V[g(X_0) + d(X_0)\upsilon_t|y_0] \\
&= V[g(X_0)|y_0] + V[d(X_0)\upsilon_t|y_0] + \mathrm{Cov}[g(X_0), d(X_0)\upsilon_t|y_0] \\
&\quad + \mathrm{Cov}[d(X_0)\upsilon_t, g(X_0)|y_0] \\
&= V[g(X_0)|y_0] + E[D(X_0)|y_0] \\
&\approx V[g(\mu_0) + J_g(\mu_0)(X_0 - \mu_0)|y_1] + E[D(\mu_0) + J_D(\mu_0)(X_0 - \mu_0)|y_0] \\
&= J_g(\mu_0)V[(X_0 - \mu_0)|y_0]J_g(\mu_0)^T + D(\mu_0) + J_D(\mu_0)E[(X_0 - \mu_0)|y_0] \\
&= J_g(\mu_0)M_0J_g(\mu_0)^T + D(\mu_0).
\end{aligned}$$

The filter then predicts the next observation. In our specific case, given linearity and additivity of the normal noise, this does not differ from the ordinary Kalman filter of the previous section. In particular

$$\begin{aligned}\alpha_1 &= E[Y_1|y_0] = E[FX_1 + \epsilon_1|y_0] = FE[X_1|y_0] \approx F\rho_1,\\ A_1 &= V[Y_1|y_0] = V[FX_1 + \epsilon_1|y_0] = FV[X_1|y_0]F^T + \Sigma_\epsilon \approx FP_1F^T + \Sigma_\epsilon.\end{aligned}$$

The filtering step is again analogous to that of the ordinary Kalman filter, i.e.

$$\begin{aligned}\mu_1 &= E[X_1|y_1] \approx \rho_1 + P_1F^TA_1^{-1}(y_1 - \alpha_1),\\ M_1 &= V[X_1|y_1] \approx P_1 - P_1F^TA_1^{-1}FP_1.\end{aligned}$$

The recursions are then repeated for $t = 2, ..., T$.

It should be noted that the Taylor expansion has been truncated at the first order. Including second order terms leads to the second order extended Kalman filter or second order nonlinear filter (SNF) (Singer, 2002; Jazwinski, 2007, Chapter 9).

2.3 Continuous-discrete filtering

When the dynamics of the hidden states are continuous in time, but the observations are collected at discrete time-intervals, the discrete filters are no longer appropriate. The Kalman-Bucy filter (Kalman and Bucy, 1961) has been developed for continuous-discrete linear normal models. We here present the extended Kalman-Bucy filter (see e.g. Singer, 2006 and 2002; Särkkä, 2007) for the nonlinear/non-normal scenario.

2.3.1 Kalman-Bucy filter

Assume a system of the form

$$\begin{aligned}Y_t &= FX_t + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon)\\ dX(t) &= GX(t)dt + \sqrt{\Sigma_d}dB(t), \end{aligned} \tag{2.6}$$

where $dB(t)$ is, as usual, a p-dimensional Wiener process. There are indeed different derivations of the Kalman-Bucy filter, we here follow one of the approaches adopted in Särkkä (2007) and Singer (2002 and 2006). Start by writing the Euler-Maruyama approximation over a time-interval δ_t of Equation 2.6, i.e.

$$X_{t+\delta_t} = X_t + GX_t\delta_t + \sqrt{\Sigma_d}\Delta B_t + o(\delta_t),$$

where $\Delta B_t \sim \mathcal{N}(0, \delta_t)$.

We first perform the prediction step over the time-step δ_t, and then move to the continuous limit. Let $\rho(t) = E[X(t)|y_0]$ and $P(t) = V[X(t)|y_0]$. For ease of notation, we now consider $\rho(t)$ and $P(t)$ as the estimates of the mean and variance conditional on all the available observations up to time t. The prediction over δ_t leads to

$$\begin{aligned}\rho_{t+\delta_t} &= E[X_{t+\delta_t}|y_0] = E[X_t|y_0] + GE[X_t|y_0]\delta_t + o(\delta_t)\\ &= \rho_t + G\rho_t\delta_t + o(\delta_t),\end{aligned}$$

and

$$\begin{aligned}P_{t+\delta_t} &= V[X_{t+\delta_t}|y_0] = V[X_t + GX_t\delta_t|y_0] + \Sigma_d\delta_t + o(\delta_t)\\ &= V[X_t|y_0] + \delta_t^2 GV[X_t|y_0]G^T\\ &\quad + \mathrm{Cov}[X_t, GX_t\delta_t|y_0] + \mathrm{Cov}[GX_t\delta_t, X_t|y_0]\\ &\quad + \Sigma_d\delta_t + o(\delta_t)\\ &= P_t + \delta_t^2 GP_tG^T + \delta_t P_t^T G^T + \delta_t GP_t\\ &\quad + \Sigma_d\delta_t + o(\delta_t). \qquad (2.7)\end{aligned}$$

Therefore

$$\begin{aligned}\rho_{t+\delta_t} - \rho_t &= G\rho_t\delta_t + o(\delta_t),\\ P_{t+\delta_t} - P_t &= \delta_t^2 GP_tG^T + \delta_t P_t^T G^T + \delta_t GP_t + \Sigma_d\delta_t + o(\delta_t).\end{aligned}$$

By dividing both sides of the mean and variance equation by δ_t and taking the limit as $\delta_t \to 0$, we obtain the moment equations

$$d\rho(t) = G\rho(t)dt, \qquad (2.8)$$

$$dP(t) = GP(t)dt + P(t)^T G^T dt + \Sigma_d dt. \qquad (2.9)$$

There are here two important considerations. First, Equations 2.8 and 2.9 can be solved as ordinary differential equations until the next observation time-point, only because we are in the framework of a linear and normal model.

Second, note that as mentioned in Singer (2006), an Euler approximation of Equation 2.9 has a lower precision than Equation 2.7. This is due to the fact that drawing the limit eliminates the terms of order $O(\delta_t^2)$.

The measurement and filtering steps are then the same as in the ordinary

Kalman filter, and new optimal estimates $\rho(1)$ and $P(1)$ are obtained. Finally, again, these steps are iterated for $t = 2, ..., T$.

2.3.2 Extended Kalman-Bucy filter

We now follow analogous steps to derive the extended Kalman-Bucy filter (EKBF) (Kulikov and Kulikova, 2014; Singer, 2002). In particular, consider a model of the form

$$\begin{aligned} Y_t &= FX_t + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon) \\ dX(t) &= g(X(t))dt + d(X(t))dB(t). \end{aligned} \tag{2.10}$$

Again, write the Euler-Maruyama approximation over a time-interval δ_t of Equation 2.10, i.e.

$$X_{t+\delta_t} = X_t + g(X_t)\delta_t + d(X_t)\Delta B_t + o(\delta_t).$$

Assume $\rho_t = E[X_t|y_0]$ and $P_t = V[X_t|y_0]$. With steps analogous to the linear case, write the prediction over δ_t as

$$\rho_{t+\delta_t} = E[X_{t+\delta_t}|y_0] = \rho_t + E[g(X_t)|y_0]\delta_t + o(\delta_t), \tag{2.11}$$

and

$$\begin{aligned} P_{t+\delta_t} &= V[X_{t+\delta_t}|y_0] = V[X_t + g(X_t)\delta_t|y_0] + E[D(X_t)|y_0]\delta_t + o(\delta_t) \\ &= V[X_t|y_0] + \delta_t^2 V[g(X_t)|y_0] + \mathrm{Cov}[X_t, g(X_t)\delta_t|y_0] \\ &\quad + \mathrm{Cov}[g(X_t)\delta_t, X_t|y_0] + E[D(X_t)|y_0]\delta_t + o(\delta_t) \\ &= P_t + \delta_t^2 V[g(X_t)|y_0] + \delta_t \,\mathrm{Cov}[X_t, g(X_t)\delta_t|y_0] \\ &\quad + \mathrm{Cov}[g(X_t), X_t|y_0]\delta_t + E[D(X_t)|y_0]\delta_t + o(\delta_t). \end{aligned} \tag{2.12}$$

Following the same steps of the Kalman-Bucy filter, we have

$$\begin{aligned} d\rho(t) &= E[g(X(t))|y_0]dt, \\ dP(t) &= \mathrm{Cov}[X(t), g(X(t))|y_0]dt + \mathrm{Cov}[g(X(t)), X(t)|y_0]dt + E[D(X(t))|y_0]dt. \end{aligned}$$

However, these are no longer ODEs, as they include the mean and covariance of a nonlinear transformation of X (Singer, 2002). In analogy with the discrete EKF, the extended Kalman-Bucy filter performs a Taylor expansion of g and D about ρ_0. By plugging the expansions 2.4 and 2.5, truncated at the first order, into Equations

2.11 and 2.12, we obtain

$$\begin{aligned}
\rho_{t+\delta_t} - \rho_t &\approx g(\rho_t)\delta_t \\
P_{t+\delta_t} - P_t &\approx \delta_t^2 J_g(\rho_t) P_t J_g(\rho_t)^T + J_g(\rho_t) P_t \delta_t + P_t^T J_g(\rho_t)^T \delta_t \\
&\quad +D(\rho_t)\delta_t.
\end{aligned}$$

As $\delta_t \to 0$ we obtain the approximate mean and variance equations

$$\begin{aligned}
d\rho(t) &\approx g(\rho(t))dt, \\
dP(t) &\approx J_g(\rho(t))P(t)dt + P(t)^T J_g(\rho(t))^T dt + D(\rho(t))dt.
\end{aligned}$$

Note that these are the same ODEs provided by the linear noise approximation (see Section 1.3.3). We highlight here again the point that moment estimates are exact only if the functions g and D are linear with respect to $X(t)$. When this is not the case, terms of order greater than one are neglected in the Taylor expansion, and consequently in the estimate of the mean and variance. This also means, in turn, that no higher moments of $\pi(x(t)|y_0)$ than the first are included in the estimate of the mean, and no higher moments than the second are included in the estimate of the variance (see e.g. Singer, 2002, and Jazwinski, 2007, Chapter 9).

Once again, we refer to the ordinary Kalman filter for the measurement and filtering steps, and the procedure is iterated over $t = 2, ..., T$.

2.4 Destructive sampling

The filtering methodologies introduced in the previous sections provide an estimate of the likelihood when measurements are assumed to come from the same process, i.e. they are independent conditional on the parameters *and* on the hidden states. When destructive sampling is employed, the measurements come instead from independent and identically distributed copies of the process at each time point. They are therefore independent conditional only on the parameters, and the likelihood reduces to (Stathopoulos and Girolami, 2013)

$$\pi(y_{0:T}|\Psi) = \prod_{t=0}^{T} \pi(y_t|\Psi),$$

where

$$\pi(y_t|\Psi) = \int_{X_{0:t}} \pi(y_t|x_t, \Psi)\pi(x_{0:t}|\Psi)dx_{0:t}$$

$$= \int_{X_{0:t}} \pi(y_t|x_t,\Psi)\pi(x_0|\Psi)\prod_{i=1}^{t}\pi(x_t|x_{t-1},\Psi)dx_{0:t}. \qquad (2.13)$$

In practice, this means that we only need to perform a 'long' prediction step up to each observation time-point, and a single measurement step to obtain the likelihood of that specific point.

A pictorial representation of the state-space structure induced by this sampling technique is provided in Figure 2.2.

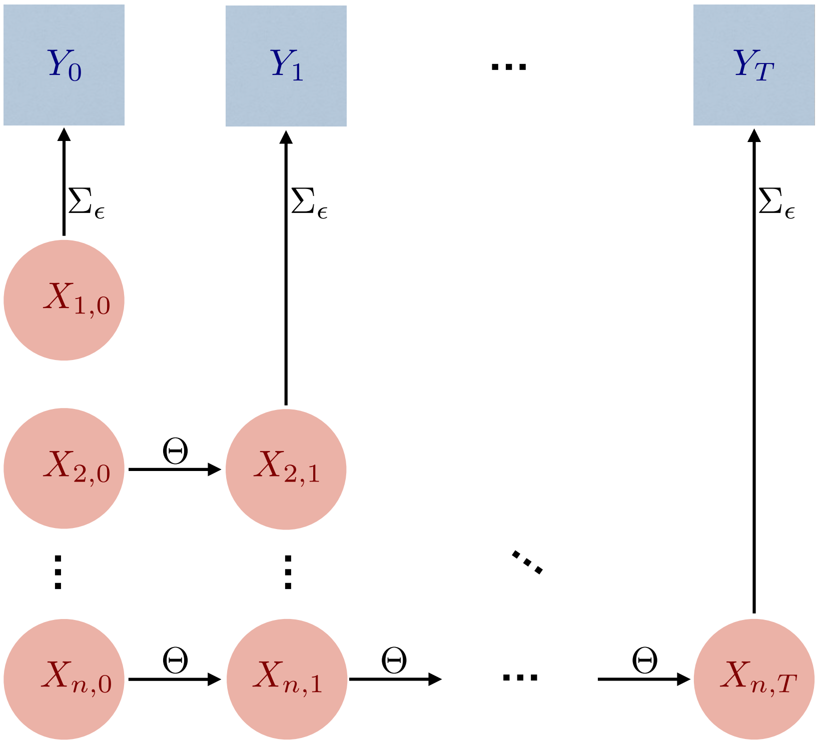


Figure 2.2: General schematic representation of the dependence structure induced by the destructive sampling. Unobserved states are indicated by X, observed by Y. Arrows indicate dependence, induced by either a linear or nonlinear transformation. Parameters involved in each transformation are superimposed to the corresponding arrows. Each unobserved state row refers to one sample, for a total of n samples. Note that $n = T + 1$.

The EKF can still be applied in order to approximate the predictive densities in the case of a non-linear/non-normal state-space model, and this allows to obtain a closed form for the integral in Equation 2.13. This approach is equivalent to Stathopoulos and Girolami (2013).

Note also that we need to know, or to include in the estimation algorithm, the mean and variance of $\pi(x_0|\Psi)$.

It is also worth deriving the form of the smoothing density in the case of independent observations. Equation 2.2 becomes, simply

$$\pi(x_t|x_{t+1}, y_{0:T}, \Psi) = \pi(x_t|y_t, \Psi),$$

where the equivalence follows from the independence of the hidden state at time t and any past observed and unobserved state of the system. In principle, since a new process generates the observed values at each time point, if the parameters are known, the past provides no information about the present unobserved state. The smoothing process then essentially reduces to filtering.

2.5 Inference for two observed transcription factors as regulators

We start by considering the scenario in which we are interested in inference about the parameters of the model, and we therefore target the posterior distribution of Equation 2.1. This distribution is analytically intractable, so inference relies on the posterior samples obtained with an appropriate Markov chain Monte Carlo (MCMC) algorithm. The key element of the MCMC algorithm is the acceptance rate, which ensures that the accepted samples, after a suitable burn-in period, come from the distribution of interest.

The algorithm is initialised with a set of parameter values Ψ, and proposes a new set of values Ψ^* from an arbitrary distribution m. In a Metropolis-Hastings scheme with random walk proposals, we have that the proposal density has the form $m(\Psi^*|\Psi)$, and the proposed values are accepted with probability (Wilkinson, 2012, Chapter 9)

$$\min\left\{1, \frac{\pi(\Psi^*)\pi(y_{0:T}|\Psi^*)m(\Psi|\Psi^*)}{\pi(\Psi)\pi(y_{0:T}|\Psi)m(\Psi^*|\Psi)}\right\}, \tag{2.14}$$

where we can drop the ratio $m(\Psi|\Psi^*)/m(\Psi^*|\Psi)$ when m is a symmetric density. Note that targeting both the unobserved states and the parameter results in a proposal of the type $(\Psi^*, x^*_{0:T})$ from an arbitrary distribution s, and an acceptance rate of the form

$$\min\left\{1, \frac{\pi(\Psi^*)\pi(y_{0:T}|\Psi^*)\pi(x^*_{0:T}|y_{0:T}, \Psi^*)s(\Psi, x_{0:T})}{\pi(\Psi)\pi(y_{0:T}|\Psi)\pi(x_{0:T}|y_{0:T}, \Psi)s(\Psi^*, x^*_{0:T})}\right\},$$

Writing $s(\Psi^*, x^*_{0:T}) = c(x^*_{0:T}|\Psi^*)m(\Psi^*)$, and choosing $c(x^*_{0:T}|\Psi^*) = \pi(x^*_{0:T}|y_{0:T}, \Psi^*)$, simplifies the above expression into the acceptance rate for the marginal posterior of the parameters of Equation 2.14. This means that a sample from the distribution of the hidden states, given the data and the parameters, can be simply obtained by drawing from the smoothing density $\pi(x^*_{0:T}|y_{0:T}, \Psi^*)$ once a new set of parameters Ψ^* has been accepted. For storage and computational time parsimony, we therefore decide to first design an appropriate MCMC algorithm for parameter estimation, and then draw posterior samples of the unobserved states given a thinned set of posterior parameter samples.

In order to obtain the evaluation of the likelihood at Ψ^*, $\pi(y_{0:T}|\Psi^*)$, we run the EKBF filter for the destructive sampling scenario. Moreover, since the Euler-Maruyama approximation of the hidden state SDE leads to discrete moment equations that are more precise than an Euler approximation of the corresponding moments ODEs (Singer, 2006), while not requiring to resort to more advanced ODE solvers, we rewrite our Model 1.15 in terms of its Euler-Maruyama approximation. Formally,

$$\begin{aligned} Y_{M_g,t} &= \kappa X_{M_g,t} + \kappa\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0,\sigma_\epsilon^2) \\ X_{M_g,t} &= X_{M_g,t-\delta_t} \\ &\quad + \left(\nu\left(x_{P_A,t-\delta_t}, x_{P_B,t-\delta_t}\right) - \mu_{M_g} X_{M_g,t-\delta_t}\right)\delta_t \\ &\quad + \sqrt{\nu\left(x_{P_A,t-\delta_t}, x_{P_B,t-\delta_t}\right) + \mu_{M_g} X_{M_g,t-\delta_t}}\,\Delta B_t. \end{aligned} \tag{2.15}$$

Note that δ_t needs generally to be chosen much smaller than the sampling interval, in order to obtain a reasonable approximation of the underlying SDE. Setting $\delta_t = 0.1\,\mathrm{h}$ seems to give reasonably accurate results in our case.

A final note is required on the two known inputs, x_{P_A} and x_{P_B}. In our simulations, the two TFs have been subjected to destructive sampling, like the child gene mRNA. However, in a scenario when both are observed, the given time-series are treated as an external input, and therefore assumed to be the same for all unobserved processes. This, in practice, has not a major influence in our case, given that we already need to assume that the TFs are close to their deterministic equilibrium in each cell (and therefore also in their aggregated values). On the other hand, it is worth keeping this in mind if the model were to be applied to non-aggregated samples, possibly more stochastic, but still undergoing destructive sampling.

2.5.1 Rescaling of the parameters

The estimation algorithm is first run on the set of simulated data in Figures 1.2 (a) and 1.2 (b), in order to test its performance. The aggregated counts are divided by their mean. Finally, to allow for the presence of measurement error, independent draws from a $\mathcal{N}(0, \sigma_\epsilon^2)$ are added to the aggregated counts at each observation time point. Rescaling of the TFs with respect to their mean levels affects only the estimate of their dissociation coefficients, while rescaling of the child mRNA affects the transcriptional rates. Recall in fact from Equation 1.5 that the mRNA transcription function is given by

$$\nu(x_{P_A,t}, x_{P_B,t}) = \left(\frac{R'_0 + R'_A \frac{x_{P_A,t}}{K_A} + R'_B \frac{x_{P_B,t}}{K_B} + R'_{A,B} \frac{1}{K_c} \frac{x_{P_A,t}}{K_A} \frac{x_{P_B,t}}{K_B}}{1 + \frac{x_{P_A,t}}{K_A} + \frac{x_{P_B,t}}{K_B} + \frac{1}{K_c} \frac{x_{P_A,t}}{K_A} \frac{x_{P_B,t}}{K_B}} \right),$$

and assume that we divide the time series of the TFs by their means. We obtain

$$\nu\left(\frac{x_{P_A,t}}{\overline{x}_{P_A}}, \frac{x_{P_B,t}}{\overline{x}_{P_B}}\right) = \left(\frac{R'_0 + R'_A \frac{x_{P_A,t}}{\overline{x}_{P_A} K'_A} + R'_B \frac{x_{P_B,t}}{\overline{x}_{P_B} K'_B} + R'_{A,B} \frac{1}{K_c} \frac{x_{P_A,t}}{\overline{x}_{P_A} K'_A} \frac{x_{P_B,t}}{\overline{x}_{P_B} K'_B}}{1 + \frac{x_{P_A,t}}{\overline{x}_{P_A} K'_A} + \frac{x_{P_B,t}}{\overline{x}_{P_B} K'_B} + \frac{1}{K_c} \frac{x_{P_A,t}}{\overline{x}_{P_A} K'_A} \frac{x_{P_B,t}}{\overline{x}_{P_B} K'_B}} \right),$$

from which is clear that $K'_A = K_A/\overline{x}_{P_A}$ and $K'_B = K_B/\overline{x}_{P_B}$.

With respect to the child gene mRNA, we assume instead the rescaling to be incorporated in the factor κ, and then move κ from the observation equation to the unobserved states, i.e. Equation 2.15 becomes

$$\begin{aligned} Y_{M_g,t} &= \tilde{X}_{M_g,t} + \kappa\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2) \\ \tilde{X}_{M_g,t} &= \tilde{X}_{M_g,t-\delta_t} \\ &\quad + \left(\tilde{\nu}\left(x_{P_A}, t-\delta_t, x_{P_B,t-\delta_t},\right) - \tilde{\mu}_{M_g}\tilde{X}_{M_g,t-\delta_t}\right)\delta_t \\ &\quad + \sqrt{\tilde{\nu}\left(x_{P_A,t-\delta_t}, x_{P_B,t-\delta_t}\right) + \tilde{\mu}_{M_g}\tilde{X}_{M_g,t-\delta_t}}\Delta B_t \end{aligned}$$

where $\tilde{X}_{M_g,t} = \kappa X_{M_g,t}$, and therefore

$$\begin{aligned} \kappa X_{M_g,t} &= \kappa X_{M_g,t-\delta_t} \\ &\quad + \left(\kappa\nu\left(x_{P_A,t-\delta_t}, x_{P_B,t-\delta_t}\right) - \mu_{M_g}\kappa X_{M_g,t-\delta_t}\right)\delta_t \\ &\quad + \sqrt{\kappa}\sqrt{\kappa\nu\left(x_{P_A,t-\delta_t}, x_{P_B,t-\delta_t}\right) + \mu_{M_g}\kappa X_{M_g,t-\delta_t}}\Delta B_t. \end{aligned}$$

This implies $\tilde{R}_0 = \kappa R'_0$, $\tilde{R}_A = \kappa R'_A$, $\tilde{R}_B = \kappa R'_B$, $\tilde{R}_{A,B} = \kappa R'_{A,B}$ and $\tilde{\mu}_{M_g} = \mu_{M_g}$. Note that the latter rescaling induces a parametrisation which is independent of the observed mean level of the data, and has a clear interpretation, as it refers to a system having on average one molecule (if $\kappa = 1/\bar{X}_{Mg}$). The stochastic kinetic parameters, relative to the actual molecules numbers, can be inferred by dividing the estimated rates by the estimated parameter κ. With respect to the TFs, the ability to infer stochastic dissociation coefficients depends on the input data: if data are provided in terms of molecule numbers, we can analogously transform back the estimated values, multiplying them by the mean levels of each TF.

2.5.2 Estimation results

We present here the MCMC results for the case in which the two TFs are observed. We assume that the TFs are known, but corrupted with measurement error, which is therefore added to their SSA simulated time-series. To handle the presence of measurement error in the TFs time-series, we perform smoothing via the *smoothing splines* function implemented in MATLAB, adopting the default smoothing bandwidth. We then perform inference via a Metropolis-within-Gibbs algorithm with single-parameter proposals and updates. Every 100 iterations, we adapt the variance of the normal distribution for the proposals in order to reach an acceptance rate of about 0.44 (Roberts and Rosenthal, 2009), regarded as optimal under some regularity conditions (Roberts and Rosenthal, 2001).

As in Roberts and Rosenthal (2009), we also implement a version of the algorithm which automatically identifies the parametrisation exploring more efficiently the support of the posterior distribution: every 10^3 iterations, we compute the mean square distance between consecutive accepted values for the current parametrisation, either with or without taking the logarithm of the parameters, and we compare it with the same quantity for the last time the algorithm has visited the alternative parametrisation. The parametrisation of the next 10^3 iterations will be the one providing the highest mean square distance. Roberts and Rosenthal (2009) also suggest to force a switch after the same parametrisation has been used for a predefined number of times, in order to avoid the possibility that the algorithm effectively gets stuck in one parametrisation. We set this value to 10^4 iterations.

Priors for all the parameters involved are set to be $Exp(100)$, with the exception of the degradation rate, for which we assume an informative prior $Ga(49.8, 0.02)$: the mean is assumed equal to the true simulation value, and the variance is in the range of those available for the *Arabidopsis thaliana* available data, as estimated by the switch tool described in Section B.2 in Appendix. The MCMC algorithm is run for 2×10^6 iterations.

Our simulation study suggests that the identification of all the parameters involved in Model 2.15 is only achievable in the presence of low measurement error, i.e. signal to noise ratio equal to 100, and assuming very frequent observations, i.e. $\Delta_t = 0.1\,\text{h}$. The cooperativity parameter K_c has in particular shown very strong correlation with almost all the parameters involved in the transcription function, and significant trade-off with the two dissociation coefficients K'_A and K'_B. We also observe trade-off between $\kappa^2\sigma^2_\epsilon$ and κ. Additional simulation results exclude a significant impact of approximate handling of the TFs inputs, as well as of aggregation. We believe that the challenging shape of the posterior density induced by correlation is the reason why the estimation procedure provides highest posterior density intervals (HPDIs, code from Vehtari, 2001) which not always contain the true parameters values, in a set of 10 independent replications of the simulated data and the estimation algorithm. The correlation matrixes for the parameters of the simulation scenarios A and B of Figure 1.2 are shown Tables 2.1 and 2.2, respectively, and we can observe pairwise correlations assuming values higher than 0.8 in several cases. Correlation is also reflected in the plots of Figures 2.3 and 2.4, for scenario A and B of Figure 1.2, respectively. At the 99 % level, the most challenging parameters are indeed K_B and K_c for scenario A, which are within the HPDIs in 6 out of 10 cases, and μ_{M_g} for scenario B, which is again within the intervals in 6 out of 10 cases. However, even when parameters do not belong to the HPDIs, their estimated values still generally capture the biological characteristics of the model - i.e. the relative changes in transcriptional rate, and the direction in the type of cooperativity. It is finally worth mentioning that in scenario B of Figure 1.2, despite the estimate of K_c belonging to the HPDI intervals in all cases at level 99%, the same intervals also include 1, i.e. the case of no cooperativity, in 6 out of 10 cases.

2.6 Discussion

To summarise the findings of our simulation study, we report the preliminary results on the transcriptional regulatory scenario with known TFs, which are observed but corrupted by measurement error. In particular, quite restrictive conditions seem to be necessary to perform inference on all of the parameters of the transcription function. The cooperativity parameter K_c is particularly problematic, and in our study can only be inferred in the presence of very frequently sampled observations, $\Delta_t = 0.1\,\text{h}$, and low signal to noise ratio, i.e. equal to 100. In our two simulation scenarios, the HPDIs (at level 99 %) do not always contain the true parameters values, possibly due to the challenging shape of the posterior density induced by strong correlations between the parameters themselves. However, the regulatory logics (induction/repression and the relative strengths) as well as the ‘direction’ of

	$\kappa R_0'$	$\frac{R_A}{R_0}$	$\frac{R_B}{R_0}$	$\frac{R_{A,B}}{R_0}$	K_c	K_A'	K_B'	μ_{M_g}	σ_ϵ^2	$\kappa E[X_{M_g}(0)]$	$\kappa V[X_{M_g}(0)]$	κ
$\kappa R_0'$	1.00	-0.48	-0.69	0.38	-0.30	0.21	0.46	0.75	-0.02	-0.40	0.11	-0.01
R_A/R_0	-0.48	1.00	0.39	-0.33	0.74	-0.90	-0.66	-0.01	-0.03	0.20	-0.02	0.04
R_B/R_0	-0.69	0.39	1.00	0.03	-0.10	-0.05	0.04	-0.42	0.02	0.35	-0.06	0.01
$R_{A,B}/R_0$	0.38	-0.33	0.03	1.00	-0.63	0.21	0.60	0.06	0.02	-0.10	0.03	-0.02
K_c	-0.30	0.74	-0.10	-0.63	1.00	-0.80	-0.81	0.08	-0.03	0.09	-0.01	0.03
K_A'	0.21	-0.90	-0.05	0.21	-0.80	1.00	0.72	-0.19	0.04	-0.09	-0.00	-0.04
K_B'	0.46	-0.66	0.04	0.60	-0.81	0.72	1.00	-0.04	0.03	-0.15	0.03	-0.03
μ_{M_g}	0.75	-0.01	-0.42	0.06	0.08	-0.19	-0.04	1.00	-0.04	-0.27	0.11	0.01
σ_ϵ^2	-0.02	-0.03	0.02	0.02	-0.03	0.04	0.03	-0.04	1.00	-0.01	0.16	-0.87
$\kappa E[X_{M_g}(0)]$	-0.40	0.20	0.35	-0.10	0.09	-0.09	-0.15	-0.27	-0.01	1.00	-0.00	0.02
$\kappa V[X_{M_g}(0)]$	0.11	-0.02	-0.06	0.03	-0.01	-0.00	0.03	0.11	0.16	-0.00	1.00	-0.18
κ	-0.01	0.04	0.01	-0.02	0.03	-0.04	-0.03	0.01	-0.87	0.02	-0.18	1.00

Table 2.1: Correlation matrix of a thinned MCMC sample from the posterior distribution of the parameters for Model 2.15, as applied to one sample dataset simulated according to scenario A of Figure 1.2. Cells containing correlations higher than 0.8 in absolute value are highlighted in grey.

	$\kappa R_0'$	$\frac{R_A}{R_0}$	$\frac{R_B}{R_0}$	$\frac{R_{A,B}}{R_0}$	K_c	K_A'	K_B'	μ_{M_g}	$\kappa^2\sigma_\epsilon^2$	$\kappa E[X_{M_g}(0)]$	$\kappa V[X_{M_g}(0)]$	κ
$\kappa R_0'$	1.00	-0.08	-0.45	-0.35	0.41	-0.29	-0.20	0.92	0.06	-0.23	0.03	-0.06
R_A/R_0	-0.08	1.00	0.80	-0.05	0.07	-0.43	-0.14	-0.02	-0.02	0.14	-0.02	0.01
R_B/R_0	-0.45	0.80	1.00	-0.19	0.22	-0.50	-0.42	-0.52	-0.04	0.11	-0.02	0.02
$R_{A,B}/R_0$	-0.35	-0.05	-0.19	1.00	-0.89	0.75	0.80	-0.08	0.01	0.20	-0.02	0.01
K_c	0.41	0.07	0.22	-0.89	1.00	-0.88	-0.90	0.06	0.01	-0.27	0.02	-0.04
K_A'	-0.29	-0.43	-0.50	0.75	-0.88	1.00	0.80	0.01	-0.01	0.18	-0.00	0.04
K_B'	-0.20	-0.14	-0.42	0.80	-0.90	0.80	1.00	0.15	-0.01	0.26	-0.02	0.04
μ_{M_g}	0.92	-0.02	-0.52	-0.08	0.06	0.01	0.15	1.00	0.05	-0.09	0.02	-0.04
$\kappa^2\sigma_\epsilon^2$	0.06	-0.02	-0.04	0.01	0.01	-0.01	-0.01	0.05	1.00	-0.01	0.28	-0.82
$\kappa E[X_{M_g}(0)]$	-0.23	0.14	0.11	0.20	-0.27	0.18	0.26	-0.09	-0.01	1.00	-0.04	0.02
$\kappa V[X_{M_g}(0)]$	0.03	-0.02	-0.02	-0.02	0.02	-0.00	-0.02	0.02	0.28	-0.04	1.00	-0.34
κ	-0.06	0.01	0.02	0.01	-0.04	0.04	0.04	-0.04	-0.82	0.02	-0.34	1.00

Table 2.2: Correlation matrix of a thinned MCMC sample from the posterior distribution of the parameters for Model 2.15, as applied to one sample dataset simulated according to scenario B of Figure 1.2. Cells containing correlations higher than 0.8 in absolute value are highlighted in grey.

K_c (attraction/repulsion in the binding), are properly identified.

We here also point out that estimation performs poorly when the dissociation coefficients of the TFs are low, most likely due to saturation: the transcription function appears flat for points in the domain which are far away from the value of the dissociation coefficient, if the region of the function support close to the dissociation coefficient is rarely or never visited. Hence, the dynamic effect of the TFs becomes less well defined, and can be partially incorporated in the basal transcrip-

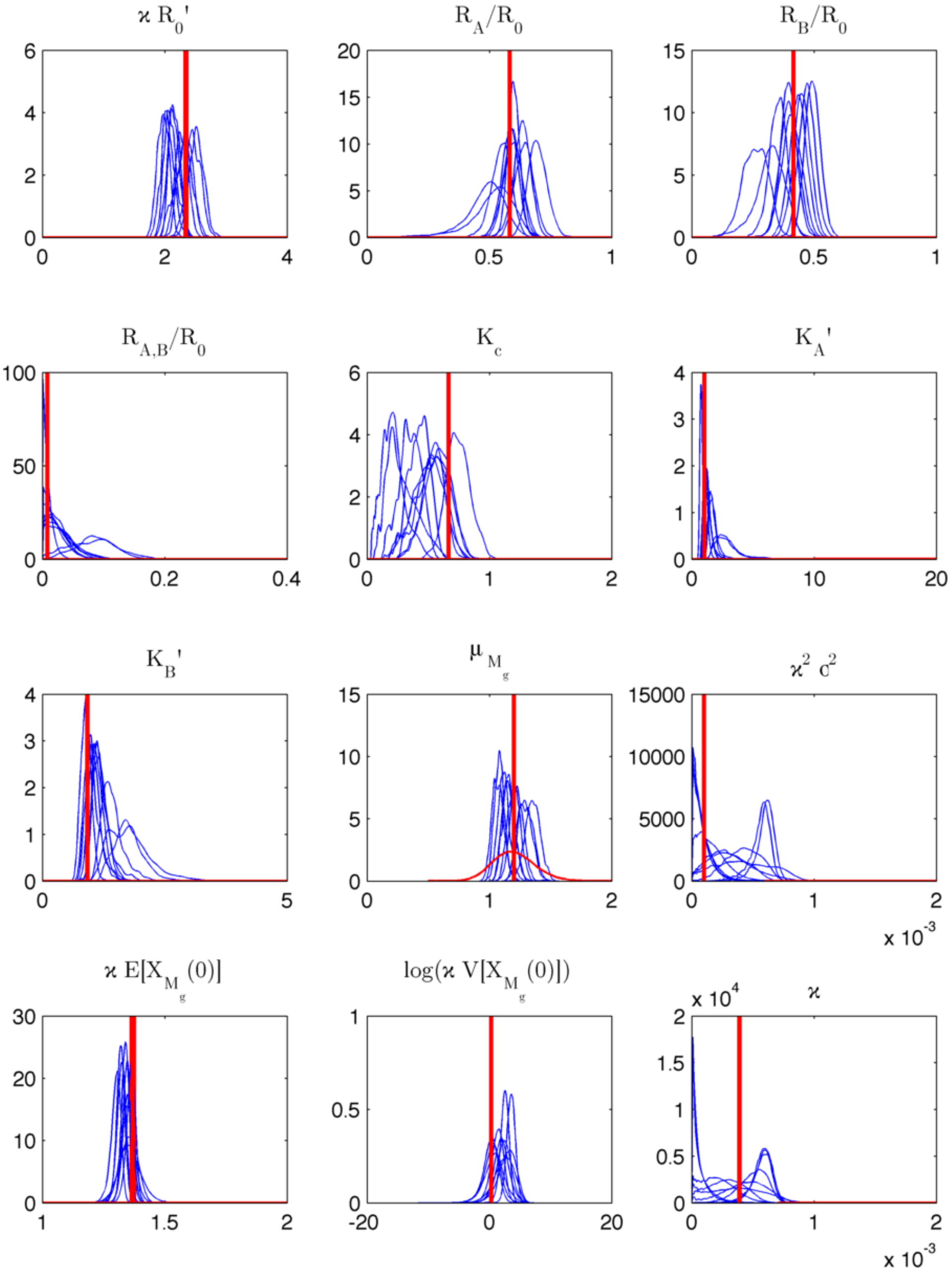


Figure 2.3: Kernel density estimate of the posterior density of the parameters. Model 2.15, as applied to data simulated according to scenario A of Figure 1.2. MCMC samples for 10 independent replications. The red vertical line is at the true value, and the prior density is also superimposed in red.

tional rate, leading to poor identifiability and therefore poor inferences. This also means, in other terms, that it is important for identifiability purposes that all the four configurations of the promoter are visited, i.e. empty, only TF A, only TF B,

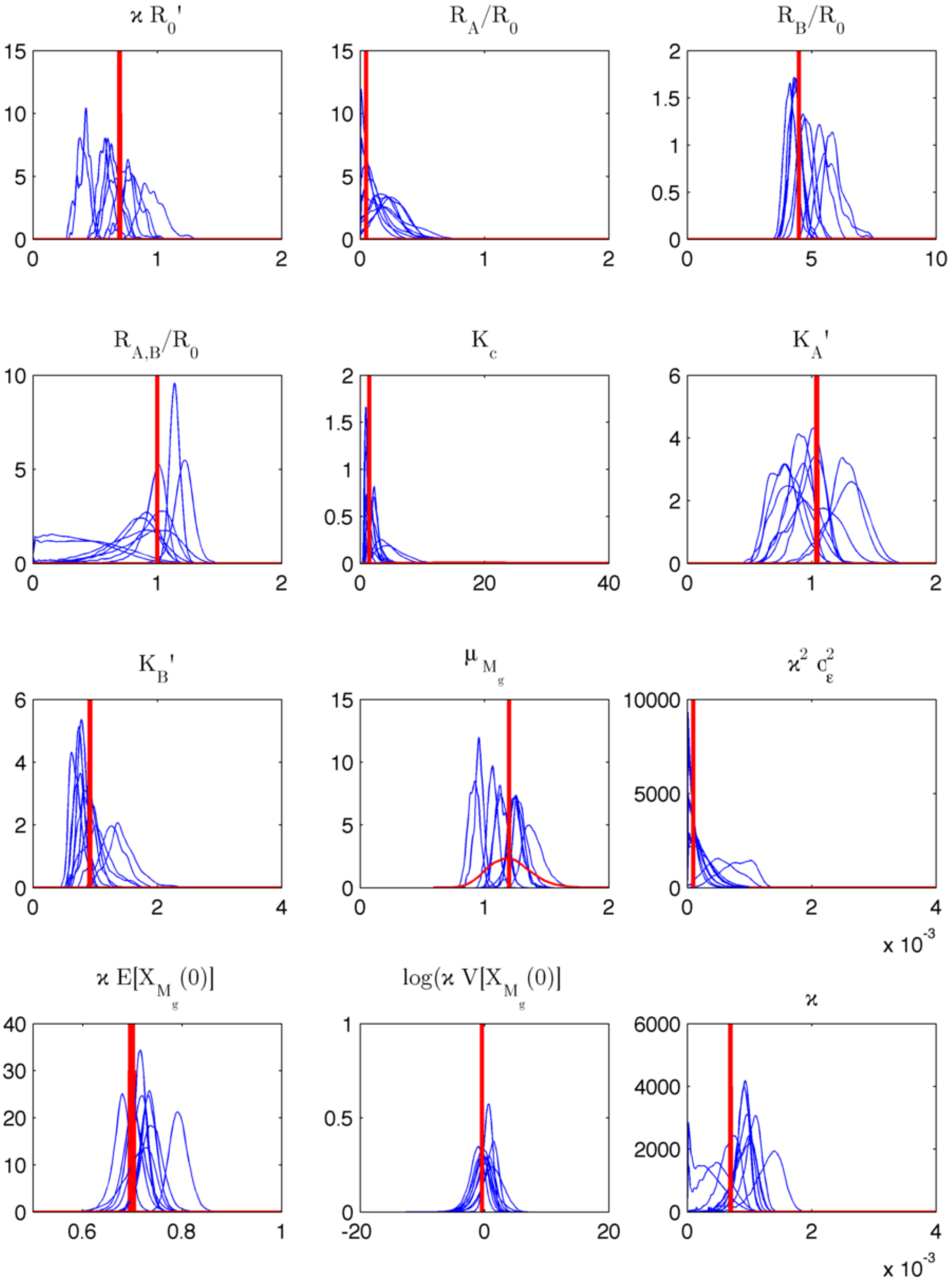


Figure 2.4: Kernel density estimate of the posterior density of the parameters. Model 2.15, as applied to data simulated according to scenario B of Figure 1.2. MCMC samples for 10 independent replications. The red vertical line is at the true value, and the prior density is also superimposed in red.

and both TF A and B bound.

Part II

Modelling transcriptional regulation in *Arabidopsis thaliana*

Chapter 3

Biological background and available data

In this chapter we present the biological framework of our work. In particular we concentrate on the concepts of transcriptional gene regulation and circadian rhythmicity, with a special focus on the *Arabidopsis thaliana* model plant, whose analysis is the aim of Chapter 4. We also introduce the *Arabidopsis thaliana* data, alongside a brief overview of the experimental techniques employed for their collection, and propose three additional models of transcriptional regulation motivated by the available data.

3.1 Transcriptional regulation

Gene regulation is the process responsible for cell differentiation in both eukaryotic and prokaryotic organisms. As stated by the central dogma of molecular biology of Crick (1958 and 1970), DNA is transcribed into RNA, and then translated into proteins, in a sequential flow of information for which only few exceptions have been observed. Since the same DNA is generally shared by all the cells of a given organism, observed differences in relative abundances of RNAs and proteins across different cell types, must be induced by a regulatory process acting at the transcriptional and post-transcriptional level (Latchman, 2005, Chapter 2).

Indeed, in Eukaryotes, mechanisms of both transcriptional and post - transcriptional regulation have been identified, although the former is believed to play the most important role (Latchman, 2005, Chapter 4).

Disruption of transcriptional regulation has been associated with a variety of diseases, such as cancer, diabetes, autoimmune and neurological diseases (Latchman, 2005, Chapter 9; for a review, see also e.g. Lee and Young, 2013), thus stressing its importance and motivating its study.

Transcriptional regulation is a complex process believed to be influenced by two main interacting players, the chromatin structure and the activity of TFs proteins (see e.g. Voss and Hager, 2014; Collingwood et al., 1999). The effect of TFs on transcription has been partially addressed in the modelling of Chapter 1.

The chromatin structure denotes the complex formed by the DNA and some nuclear proteins, called histones. A tight chromatin structure strongly impairs binding of the transcription factors to short and long distance regulatory elements, as well as the binding of the basal transcriptional complex, a structure which includes the RNA Polymerase (RNAP) enzyme and additional transcription factors, and is required to initiate transcription of the DNA into RNA (Latchman, 2005, Chapter 6).

The basal transcriptional complex is necessary in order to start transcription, but its presence is usually associated with low (baseline) transcription rates (Farnham, 2009). Binding of the TFs to the regulatory sequences seems therefore necessary to achieve high transcriptional rates, and can take place through complex spatio-temporal regulatory patterns (Farnham, 2009; Spitz and Furlong, 2012).

Short distance regulatory sequences bound by the TFs are located in a region, denoted promoter, close to the DNA sequence which has to be transcribed into RNA. TFs can also bind long distance regulatory elements, denoted enhancers, insulators, and silencers (Latchman, 2005, Chapter 7). Long range interactions have been experimentally mapped, e.g. in Sanyal et al. (2012), while the role of enhancer sequences is reviewed in Spitz and Furlong (2012). The distance of a TF binding site from the promoter region is suspected to be linked to the activity of the TF itself, as a closer binding site might imply a direct interaction with the basal transcriptional complex (Farnham, 2009). At the same time, a long-distance binding site suggests the existence of an interaction with other proteins, which could mediate the effect of the TF on the basal transcriptional complex or on the chromatin structure (Farnham, 2009). Characterisation of binding sites associated with known transcription factors is indeed an active area of research, and the main focus of Chip experiments (for a review, see Park, 2009).

A TF can either act as an activator, i.e. increase the transcription rate, or as a repressor, i.e. lower or stop transcription of a child gene (Latchman, 2005, Chapter 8). TFs can operate independently, although experiments have shown evidence of TFs binding in clusters and giving rise to complex networks of interactions (Farnham, 2009), as for example for the *Drosophila melanogaster* (Mann and Carrol, 2002). Studies have suggested that this behaviour also applies to plants (Chattopadhyay et al., 1998; Menkens et al., 1995; Michael and McClung, 2002). A TF which only influences transcription in combination with another TF is called a co-factor, and may be a co-activator or a co-repressor (Latchman, 2005, Chapter 8). Finally, binding of a TF may not have an effect on transcription, and in this case the binding is called ‘non-functional’ (see e.g. Spitz and Furlong, 2012).

Activation of transcription can be achieved by direct or mediated interaction of the TF with the basal transcriptional complex, and by opening the chromatin structure (Latchman, 2005, Chapter 8; see also Voss and Hager, 2014, for a review of interactions between TFs and chromatin structure).

Conversely, repression by a TF may be achieved by direct interaction with the basal transcriptional complex, by inhibiting the activity of activators, or by inducing a tighter chromatin structure (Latchmann, 2005, Chapter 8).

Together with the characterisation of binding sites peculiar to specific TFs, other questions are of relevance in order to enhance our understanding of transcriptional regulatory mechanisms. In this part of our work, we focus on the effect of a particular TF belonging to the *Arabidopsis Thaliana* plant, called late elongated hypocotyl (LHY). LHY belongs to the class of rhythmic genes cycling with a period of 24 hours. This type of rhythmicity is called circadian and is briefly introduced in Section 3.2.

Here we focus on experimental data aimed at elucidating the following aspects of LHY regulation

- the putative LHY target binding sequences, as provided by a Chip-seq experiment (Carré lab.);
- the effect of an increase of LHY on the transcriptional dynamics of its target genes, assessed with an induction experiment (Carré lab., see Adams et al., 2015). In this experiment, LHY protein is artificially increased, and expression levels of the regulated genes are recorded at different time points after the induction;
- the model-based inference of parameters related to transcriptional regulation, as well as the reconstruction of a putative unobserved TF, for genes which have promoters known to be bound by LHY. The unobserved TF may be either a co-factor of LHY, if LHY is consistently bound to the promoter, or a different TF, if LHY binding is non-functional. Finally, if LHY is a functional regulator, the inspection of the correlation between the reconstructed TF and LHY time-series may indicate if LHY can be assumed to be a major regulator

for the gene under study. This is the main purpose of our analysis in Chapter 4.

3.2 Circadian rhythms

Most organisms have developed a self-sustained mechanism to optimise and synchronise their biological functions in accordance with the daily transitions from light to dark, and from higher to lower temperatures. This mechanism is called the circadian clock (McClung, 2006; Harmer, 2009). A circadian oscillator, arising from a few genes linked by regulatory feedback loops, is located in each cell, and is believed to regulate mechanisms of transcriptional regulation of a wide range of downstream genes. Being generated by intracellular autonomous mechanisms, circadian rhythms persist in experimental settings with constant light and temperature, although they can be reset by a change in the environmental conditions, most notably light (McClung, 2006; Harmer, 2009). Post-transcriptional regulation, extracellular signalling, and mechanisms of synchronisation between different cells are indeed believed to play an additional important role for circadian rhythms. The latter aspects have been investigated for plants in e.g. Takahashi et al. (2015).

Among plants, a special focus has been historically applied to the study of the circadian behaviour of the model plant *Arabidopsis thaliana*. Between 5% and 40% of its genes have been shown in different studies to have a circadian rhythmic expression (Covington et al., 2008), as observed in the oscillatory behaviour of their mRNA levels, persisting under constant light and temperature.

The structure of the *Arabidopsis thaliana* central clock has been widely investigated during the recent decade, and is the focus of much ongoing research (Huang et al., 2012; Locke et al., 2006; Alabadí et al., 2001; Salome and McClung, 2004; Harmer, 2009; Adams et al., 2015). Adams et al. (2015), proposes a model comprising two main loops. In a first loop LHY/CCA1 represses itself, as well as TOC1, PRR9, PRR7 and PRR5, which in turn repress LHY/CCA1. In a second loop, TOC1 represses the Evening Complex (formed by LUX and ELF3 and ELF4) and the Evening Complex in turn represses itself, TOC1, PRR9 and PRR7.

The phases of genes exhibiting circadian behaviour in the *Arabidopsis thaliana*, seem to cover the whole circadian cycle (Harmer et al., 2000; Michael et al., 2008), and are likely to be directly linked with the function of the genes themselves in the metabolism of the plant (Harmer et al., 2000; Dodd et al., 2005). According to Michael and McClung (2002), the phase of expression is directly related to the presence of specific binding sites in the gene promoter, which are bound by TFs belonging to the central clock, and peaking at approximately the same time of the day, when activating, or in anti-phase, when repressing.

For the reasons outlined above, TFs belonging to the central clock, and their downstream genes, are of particular interest. The repressive role of TOC1 has been recently investigated (Huang et al., 2012); the aim of this work is to investigate the activity of LHY.

3.3 Data

3.3.1 Nanostring experiment

In the Nanostring experiment (Carré lab. at Warwick) the levels of mRNA of 100 *Arabidopsis thaliana* genes are sampled every two hours, for a total of 24 data-points. In addition, the level of LHY protein is recorded at the same time-points. The cells are kept under constant light for the whole duration of the experiment. The 100 genes consist of five control genes plus 95 genes which have promoters known to be bound by LHY, according to an additional Chip-seq experiment (see Section 3.3.3). Genes are chosen in order to form different groups, and in particular they are selected according to the strength of LHY binding, the phase of expression (divided into four categories), the presence of motifs and combinations of motifs (see Section 3.3.3).

The available time series mRNA measurements are aggregated over many cells in a probe. It is also worth noting that all time series measurements are recorded in relative numbers of molecules. The counts of mRNA for each species are in fact collected, and then divided by the levels of one specific transcript (UBC12), expressed approximately constantly during the experiment. We plot on the right panel of Figure 3.1 a summary of the normalised available mRNA time-series, where we can observe a wide range of phases and profiles. The left panel of Figure 3.1 gives observed LHY protein levels.

3.3.2 Induction experiment

In the Induction experiment, again performed by the Carré lab., LHY protein is induced with alcohol under constant light conditions every 4 hours, at 6 different times of the circadian day. Specifically, LHY is induced at time 0, 4, 8, 12, 16, 20 (hours) and levels of mRNA expression of the Nanostring genes, from both an induced and a control sample, are recorded 2 hours after the induction. The experiment is replicated once, meaning that two independent observations for both the

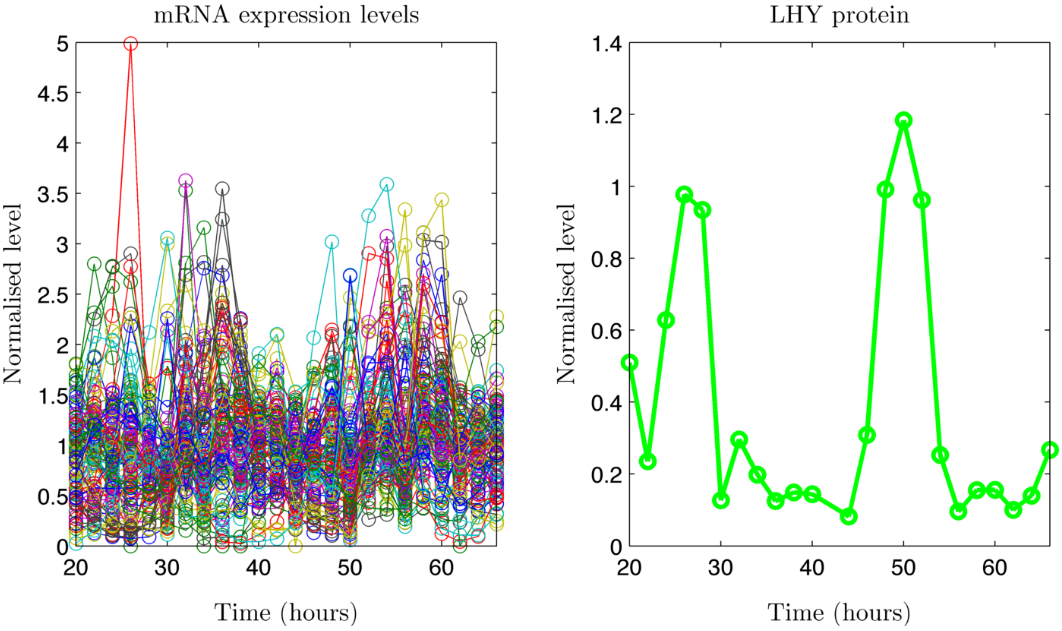


Figure 3.1: mRNA expression levels of selected *Arabidopsis thaliana* genes, rescaled by their mean level (left), LHY protein levels, in ng, taken at the corresponding time-points (right). One observation is missing at hour 42 for LHY protein. Nanostring data-set, Carré lab.

induced and the control data, are available.

Visual inspection of the distribution of the differences between the two replicates, on both the original and the logarithmic scale, allows to assess marginal normality. Results are shown in Figure 3.2. Note that we only plot absolute values as data are provided in the form of mean and standard deviation; having two replicates, it is possible to recompute the original values, but not their order. We notice that the values on the logarithmic scale seem to fulfil the normality assumption more closely than values on the original scale, in terms of kurtosis. The same visual inspection of normality on the logarithmic scale is carried out separately for each experiment, and shown in Figure 3.3. We can see that no evident differences arise between different experiments, making it sensible to assume that normality holds in all cases.

We then aim at comparing the difference between the mean expression levels of the induced and the control sample, at each time-point and for each gene, and we adopt the logarithmic transformation of the observed data to fulfil the normality requirement. If, for a given gene and for at least one time-point, a negative difference is found to be significant, then the gene is classified as repressed; if at least one difference is significantly positive, the gene is classified as induced. We be-

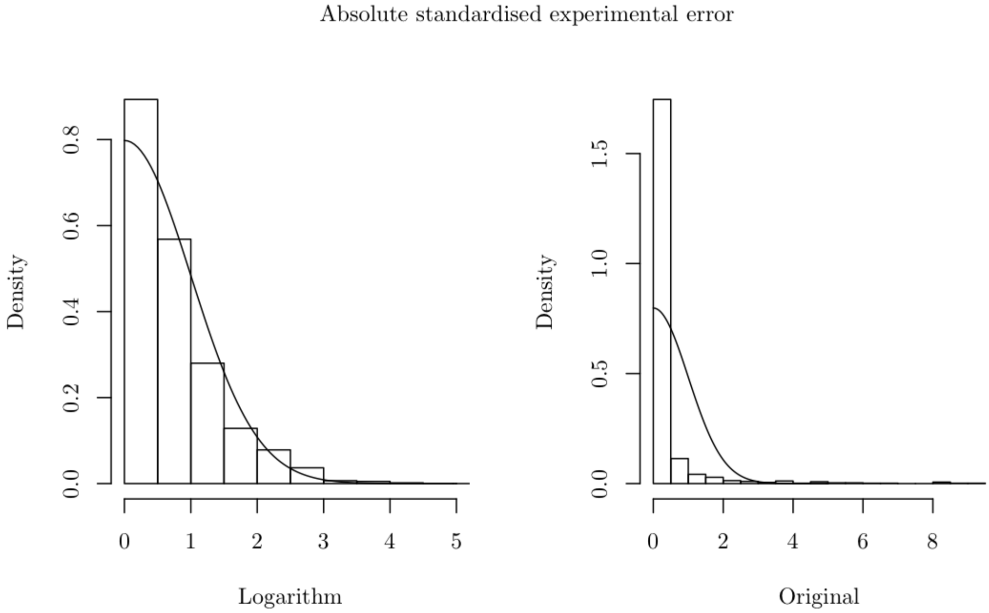


Figure 3.2: Induction experiment data-set, absolute standardised differences between two independent replicates of the same experiment (pooled across all the experiments), for the logarithmic (left) and original (right) scale values, normal distribution superimposed to the histograms. Carré lab.

lieve, moreover, that further information is provided by the number of data-points at which the gene is significantly repressed or induced: an effect of LHY at only one data-point seems to suggest the presence of additional TFs, i.e. LHY may need another protein to become abundant (or scarce) to become functional. On the other hand, a gene which is repressed by LHY at most data-points, i.e. throughout time, seems to suggest that mainly LHY is responsible for the child gene regulation. We therefore define two additional categories, namely 'consistently repressed' and 'consistently activated', if for at least five out of six time-points the gene is significantly repressed or activated, respectively. This choice allows the possibility that the gene has already a low/high transcriptional rate at one time-point, which cannot be further repressed/activated. Finally, if more than one significant difference among the six time-points is observed, but there is no agreement with respect to the sign, then neither induction nor repression is inferred.

The homoschedasticity assumption is checked by means of a Bartlett's test (Bartlett, 1937). Whenever the null hypothesis of homoschedasticity is rejected, the degrees of freedom of the t-test distribution are estimated as in Satterthwaite (1946) and Welch (1947).

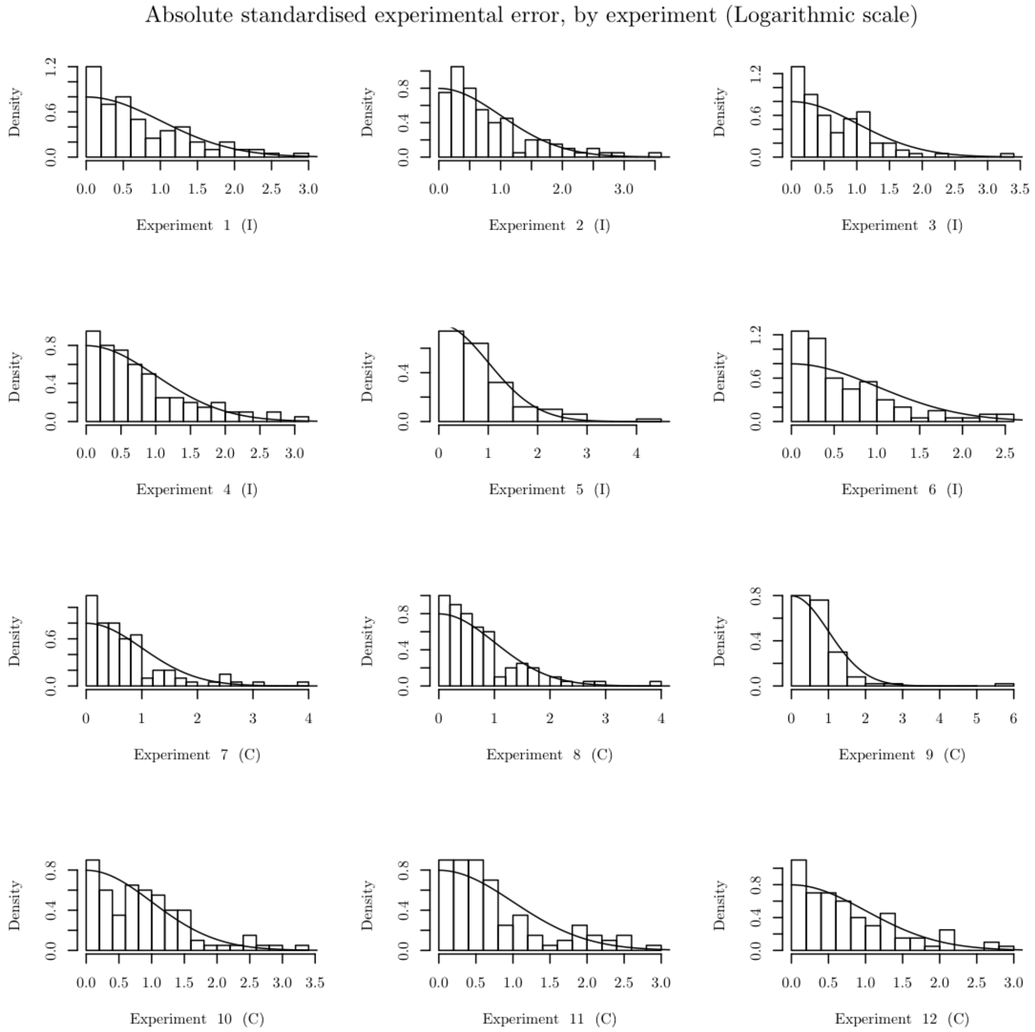


Figure 3.3: Induction experiment data-set, absolute standardised differences between two independent replicates of the same experiment, by experiment. Logarithmic scale values, normal distribution superimposed to the histograms. Carré lab.

The null hypothesis that LHY has no effect on transcription, against the alternative hypothesis that it either activates or represses transcription, is tested at six time-points, meaning that we are in the framework of multiple testing. Adopting the Bonferroni correction, the p-value for significance is then set to $\alpha/6$, where α is the level of the test.

The Bonferroni correction is aimed at controlling the so called family-wise error rate (FWER), i.e. the probability of rejecting at least once, in a set of n comparisons, a true hypothesis (Goeman and Solari, 2014). Goeman and Solari

(2014), however, show that the Bonferroni correction is conservative, i.e. the FWER is strictly smaller than α, unless all n hypotheses are true *and* there is no intersection between the events 'the p-value of experiment i is lower than α/n ' for $i = 1, ..., n$.

Table 3.1 shows a summary of the results of the induction experiment analysis for the Nanostring genes. Given the conservativeness of the Bonferroni correction, we present the results for different significance levels α. We can observe, at all significance levels, that slightly more than the 50% of the Nanostring genes are classified as 'repressed' by LHY; between the 0% and the 16%, depending on the significance level, are classified as 'consistently repressed', and between the 3% and 4 % as either 'induced' or 'significantly induced'. Overall, LHY seems therefore to influence the expression levels of approximately two-thirds of the available Nanostring genes.

Significance level	Consistently repressed	Repressed	None	Induced	Consistently induced	NA	Total
0.3 (0.05)	16	57	22	1	3	1	100
0.2 (0.03)	14	53	28	2	2	1	100
0.1 (0.02)	4	60	31	4	0	1	100
0.05 (0.01)	0	51	45	3	0	1	100

Table 3.1: Classification of the Nanostring experiment *Arabidopsis Thaliana* genes according to the induction experiment analysis, for different significance levels. Significance levels in brackets correspond to $\alpha/6$, i.e. the significance level adjusted according to the Bonferroni correction. One gene, UBC21, is used for normalisation, and it is therefore not available (NA) for the analysis. Induction experiment data-set, Carré lab.

3.3.3 Motifs

Chip-seq experiments are aimed at identifying sequences of the DNA, called motifs, that are more likely to be bound by a specific TF (for a review, see Park, 2009). Different sequences of the DNA are identified according to the presence and order of four biological compounds called nucleobases: adenosine (A), cytosine (C), guanine (G) and thymine (T). A significant DNA enrichment is statistically assessed with respect to a control. Sequences of the enriched regions are then analysed by means of motif finding algorithms, as e.g. multiple EM for motif elicitation (MEME) (Bailey et al., 2006). The motifs are finally assigned to each gene, according to their proximity to the promoter regions.

The Chip-seq experiment and analysis carried out by the Carré lab. on LHY protein identifies the following motifs:

- EE (Evening element) ((AAA)ATATCT): depending on the number of As at the beginning of the sequence, the motif is denoted as the 1A, 2A, 3A or 4A Evening Element (EE-1A, EE-2A, EE-3A or EE-4A, respectively). Association between the presence of the EE in the promoter of circadian genes and evening phases of mRNA expression has been observed (Harmer et al., 2000; Michael and McClung, 2002; Harmer and Kay, 2005; Covington et al., 2008). Moreover, experiments performed on synthetic promoters containing a luciferase reporter construct, have shown that the presence of EE motifs is sufficient to induce evening phases in the observed light intensities; on the contrary, light intensities show decreased rhythmicity when same EE motifs are mutated (Harmer and Kay, 2005). Mutation of the EE has also been observed to be linked to overall higher or lower levels of light intensity, depending on the the EEs neighbouring nucleotides sequences (Harmer and Kay, 2005);

- CBS (Circadian clock associated-1 binding site) (AAAAATCT): in the same work, Harmer and Kay (2005) challenge the earlier hypothesis that the presence of the CBS motif in the promoter of given gene, causes dawn phases of expression. Indeed, mutation of EE elements in a synthetic promoter driving evening-phased light intensities, into CBS motifs, does not significantly affect the observed phases;

- ABRE (Abscisic acid regulated element) (C/ACACGTGG/T): this motif, also known as G-Box, is bound by basic leucine zipper proteins (bZIP), and is believed to convey the effect of environmental signals on gene transcription (Menkens et al., 1995). Menkens et al. (1995) also formulate the hypothesis that the effect of the presence of the G-Box in the promoter region of a gene on its transcriptional activity, is mainly determined by the presence of additional motifs in the same promoter region, due to interactions between the corresponding TFs;

- HEX (Hexamer) (CCACGTCA or TGACGTGG): this motif is bound by two classes of leucine zipper proteins, TGA1 and GBF1, related, among the other functions, to response to light stimuli (Schindler et al., 1992).

The Chip-seq experiment provides a set of motifs that are likely to be bound by LHY. Nevertheless, Chip-seq results have their limitations, which Farnham (2009) summarises as follows:

- the assignment of a motif to the closest gene is not always accurate, as long-distance regulatory interactions may be taking place;

- the binding of a TF to a specific motif does not imply regulation, e.g. other factors may be required;
- when the effect of a TF on transcription is assessed by knocking-out the TF itself, i.e. by reducing its level, it is possible that no differences are observed in the transcription of the child gene, either because similar TFs are performing the same role, thus replacing the missing TF, or low levels of the knocked-out TF are still present and functional.

Relating the last point to the induction experiment performed on the *Arabidopsis thaliana* genes, the same reasoning can be true in the opposite scenario, i.e. if the levels of a specific TF are artificially increased: if at a given time-point the level of LHY is high in both the control and induced sample, for example because LHY is close to its peak, an additional increase in the induced sample may not have a significant influence on the child gene mRNA levels. This can be due to, for example, a low dissociation coefficient (see Section 1.3.2, for detailed quantitative explanation of this point).

Finally, Farnham (2009) also points out that a TF does not always influence transcription by binding its motifs: due to interactions with other proteins it can either bind similar motifs, or regulate transcription by interacting with other TFs, without binding to the DNA.

3.3.4 Prior information about the dissociation coefficients

Some information about the strength of LHY binding and unbinding to the EE and the CBS motifs is available. The data, calculations and interpretations of this section and Section B.1 in Appendix are based on personal communication with I. Carré. The available experimental information is summarised in Table 3.2, and seem to suggest that the overall binding strength for the EE motif increases with increasing number of As, although the unbinding is not particularly affected by the motif sequence. This is consistent with the notion that binding rates depend on the binding site, and in particular are proportional to its linear dimension (Tkačik and Walczac, 2011; Bialek and Setayeshgar, 2005).

The last column of Table 3.2 also provides a summary of the ratio between the dissociation coefficient associated with each binding site, and the average concentration of LHY. Recall that the dissociation coefficient is the ratio between the unbinding and binding rate. We can see that CBS, EE-1A and EE-2A dissociation coefficients tend to be close to LHY average levels, while the presence of EE-3A

Motif	Sequence	Binding ($M^{-1}s^{-1}$)	Unbinding (s^{-1})	Dissociation (M)	$\frac{\text{Dissociation}}{LHY}$
EE - 4A	AAAATATCT	6.6×10^5	1.7×10^{-3}	2.6×10^{-9}	0.2
EE - 3A	AAATATCT	4.2×10^5	2×10^{-3}	4.8×10^{-9}	0.37
EE - 2A	AATATCT	1.3×10^5	1.6×10^{-3}	1.2×10^{-8}	0.95
EE - 1A	ATATCT	1.2×10^5	2.1×10^{-3}	1.7×10^{-8}	1.35
CBS	AAAAATCT	7.5×10^4	1.2×10^{-3}	1.6×10^{-8}	1.24
EE-4A and EE-3A		4.8×10^6	8.6×10^{-4}	1.8×10^{-10}	1.0×10^{-2}

Table 3.2: Binding and unbinding rates for LHY protein, for selected motifs. Rates are provided in units of Moles (M) and seconds (s). I. Carré personal communication.

and EE-4A seems to correspond to a higher degree of 'stickiness', as their the corresponding dissociation coefficients are up to five times lower than the LHY average concentration. Finally, when both the EE-3A and EE-4A are present, an even stronger attraction is observed, leading to a ratio of 10^{-2}. Further details on the computation of LHY average concentration are provided in Section B.1 in Appendix.

3.4 Simulations and modeling for the *Arabidopsis Thaliana* data

In this section we present simulated data for three possible regulatory scenarios, and in particular we assume that a putative child gene can be regulated by: only LHY, LHY and an unobserved TF, and, finally, only an unobserved TF.

We assume binding and unbinding rates, as well as LHY protein molecules numbers, in the range of those provided in Section 3.3.4. Moreover, we once again reproduce destructive sampling by simulating independent and identically distributed copies of the process for each data-point.

We resort to the diffusion approximation as a simulation technique, assuming promoter equilibrium and aggregate hazards over 100 cells (see Section 1.3.2); an exact simulation of the full system is in fact computationally highly time-demanding. We refer to Section 4.1.2 for a comparison of the inferential results under the two simulation methodologies.

3.4.1 Case 1: model for known LHY as only regulator

In this simulation scenario we assume that only LHY is regulating the child gene. Although this is probably an oversimplification of the system if we consider the real data scenario, we can still postulate that LHY is the main regulator, and is therefore able to provide the observed circadian rhythmicity to at least a subset of the Nanostring genes. This model seems therefore a sensible starting point for our analysis.

We assume two possible scenarios of regulation, namely one in which LHY is a repressor, and one in which it is an activator.

Following the same steps of Section 1.3.2, the transcription function comprising only LHY as a regulator is

$$\nu\left(x_{P_{LHY},t}\right) = \frac{R'_0 + R'_{LHY}\frac{x_{P_{LHY},t}}{K_{LHY}}}{1 + \frac{x_{P_{LHY},t}}{K_{LHY}}},$$

and the state-space formulation of the model, in analogy with the Model 2.15, is

$$\begin{aligned} Y_{M_g,t} &= \kappa X_{M_g,t} + \kappa\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0,\sigma^2_\epsilon) \\ X_{M_g,t} &= X_{M_g,t-\delta_t} + \left(\nu\left(x_{P_{LHY},t-\delta_t}\right) - \mu_{M_g}X_{M_g,t-\delta_t}\right)\delta_t \\ &\quad + \sqrt{\nu\left(x_{P_{LHY},t-\delta_t}\right) + \mu_{M_g}X_{M_g,t-\delta_t}}\Delta B_t. \end{aligned} \tag{3.1}$$

Note that, in our simulations, $x_{P_{LHY}}$ is not provided as a known input. In order to obtain simulated data for LHY protein, we adopt the diffusion approximation of the system defined by reactions 14-17 in Table 1.1, which refer to transcription and translation of TF A. Substituting A with LHY, we have

$$\begin{aligned} X_{P_{LHY},t} &= X_{P_{LHY},t-\delta_t} + \left(\alpha_M X_{M_{LHY},t-\delta_t} - \mu_P X_{P_{LHY},t-\delta_t}\right)\delta_t \\ &\quad + \sqrt{\alpha_M X_{M_{LHY},t-\delta_t} + \mu_P X_{P_{LHY},t-\delta_t}}\Delta B_t \\ X_{M_{LHY},t} &= X_{M_{LHY},t-\delta_t} + \left(\nu_{LHY,i} - \mu_M X_{M_{LHY},t-\delta_t}\right)\delta_t \\ &\quad + \sqrt{\nu_{LHY,i} + \mu_M X_{M_{LHY},t-\delta_t}}\Delta B_t, \end{aligned} \tag{3.2}$$

where $\nu_{LHY,i}$, $i = 0,...,w$, denotes the transcription rate between switch time s_i and s_{i+1}, as detailed in Section B.2 in Appendix. Observed values of LHY protein are then obtained by dividing the simulated values $x_{P_{LHY}}$ by their mean level, and adding measurement error. The signal to noise ratio is set equal to $\bar{x}_{M_g}/\sigma_\epsilon = \bar{x}_{P_{LHY}}/\sigma_\epsilon = 10$.

Figure 3.4 shows two sample simulations of the system. We can see, as expected, that when LHY acts as an activator, in scenario A, roughly contemporary phases are observed for LHY and the child gene, while the repressive role of scenario B is characterised by anti-phase expression profiles.

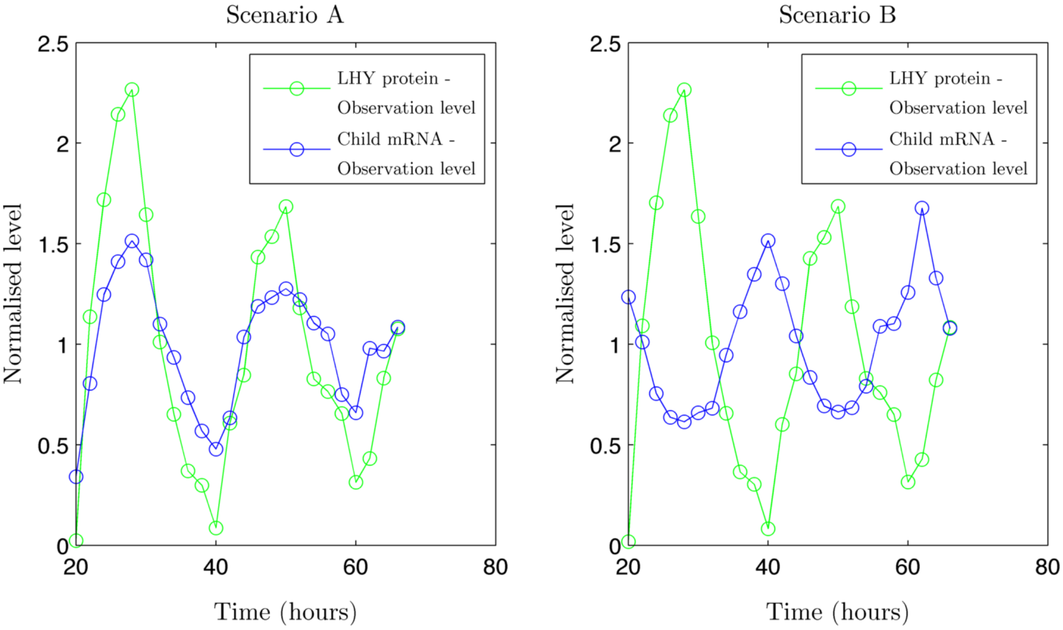


Figure 3.4: Simulated data from Model 3.1 - 3.2. Observation level, i.e. $\Delta_t = 2\,\mathrm{h}$ and levels are mean-centred and corrupted with measurement error. Signal to noise ratio ($\bar{x}_{M_g}/\sigma_\epsilon = \bar{x}_{P_{LHY}}/\sigma_\epsilon = 10$). Scenario A (left) $R'_0 = 5 \times 10^3$ molecules/h, $R'_{LHY} = 6 \times 10^4$ molecules/h; scenario B (right) $R'_0 = 5 \times 10^4$ molecules/h, $R'_{LHY} = 2.5 \times 10^3$ molecules/h. Common parameters: $K_{LHY} = 2 \times 10^6$ molecules. LHY protein is obtained from Model 3.2, assuming parameters $\nu_{LHY}(t) = [2.88 \times 10^4, 8.7 \times 10^2, 1.68 \times 10^4, 2.16 \times 10^3, 1.26 \times 10^4]$ (in molecules per hour) with switch times $Swt_A = [27, 40, 50, 61]$ (in hours), $\mu_M = 0.5\,\mathrm{h}^{-1}$, $\alpha_M = 40\,\mathrm{h}^{-1}$, $\mu_P = 0.34\,\mathrm{h}^{-1}$, $\mu_{M_{M_g}} = 1.2\,\mathrm{h}^{-1}$. Aggregate hazards for 100 cells.

3.4.2 Case 2: model for known LHY and one unknown TF as regulators

Here we provide a simulation scenario which comprises both LHY and an unobserved TF B, and where both transcription factors are dynamically influencing the transcription of the child gene. We maintain the same regulatory logics assumed in Section 1.2.1, but adapt the dissociation coefficients and system size to those provided in Section 3.3.4, which have become available only at a later stage of the project.

The transcription function $\nu(\cdot)$ has the form of Equation 1.5, where TF A is replaced by LHY, i.e.

$$\nu(x_{P_{LHY},t}, x_{P_B,t}) = \frac{R'_0 + R'_{LHY}\frac{x_{P_{LHY},t}}{K_{LHY}} + R'_B\frac{x_{P_B,t}}{K_B} + R'_{LHY,B}\frac{x_{P_{LHY},t}x_{P_B,t}}{K_{LHY}K_BK_c}}{1 + \frac{x_{P_{LHY},t}}{K_{LHY}} + \frac{x_{P_B,t}}{K_B} + \frac{x_{P_{LHY},t}x_{P_B,t}}{K_{LHY}K_BK_c}}. \quad (3.3)$$

The full simulation model is given by

$$
\begin{aligned}
Y_{M_g,t} &= \kappa X_{M_g,t} + \kappa \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2) \\
X_{M_g,t} &= X_{M_g,t-\delta_t} + \delta_t \left(\nu\left(X_{P_{LHY},t-\delta_t}, X_{P_B,t-\delta_t}\right)\right) - \mu_{M_g} X_{M_g,t}\big) \\
&\quad + \sqrt{\nu\left(X_{P_{LHY},t-\delta_t}\right), X_{P_B,t-\delta_t}) + \mu_{M_g} X_{M_g,t-\delta_t})}\Delta B_t \\
X_{P_{LHY},t} &= X_{P_{LHY},t-\delta_t} + \left(\alpha_M X_{M_{LHY},t-\delta_t} - \mu_P X_{P_{LHY},t-\delta_t}\right)\delta_t \\
&\quad + \sqrt{\alpha_M X_{M_{LHY},t-\delta_t} + \mu_P X_{P_{LHY},t-\delta_t}}\Delta B_t \\
X_{M_{LHY},t} &= X_{M_{LHY},t-\delta_t} + \left(\nu_{LHY,i} - \mu_M X_{M_{LHY},t-\delta_t}\right)\delta_t \\
&\quad + \sqrt{\nu_{LHY,i} + \mu_M X_{M_{LHY},t-\delta_t}}\Delta B_t \\
X_{P_B,t} &= X_{P_B,t-\delta_t} + \left(\alpha_M X_{M_B,t-\delta_t} - \mu_P X_{P_B,t-\delta_t}\right)\delta_t \\
&\quad + \sqrt{\alpha_M X_{M_B,t-\delta_t} + \mu_P X_{P_B,t-\delta_t}}\Delta B_t \\
X_{M_B,t} &= X_{M_B,t-\delta_t} + \left(\nu_{B,i} - \mu_M X_{M_B,t-\delta_t}\right)\delta_t \\
&\quad + \sqrt{\nu_{B,i} + \mu_M X_{M_B,t-\delta_t}}\Delta B_t,
\end{aligned}
\tag{3.4}
$$

which is the diffusion approximation of the model based on the set of reactions in Table 1.1, assuming the QSSA for the promoter states.

For completeness, we provide again in Figure 3.5 simulations from the two regulatory scenarios assumed in Section 1.2.1, but under the new dissociation coefficient values and system size. We can see that both the TFs are affecting the dynamics of the child gene, as it is evident e.g. by the temporal distribution of the phases. We refer to Section 1.2.1 for a more detailed comment of the regulatory logics induced by the chosen parameters.

As TF B is not observed in the real data scenario, our aim is now to propose an approximate model, which is able to infer the expression profile of TF B, while being parsimonious in the number of parameters.

Unobserved TF approximate model

In the context of circadian genes, TFs must be circadian in order to influence the dynamics. A quite general model for a periodic time series is provided by the Fourier series.

We provide in Section B.3 in Appendix details about the Fourier series exact representation of a sequence of real numbers of arbitrary length. In practice, however, a 'perfect' reconstruction of the unobserved TF time-series at each of the $n = 24$ mRNA data-points, would require $n/2 = 12$ parameters. This seems too demanding, given the available information, and possibly unnecessary, given the

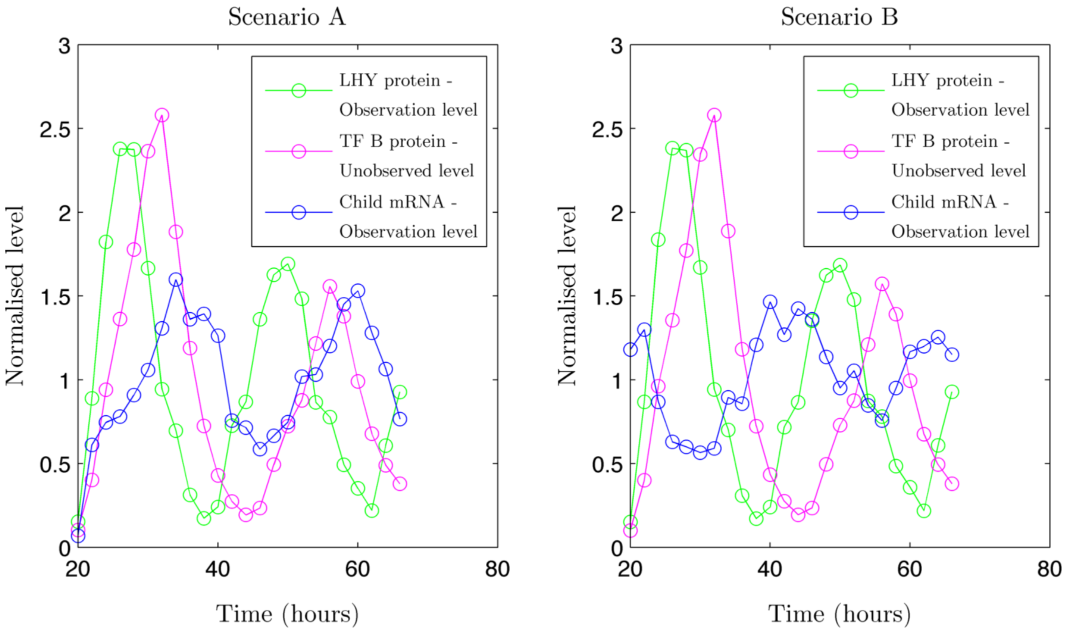


Figure 3.5: Simulated data from Model 3.4. Observation level, i.e. $\Delta_t = 2\,\mathrm{h}$ and levels of the child mRNA and LHY are mean-centred and corrupted with measurement error. Signal to noise ratio ($\bar{x}_{M_g}/\sigma_\epsilon = \bar{x}_{P_{LHY}}/\sigma_\epsilon = 10$). Scenario A (left) $R'_0 = 5 \times 10^4$ molecules/h, $R'_{LHY} = 2.5 \times 10^3$ molecules/h, $R'_B = 2 \times 10^5$ molecules/h,$R'_{LHY,B} = 5 \times 10^4$ molecules/h, $K_c = 1.5$; scenario B (right) $R'_0 = 6 \times 10^4$ molecules/h, $R'_{LHY} = 3.5 \times 10^4$ molecules/h, $R'_B = 2.510^4$ molecules/h, $R'_{LHY,B} = 5 \times 10^2$ molecules/h, $K_c = 0.66$. Common parameters: $K_{LHY} = 2 \times 10^6$ molecules, $K_B = 2 \times 10^6$ molecules. LHY and TF B are simulated assuming parameters $\nu_{LHY}(t) = [2.88 \times 10^4, 8.7 \times 10^2, 1.68 \times 10^4, 2.16 \times 10^3, 1.26 \times 10^4]$ (in molecules per hour) with switch times $Swt_A = [27, 40, 50, 61]$ (in hours), $\nu_B(t) = [2.1 \times 10^3, 1.71 \times 10^4, 2.91 \times 10^4, 9 \times 10^2, 9 \times 10^3, 1.68 \times 10^4, 2.1 \times 10^3]$ (in molecules per hour) with switch times $Swt_B = [21, 28, 45, 52, 56]$ (in hours), respectively. $\mu_M = 0.5\,\mathrm{h}^{-1}$, $\alpha_M = 40\,\mathrm{h}^{-1}$, $\mu_P = 0.34\,\mathrm{h}^{-1}$, $\mu_{M_{M_g}} = 1.2\,\mathrm{h}^{-1}$. Aggregate hazards for 100 cells.

purposes of our analysis. In analogy with harmonic regression (Prado and West, 2010, Chapter 3), we then restrict our model to fewer harmonics. We aim for a model which describes the main interesting features of the unobserved TF, while being parsimonious in the number of parameters.

A crucial choice concerns therefore the number of harmonics. We are mainly interested in the phase of the unobserved TF, and in the relative amplitudes of the cycles. By considering a total observation time of two-cycles, which corresponds to the Nanostring data scenario, this can be achieved by retaining only the first two harmonics. The first harmonic has a period of 48 hours, and models the difference in amplitude between the two cycles, while the second harmonic has a period of 24 hours, and is responsible for circadian rhythmicity. The fit is further improved by introducing additional harmonics, and in particular we have found that including the fourth harmonic can improve mean model fit for the available LHY protein time-series. We hence choose a Fourier model with three harmonics, namely 1, 2, and 4. A more general model may be attained by introducing the period length of the harmonics as an additional parameter, to be estimated. However, due to the small sample size, we do not consider this extension here.

Furthermore, higher flexibility can be achieved by moving from a deterministic model, to a stochastic one, by introducing variability in the coefficients. The Fourier coefficients can indeed be considered as additional unobserved states in our model for the child mRNA, accounting for the unobserved TF time-evolution and having their own state-space representation. Focusing on the model with three harmonics for the unobserved TF, we have the following state-space representation (Prado and West, 2010, Chapter 4)

$$\begin{aligned} X_{P,t} &= a_0 + AX_{har,t} \qquad (3.5)\\ X_{har,t} &= BX_{har,t-\delta_t} \end{aligned}$$

where a_0 is the mean level of X_P, X_{har} is a 6×1 vector accounting for the time evolution of the harmonics, $A = [1, 0, 1, 0, 1, 0]$ and B=blockdiag(H_1,H_2,H_4), with

$$H_j = \begin{bmatrix} \cos(\alpha j) & \sin(\alpha j) \\ -\sin(\alpha j) & \cos(\alpha j) \end{bmatrix}.$$

The parameter α is given by $2\pi/n$ and corresponds to the frequency of the first harmonic. The observation equation for $X_{P,t}$ does not include measurement error, as this is just a building block of the full model including the child gene mRNA. The destructive sampling induced dependence structure implies that no update of the mean and variance of the unobserved states is performed as new observations become available. Moreover, we do not assume any additive noise term in the evolution of the states X_{har}, as, we do not assume that the Fourier coefficients may be changing over time.

We check the suitability of the assumed Fourier model by fitting Model 3.5 to the values simulated according to the mechanistic Model 3.4. Parameters are estimated by means of an MCMC algorithm, assuming $\mathcal{N}(0, 20^2)$ priors for the variance of the Fourier coefficients initial conditions, on the logarithmic scale. As for the the mean of the initial conditions, we bound the parameter space so that any combination of Fourier coefficients giving rise to any negative predicted mean-point, is rejected; hence, we effectively induce a uniform multivariate prior, whose boundaries are hard to define analytically ‘a priori’ . This choice is motivated as follows.

Our ultimate goal is to incorporate the unobserved TF into the full model comprising the child gene mRNA. In the full model, the unobserved TF mean prediction therefore serves as an input for the child mRNA mean and variance predictions; a negative proposed value may induce a negative transcription function, and therefore a negative variance for the child gene mRNA, for some parameters combinations. A negative variance is statistically meaningless, and causes computational issues. A negative mean for the levels of the TF and the child mRNA is, moreover, not biologically meaningful. We note, however, that the assumed constraint does not completely rule out negative values for the unobserved TF profile, due to the variance. This seems however a minor drawback, common also to more refined modelling approaches, as e.g. the diffusion approximation of the immigration and death process itself (Wilkinson, 2012, Chapter 5).

The predictive fit of simulated TF B is shown in Figure 3.6. Table 3.3 provides median estimates and HPDIs. For details on the computation of the likelihood, we refer to Chapter 2.

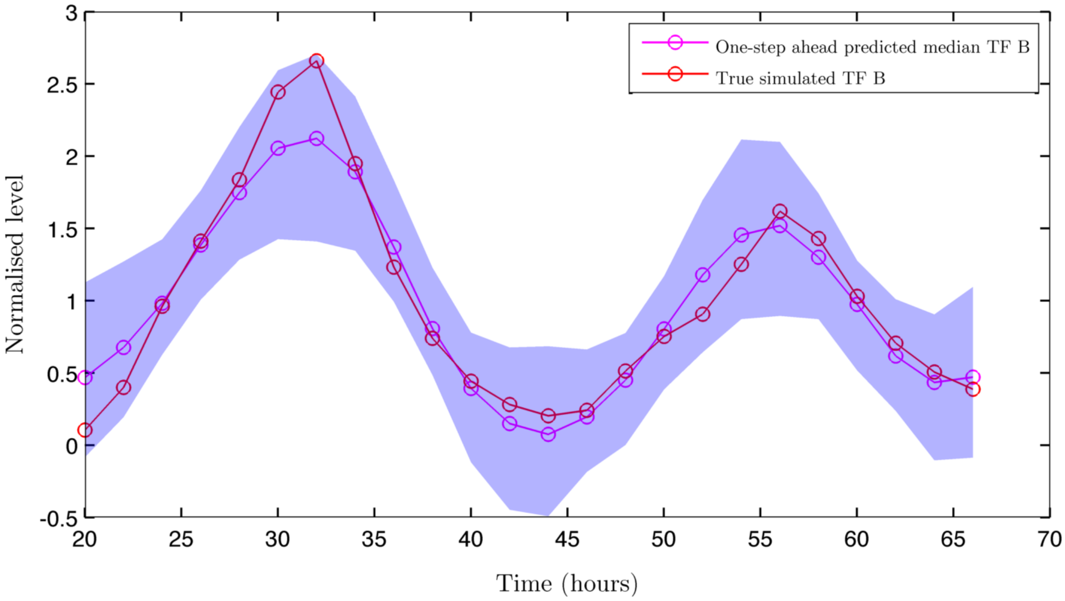


Figure 3.6: One-step ahead predictive density (mean and 95% HPDIs) for Model 3.5, as applied to one sample simulation of TF B according to Model 3.4, which defines a fully mechanistic model. True simulated TF B is superimposed in red. Simulation parameters are set to $\nu_B(t) = [2.1 \times 10^3, 17.1 \times 10^3, 29.1 \times 10^3, 9 \times 10^2, 9 \times 10^3, 16.8 \times 10^3, 2.1 \times 10^3]$ (in molecules per hour) with switch times $Swt_B = [21, 28, 45, 52, 56]$ (in hours). $\mu_M = 0.5\,\mathrm{h}^{-1}$, $\alpha_M = 40\,\mathrm{h}^{-1}$, $\mu_P = 0.34\,\mathrm{h}^{-1}$, $\mu_{M_{M_g}} = 1.2\,\mathrm{h}^{-1}$.

We note from the estimated parameters and fit that the model seems to capture the dynamics of the unobserved TF, to a reasonable degree. A median

Parameter	Median	95% HPDI
$V[X_{har,1}(0)]$	-15.78	[-44.34, -3.63]
$V[X_{har,2}(0)]$	-9.93	[-40.07, -1.83]
$V[X_{har,3}(0)]$	-2.79	[-24.23, -0.84]
$V[X_{har,4}(0)]$	-14.28	[-37.88, -3.00]
$V[X_{har,5}(0)]$	-8.86	[-30.61, -1.23]
$V[X_{har,6}(0)]$	-13.65	[-39.39, -3.22]
$E[X_{har,1}(0)]$	0.21	[7.42×10^{-2}, 0.31]
$E[X_{har,2}(0)]$	-0.76	[-0.88, -0.65]
$E[X_{har,3}(0)]$	4.59×10^{-2}	[-4.62×10^{-2}, 0.15]
$E[X_{har,4}(0)]$	0.30	[0.20, 0.46]
$E[X_{har,5}(0)]$	1.82×10^{-3}	[-0.10, 9.34×10^{-2}]
$E[X_{har,6}(0)]$	4.62×10^{-2}	[-6.06×10^{-2}, 0.16]

Table 3.3: Medians and 95% HPDIs for the parameters of Model 3.5, as applied to one sample simulation of TF B according to Model 3.4, which defines a fully mechanistic model. Simulation parameters are set to $\nu_B(t) = [2.1 \times 10^3, 1.71 \times 10^4, 2.91 \times 10^4, 9 \times 10^2, 9 \times 10^3, 1.68 \times 10^4, 2.1 \times 10^3]$ (in molecules per hour) with switch times $Swt_B = [21, 28, 45, 52, 56]$ (in hours). $\mu_M = 0.5\,\mathrm{h}^{-1}$, $\alpha_M = 40\,\mathrm{h}^{-1}$, $\mu_P = 0.34\,\mathrm{h}^{-1}$, $\mu_{M_{M_g}} = 1.2\,\mathrm{h}^{-1}$.

underestimation of the first peak is observed in Figure 3.6, although compensated by variability. Moreover, we observe in Table 3.3 that the second harmonic has the largest estimate for the mean of the initial condition (median equal to -0.76), confirming the predominance of circadian periodicity.

As a final remark, the Fourier series model with one harmonic is equivalent, i.e. has the same prediction function, to an autoregressive model of order 2, AR(2), where the first autoregressive coefficient has been set to -1, and the second is given by $2\cos(\alpha)$, α being the frequency of the chosen harmonic. Moreover, in analogy with the Fourier representation, an increasing number of harmonics would correspond to additional autoregressive terms (see Prado and West, 2010, Chapter 4).

The final model including the child mRNA and the unobserved TF B has then the usual state-space representation

$$\begin{aligned}
Y_{M_g,t} &= \kappa X_{M_g,t} + \kappa\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2) \qquad (3.6)\\
X_{M_g,t} &= X_{M_g,t-\delta_t} + \delta_t \left(\nu\left(x_{P_{LHY},t-\delta_t}, a_0 + AX_{har,t-\delta_t}\right) - \mu_{M_g} X_{M_g,t-\delta_t}\right)\\
&\quad + \sqrt{\nu\left(x_{P_{LHY},t-\delta_t}, a_0 + AX_{har,t-\delta_t}\right) + \mu_{M_g} X_{M_g,t-\delta_t}}\,\Delta B_t\\
X_{har,t} &= BX_{har,t-\delta_t},
\end{aligned}$$

Note that, although the Fourier representation is linear and normal, and therefore has a normal transition density for any arbitrarily large time-interval, we still need to obtain an input value for the unobserved TF at a fairly short time-interval, the same adopted for the Euler-Maruyama approximation of the child mRNA SDE.

3.4.3 Case 3: model for one unknown TF as only regulator

Finally, we propose a simulation scenario which assumes that both LHY and TF B are binding the promoter of the child gene, but there is no dynamical influence of LHY on transcription. We obtain simulated data for this scenario by adopting Model 3.4, and letting the dissociation coefficient of LHY be very small, with respect to its mean level and to the dissociation coefficient of TF B. This effectively induces the limiting transcription function of Equation 1.8, which we recall has the form

$$\nu(x_{P_B,t}) \quad = \quad \frac{R'_{LHY} + R'_{LHY,B}\frac{x_{P_B,t}}{K_B K_c}}{1 + \frac{X_{P_B,t}}{K_B K_c}}.$$

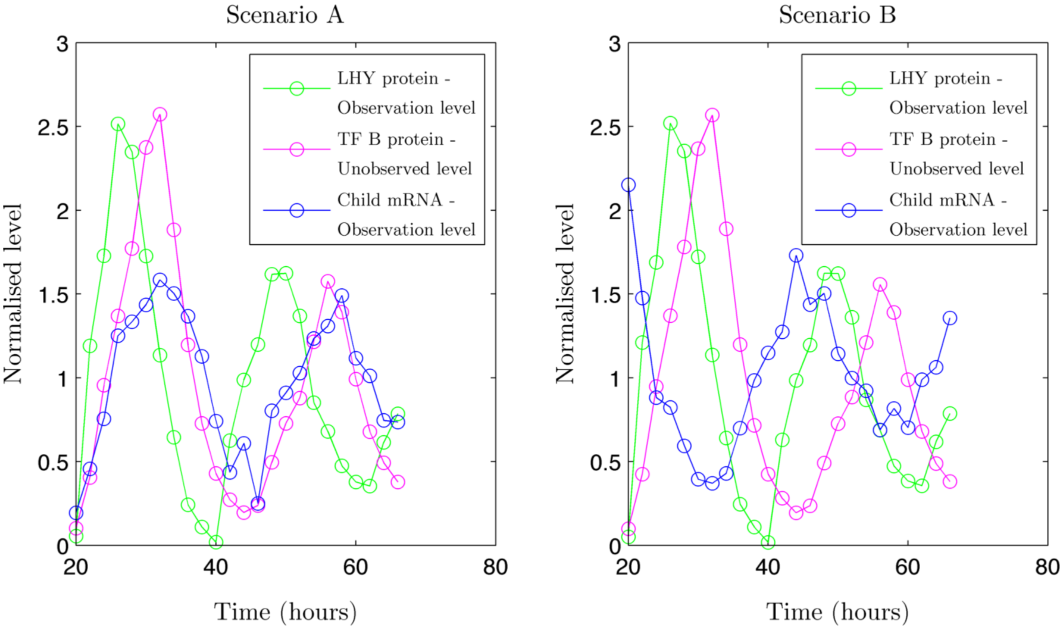
The assumed dissociation coefficient of LHY is in this scenario well below the range provided in Section 3.3.4. However, the values provided may not be very reliable, due to the lack of experimental replicates, and they do not consider the case of multiple binding sites: if cooperativity in the binding between molecules of the same TF is low, and in particular if repulsion is in place, the resulting model would be equal to a model for one binding site, but with a lower dissociation coefficient. To see this, replace TF B with LHY in Equation 3.3, and let $K_c \to \infty$. It has to be noted, however, that attractive cooperativity between molecules of the same TF is generally assumed, which leads to a Hill-type transcription function (see Section 1.4.1). In the case of strong cooperativity, therefore, the dissociation coefficient would be unaffected.

This scenario serves also as a general example for the case in which LHY has no dynamical effect on the child gene, either because the dissociation coefficient is low with respect to LHY mean level, or because LHY binding is non-functional, in which case we would obtain the limiting transcription function of Equation 1.9, which we recall is given by

$$\nu(x_{P_B,t}) \quad = \quad \frac{R'_0 + R'_B\frac{X_{P_B,t}}{K_B}}{1 + \frac{X_{P_B,t}}{K_B}}.$$

Note that the functions of Equations 1.8 and 1.9 only differ in the interpretation of the parameters. A situation in which LHY has no dynamical effect on the child gene is also suggested by the induction experiment result for a subset of the available genes (see Section 3.3.2).

Finally, a model assuming only one unobserved TF is generally applicable in situations in which the main regulating TF is not available or known. After fitting the model to the available data, reconstructed profiles can be compared with a set of available candidates. We indeed provide a study of the correlation between the inferred TF B and LHY in the Nanostring data-set in Section 4.2.2.


Scenario A

Scenario B

Figure 3.7: Simulated data from Model 3.7. Observation level, i.e. $\Delta_t = 2\,\text{h}$ and levels of the child mRNA and LHY are mean-centred and corrupted with measurement error. Signal to noise ratio ($\bar{x}_{M_g}/\sigma_\epsilon = \bar{x}_{LHY}/\sigma_\epsilon = 10$). Scenario A (left) $R'_0 = 5 \times 10^4$ molecules/h, $R'_{LHY} = 2.5 \times 10^3$ molecules/h, $R'_B = 2 \times 10^5$ molecules/h, $R'_{LHY,B} = 5 \times 10^4$ molecules/h; scenario B (right) $R'_0 = 6 \times 10^4$ molecules/h, $R'_{LHY} = 3.5 \times 10^4$ molecules/h, $R'_B = 2.5 \times 10^4$ molecules/h, $R'_{LHY,B} = 5 \times 10^2$ molecules/h. Common parameters: $K_{LHY} = 6 \times 10^2$ molecules, $K_B = 2 \times 10^6$ molecules. LHY and TF B proteins are obtained from model 3.2 assuming parameters $\nu_{LHY}(t) = [2.88 \times 10^4, 8.7 \times 10^2, 1.68 \times 10^4, 2.16 \times 10^3, 1.26 \times 10^4]$ (in molecules per hour) with switch times $Swt_A = [27, 40, 50, 61]$ (in hours), $\nu_B(t) = [2.1 \times 10^3, 1.71 \times 10^4, 2.91 \times 10^4, 9 \times 10^2, 9 \times 10^3, 1.68 \times 10^4, 2.1 \times 10^3]$ (in molecules per hour) with switch times $Swt_B = [21, 28, 45, 52, 56]$ (in hours), respectively. $\mu_M = 0.5\,\text{h}^{-1}$, $\alpha_M = 40\,\text{h}^{-1}$, $\mu_P = 0.34\,\text{h}^{-1}$, $\mu_{M_{M_g}} = 1.2\,\text{h}^{-1}$. Aggregate hazards for 100 cells.

Figure 3.7 shows two sample simulations from the scenario described: one in which TF B is an activator showing, as in the case with only LHY, a concordance in phase with the child gene mRNA, and a case in which TF B is a repressor, showing the typical anti-phase behaviour. Note that the extremely low dissociation coefficient assumed for LHY makes it uninfluential for the child gene dynamics. We return to this point in Section 4.1.3, and show that the dynamics of the child gene are indeed well fitted by the reduced model which assumes only TF B as a regulator, namely

$$\begin{aligned} Y_{M_g,t} &= \kappa X_{M_g,t} + \kappa \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2) \\ X_{M_g,t} &= X_{M_g,t-\delta_t} + \delta_t \left(\nu \left(a_0 + A X_{har,t-\delta_t} \right) - \mu_{M_g} X_{M_g,t-\delta_t} \right) \\ &\quad + \sqrt{\nu \left(a_0 + A X_{har,t-\delta_t} \right) + \mu_{M_g} X_{M_g,t-\delta_t}} \Delta B_t \\ X_{har,t} &= B X_{har,t-\delta_t}, \end{aligned} \tag{3.7}$$

where we again adopt for TF B the approximate Fourier modelling introduced in the previous section.

Chapter 4

Inference and results for *Arabidopsis Thaliana*

In this chapter we present the inferential results for both the parameters generating the artificial data introduced in Section 3.4, and the *Arabidopsis thaliana* real data introduced in Section 3.3.1. Inference is carried out in a Bayesian framework, following the methodology introduced in Chapter 2.

4.1 Inference validation on simulated data

In this section we focus on the three models of transcriptional regulation of Section 3.4, and infer the parameters that produced the synthetic data. When the assumed model requires estimation of the unobserved TF, we compare the profile inferred assuming the Fourier representation of Section 3.4.2, with the path for the unobserved TF simulated according to the 'true' mechanistic model.

The most interesting finding of our simulation study is the presence of two main modes in the posterior density of the parameters, when the model comprises the unobserved TF. In one mode, the unobserved TF acts as a repressor, while in the other mode it acts as an activator. The two modes, which seem to be approximately equally likely, correspond to inferred profiles of the TF that are in anti-phase. Further details about this aspect are provided in Section 4.1.2.

For all the three modelling scenarios of Section 3.4, we assume availability of prior information on the degradation rate of the child mRNA. In the real data application, this information is provided by fitting the switch model discussed in Section B.2 in Appendix to mRNA time-series of the Nanostring experiment genes, collected in an additional microarray experiment (Carré lab.).

Finally, recall that in our simulations both the child gene mRNA and LHY protein are divided by their mean value, corrupted with simulated measurement error, and thinned by recording values at $\Delta_t = 2\,\mathrm{h}$, in order to mimic the real data scenario. We refer to Section 2.5.1 for the effect of mean-centering on parameter estimates. The measurement error standard deviation is set so that the signal to noise ratio is equal to 10. The signal to noise ratio estimated in the real data tends to vary widely from gene to gene, but the simulation value of 10 lies within the observed range.

4.1.1 Case 1: inference for known LHY as only regulator

In this first scenario, we perform inference on the parameters of Model 3.1, i.e. assuming that there is one active TF only, which is observed and in our case behaves as LHY. The model is applied to simulated data, whose parameters values are assumed as in Figure 3.4. Recall that Figure 3.4 comprises two simulation scenarios, corresponding to two regulatory roles of LHY, namely one in which it is an activator and one in which it is repressor of the child gene mRNA transcription. Our aim is to infer the transcription function parameters, the mean and variance of the initial condition of the child mRNA, and the noise and scale parameters σ_ϵ^2 and κ.

We assume biologically sensible ranges for the parameters involved, and in particular we set a $\mathcal{N}(0, 10^2)$ prior for $\log(\kappa R_0')$, $\log(K_{LHY}')$ and $\log(\kappa E[X_{M_g}(0)])$, a half-normal distribution, obtained by folding a $\mathcal{N}(0, 10^2)$ about zero, for $\log(R_{LHY}/R_0)$, a $\mathcal{N}(0, 20^2)$ for $\log(\kappa V[X_{M_g}(0)])$ and $\log(\kappa^2 \sigma_\epsilon^2)$, and, finally, a $\mathcal{N}(0, 50^2)$ for $\log(\kappa)$, to allow for very large molecules counts.

Note that we sample all the parameters in the logarithmic space except for the degradation rate, for which a gamma prior can be formulated on the basis of the results from fitting the switch model (see Section B.2 in Appendix) to additional microarray mRNA expression profiles. This parametrisation allows to set an easily interpretable prior on the parameter $\log(R_{LHY}/R_0)$, as a negative support implies repression, while a positive one activation. Moreover, if no prior information is available, a prior centred at zero would be conservative with respect to the null hypothesis of no regulatory effect. We here set a half-normal prior, whose support, negative or positive, is provided by the induction experiment result (see Section 3.3.2). The logarithmic parametrisation seems also to generally help exploration of the posterior density.

We adopt an adaptive MCMC algorithm, where we propose jointly the parameters belonging to the transcription function $\log(\kappa R_0')$, $\log(R_{LHY}/R_0)$ and $\log(K_{LHY}')$, by adapting the covariance function of the proposal density according to the empirical covariance function of the chain accepted values (see Roberts and Rosenthal, 2009). For the remaining parameters, i.e. μ_{M_g}, $\log(\kappa E[X_{M_g}(0)])$, $\log(\kappa V[X_{M_g}(0)])$, $\log(\kappa^2\sigma_\epsilon^2)$ and $\log(\kappa)$, we adopt a single-parameter adaptive scheme, where the proposal variance of each component update is adapted in order to reach an acceptance rate of 0.44 (Roberts and Rosenthal, 2009).

The algorithm is run for a total of 3.4×10^5 iterations, and values are thinned by retaining one sample every 100 iterations, after discarding a burn-in of 5×10^4 iterations. Initial conditions for all the parameters chains are randomly drawn from the prior densities, and convergence is monitored by visual inspection of the trace-plots. Figure 4.1 shows a sample set of trace-plots, after thinning and discard of the burn-in. The mixing of the chains tend to vary between parameters and simulation runs, but we notice that the parameters belonging to the transcription function tend to generally show a slower mixing.

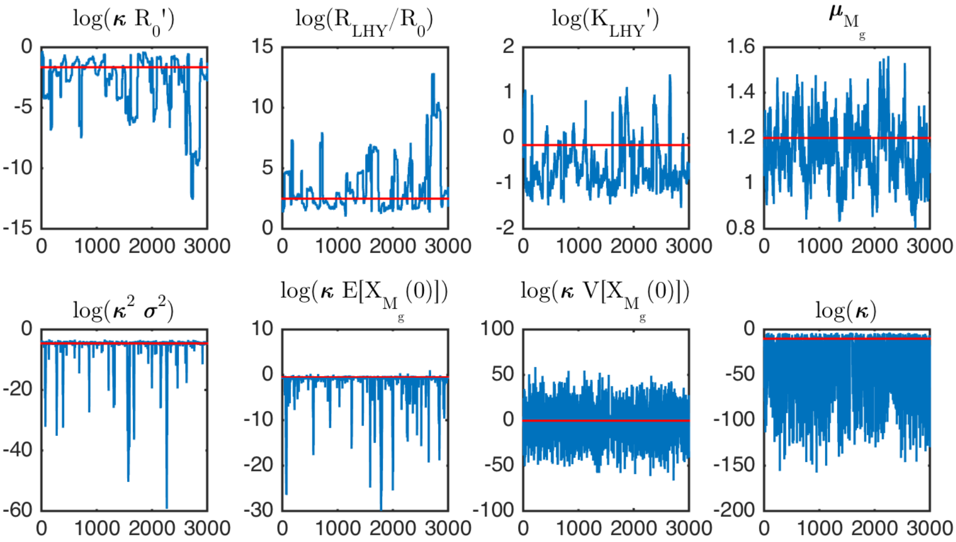


Figure 4.1: Trace-plots of the MCMC algorithm targeting the posterior densities of the parameters. Model 3.1, as applied to data simulated according to the scenarios of Figure 3.4, Scenario A. The red horizontal line is at the true value, and values are thinned by retaining one sample every 100 iterations, after discarding a burn-in of 5×10^4 iterations. Smoothed LHY input.

We first apply our estimation algorithm by assuming the LHY input to be fully observed, i.e. with no measurement error and sampled at frequency $\Delta_t = 0.1\,\mathrm{h}$. We provide in Figure 4.2 the median and 95% HPDIs for the smoothing density of the child mRNA, showing that the true simulated path is generally included in the 95 % HPDIs. A minor exception is represented by the first time-point, included in 8 cases out of 10. This is possibly due to the very high variability of the posterior density of the variance of the initial condition, as can be seen in Figures 4.3 (a) and (b), where we provide the posterior densities of all parameters. Figures 4.3 (a) and (b) also show that the model is able to reliably estimate all the parameters involved, as the true values are not included in the 95 % HPDIs in a maximum of one out of 10 cases.

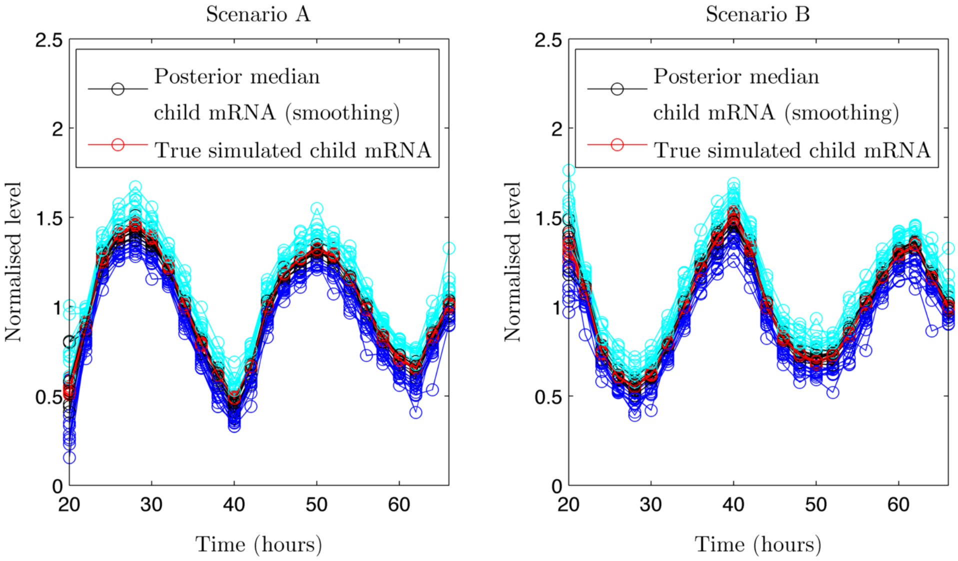
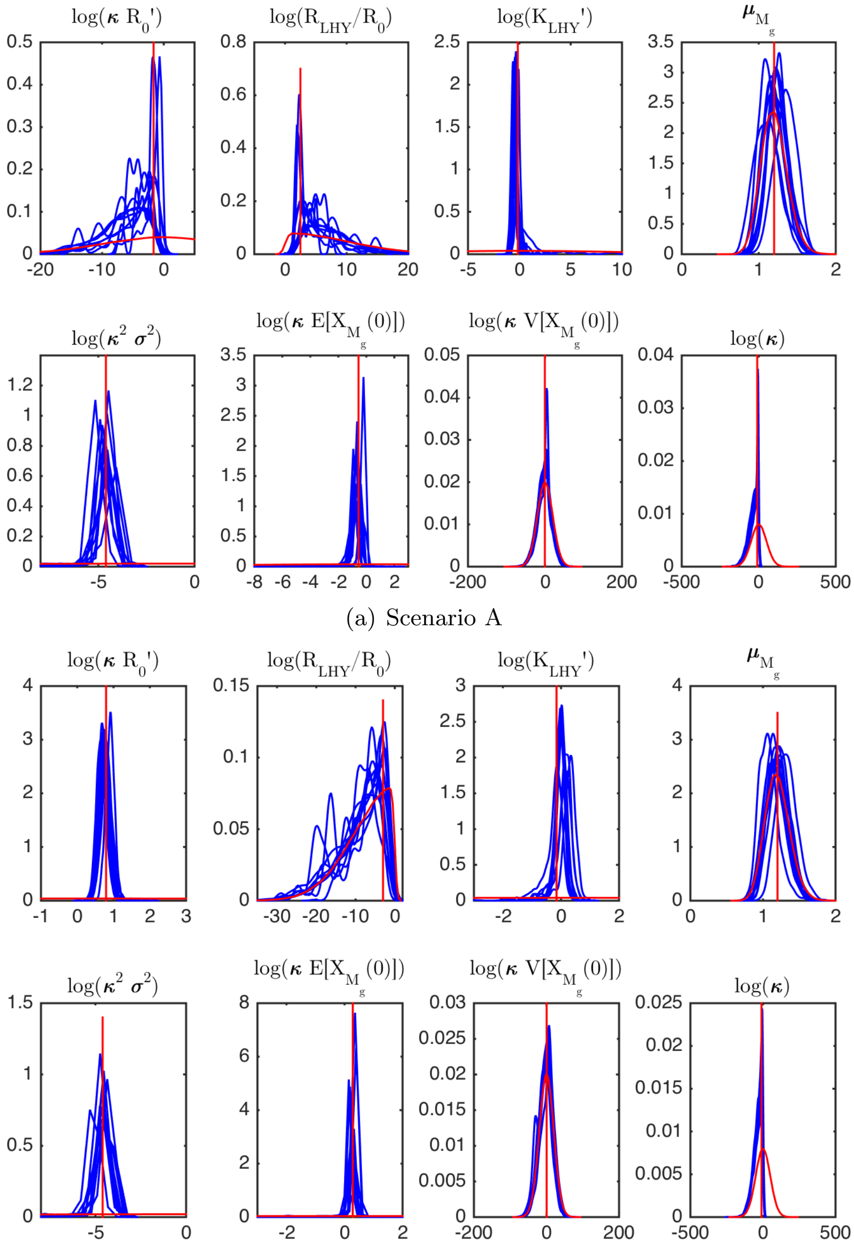


Figure 4.2: Unobserved child mRNA inference (smoothing): posterior median (black) and 95 % HPDIs (lower: blue; upper:cyan). True simulated child mRNA superimposed (red). MCMC samples for 10 independent simulations from Model 3.1, as applied to simulated data according to the scenario A (left) and B (right) of Figure 3.4. LHY input known.

In the real data scenario, however, observations of both the child gene mRNA and LHY protein are only available every two hours, and corrupted with measurement error. In order to partially eliminate the impact of measurement error, and to obtain inputs at a grid fine enough so that the Euler-Maruyama approximation for the unobserved child mRNA state holds, we perform smoothing of the corrupted LHY time-series via the *smoothing splines* function implemented in MATLAB, adopting the default smoothing bandwidth.

Figures 4.4 (a) and (b) provide the posterior densities of the parameters when adopting a smoothed LHY input. We notice that scenario A remains substantially unchanged, while scenario B seems to incorporate a lower bias in the dissociation coefficient estimate when LHY input is adopted in its smoothed form, rather than when it is fully known; we also notice, particularly in scenario B, a correlation between the parameters $\log(R_{LHY}/R_0)$ and $\log(K'_{LHY})$. It is possible that the rougher LHY input mitigates the correlation between the two parameters, making

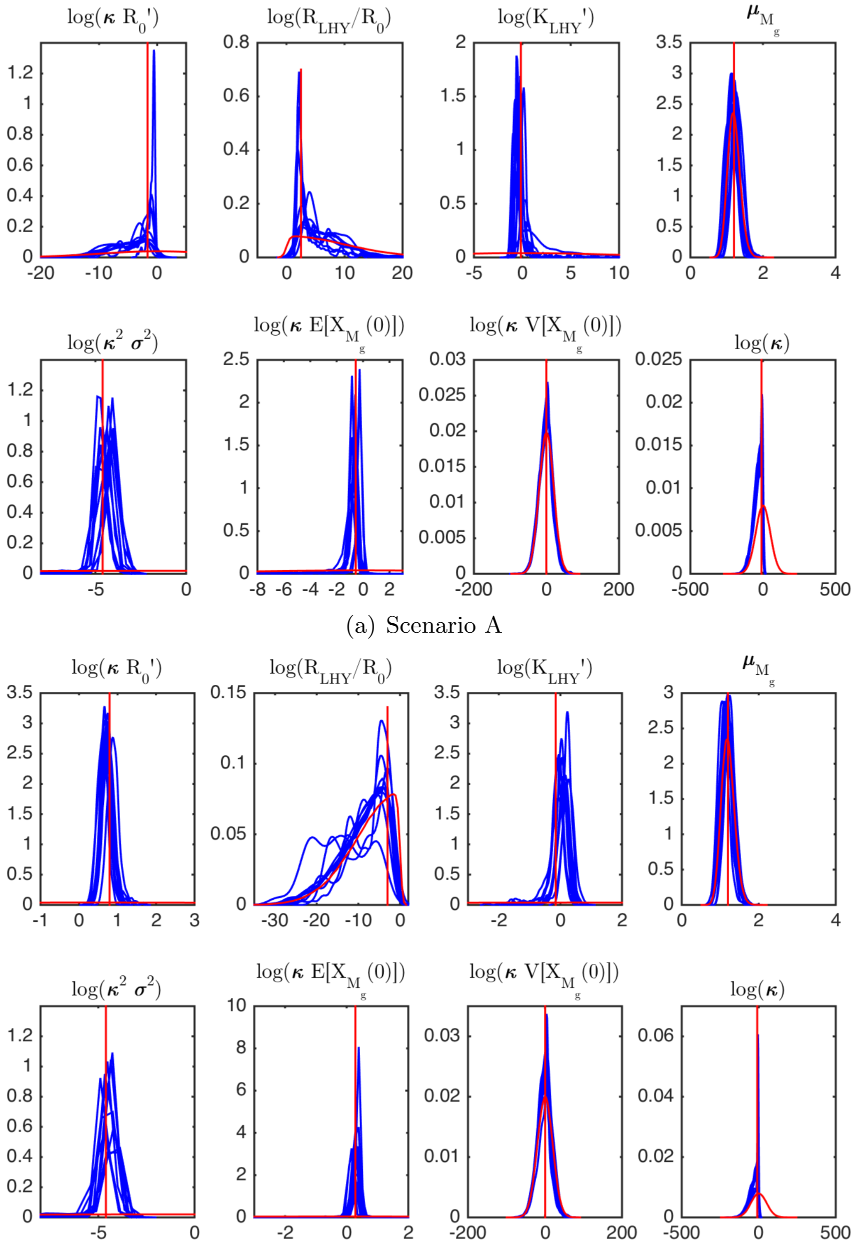


(a) Scenario A

(b) Scenario B

Figure 4.3: Kernel density estimates of the marginal posterior densities of the parameters. Model 3.1, as applied to data simulated according to the scenarios of Figure 3.4. MCMC samples for 10 independent replications for each scenario. The red vertical line is at the true value, and the prior density is also superimposed in red. LHY input known.

it easier to obtain samples from the peak region of the posterior density.



(a) Scenario A

(b) Scenario B

Figure 4.4: Kernel density estimates of the marginal posterior densities of the parameters. Model 3.1, as applied to data simulated according to the scenarios of Figure 3.4. MCMC samples for 10 independent replications for each scenario. The red vertical line is at the true value, and the prior density is also superimposed in red. Smoothed LHY input.

4.1.2 Case 2: inference for known LHY and one unknown TF as regulators

In this section we consider the more complex regulatory scenario which comprises both LHY and TF B as regulators of the child gene. We validate inference for Model

3.6, by applying it to 10 i.i.d. simulations generated according to the scenarios of Figure 3.5. Again, recall that we focus on two regulatory sub-scenarios, namely A and B, as shown in Figure 3.5.

When the unobserved TF B is introduced into the model, interest lies in both the transcription function, initial condition, scale and noise parameters, and in the reconstruction of the unobserved TF B. Moreover, the regulatory logics are identified by two further ratios, compared to the case comprising only one observed TF, namely (R_B/R_0), and $(R_{LHY,B}/R_B)$. Finally, the transcription function additionally includes the dissociation coefficient K'_B, and the cooperativity parameter K_c. This is clearly a very ambitious inferential framework for the data available.

Our study suggests, however, that inference is feasible if priors on the sign of $\log(R_{LHY}/R_0)$ and $\log(R_{LHY,B}/R_B)$, as well as of $\log(K'_{LHY})$ are available, and K_c is set to a predefined value. For our simulation parameters, setting $K_c = 1$, i.e. the case of independent binding, seems to not influence the posterior estimates of the remaining parameters.

The desired priors may be available from experimental results of the type introduced in Chapter 3. In particular, the dissociation coefficient of LHY for selected binding sites is provided in Section 3.3.4, while the induction experiment can inform on the underlying regulatory logics. We expect in fact that, if TF B is dynamically contributing to the transcriptional dynamics of the child gene, and an increase of LHY protein always leads to, for example, a decrease of transcriptional activity of the child gene, then the interaction effect should be repressive. The underlying reasoning is the following. Recall that we obtain a negative derivative of the transcription function in Equation 1.5 with respect to LHY, assuming $K_c = 1$, when

$$R'_{LHY} - R'_0 < (R'_B - R'_{LHY,B})(X_B(t)/K_B).$$

with TF B being circadian. As $X_B(t) \to 0$, the derivative is negative if $R'_{LHY} - R'_0 < 0$, which translates into $\log(R_{LHY}/R_0) < 0$; on the other hand, when levels of TF B are high we have that as $X_B(t) \to \infty$, the derivative is negative if $R'_B - R'_{LHY,B} > 0$, which translates into $\log(R_{LHY,B}/R_B) < 0$. Clearly, these are only limiting relationships, but we can postulate that a highly significant repression observed at all, or almost all, time points, well motivates prior assumptions concerning the sign of these parameters.

Despite the theoretical possibility of obtaining the priors of interest, there are no cases among the Nanostring rhythmic genes in which only one EE or CBS binding site is present in the promoter region of a given gene (thus providing a prior for the dissociation coefficient of LHY), and the same gene is consistently repressed or activated by LHY at level $\alpha = 0.1$ - the cases we classify as '-2' and '2' in Section 3.3.2, respectively. We therefore study this scenario exclusively at a simulation level.

We maintain the same priors of Section 4.1.1, for the common parameters, with the exception of $\log(K'_{LHY})$, for which we set a normal prior with mean at the true simulation value and standard deviation equal to $\log(2)/2$, following expert judgement. For the mean of the initial condition of the Fourier coefficients, as outlined in Section 3.4.2, we drop any proposed value which leads to a negative mean prediction of the unobserved TF, and assume a uniform prior on the allowed domain. We assume half-normal priors, obtained by folding a $\mathcal{N}(0, 10^2)$ distribution about 0, for $\log(R_{LHY}/R_0)$ and $\log(R_{LHY,B}/R_B)$, for the same reasons mentioned above. We assume a $\mathcal{N}(0, 20^2)$ prior for $\log(R_B/R_0)$, based on the consideration that slightly less weight is given to more extreme values than in the half-normal case, if we were assuming the same standard deviation. Finally, we assume a $\mathcal{N}(0, 10^2)$ for $\log(K'_B)$. Recall also that K_c is set to 1.

The MCMC algorithm is run for 4×10^4 pilot iterations and 1.6×10^5 additional iterations, of which we discard 10^4 iterations as burn-in. The posterior samples are thinned by recording one sample every 100 iterations, and initial conditions are randomly drawn from the prior distributions. Convergence is monitored via visual inspection of the trace plots.

We note at this point that the unobserved TF mean level is not identifiable, as the value is absorbed by the dissociation coefficient K'_B. We therefore aim at inferring the relative amplitude of the two cycles, as well as their phase. Recall that we retain for this purpose the harmonics 1, 2 and 4, with harmonic 1 having period length equal to the total observational time.

A comment about parameter rescaling is also required. As we note in Section 2.5.1, it is possible to infer the stochastic basal transcription rate by dividing the estimated value by the estimated parameter κ. When the TFs are known in units of molecule numbers, the stochastic dissociation coefficients can be inferred by multiplying the estimated dissociation coefficients by the mean levels of the TFs. For the unobserved TF considered here, this is unfortunately not possible, as we are not adopting a mechanistic model for its dynamics. Although by introducing variability in the Fourier representation, we can still disentangle mean and variance of the unobserved TF, the variance introduced by our approach is likely to account for model mismatch, rather than for intrinsic noise, and therefore an accurate estimation of molecule counts is not possible.

Moreover, it turns out that due to the cyclical nature of the oscillations, the model is also not able to significantly discriminate whether the unobserved TF is a repressor, or an activator peaking 12 hours later. We now discuss this issue, and propose a first solution.

We find that the adaptive MCMC scheme developed for the observed TF scenario is unable to properly explore the bimodal posterior distribution, as once it gets ‘stuck’ into one of the two modes, it gradually adapts the proposal distributions according to the local mode, making it less and less likely to escape and explore the other mode. A non-adaptive scheme, on the other hand, is highly inefficient, as the chain proposes a very large number of samples in low density areas.

One popular approach which deals with multimodal posterior densities is represented by the so-called tempering (see e.g. Neal, 1996; Marinari and Parisi, 1992). The main idea underlying the different tempering techniques is to gradually ‘flatten’ the target distribution, so that the algorithm can more easily move between high density areas. However, tempering algorithms can be computationally demanding, and hard to tune. Moreover, in our scenario a biological understanding of the source of bimodality is available, in that each mode is linked to a regulatory mechanism of circadian oscillatory genes. One mode represents the effect of an activator on transcription, where we observe that the levels of the child gene increase approximately at the same time at which the TF is increasing. Alternatively, the child gene dynamics can be induced by a TF which is in perfect anti-phase to this, and hence is found to act as a repressor, whose levels are decreasing approximately at the same time at which the child mRNA is increasing.

Although both the biological understanding of the system and our simulation study strongly support the hypothesis of this duality, it is non-trivial to derive the exact analytic relationship between the parameters of the two modes, starting from the assumed transcription function. It is indeed possible that the two modes are not *exactly* equally likely, and therefore no closed-form relationship is available, but the second mode has still a non-negligible probability, and we therefore wish to sample from both.

We first study this issue in an ‘exploratory’ approach, which takes advantage of a pilot run on a flattened posterior distribution target to locate the two modes. A second solution, based on the biological understanding of the source of bimodality, and adopted in the subsequent real data analysis, is provided in Section 4.1.3.

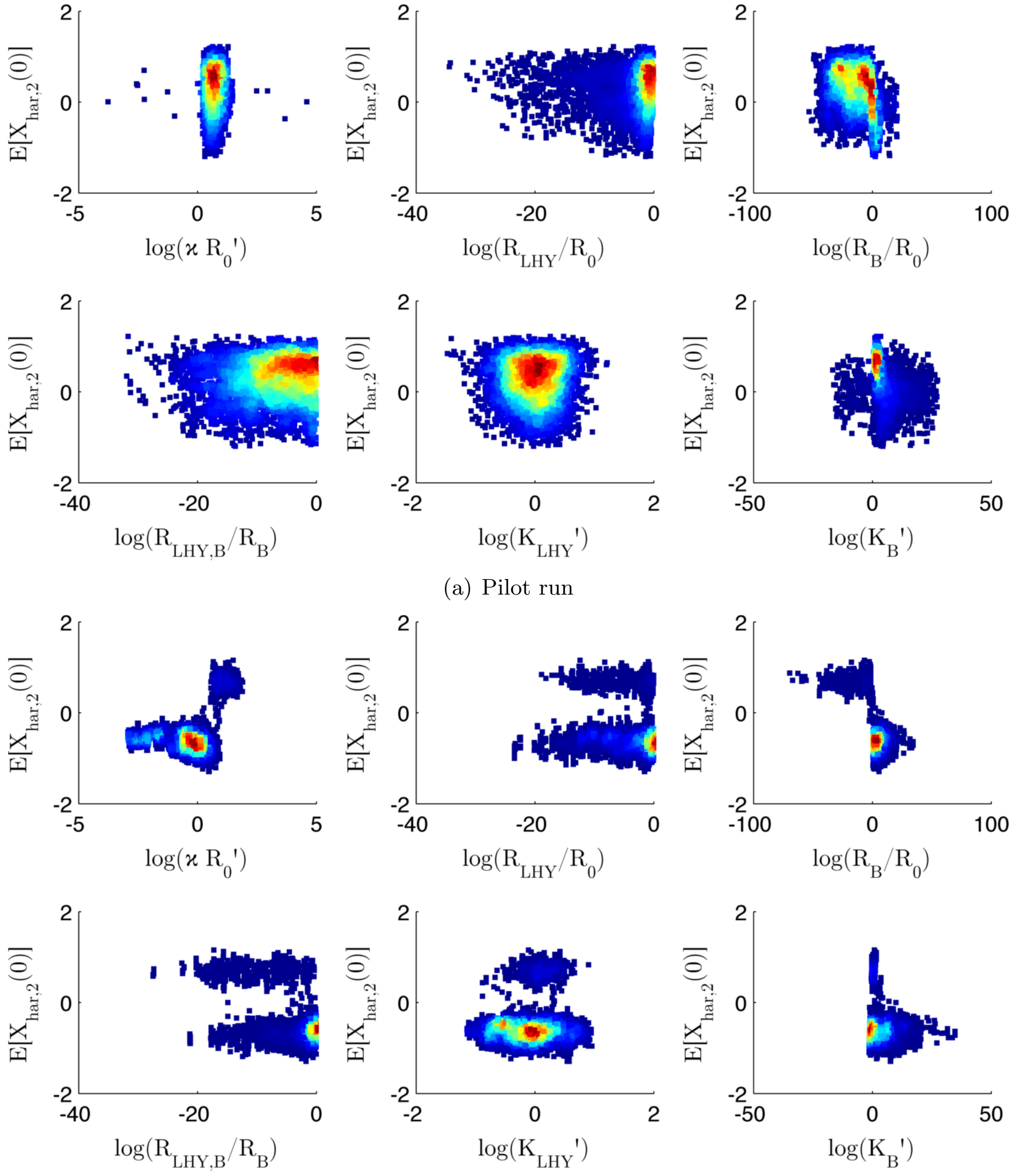
We have observed that, if the variance of the initial condition of the unobserved TF Fourier coefficients is high, the overall posterior density tends to be flatter, making it easier to locate the two modes. The idea of this first approach is therefore to specify a number of pilot iterations, in which we set informative priors on the logarithm of the variance of the initial condition of the unobserved TF

Fourier coefficients, namely a $\mathcal{N}(0, 1)$; at the end of the pilot run, the locations of the two modes can be identified by means of a clustering algorithm on the chain samples of the mean of the initial condition of the Fourier coefficients, which is aimed at identifying two anti-phase profiles, namely one for an activator and one for a repressor. We then set $\mathcal{N}(0, 20^2)$ priors on the logarithm of the variances of the initial conditions of the Fourier coefficients, and start two parallel chains from the MCMC sample of the pilot run which provides the maximum value of the likelihood in each cluster.

This approach is then combined with a locally adaptive scheme as in Roberts and Rosenthal (2009), to deal with the fact that, although infrequently, jumps between the two modes, within the same chain, can still be observed after the split. At each iteration we compute the sum of the squared distance of the accepted mean of the initial condition of the Fourier coefficients, from the cluster centroids estimated at the end of the pilot run, and assign the sample to the closest mode. Separately for each mode, and for the two parallel chains, the variances and covariances of the proposal densities are then adapted according to the previously accepted samples. This scheme allows to sample more efficiently within either of the two modes, particularly if they have different shapes. In the pilot run we adopt a mixture of independence sampler, adaptive single-component and block updates, to help exploration of the posterior density. In the second part of the algorithm run, we sample individually the degradation, measurement error variance, scale and initial condition parameters. Three blocks of parameters are then defined by: the transcription function parameters, the mean of the initial condition of the Fourier coefficients, and the variance of the initial condition of the Fourier coefficient.

To illustrate visually the presence of the two modes, we show the MCMC output of the estimation process performed on one simulation replicate of scenario A of Figure 3.5, by assuming Model 3.4. We plot in Figure 4.5 (a) pairwise scatter plots of $E[X_{har,2}(0)]$, which appears to be the Fourier parameter with the greatest weight, and the parameters of the transcription function, in the pilot run. Figure 4.5 (b), shows the same plots for the samples obtained in one of the final parallel chains. We can see that the pilot run tends to explore more widely the posterior density, while bimodality becomes more evident as the variance of the initial condition of the Fourier coefficients decreases. Finally, we observe in Figure 4.6 that in the first parallel chain, after a few iterations in the alternative mode, the chain jumps to the true mode, sampling the values of Fourier coefficients initial conditions from approximately the same region as the second parallel chain.

Now, we present the full MCMC simulation results for all the 10 simulation



(a) Pilot run

(b) First parallel chain (chain in which the jump is observed)

Figure 4.5: Scatter plots of the MCMC samples of the Fourier coefficient $E[X_{har,2}(0)]$ and the parameters of the transcription function. Model 3.6, as applied to data simulated according to scenario A of Figure 3.5. Plot code from Henson (2005).

replicates. Figure 4.7 shows the posterior densities for the transcription function, noise, scale and degradation parameters, for the true and alternative mode induced by the model when applied to data simulated according to scenario A of Figure 3.5,

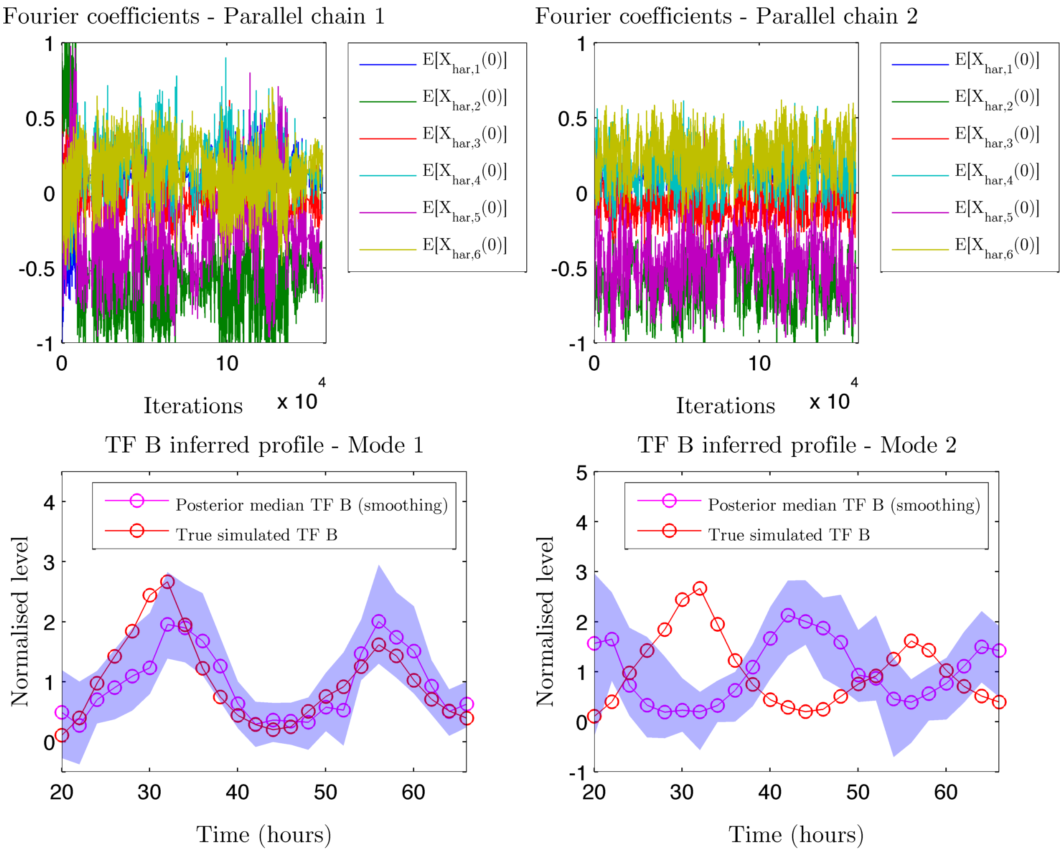
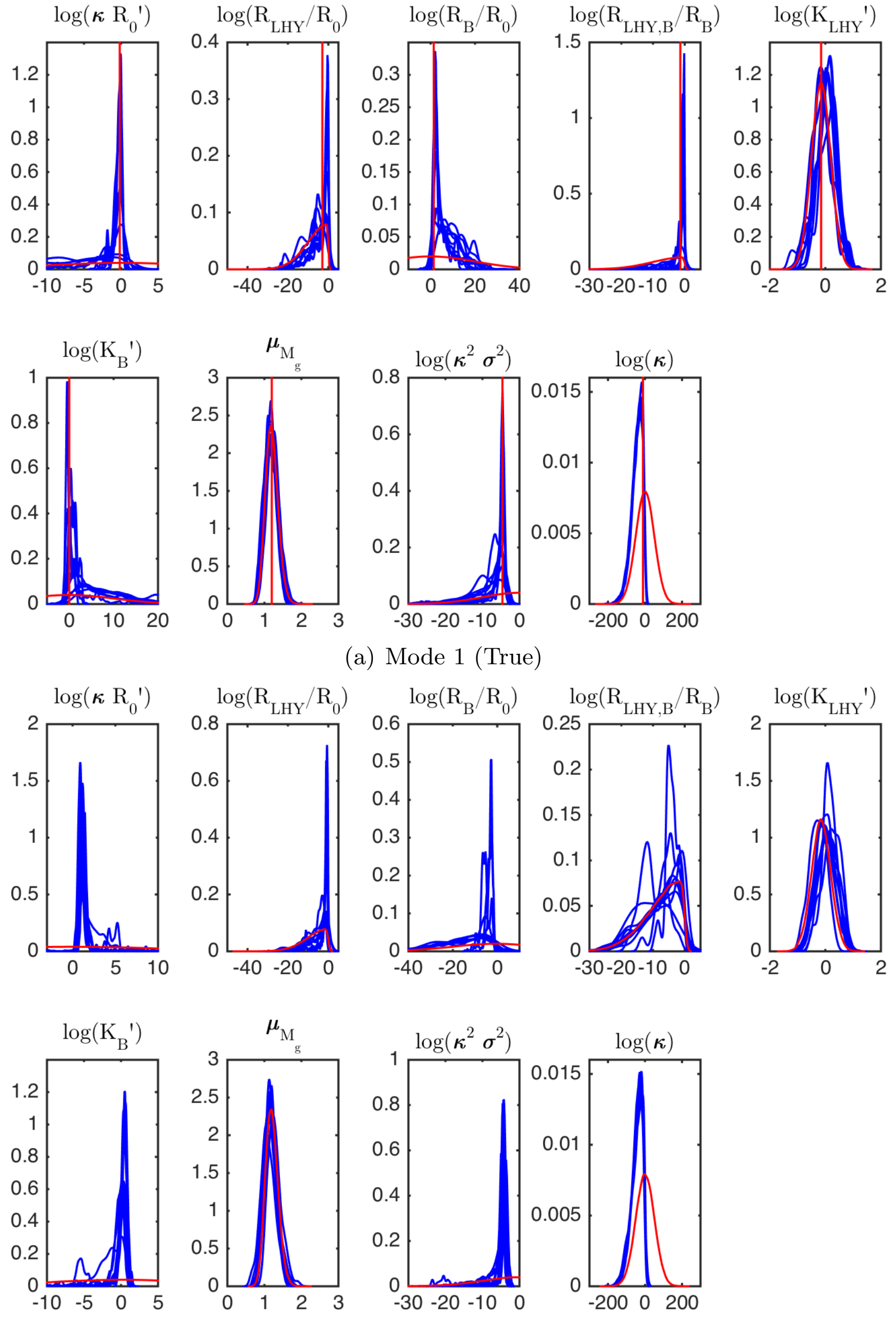


Figure 4.6: Parallel MCMC chains trace-plots of the mean of the initial condition of the Fourier coefficients, top panels. Unobserved TF B inference (smoothing): posterior median (magenta) and 95% HPDI (shaded blue), bottom panels. True simulation mode (left) and alternative mode (right). Model 3.6, as applied to data simulated according to scenario A of Figure 3.5.

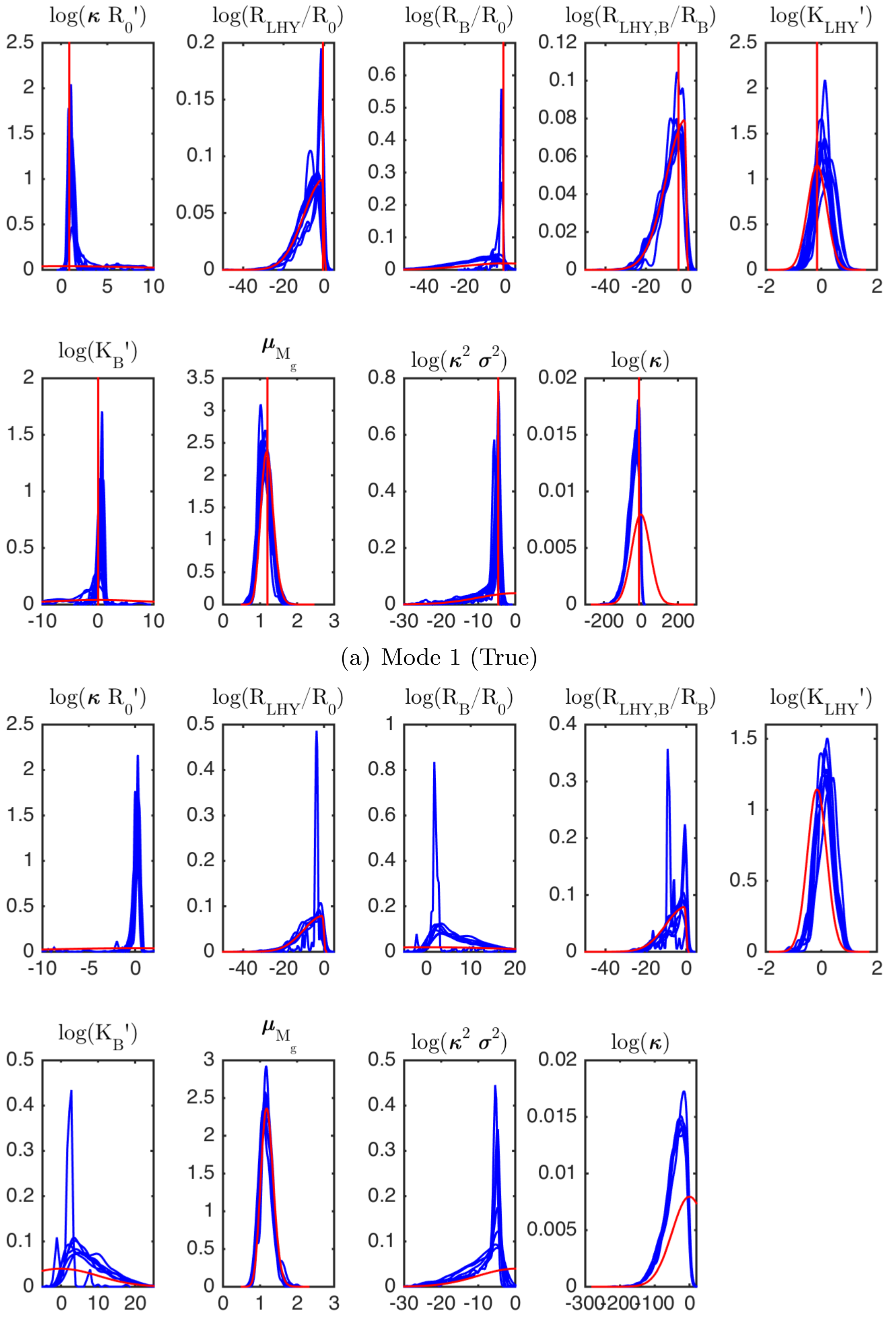
while the same plots for simulation scenario B of Figure 3.5 are provided in Figure 4.8. Note, however, that the algorithm does not always locate the two modes, so, in scenario A, the true mode is visited in all the 10 simulations, while the alternative mode in 7 out of 10 cases. As for simulation scenario B, the true mode is visited in 9 out of 10 simulations, while samples from the alternative mode are available from all the 10 simulations. We can observe that each of the two modes corresponds, as expected, to a different role of the unobserved TF B on its own, namely a repressor or an activator. In simulation scenario A, we are generally able to reliably estimate the parameters, we only report some bias for $\log(R_{LHY}/R_0)$ and $\log(R_B/R_0)$, which tend to be sightly underestimated, having 95 % HPDIs which do not include the true value in 4 out of 10 cases. As for simulation scenario B, we only report some difficulty for $\log(R_B/R_0)$, whose 95 % HPDIs so not contain the true value in 3 out of 9 cases.



(a) Mode 1 (True)

(b) Mode 2 (Alternative)

Figure 4.7: Kernel density estimates of the marginal posterior densities of the transcription function, noise, scale and degradation parameters. Model 3.6, as applied to data simulated according to scenario A of Figure 3.5. MCMC samples for 10 independent replications. The red vertical line is at the true value, and the prior density is also superimposed in red.



(a) Mode 1 (True)

(b) Mode 2 (Alternative)

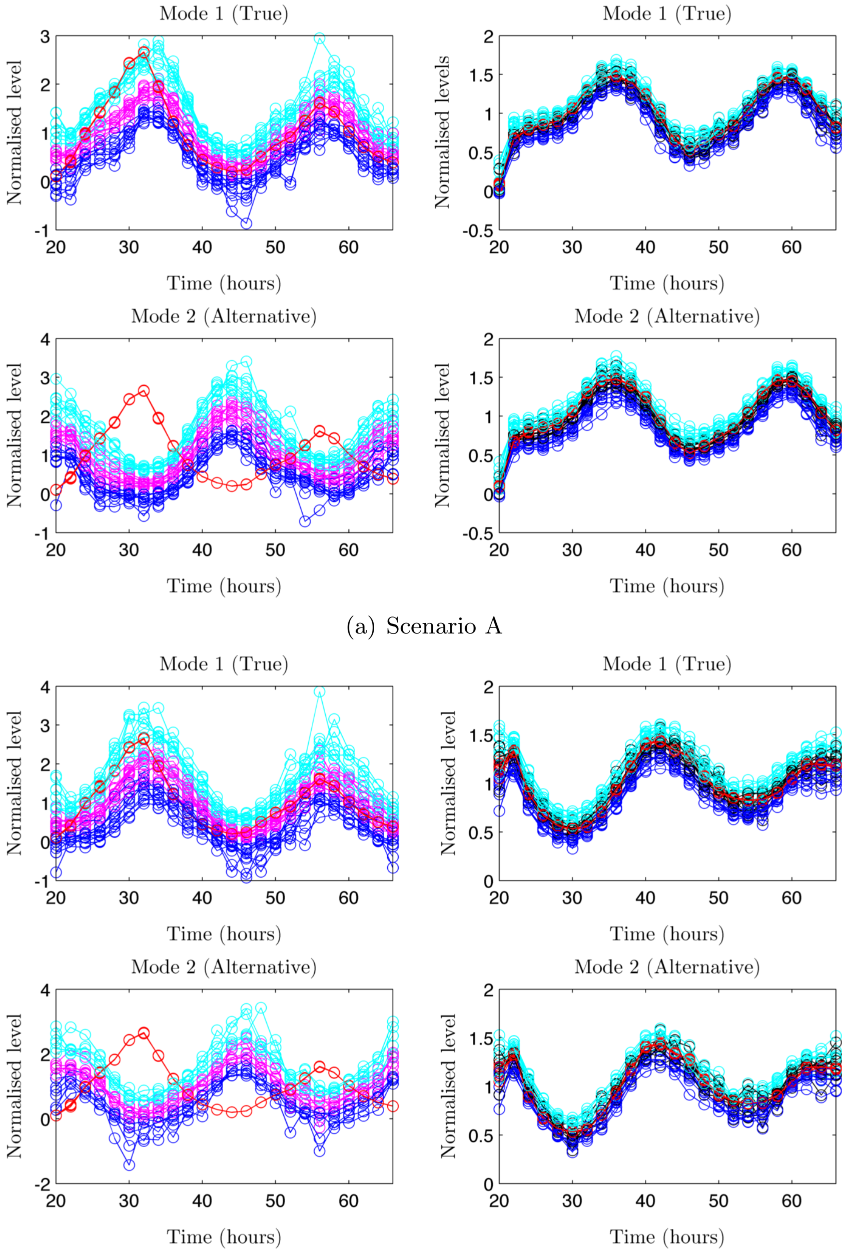
Figure 4.8: Kernel density estimates of the marginal posterior densities of the transcription function, noise, scale and degradation parameters. Model 3.6, as applied to data simulated according to scenario B of Figure 3.5. MCMC samples for 10 independent replications. The red vertical line is at the true value, and the prior density is also superimposed in red.

We provide in Figures 4.9 (a) and (b) the posterior smoothing densities for the unobserved TF B and the child mRNA profiles of simulation scenario A and B, respectively. Focussing first on simulation scenario A, we observe that the true simulated profile of the unobserved mRNA is generally included in the 95 % HPDIs of the smoothing density: considering all time-points, it is always included in at least 7 out of 10 cases, and for most time-points in 9 or 10 out of 10 cases. As for the unobserved TF B smoothing, the most notable mean mismatch is observed in its first peak, and, in particular, the mean of the first cycle seems to be estimated with a delay of about 1-2 hours. In most cases however, the true simulated time-series is still included in the 95 % HPDI, although we note that this is not the case for the third and fourth time-point in half of the cases. The same delay is observed in simulation scenario B, leading to one time-point in the first cycle to be always excluded from the 95 % HPDIs. A, possibly related, low empirical coverage for simulation scenario B is also observed in the unobserved mRNA, at the time-points corresponding to the first peak of the unobserved TF. We note, however, that the most biassed parameter in simulation scenario B is $\log(R_B/R_0)$: we can postulate that the correlation in the posterior density plays a role here. A second relevant source of mismatch is highly likely to be in the approximate handling of the unobserved TF.

A comparison of the log-likelihood values of the two modes (not shown), in both scenario A and B, has revealed that simulation scenario A seems to slightly favour the alternative mode, while simulation scenario B slightly favours the true one. This result suggests that the likelihood tends to prefer the scenario in which the unobserved TF acts as an activator, again possibly due to the approximate modelling of the unobserved TF B, and the posterior density correlation structure.

Therefore, in order to assess whether an increase in the number of data-points would provide a significant advantage in terms of discrimination between the two modes, we run the MCMC algorithm for two simulation replicates in scenario A and two simulation replicates in scenario B, assuming $\Delta_t = 1\,\text{h}$. The increasingly peaky likelihood, induced by the higher number of observations, makes the approach adopted so far to locate the two modes more challenging. To make sure that both modes are visited, we run two parallel chains from the beginning of the algorithm, one assuming a $HN(10^2)$ prior with negative support, and the second a $HN(10^2)$ prior with positive support for $\log(R_B/R_{A,B})$. More details about this approach are provided in Section 4.1.3.

We can see in Figure 4.10 that the two modes are approximately equally likely, suggesting that an increased number of observations is not leading to a



(a) Scenario A

(b) Scenario B

Figure 4.9: Unobserved TF B inference (smoothing), left panels: posterior median (magenta) and 95 % HPDIs (lower: blue; upper:cyan). True simulated child unobserved TF B (red). Unobserved child mRNA inference (smoothing), right panels: posterior median (black) and 95 % HPDIs (lower: blue; upper:cyan). True simulated child mRNA (red). True mode (top panels in each subfigure), and alternative mode (bottom panels in each subfigure). MCMC samples for 10 independent simulations from Model 3.6, with parameters as in Figure 3.5.

stronger discrimination between the two modes.

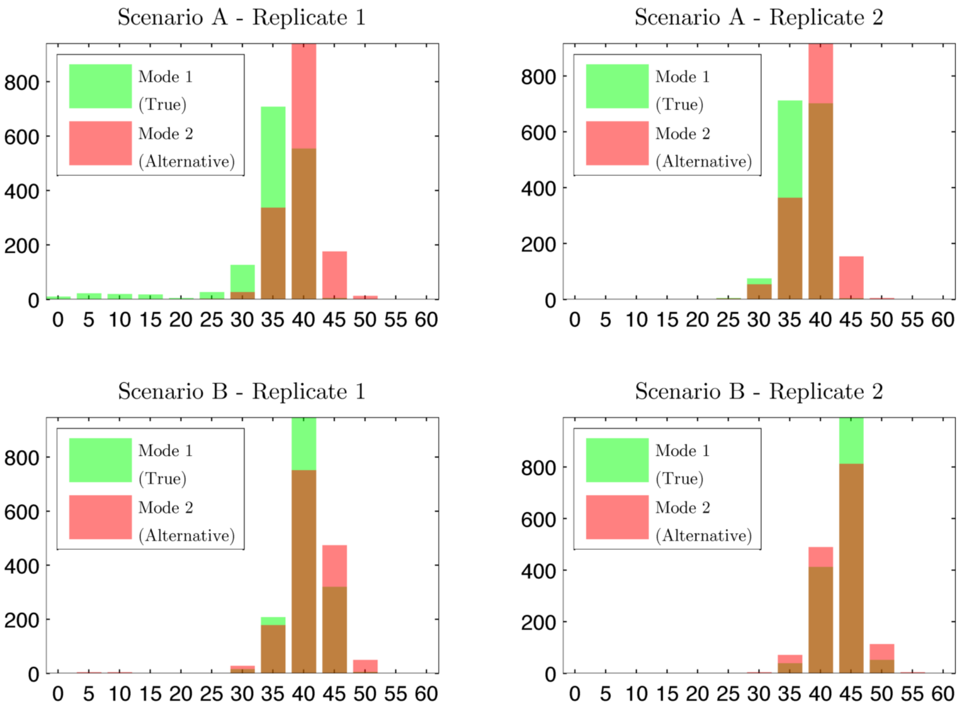


Figure 4.10: Log-likelihood samples for four independent MCMC runs on Model 3.6, as applied to data simulated according to the scenarios of Figure 3.5. $\Delta_t = 1\,\text{h}$. Comparison between mode 1 (True) and 2 (Alternative) for each simulation replicate. Plot code from Greene (2014a and 2014b).

We also observe, however, that the coverage for the unobserved TF B tends to worsen. There is a reduced variability in the unobserved TF smoothing density, as expected with a higher number of observations, and the chains tend to sample the mean of the initial condition of the Fourier coefficients in a sub-region of the $\Delta_t = 2\,\text{h}$ case; this sub-region can include the 'true' parameters values, however, the reduced variance contributes in inducing a lower coverage. Figure C.1 in Appendix shows the comparison for the inferred smoothing profiles of TF B, assuming $\Delta_t = 2\,\text{h}$ and $\Delta_t = 1\,\text{h}$. Our interpretation is that the likelihood is increasingly concentrated, and therefore the algorithm tends to struggle in moving within the high density area of the posterior, which has now possibly additional sub-peaks. In this scenario, the assumed form of the unobserved TF - i.e. the Fourier series - starts to become 'too rough', and model mismatch can have a more significant influence. Our suggestion is therefore to resort to alternative, possibly mechanistic, modelling techniques if more data-points are available.

As a final check, we fit our model to one SSA simulation of the full system, for each simulation scenario. Recall that we have so far resorted to the diffusion approximation, which further assumes the aggregate hazards and the QSSA, given the significant speed-up in simulations. Figure 4.11 shows a comparison of one SSA simulation for scenario A and B, and 50 diffusion approximation simulations. We note a generally higher variability in the SSA simulation than in the diffusion approximation ones. We therefore run the estimation algorithm in order to assess whether the observed difference has any significant impact on inference. Both the parameters estimates, and the unobserved TF profile, do not seem to significantly differ with respect to the case where the diffusion approximation simulations are employed. It seems that inference under the two simulation methodologies mostly differ in the variance of the unobserved TF, having wider HPDIs. The comparison of the unobserved TF smoothing densities and the parameters posterior densities are shown in Figures C.2, and C.3 (a) and (b) in Appendix, respectively. The high molecules counts of the TFs imply tenability of the assumption required for aggregation of the hazards, and the high molecules counts of the child mRNA justifies the diffusion approximation of the underlying birth and death process. Any observed mismatch is then likely to be due to the QSSA.

4.1.3 Case 3: inference for one unknown TF as only regulator

We finally validate the inferential process for Model 3.7, as applied to data simulated as in Figure 3.7. In this scenario, only the unobserved TF B is assumed to dynamically influence transcription of the child gene.

Here we tackle the bimodality issue by performing inference for the two modes independently. In particular, we run two parallel chains, one adopting a $HN(10^2)$ prior with support $[0, \infty)$, and the second chain adopting a $HN(10^2)$ prior with support $(-\infty, 0]$, for $\log(R_B/R_0)$. The two chains cover the full parameter support, and we are also guaranteed that both modes are visited. The prior densities and the MCMC scheme are specified as in the case which assumes both LHY and the unobserved TF B as regulators (see Section 4.1.2).

The algorithm is run for 2.5×10^5 iterations, we discard a burn-in of 10^5 iterations, and thin the posterior samples by recording one sample every 200 iterations. Again, initial conditions for all the parameters are randomly drawn from the prior densities, and convergence is monitored via visual inspection of the trace plots.

Figures 4.12 (a) and (b) show the parameter posterior densities, as estimated on 10 i.i.d. replications of simulated data from scenario A of Figure 3.7, and for the true and alternative mode, respectively. Figures 4.13 (a) and (b) show analogous plots for scenario B of Figure 3.7. Focusing first on simulation scenario A, the

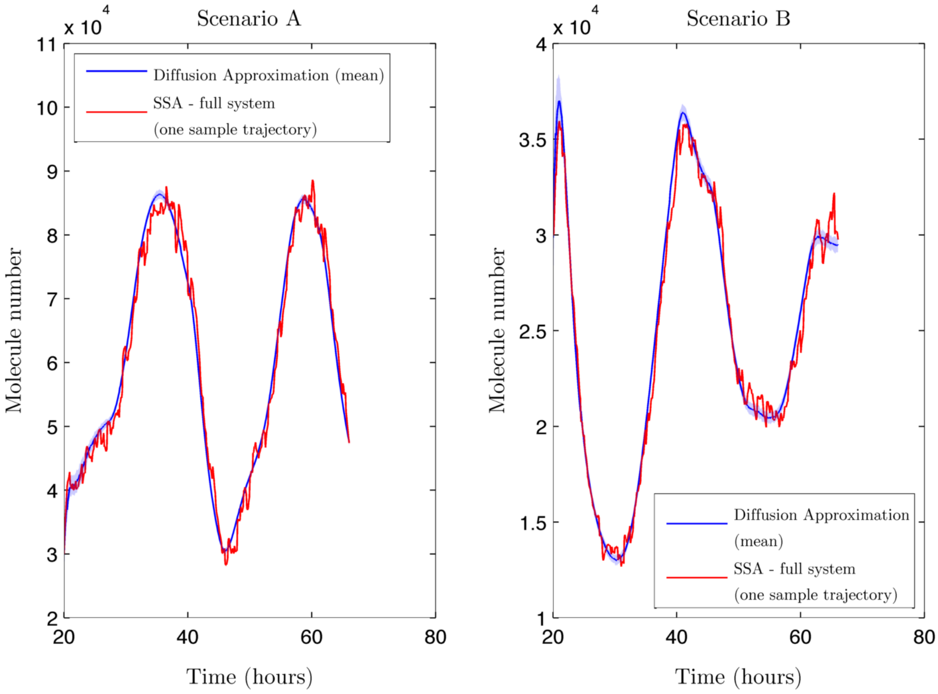


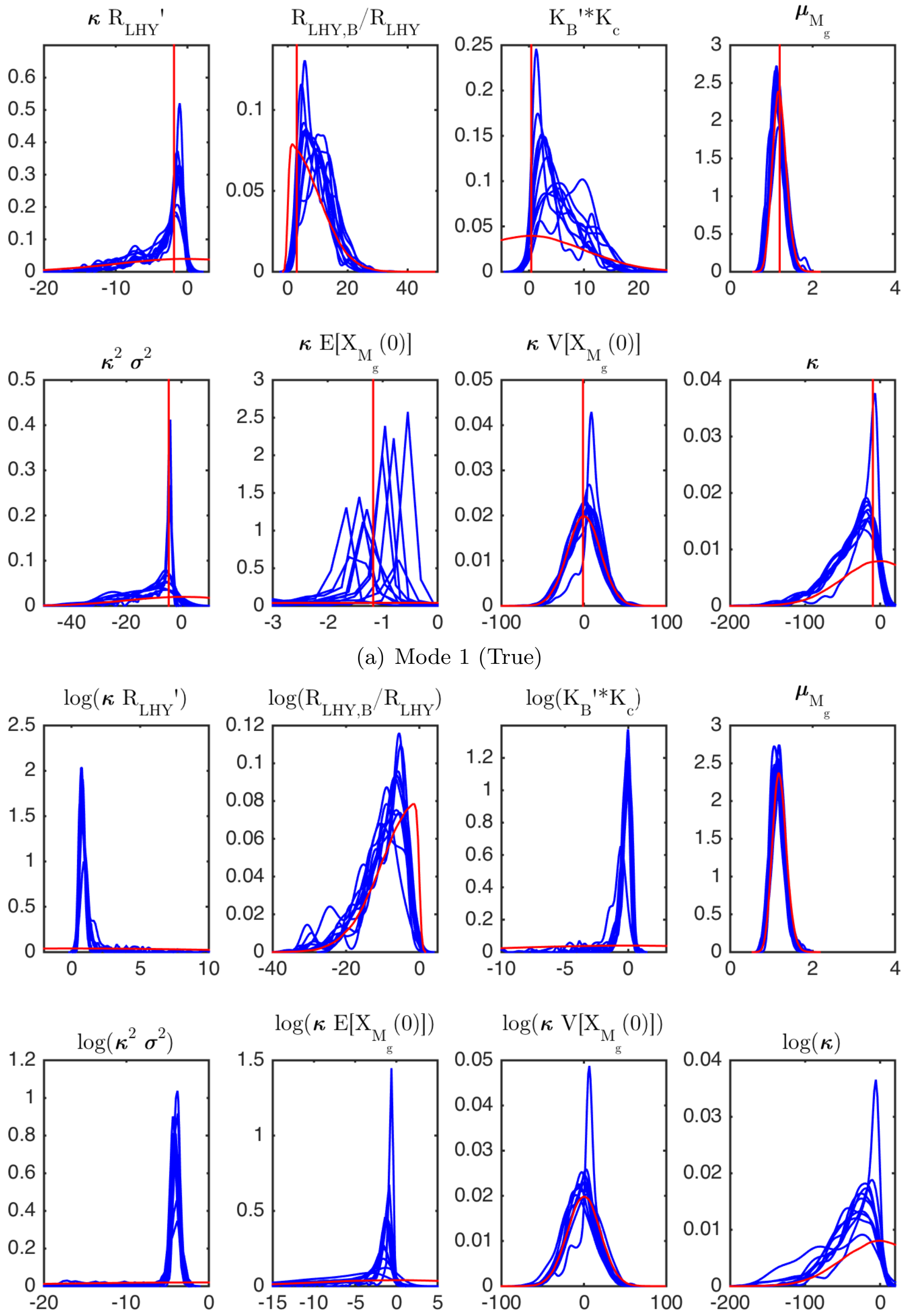
Figure 4.11: Comparison of one SSA simulation from the full set of reactions of Table 1.1 (red), and 50 diffusion approximation simulations according to Model 3.4, mean (blue) ±2 SD (shaded blue). Parameters for the SSA are set equal to: $R_0 = 50\,\text{molecules/h}$, $R_{LHY} = 2.5\,\text{molecules/h}$, $R_B = 2 \times 10^2\,\text{molecules/h}$,$R_{LHY,B} = 50\,\text{molecules/h}$, $k_{+c} = 0.8$, $k_{-c} = 1.2$ (scenario A, left); $R_0 = 60\,\text{molecules/h}$, $R_{LHY} = 35\,\text{molecules/h}$, $R_B = 2.5\,\text{molecules/h}$, $R_{LHY,B} = 0.5\,\text{molecules/h}$, $k_{+c} = 1.2$, $k_{-c} = 1.2$ (scenario B, right). Common parameters: $k_{+LHY} = k_{+B} = 3 \times 10^{-4}\,\text{molecules}$, $k_{-LHY} = k_{-B} = 6\,\text{molecules}$. LHY and TF B are simulated assuming parameters $\nu_{LHY}(t) = [2.88 \times 10^2, 8.7, 1.68 \times 10^2, 21.6, 1.26 \times 10^2]$ (in molecules per hour) with switch times $Swt_A = [27, 40, 50, 61]$ (in hours), $\nu_B(t) = [21, 1.71 \times 10^2, 2.91 \times 10^2, 9, 90, 1.68 \times 10^2, 21]$ (in molecules per hour) with switch times $Swt_B = [21, 28, 45, 52, 56]$ (in hours), respectively. $\mu_M = 0.5\,\text{h}^{-1}$, $\alpha_M = 40\,\text{h}^{-1}$, $\mu_P = 0.34\,\text{h}^{-1}$, $\mu_{M_{M_g}} = 1.2\,\text{h}^{-1}$ and $X_{RNAPc} = 10\,\text{molecules}$. Values are summed over 100 cells. Parameters for the diffusion approximation simulations are set as in Figure 3.5.

true parameters values are generally included in the 95% HPDIs, although we find poor mixing of the chain for $\log(\kappa E[X_{M_g}(0)])$. Considering simulation scenario B, estimation seems to be more challenging, and in particular the rescaled dissociation coefficient parameter $\log(K'_B K_c)$ falls inside the HPDIs in 4 cases out of 10 at level 95%, and 6 cases out of 10 at level 99%. We also note that the true value of the log ratio $\log(R_{LHY,B}, R_B)$ is included in the HPDIs in 8 out of 10 cases at level 95%. We believe that this result is due to the posterior correlation structure, and indeed we notice in Figure 4.14 (b) that the first peak of the unobserved TF B smoothing density median tends to be underestimated, leading to two simulation data-points not included in any HPDI.

Comparing this situation with scenario A, we can see in Figure 4.14 (a) that the first peak is better estimated, but the second cycle seems to be preceded by the median fit by about 1-2 hours. This mismatch is however generally compensated by the variability, and we observe that in scenario A the smoothing density of the unobserved TF does not include the true value at worst in 6 out of 10 cases, at four time-points. In one replicate we have also obtained very wide HPDIs for the unobserved TF (not shown for plotting purposes).

The remaining plots of Figures 4.14 (a) and (b), show that, as expected, the two modes provides anti-phase smoothing profiles for the unobserved TF, as well as that scenario B tends to perform better than scenario A in terms of inclusion of the unobserved mRNA true simulation profile in the 95 % HPDIs of the smoothing density. We again attribute this mismatch to the approximate handling of the unobserved TF.

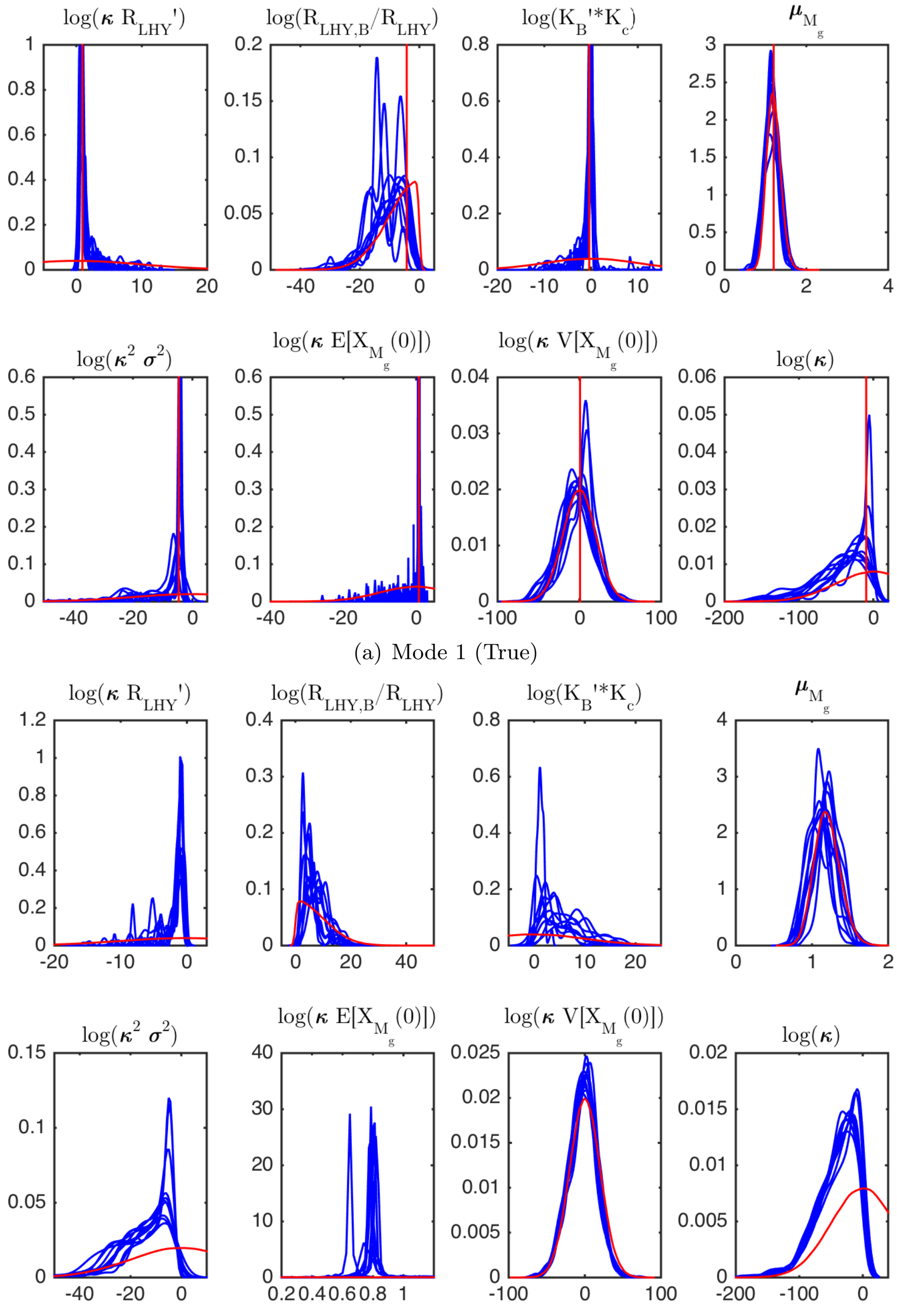
Despite the fact that the 95 % HPDIs do not always provide the expected empirical coverage, the model offers the possibility to infer, at least approximately, the phase and relative amplitudes of the two cycles of the unobserved TF, which is our main objective.



(a) Mode 1 (True)

(b) Mode 2 (Alternative)

Figure 4.12: Kernel density estimates of the marginal posterior densities of the model parameters posterior densities, excluding the mean and variance of the initial condition of the Fourier coefficients. Model 3.7, as applied to data simulated according to scenario A of Figure 3.7. MCMC samples for 10 independent replications. The red vertical line is at the true value, and the prior density is also superimposed in red.

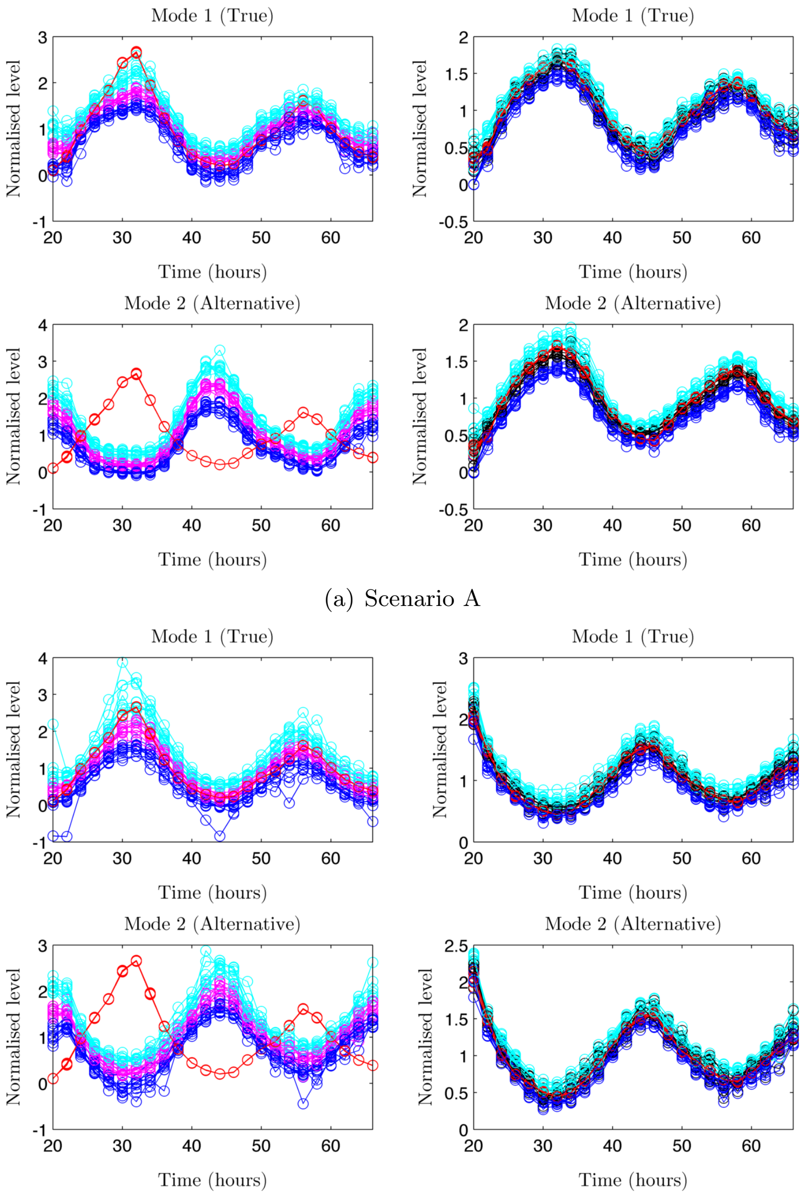


(a) Mode 1 (True)

(b) Mode 2 (Alternative)

Figure 4.13: Kernel density estimates of the marginal posterior densities of the model parameters, excluding the mean and variance of the initial condition of the Fourier coefficients. Model 3.7, as applied to data simulated according to scenario B of Figure 3.7. MCMC samples for 10 independent replications. The red vertical line is at the true value, and the prior density is also superimposed in red.

.



(a) Scenario A

(b) Scenario B

Figure 4.14: Unobserved TF B inference (smoothing), left panels: posterior median (magenta) and 95 % HPDIs (lower: blue; upper:cyan). True simulated child unobserved TF B superimposed (red). Unobserved child mRNA inference (smoothing), right panels: posterior median (black) and 95 % HPDIs (lower: blue; upper:cyan). True simulated child mRNA superimposed (red). True mode (top panels in each subfigure), and alternative mode (bottom panels in each subfigure). MCMC samples for 10 independent simulations from Model 3.7, as applied to the simulation scenarios of Figure 3.7, with the exception of the unobserved TF smoothing profile for the true mode in scenario A: one posterior profile has extremely wide HPDIs, and is thus excluded for plotting purposes.

4.2 Data analysis for *Arabidopsis thaliana*

In this section we provide the data analysis of the *Arabidopsis thaliana* genes mRNA, as measured by the Nanostring experiment introduced in Section 3.3.1.

We first select a subset of the 100 Nanostring genes by requiring rhythmicity in their expression, as it seems reasonable to assume that non-circadian dynamics cannot be influenced by circadian ones, in a causal framework. We start our analysis with preliminary statistics related to the classification of the Nanostring rhythmic genes according to the presence of binding sites in their promoter region, and the relationship between binding sites groups and the induction experiment results of Section 3.3.2, as well as amplitude and phase.

We then inspect the parameter estimates obtained by fitting the model which assumes only one unobserved TF to the Nanostring rhythmic genes, and check normality and periodicity of the residuals. Finally, we investigate the correlation between the inferred unobserved TF profiles and LHY observed protein, as well as synchrony among unobserved mRNA profiles.

It is worth noting that a first attempt was made at fitting the model comprising only the observed LHY as a regulator, introduced in Section 3.4.1. However, this approach seems not able to satisfactorily fit the available data. Multiple explanations can be put forward. It is possible that LHY binding is non-functional, at least for a subset of the Nanostring genes, as the induction experiment result points out. It is also possible that LHY binding is functional, but it requires the presence of additional TFs to influence transcription of the child genes. The model with only one unobserved TF provides an advancement in this direction, representing a more flexible model, which allows to compare *a posteriori* the reconstructed TF profile with the available LHY time-series: a correlated result may point in the direction of LHY being nevertheless an important regulator for the child gene, while a completely uncorrelated TF points in the direction of a non-functional binding of LHY.

A further advancement may be represented by the model introduced in Section 3.4.2, which assumes both LHY and an unobserved TF as functional regulators. As anticipated, however, inference would require in this case availability of prior information concerning the dissociation coefficient of LHY, and its consistent induction effect, as outlined and validated with the simulation study of Section 4.1.2. Unfortunately there are no rhythmic genes in the Nanostring data-set having exactly one binding site, among those bound by LHY listed in 3.3.3, as well as a consistent induction by LHY at level $\alpha = 0.1$.

4.2.1 Preliminary analysis of the Nanostring mRNA data

We first select 91 Nanostring genes, by excluding the five control genes, and by retaining among the remaining 95, the ones for which a prior concerning degradation rate is available by fitting the switch tool of Section B.2 in Appendix on additional mRNA time-series data for the same genes from a microarray experiment.

A first classification of the Nanostring genes is performed by assessing their circadian rhythmicity. This characteristic can be statistically evaluated by means of spectral analysis. From Section B.3 in Appendix, each element of the observed time series $Y_1, ..., Y_n$ of the child mRNA can be written as

$$Y_i = a_0 + \sum_{q=1}^{h} H_q(i),$$

with

$$H_q(i) = a_q \cos(\alpha q i) + b_q \sin(\alpha q i), , \tag{4.1}$$

where h is in our case equal to $n/2 = 12$, and $\alpha = 2\pi/24$. If we consider only the first 11 harmonics, we have

$$\begin{aligned} a_q &= \frac{2}{n}\sum_{i=1}^{n} Y_i \cos(\alpha q i), \\ b_q &= \frac{2}{n}\sum_{i=1}^{n} Y_i \sin(\alpha q i) \quad 1 \leq q < n/2. \end{aligned}$$

We refer to Section B.3 in Appendix for details about the remaining quantities involved in Equation 4.1. Each harmonic H_q is responsible for a particular periodicity in the data. A useful tool for a statistical analysis of the underlying periodicities is the periodogram (Prado and West, 2010, Chapter 3), namely

$$I(\omega_q) = \frac{n}{2}\left(a_q^2 + b_q^2\right),$$

where ω_q denotes the frequency of harmonic H_q in the time-units of the observed data. The periodogram provides an estimate of the relative importance of each frequency, or equivalently harmonic, for the overall signal $Y_{1:n}$. However, in this form, it is difficult to draw any inferential conclusion about the estimated values of the periodogram at each frequency ω_q, and in particular to test the hypothesis that the series $Y_{1:n}$ can be generated by a white noise process, against the hypothesis that a significant periodicity is present. A more useful quantity in this direction is the normalised periodogram of Scargle (1982), which has the form

$$I(\omega_q) = \frac{1}{2}\left[\left(\frac{a_q}{\sqrt{\frac{4}{n^2}\sum_{i=1}^{n}\cos(\alpha q i)^2}}\right)^2 + \left(\frac{b_q}{\sqrt{\frac{4}{n^2}\sum_{i=1}^{n}\sin(\alpha q i)^2}}\right)^2\right].$$

The rationale is the following (see Horne and Baliunas, 1986). Assume, under the null hypothesis, that $Y_1, ..., Y_n$ are i.i.d $\mathcal{N}(0, 1)$, i.e. the signal is standard normal white noise. Therefore

$$\begin{aligned} a_q &\sim N\left(0, \frac{4}{n^2}\sum_{i=1}^{n}\cos(\alpha q i)^2\right) \\ b_q &\sim N\left(0, \frac{4}{n^2}\sum_{i=1}^{n}\sin(\alpha q i)^2\right). \end{aligned}$$

It follows that the value assumed at each frequency by the normalised periodogram is a $\chi^2_2/2$, or, equivalently, a $Ga(1, 1)$, or an $Exp(1)$. Note that extreme observed estimates of the normalised periodogram can either indicate that the signal has a periodic component, or that it deviates from normality.

Since our goal is to assess circadian periodicity, and we observe the data for a total of two circadian cycles, our main interest is in the second harmonic. We therefore consider a gene to be rhythmic when the estimated value of the normalised periodogram, evaluated at the frequency $\omega_2 = 1/24$ cycles/h, is higher than 2.99, which is the 95% quantile of a standard exponential. This procedure selects 70 genes out of the 91 tested.

A further relevant characteristic is the presence of binding sites and binding sites combinations in the promoters, as well as the genes response to an increase in LHY. A summary of the Nanostring rhythmic genes, with respect to their binding sites group and induction experiment result is provided in Table 4.1. We can see that the EE and ABRE binding sites are the most represented among the Nanostring rhythmic genes, as well as the category of genes repressed by LHY. To test a possible association between the presence of binding sites and induction by LHY we perform Fisher's exact test (Fisher, 1922; Agresti, 1992), leading to a p-value equal to 0.62. There seems therefore to be no association between the two variables, which can suggest either that LHY acts through different binding site combinations, or alternative mechanisms, for example binding of additional transcription factors to different motifs, is required.

Another important preliminary analysis, since we are dealing with circadian genes, concerns their amplitude and phase. The distribution of the phases and

Binding sites	Induction				Total
	-2	-1	0	1	
None	2	8	6	1	17
CBS only	0	0	1	0	1
ABRE only	0	5	2	1	8
EE only	0	15	5	0	20
CBS + ABRE	0	1	0	0	1
CBS + HEX	0	1	0	0	1
CBS + EE	0	1	1	0	2
HEX + ABRE	0	2	0	0	2
ABRE + EE	0	8	1	0	9
HEX + EE	0	4	1	0	5
CBS + ABRE + HEX	0	0	1	0	1
CBS + ABRE + EE	0	0	1	0	1
ABRE + HEX + EE	0	1	0	0	1
CBS + ABRE + HEX + EE	0	1	0	0	1
Total	2	47	19	2	70

Table 4.1: Rhythmic Nanostring genes by presence of binding sites in the promoter region and induction experiment result. A binding site is defined as present if there is at least one binding site of the corresponding type in the promoter. Induction is assessed at significance level $\alpha = 0.1$ (-2 indicates consistent repression, -1 repression, 0 no effect, 1 activation, no rhythmic genes belong to the case of consistent activation, 2; see Section 3.3.2 for a more detailed explanation of the categories). Fisher's exact test for association has p-value 0.62. Nanostring data-set, Carré lab.

amplitudes of expression of the Nanostring rhythmic genes, categorised according to their binding sites group, is provided in Figures 4.15 and 4.16, respectively. The phase is computed by locating the peak time of the second harmonic, between hours 24 and 46 of the observed time-series (recall that the experiment starts at time 20). The amplitude is computed on the mean-centred time-series.

We can see that the distribution of the phases spans across the whole of the 24 hour interval, with a predominance of evening phases. This is consistent with the result of the induction experiment that identifies LHY, which is peaking in the morning, as a repressor of transcription for a significant portion of the Nanostring genes. Regarding the amplitude, it seems that the presence of the EE and HEX binding sites favours a higher amplitude, while the group with none of the known binding sites has generally lower amplitude levels. We finally remark that each binding site can be present in multiple copies in the same gene promoter. However, an increase in the number of binding sites categories is in this case not advisable, given the small number of available genes.

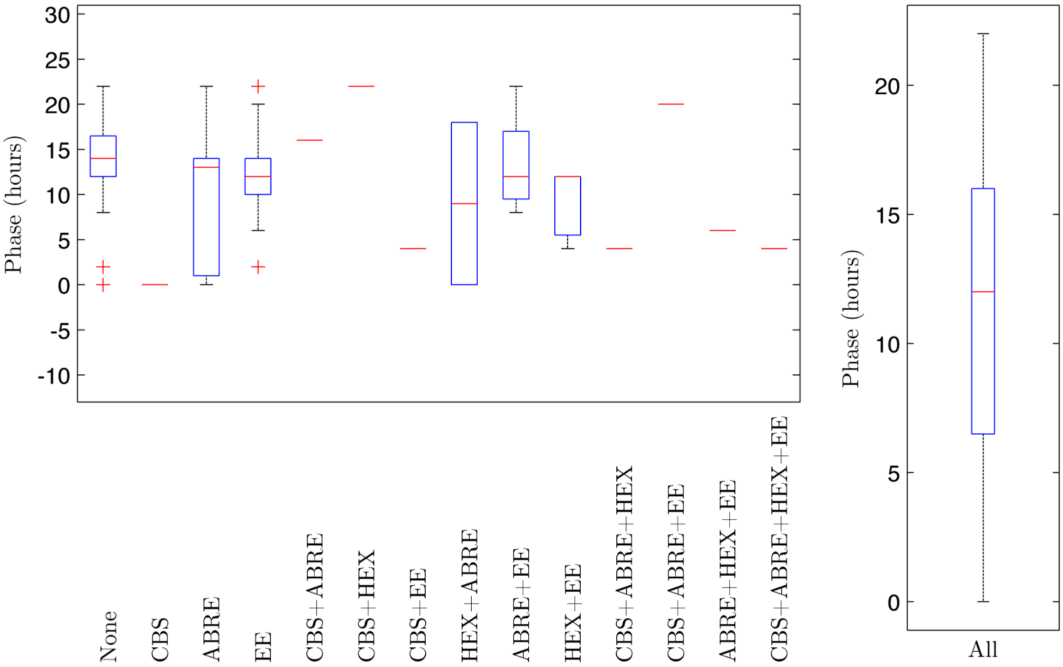


Figure 4.15: Box plots of phases of expression of the Nanostring rhythmic genes, by binding sites group. A binding site is defined as present if there is at least one binding site of the corresponding type in the promoter. Phase corresponding to the peak of the 24 hour period harmonic. Groups have different sizes, and some comprise only one or two observations, hence a single observed value is plotted as a horizontal line. Crosses correspond to outliers. Nanostring data-set, Carré lab.

4.2.2 Inference for the Nanostring mRNA data

Parameter inference

Model 3.7, which we recall comprises only one unobserved TF as transcriptional regulator of the child gene, is fitted to the Nanostring rhythmic genes data. Recall that a bimodal posterior parameter distribution is induced by this model, due to the presence of the unobserved TF; each mode corresponds to a different role of the unobserved TF on transcription of the child gene, namely that of an activator and that of a repressor.

Figure 4.17 provides median and HPDI estimates for the transcription function parameters $\log(\kappa R_0')$, $\log(R_B/R_0)$ and $\log(K_B')$. Estimates of $\log(\kappa R_0')$ have a median of the medians estimate equal to -0.81 in mode 1, and -5.4 in mode 2, with a standard deviation about these estimates of about two points on the logarithmic scale in both cases. A higher between-genes variability is observed for $\log(R_B/R_0)$, which assumes in some cases relatively extreme negative and positive values, close to

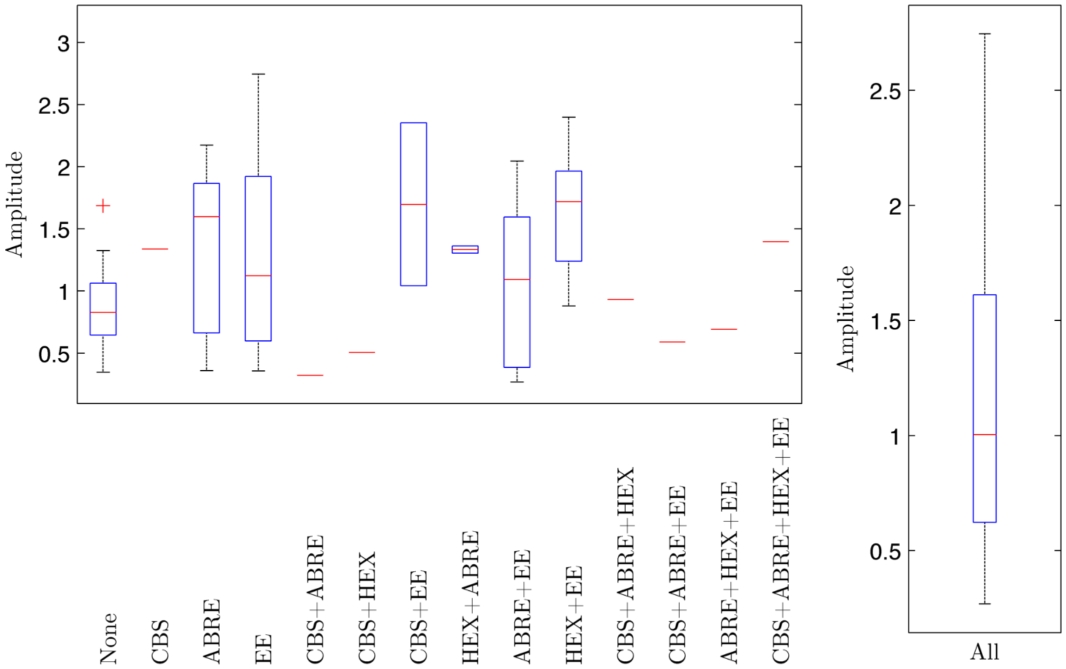


Figure 4.16: Box plots of amplitudes of the Nanostring rhythmic genes, by binding sites group. A binding site is defined as present if there is at least one binding site of the corresponding type in the promoter. Amplitude of the 24 hour period harmonic. Groups have different sizes, and some comprise only one or two observations, hence a single observed value is plotted as a horizontal line. Crosses correspond to outliers. Nanostring data-set, Carré lab.

the prior tails. In mode 2, for example, the median estimates of $\log(R_B/R_0)$ reaches a maximum of 5×10^5, on the original scale. This seems a rather large increase in the transcriptional rate, and we believe that this parameter may indeed be absorbing the effect of model misspecification, for example due to the presence of multiple binding sites for TF B, or cooperativity with LHY. A sensitivity analysis for prior specification can be appropriate, although it has not been performed at present. The parameter $\log(K'_B)$, finally, exhibits similar median estimates across the genes, particularly in mode 2, although with variable widths for the corresponding HPDIs.

In Figure 4.17 we observe median and HPDI estimates for μ_{M_g}, $\log(\kappa^2\sigma^2)$ and $\log(\kappa)$. The posterior densities of μ_{M_g} have median of the medians estimates equal to 0.25 in mode 1, and 0.24 in mode 2; we also note that consistently low values are estimated for κ, the median of the medians estimate being in this case equal to approximately -33 in both modes, which indicate the presence of high molecule counts. A low estimate was indeed expected, due to the presence of sample aggregation over several cells. Finally, the median of the medians of $\log(\kappa^2\sigma^2)$

posterior densities is equal to -4.8 in mode 1, and -8.4 in mode 2, although it exhibits a relatively high standard deviation in both modes, equal to about five points on the logarithmic scale.

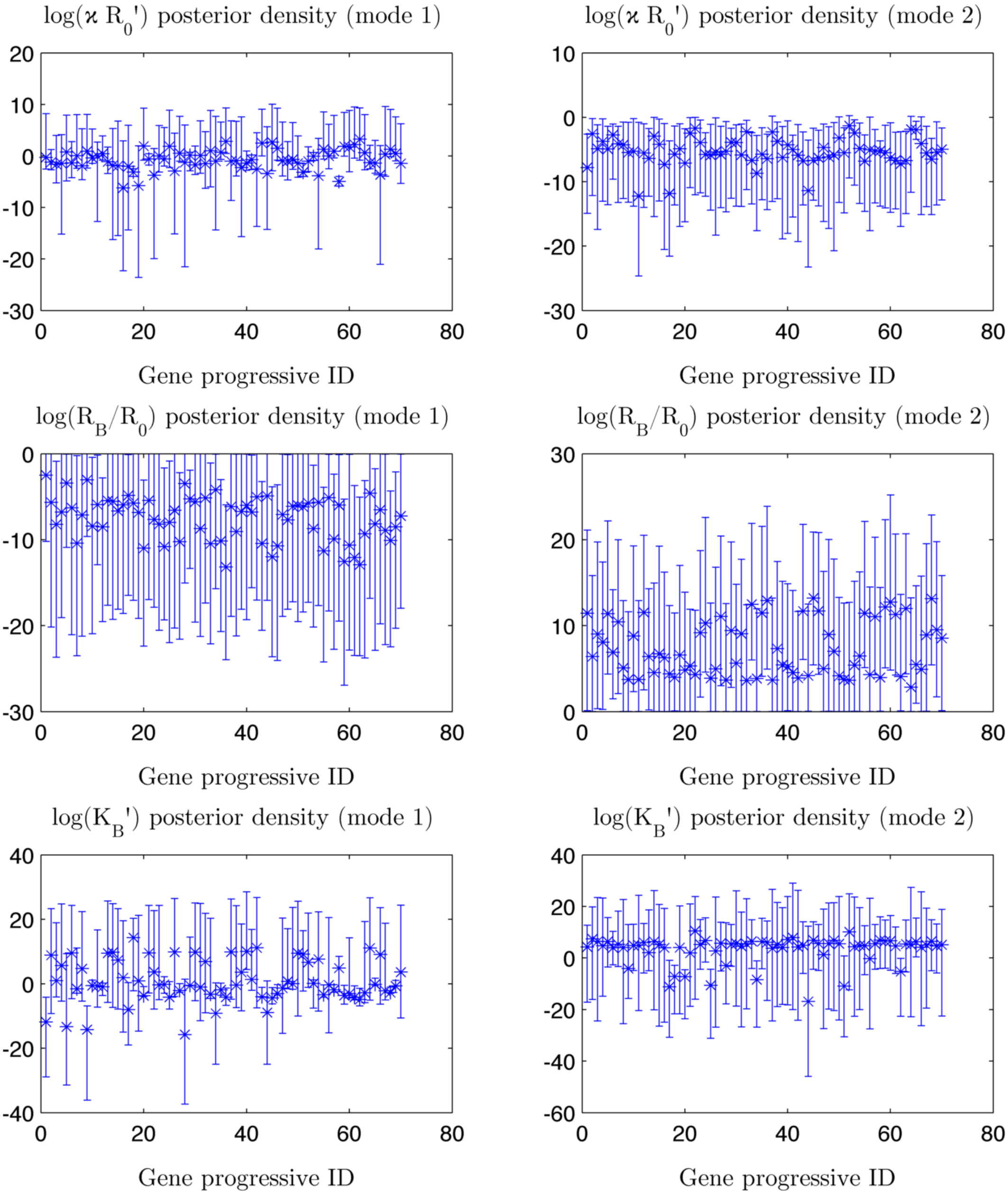


Figure 4.17: Posterior densities for the transcription function parameter of Model 3.7, as applied to the Nanostring rhythmic genes: median (star) and 95 % HPDIs (bars), by gene. Nanostring data-set, Carré lab.

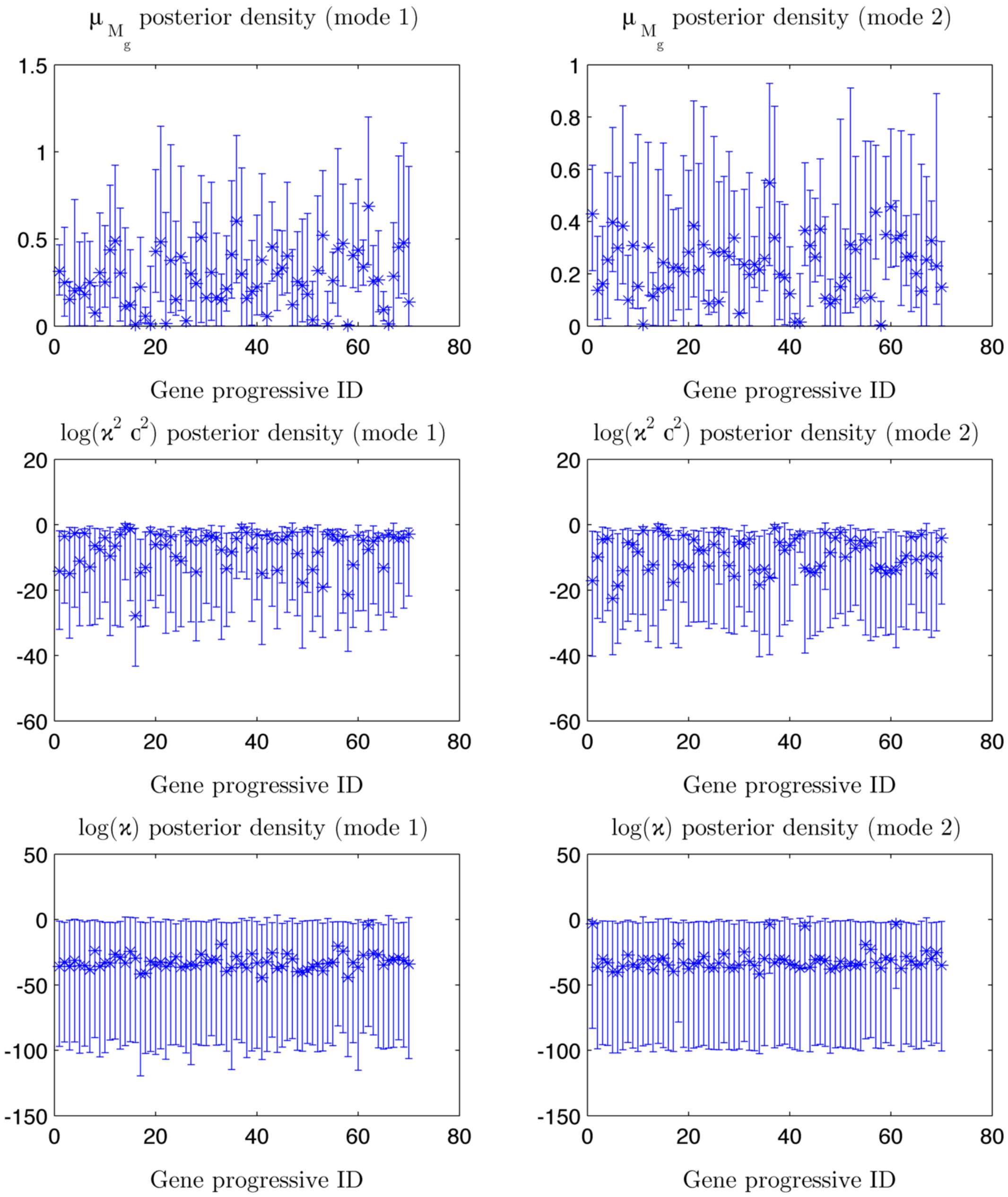


Figure 4.18: Posterior densities for the degradation, scale and noise parameter of Model 3.7, as applied to the Nanostring rhythmic genes: median (star) and 95 % HPDIs (bars), by gene. Nanostring data-set, Carré lab.

Diagnostics

In this section we perform a comprehensive study of model fit for the Nanostring rhythmic genes. A sample of standardised residuals is obtained for each thinned MCMC sample, and the Shapiro-Wilk test statistics, as well as the normalised periodogram, are then computed. An empirical distribution for each test statistic, for all the analysed genes, and for both modes, is thus available.

The Shapiro-Wilk test is aimed at assessing normality and has critical value at level $1 - \alpha = 95\%$, for a sample of 24 data-points, equal to 0.916 (Shaphiro and Wilk, 1965). The Shapiro-Wilk test is here computed with the *swtest* function in MATLAB (Saida, 2007), which performs either the Shapiro-Wilk or the Shapiro-Francia test based on the sample kurtosis. In particular, medians and HPDIs of the Shapiro-Wilk test statistics distributions, for all the genes and for the two modes, are shown in the top panels of Figure 4.19. We can observe that the assumption of normality is generally adequate, being most of the mass of the posterior distributions above the threshold level.

The normalised periodogram introduced in Section 4.2.1 enables the investigation of residual periodicities. The central panels of Figure 4.19 show the medians and HPDIs of the normalised periodogram estimate densities for the frequency 1/24 cycles/h, corresponding to the circadian periodicity. We notice that, for some genes, the 24 hour periodicity is explained to a reasonable degree, as a significant mass of the test distributions below the 95 % threshold level of 2.99 demonstrates. For other genes, the result is more ambiguous: while the threshold level belongs to an area of non-negligible density, we can see that most of the density mass is above the threshold level itself. This suggests that, although part of the density of the 24 hour periodogram estimate falls in a region where no residual circadian rhythmicity is present, for most of its mass this is not the case, and in such cases we would be skeptical about the goodness of fit.

As a general comment relative to the 24 hour periodicity fit, we have generally noticed that our model encounters difficulties when the two cycles are substantially different, and in particular in the presence of sharp peaks and abrupt changes, which cannot be explained by measurement error only.

Finally, the bottom panels Figure 4.19 show the median and HPDI of the normalised periodogram estimate densities for the frequency 1/12 cycles/h, which can also be of interest for circadian genes. In this case, we see that there is no evidence against the hypothesis that no 12 hour periodicity is present in the residuals, for all the analysed genes.

Correlation with LHY

An important analysis for the purposes of our study concerns the possible relationship between the reconstructed TF and the observed LHY protein, as one of the aims of our analysis is to put forward the hypothesis that the unobserved TF is

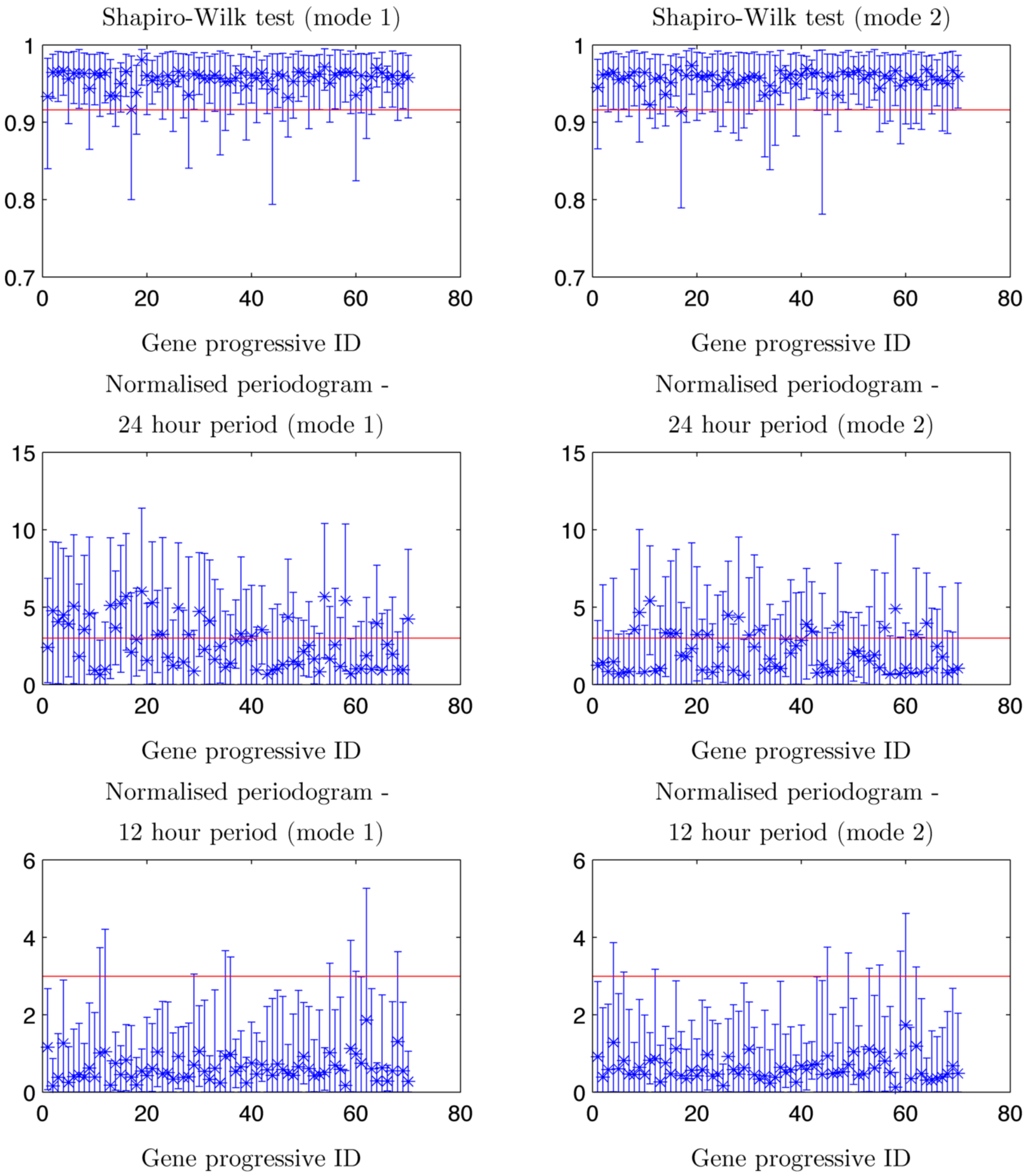


Figure 4.19: Diagnostics plots for model fit: Shapiro-Wilk test (top), normalised periodogram estimate for the 24 hour period (centre), normalised periodogram estimate for the 12 hour period (bottom), computed on the standardised residuals for Model 3.7, as applied to the Nanostring rhythmic genes. Median (star) and 95 % HPDIs (bars), red line superimposed at the 95 % significance value, under the null hypothesis that samples are i.i.d. from a $\mathcal{N}(0, 1)$. Nanostring data-set rhythmic genes, Carré lab.

'close' to LHY itself. We assess this relationship by computing, for each gene, a sample of correlation coefficients between the inferred unobserved TF profiles, and the LHY protein time-series.

To ensure the accuracy of our analysis, we retain only the genes for which a satisfactory fit of the circadian rhythmicity is achieved by the model. A good fit is assumed under two conditions: first, there is a reliable identification of the unobserved TF, i.e. if the variability associated with the inferred smoothing density is such that a constant time series has negligible probability (less than 5%); second, the median normalised periodogram estimate for the 24 hour period of the standardised residuals is below 2. The latter seems a reasonable threshold, which additionally allows to control for mRNA model fit. It is also a more restrictive criterion than assessing whether the standardised residuals of the median model fit have a normalised periodogram estimate for the 24 hour period which falls below the 95% threshold 2.99, but less restrictive than assuming a fulfilment of the standard exponential distribution fit for the full density of 24 hour periodogram estimates of the residuals. Any statistical approach employed to assess the latter, e.g. the Kolmogorov-Smirnov test, is doomed to result significant against the null hypothesis in the majority of cases, due to the high, and potentially infinite, sample size provided by the number of MCMC samples. We believe, on the other hand, that even if the 24 hour rhythmicity is still present in a minor proportion of the residuals, if this proportion is small and the model has achieved the identification of an unobserved TF, it is sensible to assume that the model is providing useful information about the unobserved regulator and the transcriptional dynamics.

A posterior sample from the density of the unobserved states, given the observations and the parameters, is provided by a draw from the smoothing distribution introduced in Section 2.4. The mean and variance of the normal smoothing distribution is computed for a thinned set of MCMC samples, and then for each thinned MCMC sample, a posterior sample for the unobserved TF and mRNA profiles is drawn.

We then apply the following procedure: we start by building N matrices, one for each analysed gene, containing in each row a sampled posterior smoothing time-series of the unobserved TF. We then compute the correlation between the LHY observed, smoothed, profile, and the samples in each of the N matrices. This procedure provides a sample of correlation values between the LHY profile and the unobserved inferred TF of each gene, the number of samples being equal to the number of MCMC iterations retained.

When we achieve a reliable identification of the unobserved TF in both modes, we have a bimodal distribution for the pairwise correlations, i.e. between LHY and the unobserved TF, if the correlation is significantly different from zero.

A sensible approach for plotting the results is, for example, by means of violin plots (Dorn, 2009): the width of the plot is approximately proportional to the density at the specified point of the y axis, obtained as a kernel smoothing estimate of the data histogram. We adopt the Fisher transformation of the correlation coefficient χ, which is given by $\log[(1+\chi)/(1-\chi)]/2$ (see e.g. Pace and Salvan, 1997, Chapter 8), in order to avoid boundary effects. The Fisher transformation has also the advantage of improving normality, when computing the correlation on i.i.d. pairs from a bivariate normal distribution. The transformation has been shown to reduce skewness and stabilise the variance (see Pace and Salvan, 1997, Chapter 8, and references therein). The gain is however less significant when observations come from time-series data, as investigated in Thompson and Fransson (2016).

Figures 4.20 show such violin plots of the distribution of the correlation between the reconstructed TF and LHY for the genes with the highest correlation, in absolute value. Plots for the remaining genes are provided in Figure D.1 in Appendix. We have sorted the genes so that on the left-hand side of the plot we have the genes with the highest median correlation, in absolute value. We recognise in the high-correlation group genes which are known to belong to the central clock of the *Arabidopsis Thaliana*, and to be repressed by LHY (Adams et al., 2015), namely ELF3, PRR9, CAB1, CCA1, TOC1, ELF4, and LUX. The observed correlation is however not always close to 1, pointing in the direction of possible additional regulators.

It is worth remarking that the correlation coefficient is aimed at assessing a linear relationship. We may be losing something in terms of sensitivity of our analysis if a non-linear relationship is present (see Quian Quiroga, 2009); on the other hand, the presence of bimodality in the unobserved TFs profiles motivates this choice, as a first simple and easily interpretable exploratory approach.

mRNA clustering

A further interesting point is the correlation between unobserved mRNAs. Given the assumed model, it is sensible to expect that a correlation between the unobserved TFs of two different genes, reflects a correlation between the unobserved mRNAs of the same genes. It is therefore of interest to analyse the posterior smoothing profiles of the child mRNAs by identifying homogenous clusters of expression. These clusters can then be compared according to the presence of binding sites and the result of the induction experiment.

We form N_c matrices, each containing one posterior mRNA profile for each gene. Note that either some mRNA samples are used more than once, or a sub-

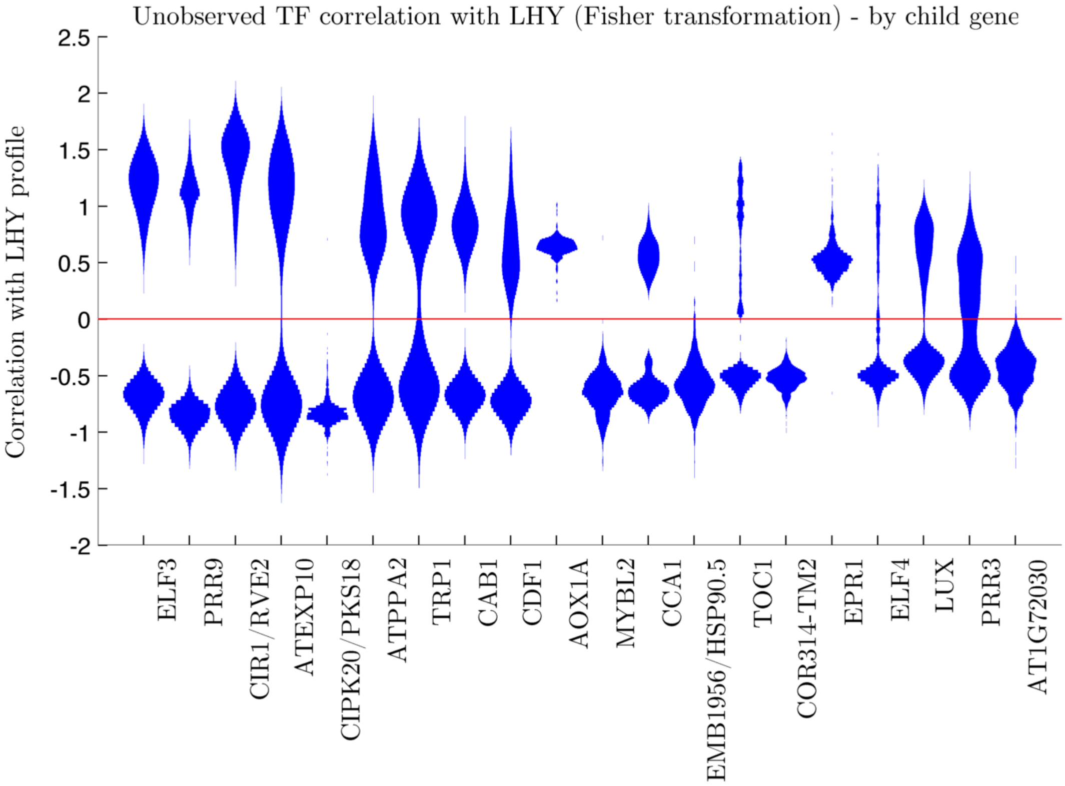


Figure 4.20: Violin plots for the correlation of the posterior unobserved TFs profiles and smoothed observed LHY. Fisher transformation of the correlation coefficient. Genes between positions 1 and 20, in order of median posterior correlation, in absolute value. Note that for a limited number of genes only one mode satisfies the model fit requirements. Rhythmic Nanostring genes, with satisfactory explained circadian rhythmicity, Carré lab. Plot code from Dorn (2009).

sample of them is employed, if the chains have a different number of iterations required for convergence for different genes. We adopt the second approach, and apply the k-means clustering algorithm (MacQueen, 1967) to each of the N_c matrixes, as implemented in MATLAB. Recall that the algorithm is aimed at identifying, for a given number of clusters, an optimal partition of the units, according to one or more variables of interest, time-points in our case. By optimal, it is meant that it minimises the sum of 'within cluster' deviance, i.e. the sum, over all clusters, variables and observations, of the squared differences between the observations and their assigned cluster centre (Fabbris, 1997, Chapter 8).

A crucial choice concerns therefore the number of clusters k. We run the clustering algorithm for each of the N_c matrixes and for an increasing number of clusters, i.e. from 1 to 33 (the total number of genes is 34), by setting the number of replicates to 100. Replication is required in order to ensure that the best classification is achieved, and therefore the total 'within cluster' sum of deviances is monotonically decreasing as the number of clusters increases. An appropriate test statistics to assess whether an additional cluster provides a significant drop of total within cluster deviance is provided in Beale (1969), cited in Everitt et al. (2011) and Fabbris (1997, Chapter 8). Figure D.2 (a) in Appendix shows the progressive decrease in the 'within cluster' deviance, as we increase the number of clusters. We observe a slight elbow in the plot of the deviances at 5, indicating that additional increases in the number of clusters are likely to mostly explain variability due to noise. Moreover, Figure D.2 (b) in Appendix shows the value assumed by Beale's F statistics for an increasing number of clusters, and for all N_c samples, along with the significance threshold at level $1 - \alpha = 95\%$. A clear peak in Beale's F statistics is observed at 4, indicating that the increase from four to five clusters has the most beneficial effect in terms of reducing the 'within cluster' deviance. We therefore opt for five clusters.

The k-means algorithm, applied to the N_c matrices with number of clusters set to 5, identifies a total of 38 different partitions. Each partition occurs with a particular frequency among the N_c samples, and thus a measure of their probability is easily obtained. The three most probable partitions have frequencies equal to approximately 0.21, 0.17 and 0.15, hence representing about the 53% of the total samples. Note that running the clustering algorithm on the median mRNA posterior profiles does *not* identify the most probable partition in our case.

Now we focus on the most probable partition. We provide in Figure 4.21 the cluster centres, which translate into five mean profiles. We can see that the five clusters basically identify four phase groups, having peaks approximately equally spaced across the circadian day, and an additional cluster which shows a bimodal peak in both cycles.

We observe in Tables 4.2 and 4.3 the distribution of the genes according to the five clusters identified by the clustering procedure, and binding site group and induction experiment result, respectively. Table 4.2 has a p-value for the Fisher's exact test for association equal to 0.17, pointing in the direction of an association between cluster group and presence of binding sites. Due to the low sample size, we have also performed the same test assuming four clusters; each cluster represents in this case approximately a phase-group, as the less important cluster among the initial five, is cluster 5, whose centre has the doubly peaked profile in Figure 4.21. We obtain in this case a p-value equal to 2.4×10^{-2}. The results overall suggest that the presence of binding sites in the promoter of a gene has an influence on the observed expression profile - note that causality is not implied by the test, but an

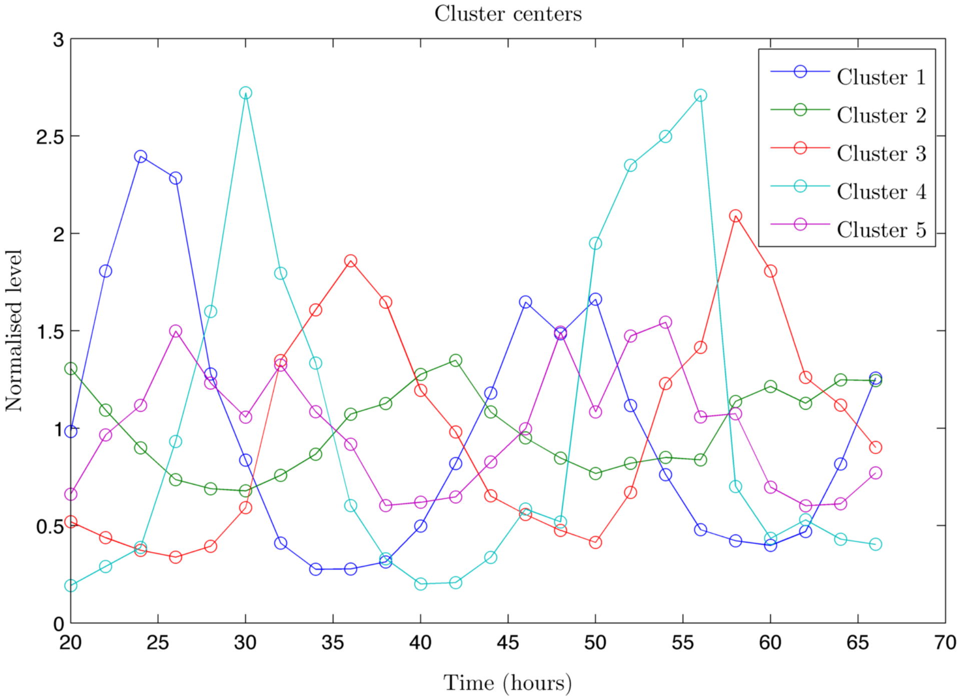


Figure 4.21: Centres of the clusters identified by the k-means algorithm, as applied to the samples posterior profiles of the child mRNA. Most probable partition. Rhythmic Nanostring genes, with satisfactory explained circadian rhythmicity. Nanostring data-set, Carré lab.

implication of the type of variables involved.

On the other hand, Table 4.3 has a p-value equal to 0.55 for the association between cluster group and induction by LHY, pointing in the direction of a not significant effect. The p-value if four clusters are assumed is equal to 0.37, thus still indicating no association. One possibility is that LHY is an important regulator, but it is not sufficient to explain the observed profiles. This result, as well as the comparison of the unobserved TFs profiles with LHY, suggests a more complex form of regulation.

4.3 Discussion

In this fourth Chapter, we have studied three modelling approaches of transcriptional regulation, validated through a simulation study. We have applied the model which assumes only one observed regulator, in this case LHY, to the Nanostring rhythmic data, although we note that it is generally not able to fit the available data. The model comprising one unobserved TF is more flexible, and allows to infer

Binding sites	Cluster					Total
	1	2	3	4	5	
None	3	1	1	0	0	5
CBS only	0	1	0	0	0	1
ABRE only	0	2	0	1	0	3
EE only	2	0	2	5	2	11
CBS + ABRE	1	0	0	0	0	1
CBS + EE	0	0	1	0	0	1
ABRE + HEX	1	1	0	0	0	2
ABRE + EE	2	0	0	1	0	3
HEX + EE	0	0	1	2	1	4
CBS + ABRE + EE	1	0	0	0	0	1
ABRE + HEX + EE	0	0	1	0	0	1
CBS + ABRE + HEX + EE	0	0	1	0	0	1
Total	10	5	7	9	3	34

Table 4.2: Rhythmic Nanostring genes, with satisfactory explained circadian rhythmicity, by presence of binding sites in the promoter region and cluster group. A binding site is present if there is at least one binding site of the corresponding type in the promoter. Centres of the five cluster groups are shown in Figure 4.21. Fisher's exact test for association has p-value 0.17. Nanostring data-set, Carré lab. at Warwick.

Induction	Cluster					Total
	1	2	3	4	5	
-2	0	0	1	0	0	1
-1	6	3	4	8	2	23
0	4	1	2	1	1	9
1	0	1	0	0	0	1
Total	10	5	7	9	3	34

Table 4.3: Rhythmic Nanostring genes, with satisfactory explained circadian rhythmicity, by induction experiment result and cluster group. Induction is assessed at significance level $\alpha = 0.1$ (-2 indicates consistent repression, -1 repression, 0 no effect, 1 activation). Centres of the five cluster groups are shown in Figure 4.21. Fisher's exact test for association has p-value 0.55. Nanostring data-set, Carré lab. at Warwick.

a distribution of profiles for the unobserved TF, which is then compared with LHY, to assess whether LHY itself can still play a major role in the regulation of the available genes. A third model, which comprises both LHY and an unobserved TF requires unfortunately prior information which is only partially available, and can therefore not be applied to the Nanostring data.

A preliminary analysis of the Nanostring rhythmic genes and available prior experimental information reveals no significant association between the presence of binding site combinations, and induction by LHY. On the other hand, a possible relationship between presence of binding sites and amplitude and phase is suggested by the corresponding box plots.

The application of the chosen model to the Nanostring data provides a mechanistic description of the transcriptional process associated to the given genes; it gives a posterior density for the unobserved mRNA and TF profiles of each gene. The availability of a distribution of profiles, rather than a single point-estimate, results in a significant advantage, as it allows to compute a distribution for any desired synchrony index, exemplified in our case by the computation of the correlation coefficient between the unobserved TF and LHY. Moreover, it has allowed to identify the most probable clustering of the unobserved mRNA of the available genes.

We have observed a high correlation between the unobserved TF and LHY profile for several genes, among which we recognise known components of the *Arabidopsis Thaliana* central clock, namely ELF3, PRR9, CAB1, CCA1, TOC1, ELF4, and LUX. Clustering of the unobserved mRNA profiles reveals a possible association between cluster group and presence of binding sites, p-value equal to 0.17, which decreases to 2.4×10^{-2} when four phase-groups are assumed, supporting the hypothesis that binding sites play an important role in defining the expression profile of a putative child mRNA. On the other hand, the relationship between cluster group and induction by LHY is not significant, p-value equal to 0.55. The data analysis results seem to indicate that, although LHY is known to probably be an important regulator for the child genes, it is not enough to explain the observed dynamics, and its effect is not strongly linked to the presence of binding sites. It is therefore highly likely that additional factors and mechanisms are playing a role.

With respect to the modelling approach, several extensions and new directions may be proposed, in particular, the Fourier model for the unobserved TF may be too simple. The approximation of an unknown TF may be improved by estimating the period as an additional parameter, or, in a more biologically-interpretable model, by modelling both the TF mRNA and protein levels with, for example, a transcriptional switch model for its mRNA, and the subsequent translation of the TF mRNA into protein. On the other hand, these extensions would require an increase in the complexity of the model, which already comprises 20 parameters for the 24 observations available for each gene. Our simulation study suggests that more refined models can, and indeed should, be considered if more data-points are available.

Part III

Modelling transcriptional regulation of the mammalian clock in the SCN

Chapter 5

Modelling and methods for mice SCN circadian dynamics

5.1 The mammalian clock

Circadian rhythms - i.e. rhythms that have the characteristics described in Chapter 3 - are observed also in mammals. A self sustained, highly synchronised and light-entrained clock is located in a region of the brain called suprachiasmatic nucleus (SCN) (see e.g. Dibner et al., 2010; Colwell, 2011; Dibner and Schibler, 2015; Hastings et al., 2008). The SCN also coordinates several peripheral clocks, observed in the major organs and responsible for the production of tissue-specific proteins (see e.g. Hastings et al., 2014; Dibner et al., 2010).

Robust oscillations are achieved thanks to two main interlocked transcriptional and translational feedback loops (TTFL) (see e.g. Hastings et al., 2008; Dibner and Schibler, 2015). In the first loop, the genes *Per* and *Cry* are activated through the binding of the CLOCK/BMAL protein complex to their promoters during the circadian morning, and then are auto-repressed by their own protein products in the evening. In the second loop, ROR and REV-ERB proteins regulate transcription of Bmal, whose protein represses in turn *Ror* and *Rev-Erb* mRNA (see e.g. Fuhr et al., 2015; Hastings et al., 2008; Dibner and Schibler, 2015). A comprehensive picture of the TTFL, as well as a detailed mathematical modelling with a set of 20 ODEs is provided in Relógio et al. (2011). Due to the availability of experimental data concerning *Per* and *Cry* genes, here we focus on the former loop. A pictorial representation containing approximate timescales of activation and auto-repression is presented in Hastings et al. (2008), and here schematically reproduced in Figure 5.1.

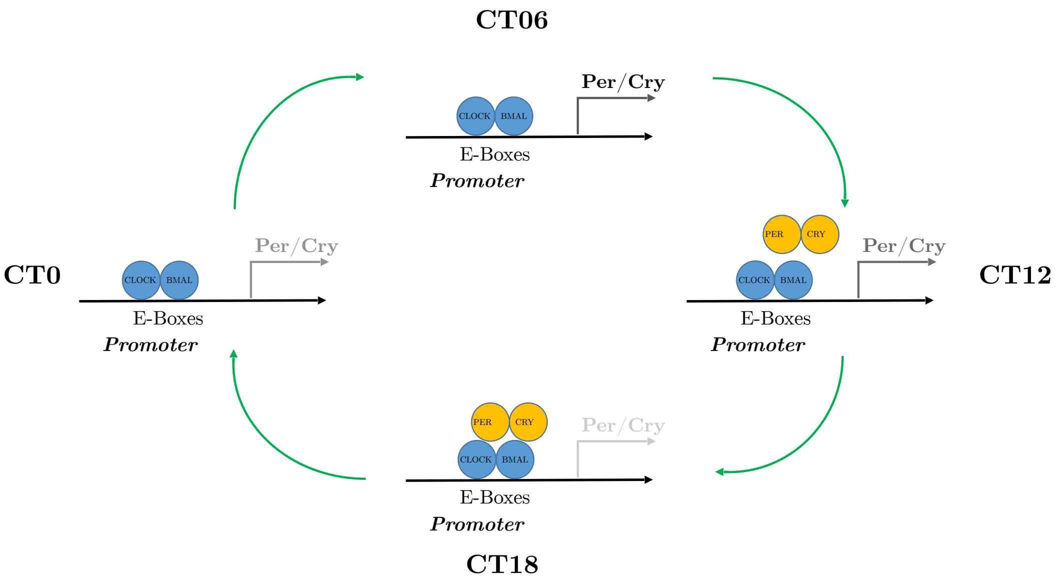


Figure 5.1: Schematic representation of the Per/Cry circadian cycle. The level of Per/Cry transcription is approximately reflected by the grey scale of the Per/Cry symbol, i.e. a darker colour corresponds to a higher transcriptional level. CLOCK/BMAL protein is bound to the enhancer box motifs (E-Box, CACGTG) present in the Per and Cry promoters, activating their transcription. PER/CRY protein complexes, resulting from translation of Per and Cry mRNA, peak in the circadian evening, and inhibit in turn transcriptional activation of CLOCK/BMAL. Adapted from Hastings et al. (2008).

However, experimental evidence suggests that induction of *Per* and *Cry* genes is due partly to the TTFL mechanisms, and partly to cytosolic signalling factors, including Calcium (Hastings et al., 2008; Colwell, 2011). It is currently believed that Calcium plays an important role in the mechanisms which allow different cells in the SCN to 'communicate' and synchronise their circadian oscillations with respect to environmental signals such as light (Brancaccio et al., 2013; DeWoskin et al., 2015; Hastings et al., 2014). The current picture links, in a causal fashion, light stimuli coming from the retina to an increase in electrical firing in the SCN, leading to an increase in Calcium levels and induction of so-called Calcium/cAMP-responsive elements (CREs), and, eventually, induction of Per and Cry genes (Brancaccio et al., 2013; Dibner and Schibler, 2015). Additionally, a key element of overall synchronisation is thought to be represented by the vasoactive intestinal peptide (VIP) (Maywood et al., 2006; Colwell, 2010; DeWoskin et al., 2015). VIP is only produced by neurons belonging to the ventral (core) region of the SCN, but the effects of its release act on the whole SCN by its binding to VPAC2 G-coupled receptors, present in all SCN cells (An et al., 2012). These receptors activate in turn the so-called Gq signalling, which again leads to an increase in Calcium levels (Brancaccio et al., 2013; Hastings et al., 2014).

In Brancaccio et al. (2013), timing and relevance of these events for circadian rhythmicity are experimentally tested. Reprogramming of circadian rhythms of Calcium, CRE and *Per*/*Cry* genes in the SCN, as a consequence of Gq signalling induction in neurons expressing VIP (but not in a VIP null host SCN), is also reported (Brancaccio et al., 2013). Moreover, the hypothesis that the temporal pattern of *Per* and *Cry* genes phases may be explained by the different sequences of binding sites present in their promoter regions, is put forward (Brancaccio et al., 2013). In particular, *Per*1 and *Per*2 both carry enhancer box motifs (E-Boxes) and a CRE, but in the former their responsiveness is higher (Brancaccio et al., 2013; Travnickova-Bendova et al., 2002). On the other hand, *Cry*1 carries only E-Boxes. In Brancaccio et al. (2013) the peak of Calcium is observed at circadian time 7 (CT07), while *Per*1-*luc*, PER2:LUC and *Cry*1-*luc* (where the symbols '-' and ':' denote a fusion construct of the two genes or proteins, and *luc* is the Luciferase gene) phases are observed approximately 2.6, 4.8 and 5.5 hours later. Impairment of Calcium has also been observed to generate arhythmic expression of *Per*1-*luc* and PER2:LUC (Colwell, 2011). The CRE element is therefore believed to convey the effect of Calcium on transcription of the Per genes, through binding of the protein CREB (DeWoskin et al., 2014). In the CRE-lacking *Cry*1 case, instead, the later peak of expression may be only indirectly related to Calcium, being a consequence of PER/CRY protein repression of CLOCK/BMAL (Brancaccio et al., 2013). A schematic summary of the available literature concerning the promoter regions of *Per*1, *Cry*1 and CRE is provided in Table 5.1.

Gene	Number of CREs (TGACGTCA)	Number of E-Boxes (CACGTG)
CRE (synthetic promoter)	2	0
*Per*1	1 (HR)	3 (HR)
*Cry*1	0	E-Box and E'-Box (CACGTT)

Table 5.1: Binding sites of the promoter regions of CRE, *Per*1, and *Cry*1, and responsiveness: high (HR) and low (LR). For the CRE synthetic promoter, refer to Brancaccio et al. (2013). Characterisation of *Per*1 promoter region is provided in Travnickova-Bendova et al., 2002, while *Cry*1 is studied in Fustin et al., 2009.

With respect to the overall synchronization of the SCN, it is important to note that phases of circadian expression of core clock genes such as *Per*2, *Cry*1 and *Bmal*1 follow a waveform trajectory starting from the dorsal (external) area of the SCN and spreading towards the ventral (core) area (Yamaguchi et al., 2003; Maywood et al., 2013; Myung et al., 2015; Evans at al., 2013). Cytosolic signalling, the aforementioned Calcium and VIP, as well as the γ-aminobutyric acid (GABA), are believed to play a crucial role in extracellular communication, but the exact mechanisms represent an active area of research (Hastings et al., 2014; Hastings et al., 2008; DeWoskin et al., 2014).

At the organism level, finally, metabolic and endocrine signals propagate from the SCN to other regions of the brain and to peripheral clocks, thus allowing a coherent and efficient temporal organisation of the overall metabolism (see e.g. Dibner et al., 2010).

5.2 Motivation, available data and background

As described in the brief biological introduction of the previous section, a relatively detailed picture of the single-cell circadian clock in mammals is now available, but the overall synchronisation and coordination between different cells still represents an open area of research. Recent work has focused on mathematical modelling of cellular clocks coupling, e.g. Gonze et al. (2005), Ananthasubramaniam et al. (2014), DeWoskin et al. (2015). The proposed approaches are based on deterministic dynamics, and do not perform parameter estimation. The modelling of DeWoskin et al. (2015) is however based on a previous work by Kim and Forger (2010), where parameters are estimated by sequentially minimising, via simulated annealing, two cost functions: the first cost function is a function of the squared difference between measured expression levels of the genes involved, and trajectories simulated according to the proposed model; the second cost function additionally incorporates the square of the ratio between the simulated and the experimentally measured period of phenotypes of mutations, minus one. The model of Kim and Forger (2010) is again based on deterministic dynamics, and confidence or credible intervals for the parameters are not provided.

We build a stochastic model which can take advantage of the observed spatio-temporal luminescence levels of Calcium, as well as $Per1$, $Cry1$ and CRE reporters, to describe the transcriptional dynamics of $Per1$, $Cry1$ and CRE and to link in a causal relationship the effect of Calcium on $Per1$ and CRE transcription. We also propose a methodology which can be readily applied to perform parameter inference in a Bayesian framework (although it is in principle not restricted to a Bayesian approach), where relevant prior information can be incorporated.

Due to time constraints we perform inference only for the parameters related to $Cry1$, as a first simple and motivating example. Prior information relative to the dissociation coefficient associated with the E-Box elements obtained from fitting the model to $Cry1$, can be incorporated at a later stage in the $Per1$ model (we assume it to be a reasonable proxy for the dissociation coefficient of the PER/CRY complex binding to the E-Box motif), along with an analogous prior which could be ideally obtained from the CRE data. The availability of prior information concerning the dissociation coefficient of $Cry1$ to the E-Box and of Calcium to CRE, would help parameter identifiability in the $Per1$ model; the analysis, once completed, has the potential to provide insight in the auto-regulatory process of *Per* and *Cry* genes, their responsiveness to Calcium, and the mechanisms by which synchronisation is achieved across the SCN.

5.2.1 Mice SCN available data

The available data, from the Hastings lab. at MRC Cambridge, comprise time series of Calcium and $Per1$-*luc*, $Cry1$-*luc* and CRE-luc in a spatial fashion across the SCN. Observations consist of light intensities recorded every 0.5 hour, for 4-5 days (length may vary across experimental replicates) on SCN slices. We provide a representation of the available data, together with different levels of spatial aggregation for $Cry1$-*luc* in Figure 5.2. $Per1$-*luc* and Calcium data are shown in Figure 5.3. Two additional experimental replicates of $Cry1$-*luc* data are available (not shown).

In the case of Calcium, signals are obtained by inserting a fluorescent protein, GcAMP3, containing a sequence responsive to Calcium, via viral transduction, i.e. by injecting a virus (Brancaccio et al., 2013; Tian et al., 2009). When Calcium binds its target sequence, light is emitted by the protein. The fluorescence observed is therefore proportional to real time Calcium levels.

$Per1$-*luc*, $Cry1$-*luc* and CRE-luc signals are obtained via a transcriptional luciferase reporter construct. This means that the gene encoding *Luc* is inserted near the promoter region of $Per1$, $Cry1$ and CRE, and can be assumed to be activated at the same time as the respective genes. *Luc* mRNA is then translated into protein. Finally, LUC protein reacts with a luciferin substrate, an enzyme which causes luminescence, thus emitting a light intensity which is then integrated and recorded by a special camera every 0.5 hour (further details are provided in the supplementary material of Brancaccio et al., 2013).

Light intensities are recorded per pixel, whose size is consistently smaller than that of an average cell. The number of pixels vary across experiments and experimental replicates, but is on the order of approximately $7 - 8 \times 10^4$ pixels per experiment. From personal communication with M. Hastings' group, a cell should indeed be covering a square of roughly 8 × 8 pixels. It is in theory possible for a pixel to be at the intersection of two or more cells, and we therefore need to assume that the condition for aggregation introduced in Chapter 2 holds. We return to this point in Section 5.3. If we accept the aggregation assumption for neighbouring cells, we can choose a suitable aggregation level for the data. Ideally, we want to avoid dealing with very low counts, as the normal approximations crucial to most of the available inferential methods may not hold. On the other hand, we also wish to observe possible spatial variations of the regulatory dynamics across the SCN, as well as avoid aggregation of significantly different cells. A good compromise seems to focus on 2 × 2 pixel boxes, based on confidential data regarding the number of PER2 protein molecules per cell observed in fibroblasts; we also take into account that the number of mRNA molecules for a given gene is generally believed to be much smaller than the number of protein molecules (Suter et al., 2011).

Two main possible issues can be observed in the data. One is the effect of saturation, which means that signals having intensity higher than a set threshold level, are all measured as equal to the threshold level itself: the saturation effect is most likely present when 'flat' peaks are observed. This feature may cause significant problems during the inferential process, and thus may require further *ad hoc* modelling. However, we notice that saturation in *Cry*1-*luc* levels is mostly restricted to the first cycle, which does not have a crucial importance in terms of parameter inference, as we model it only approximately to serve as a delayed input for the exact model of the following cycle (see Section 5.6 for further details).

A second feature is represented by an upwards trend in Calcium levels, and a decreasing trend particularly in the amplitudes of *Cry*1-*luc*. Both trends are due to experimental procedures (M. Hastings personal communication). In particular, the trend of *Cry*1-*luc* is generated by consumption of the luciferin substrate over time. As these features of the data are known to be due to the experimental process, we consider it sensible to de-trend the data.

Here we deal with the trend in *Cry*1-*luc*, as its model is the focus of our inference in Chapter 6. We divide the observations by a time-varying proportionality factor, as measured in Maywood et al. (2013). In particular, a mean decrease of approximately 30% over 4 days is shown. For simplicity, we here assume a linear decay. The adopted solution is of course relatively rough, given also the wide variability in the rate of decay across different locations, but it seems to perform on average sufficiently well, as it can be observed in Figure 5.4.

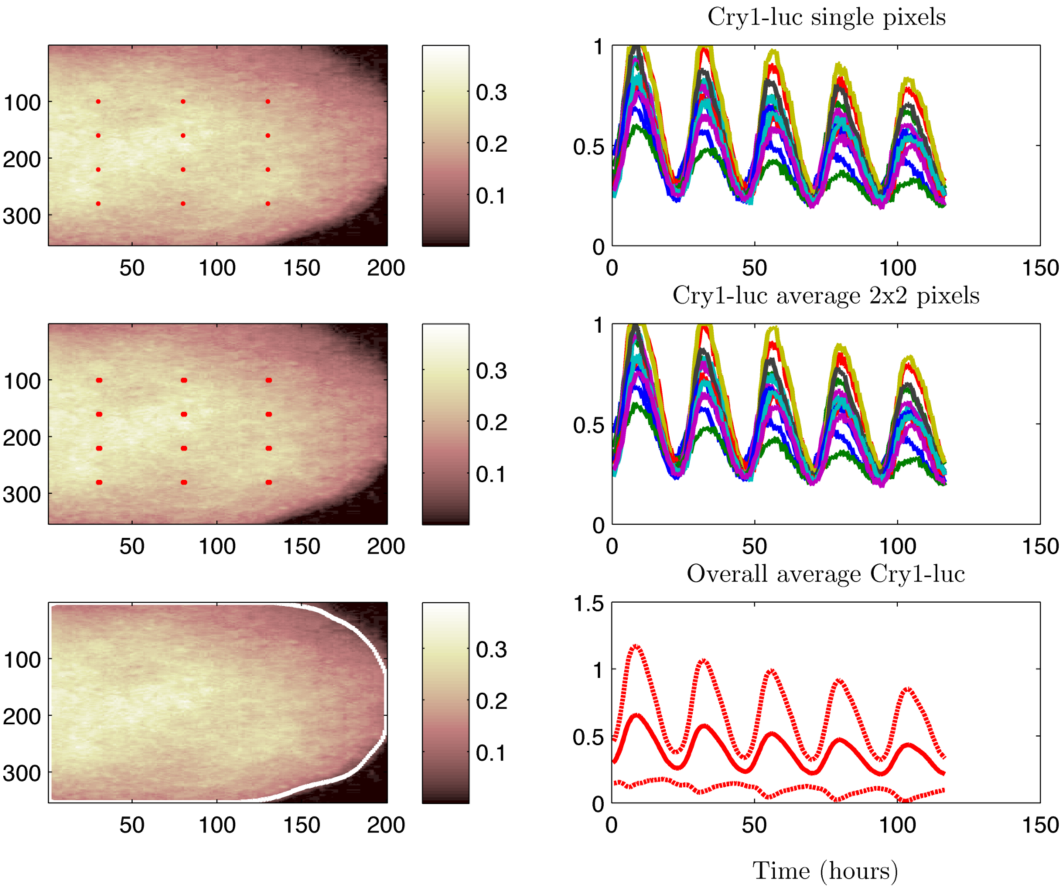


Figure 5.2: For columns from left to right: image of the mouse SCN at observation time 1 for *Cry*1-*luc*, and 12 arbitrary locations (red). Time-series for *Cry*1-*luc* for the 12 arbitrary locations in white. Results for 1 pixel at the 12 locations (top), averaged over 2 × 2 pixels at the 12 locations (center), and averaged over the whole SCN (bottom). Data from Hastings lab. at MRC, Cambridge. Code partially provided by K. Hey.

5.2.2 Mathematical modelling of mammalian clock gene dynamics: some background

In this section we provide a brief overview of some existing mathematical modelling of the mammalian clock. We refer in particular to the work of Relógio et al. (2011), Korenčič et al. (2012), Gonze et al. (2005) and Ananthasubramaniam et al. (2014).

Existing modeling approaches are mainly restricted to the deterministic case. A good starting point is the work of Relógio et al. (2011), where a set of 20 ODEs describes the dynamical evolution of the two main feedback loops, namely the *Per*/*Cry*, and the *Ror*/*Rev*-*Erb* loop. We focus on the main loop, comprising only *Per* and *Cry*. The ODE for *Per* mRNA in the Relógio et al. model is given

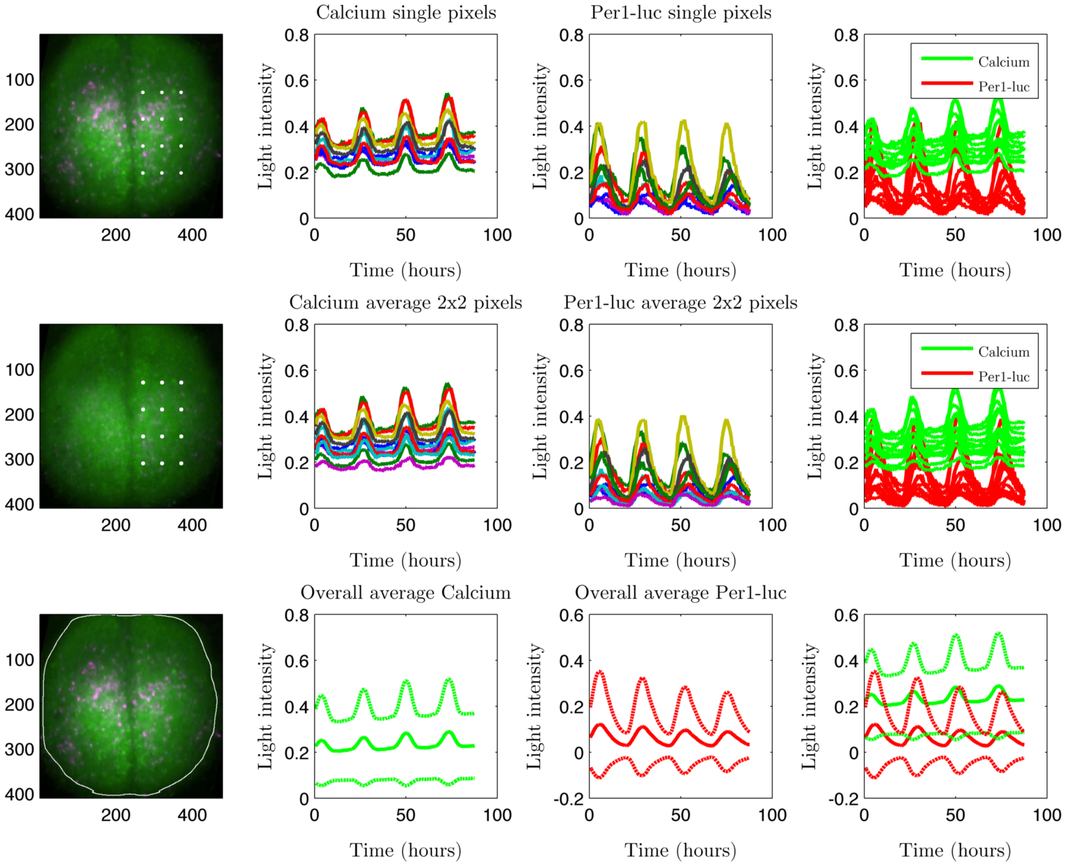


Figure 5.3: For columns from left to right: image of the mouse SCN at observation time 1 (green luminescence represents Calcium, magenta *Per*1-*luc*) and 12 arbitary locations (red), time-series for Calcium, for *Per*1-*luc*, and Calcium and *Per*1-*luc*, for the 12 arbitrary locations in white. Results for 1 pixel at the 12 locations (top row), averaged over 2×2 pixels at the 12 locations (central row), and averaged over the whole SCN (mean and ±2 SD intervals, bottom row). Data from Hastings lab. at MRC, Cambridge. Code partially provided by K. Hey.

by

$$\frac{dPer(t)}{dt} = V_{1max} \frac{1 + a\left(\frac{CB(t)}{k_{t1}}\right)^b}{1 + \left(\frac{CB(t)}{k_{t1}}\right)^b + \left(\frac{CB(t)}{k_{t1}}\right)^a b \left(\frac{PC(t)}{K_{i1}}\right)^c} - d_{Per}Per(t),$$

where CB and PC represent the CLOCK/BMAL and the PER/CRY protein complexes, respectively (see Figure 5.1), and we have followed the original paper notation for the remaining parameters, namely: V_{1max} is defined as the transcriptional rate of Per, a as the transcription fold activation, b and c as the Hill coefficients, d_{Per} as Per degradation rate, and k_{t1} and k_{i1} as Per activation and inhibition rate, respectively. The equation assumes the transcription of Per as induced by CLOCK/BMAL

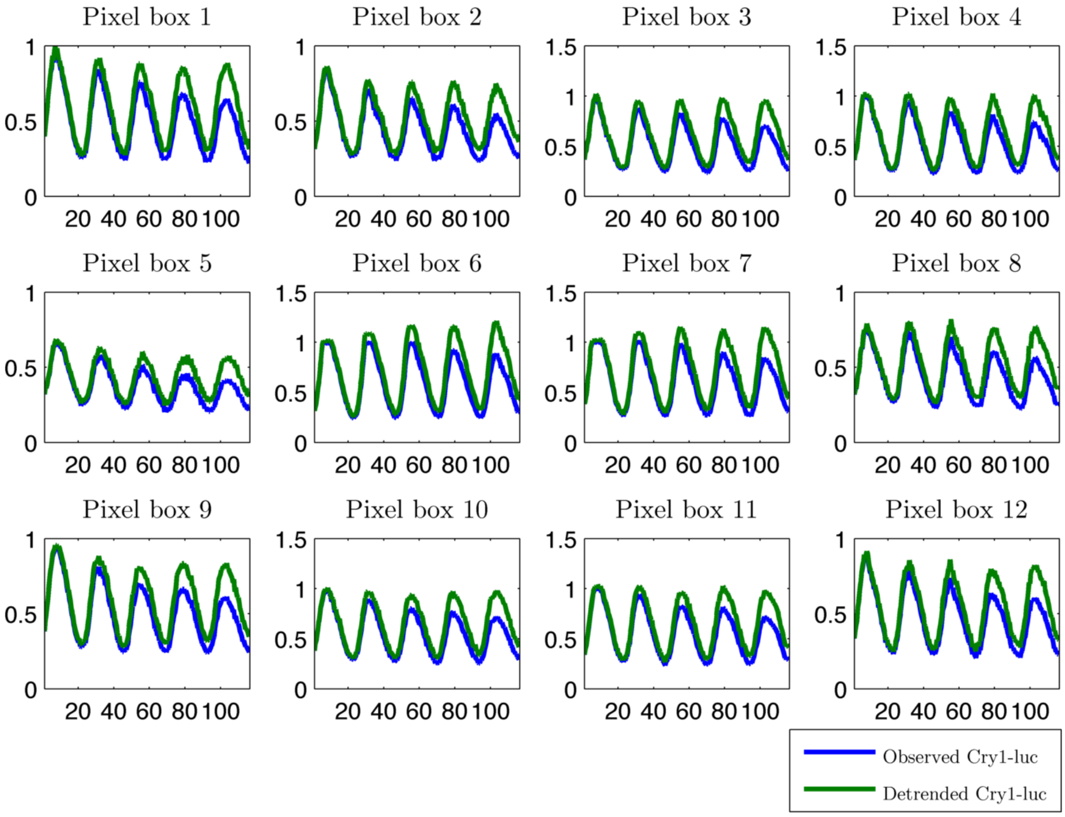


Figure 5.4: Observed *Cry-luc* (blue) and de-trended *Cry-luc* (green), for 12 pixel boxes, at the same locations across the SCN as Figure 5.3. Data from Hastings lab. at MRC, Cambridge.

protein complex, and repressed by PER/CRY protein complex. The parameters a, b, K_{t1} and K_{i1} tune the effect of the regulatory proteins, in the sense described in Chapter 1. The Relógio et al. model additionally contains a detailed description of the PER/CRY complex formation, comprising eight ODEs, accounting for nuclear export, translation, complex assembly and nuclear import. Further parts of their model concerns complex formation of CLOCK/BMAL, as well as a part of the *Ror*/*Rev-Erb* loop. Parameters in the model are set according to literature sources, when available, or in order to match phases and amplitudes observed in data.

The full Relógio et al. model is too complex to be applicable in our inferential framework. Further simplifications are required, and a step forward in this direction is provided in Korenčič et al. (2012). The authors focus on *Per*2, and apply two approximations. First, CLOCK/BMAL is assumed to be constant, and therefore its effect is incorporated in the basal transcriptional rate (R_0 in our usual notation); secondly, PER2 protein is represented as a delayed *Per*2. In this way, the *Per*2 mRNA equation can indeed represent the entire auto-repressive feedback loop. The equation becomes, with minor rearrangements, (Korenčič et al., 2012)

$$\frac{dPer2(t)}{dt} = \frac{\left(\frac{c}{ck}\right)^2}{1 + 2\frac{Per2(\tau_{Per2})}{ck} + \left(\frac{Per2(\tau_{Per2})}{ck}\right)^2} - d_{Per2}Per2(t),$$

where $Per2(\tau_{Per2})$ represents the $Per2$ input, delayed by time τ_{Per2}, and, in the authors' notation, ck can be interpreted as the dissociation coefficient for $Per2$, and $(c/ck)^2$ is equivalent to R_0, in our usual notation. Moreover, note that the exponent c of Relógio et al. (2011), is now substituted by 2, i.e. the number of E-Box like elements in the promoter region of $Per2$.

Another important aspect investigated by Korenčič et al (2012) is the range of delay values τ_{Per2} that gives rise to oscillations. Under the assumed set of parameter values, the delay must be greater than 5.3 hours in order to generate cyclic behaviour in the mRNA levels of $Per2$. Lee at al. (2001), cited in Korenčič et al (2012), report an experimental value of about eight hours for this delay.

The models introduced so far focus mainly on describing the molecular clock at a single-cell level. There is increasing interest in uncovering the mechanisms of synchrony and coupling of the individual cell clocks in the SCN. A step in this direction is provided by the model of Gonze et al. (2005). In their model, the delayed mRNA is replaced by two additional ODEs, accounting for protein translation, and nuclear export/import, respectively. We do not provide further details about this model, as the way of implementing network connectivity is reproduced in Ananthasubramaniam et al. (2014), which we briefly review next, and which is closer to our proposed model.

The model of Ananthasubramaniam et al. (2014) aims at modelling synchrony and entrainment of the clock as a consequence of VIP signalling. In this sense, activation of a putative Per gene, is achieved through both auto-repression, and VIP induction. The proposed mathematical formulation comprises both an 'AND' and an 'OR' gate, which we now describe.

In the 'OR' gate, activation is achieved if Per is low or VIP is high, their effect being additive on the overall transcription rate; the corresponding ODE formulation for Per mRNA expression is

$$\frac{dPer^i(t)}{dt} = \frac{1}{(c + Per^i(\tau_1))^2} + R_T(Per^i(\tau_3))\frac{\sum_j a_{i,j}Per^j(\tau_2)}{K_D + \sum_j a_{i,j}Per^j(\tau_2)} - d_{Per}Per^i(t), \tag{5.1}$$

for $i = 1, ..., N$, where, in the authors' notation, Per^i represents the amount of

Per in cell i, τ_1, τ_2 and τ_3 are delays employed in order to build proxies for PER protein, VPAC2R and VIP, respectively, d_{Per} represents the degradation rate of Per, $R_T(\cdot)$ is a function accounting for VPAC2R expression (in the form of 'a weighted sum of a constant and circadian expression with a constant mean', tuned by an additional parameter which moves the function more towards a circadian or a constant expression), and c and K_D are the dissociation coefficients of Per and total incident VIP, respectively. Finally, $a_{i,j}$ is the element of the matrix A containing the contribution of *'VIP released from neuron j which binds at neuron i'* (where it is also assumed that the sum over i, for a fixed j, is equal to 1). In this way, the second additive term of Equation 5.1 incorporates the effect of the available VIP - after reacting with VPAC2R - on Per transcription.

The 'AND' gate is formulated as

$$\frac{dPer^i(t)}{dt} = \frac{1}{(c + Per^i(\tau_1))^2} \left[1 + R_T(Per^i(\tau_3)) \frac{\sum_j a_{i,j} Per^j(\tau_2)}{K_D + \sum_j a_{i,j} Per^j(\tau_2)} \right] - dPer^i(t),$$

for $i = 1, ..., N$.

However, we claim that a pure 'AND' logic is achieved through a slightly different formulation, i.e.

$$\frac{dPer^i(t)}{dt} = \frac{1}{(c + Per^i(\tau_1))^2} \left[R_T(Per^i(\tau_3)) \frac{\sum_j a_{i,j} Per^j(\tau_2)}{K_D + \sum_j a_{i,j} Per^j(\tau_2)} \right] - dPer^i(t), \tag{5.2}$$

where high levels of transcription are achieved *only* if PER protein is low, and VIP/VPAC2R is high.

It is indeed possible to reformulate Equations 5.1 and 5.2 in a more familiar form, and in our usual notation from Chapter 1. Focusing on a single cell, assuming VPAC2R levels to be constant over time, and VIP to be observed, we obtain from rearrangement of the 'OR' gate of Equation 5.1

$$\frac{dPer(t)}{dt} = \frac{\frac{1}{c^2} + \left(\frac{1}{c^2} + R_T\right) \frac{VIP}{K_D} + R_T \frac{VIP}{K_D} \left(\frac{Per(\tau_1)}{c}\right)^2 + 2R_T \frac{VIP}{K_D} \frac{Per(\tau_1)}{c}}{1 + \frac{VIP}{K_D} + \frac{VIP}{K_D} \left(\frac{Per(\tau_1)}{c}\right)^2 + 2\frac{Per(\tau_1)}{c} + 2\frac{Per(\tau_1)}{c} \frac{VIP}{K_D}} - \mu Per(t).$$

As for the 'AND' gate, we have that Equation 5.2 can be written as

$$\frac{dPer(t)}{dt} = \frac{\left(\frac{1}{c^2} R_T\right) \frac{VIP}{K_D}}{1 + \frac{VIP}{K_D} + \frac{VIP}{K_D} \left(\frac{Per(\tau_1)}{c}\right)^2 + 2\frac{Per(\tau_1)}{c} + 2\frac{Per(\tau_1)}{c} \frac{VIP}{K_D}} - \mu Per(t).$$

The two models can be ultimately unified, resulting in the form presented in Chapter 1, namely

$$\begin{aligned}\frac{dPer(t)}{dt} &= \frac{R_0 + R_{VIP}\frac{VIP}{K_D} + R_{Per,VIP}\frac{VIP}{K_D}\left(\frac{Per(\tau_1)}{c}\right)^2 + 2R_{Per,VIP}\frac{VIP}{K_D}\frac{Per(\tau_1)}{c}}{1 + \frac{VIP}{K_D} + \frac{VIP}{K_D}\left(\frac{Per(\tau_1)}{c}\right)^2 + 2\frac{Per(\tau_1)}{c} + 2\frac{Per(\tau_1)}{c}\frac{VIP}{K_D}} \\ &\quad -\mu Per(t). \end{aligned} \tag{5.3}$$

Note that the ‘AND’ gate is obtained when R_0 and $R_{Per,VIP}$ are set equal to 0. We have hence shown that a general transcription function of the form introduced in Chapter 1, can summarise the two regulatory logics considered by the model of Ananthasubramaniam et al., which also accounts for network connectivity between neurons by means of VIP signalling. We propose in the following section a model for $Per1$ which incorporates most of the reviewed literature, and takes advantage of measured Calcium levels to account for extra-cellular signalling.

5.3 Proposed model: derivation

We start from the transcription function introduced in Chapter 1, and take advantage of the current biological knowledge of the process. The main modelling assumptions are the following:

1. We assume as in Korenčič et al. (2012) and Ananthasubramaniam et al. (2014), that $Per1$ and $Cry1$ are repressing their own transcription after a random delay τ_p. The delay accounts for nuclear export, protein synthesis and nuclear import. This is clearly a simplified view of the system. On the other hand, it can still provide a good indication about the spatial variation of the kinetic parameters involved, and therefore of the underlying mechanistic dynamics.

2. Auto-repression is implemented by setting the transcription rate for the promoter bound only by the delayed $Per1$ or $Cry1$ mRNA equal to zero. This is consistent with all the literature here considered.

3. Activation by CLOCK/BMAL is assumed to be constant, and incorporated in the basal transcriptional rate R_0. This is again a simplification, as CLOCK/ BMAL is likely to have circadian dynamics as well. On the other hand, there is evidence that it is bound to the promoter throughout the whole circadian cycle (Lee et al., 2001); circadian dynamics should be therefore mostly induced

by rhythmic variation of PER/CRY. This assumption is consistent with the work of Korenčič et al. (2012) and Ananthasubramaniam et al. (2014).

4. We assume Calcium to be an activator ($R_{Ca^{++}} > R_0$), acting with a random delay τ_c.
5. Activation by delayed Calcium can be either obtained independently of delayed Per drop ('OR' gate), or we can observe an interaction effect ('AND' gate).

To motivate assumption 1, we refer to what is known in the literature as the 'linear chain trick' (Smith, 2011).

Assume that the full system, accounting for transcription, nuclear export, translation, complex formation and nuclear import, can be described by the following set of ODEs, also known as the 'Goodwin oscillator' (Goodwin, 1965),

$$\begin{aligned}\frac{dM_g(t)}{dt} &= \nu(P_p(t)) - \mu_{M_g} M_g(t) \\ \frac{dP_1(t)}{dt} &= a[M_g(t) - P_1(t)] \\ &\vdots \\ \frac{dP_p(t)}{dt} &= a[P_{p-1}(t) - P_p(t)],\end{aligned} \tag{5.4}$$

where $\nu(\cdot)$ is the assumed transcription function, which has in our case a Hill form. It is shown in Smith (2011) that the system in Equation 5.4 with initial condition

$$\begin{aligned}M_g(0) &= \phi(0) \\ P_j(0) &= \int_0^\infty \phi(-s) K_{j,a}(s) ds, \quad j = 1, ..., p\end{aligned}$$

where $\phi : (-\infty, 0] \to \mathbb{R}$ is bounded and continuous, is equivalent to

$$\frac{dM_g(t)}{dt} = \nu\left(\int_0^\infty M_g(t-s) K_{p,a}(s) ds\right) - \mu_{M_g} M_g(t), \tag{5.5}$$

with initial condition $M_g(\theta) = \phi(\theta)$, for $\theta \leq 0$, and

$$K_{p,a}(s) = \frac{a^p s^{(p-1)} e^{-as}}{(p-1)!}, \tag{5.6}$$

is the probability density function of a $Ga(p, a)$, evaluated at s. We partially follow El Cheikh et al. (2012) for the proof.

The solution for $P_1(t)$ is given by

$$P_1(t) = ae^{-at}P_1(0) + \int_{-\infty}^{t} ae^{a(t-u)}M_g(u)du.$$

As $t \to \infty$, the contribution of the initial condition tends to 0, and the remaining integral can be seen as a convolution between M_g and a $Ga(1, a)$.
Consider then an arbitrary P_j, for $j \in \{2, ..., p\}$, and suppose that $P_{j-1}(t)$ is the convolution between M_g and a $Ga(j-1, a)$. The solution for $P_j(t)$ is equivalently given by

$$P_j(t) = ae^{-at}P_j(0) + \int_{-\infty}^{t} ae^{a(t-u)}P_{j-1}(u)du,$$

where again, neglecting the initial condition, we obtain the convolution between P_{j-1} and a $Ga(1, a)$. By induction, the additive property of the convolution, and the fact that the convolution of p independent $Ga(1, a)$ is a $Ga(p, a)$, we can then conclude that, as $t \to \infty$,

$$P_p(t) = (M_g * K_{p,a})(t),$$

where K is defined as in Equation 5.6. By plugging the result into the mRNA equation, we obtain indeed Equation 5.5.

Note that the mean of a $Ga(p, a)$ is given by p/a, while its variance is given by p/a^2. This provides a good insight into the properties of the distributed delay: both the mean and variance of the delay increase with an increasing number of intermediate states p. However, as a increases, the dynamics of the intermediate states become faster, and the variance decreases at a faster speed than the mean.

Two main assumptions are required for the previous result: first, all the translation and degradation rates of the intermediate states are assumed to be equal, and second, it is based on deterministic dynamics. If the degradation rates are assumed to be different, we do not recover the closed gamma form for the distribution of the delay. Indeed, we have a convolution of gamma densities having different rate parameters. However, it is proposed in Stewart et al. (2007) that a single gamma density can reasonably approximate the more complex distribution arising from the convolution of multiple gammas with different rates, by matching the exact mean and variance of the sum. The resulting approximation of e.g. the 0.95 percentile, is between 0.94 and 0.96, when the shape parameter is not below 0.1 and the rate parameters do not differ by more than a factor of 10 (Stewart et al., 2007). The result also improves as the number of densities involved increases.

As for the deterministic form, we can postulate that most of the stochasticity is indeed generated by the mRNA state, given that cellular protein levels are generally much higher than mRNA counts (see e.g. Suter, 2011).

The stochastic model for the mRNA is straightforwardly derived by assuming an immigration and death process, whose macroscopic rate equation is given by Equation 5.5. Our reaction network then reduces to the following two reactions,

$$\begin{aligned} R_1 &: \emptyset \overset{\nu(\int_{-\infty}^{t} M_g(s)K(t-s)ds)}{\rightarrow} M_g \\ R_2 &: M_g \overset{\mu_{M_g}}{\rightarrow} \emptyset. \end{aligned} \tag{5.7}$$

As for the transcription function ν, assumptions 2-5 imply the following form for $Per1$

$$\begin{aligned} &\nu\left(Per1(\tau_p), Ca^{++}(\tau_c)\right) = \\ &\frac{R_0 + R_{Ca^{++}}\left(\frac{Ca^{++}(\tau_c)}{K_{cre}}\right)^{n_c}}{1+\left(\frac{Per1(\tau_p)}{K_{pc}}\right)^{n_p}+\left(\frac{Ca^{++}(\tau_c)}{K_{cre}}\right)^{n_c}+\left(\frac{Ca^{++}(\tau_c)}{K_{cre}}\right)^{n_c}\left(\frac{Per1(\tau_p)}{K_{pc}}\right)^{n_p}} \\ &+\frac{+R_{Per1,Ca^{++}}\left(\frac{Ca^{++}(\tau_c)}{K_{cre}}\right)^{n_c}\left(\frac{Per1(\tau_p)}{K_{pc}}\right)^{n_p}}{1+\left(\frac{Per1(\tau_p)}{K_{pc}}\right)^{n_p}+\left(\frac{Ca^{++}(\tau_c)}{K_{cre}}\right)^{n_c}+\left(\frac{Ca^{++}(\tau_c)}{K_{cre}}\right)^{n_c}\left(\frac{Per1(\tau_p)}{K_{pc}}\right)^{n_p}}, \end{aligned} \tag{5.8}$$

where n_c and n_p are the Hill coefficients of Calcium (Ca^{++}) and Per, respectively, K_{pc} and K_{cre} the dissociation coefficients of PER/CRY binding to the E-Box motifs and of Calcium 'binding' to the CRE motifs, respectively, and

$$\begin{aligned} Per1(\tau_p) &= \int_{-\infty}^{t} Per1(s)K_{Per1}(t-s)ds \\ Ca^{++}(\tau_c) &= \int_{-\infty}^{t} Ca^{++}(s)K_{Ca^{++}}(t-s)ds. \end{aligned}$$

The delay distribution for Calcium can be motivated by introducing an additional intermediate state to the system of the form

$$\frac{dCREB(t)}{dt} = a[Ca^{++}(t) - CREB(t)],$$

and by deriving, analogously, its solution

$$CREB(t) = ae^{-at}CREB(0) + \int_{-\infty}^{t} ae^{a(t-u)}Ca^{++}(u)du.$$

We then have a convolution between Calcium and a $Ga(1, a)$.

Note that the proposed $Per1$ transcription function is equivalent to Equation 5.3, when substituting VIP with Calcium, and assuming strong cooperativity in the binding between $Per1$ and 'Calcium' (or, more precisely, the proteins they are a proxy for) molecules (see Chapter 1 for a more extensive explanation). However, although n_p and n_c are analytically shown to represent the number of binding sites under the assumption of strong cooperativity, we have to take into account that not all the binding sites have been observed to have the same responsiveness, and the degree of cooperativity between molecules of the same protein is indeed not known. Moreover, Calcium is only a proxy for the levels of an unobserved transcription factor, and PER is believed to bind to CRY protein to form the repressor complex. Our formulation is therefore just an approximation of a more complex real process, and additional flexibility is retained by allowing n_p and n_c to assume positive real values, representing, more broadly, the responsiveness of the promoter to the proxies Calcium and $Per1$. This justifies simulation parameter values which do not match the values provided in Table 5.1, but seem to approximately reproduce the behaviour of the observed data. Note also that a high 'Hill' coefficient is generally required to reproduce cyclicity in models based on a three-states Goodwin oscillator, although it decreases with an increasing number of intermediate states (Kim et al., 2014).

Although developed in order to model $Per1$ dynamics, the function in Equation 5.8 can be easily adapted to the $Cry1$ and CRE scenario. In particular the transcription functions are

$$\nu(Ca^{++}(\tau_c)) = \frac{R_0 + R_{Ca^{++}}\left(\frac{Ca^{++}(\tau_c)}{K_{cre}}\right)^{n_c}}{1 + \left(\frac{Ca^{++}(\tau_c)}{K_{cre}}\right)^{n_c}},$$

for CRE, where only CRE binding sites are present, and, analogously, for $Cry1$

$$\nu(Cry1(\tau_{cr})) = \frac{R_0}{1 + \left(\frac{Cry1(\tau_{cr})}{K_{pc}}\right)^{n_{cr}}},$$

where no CREs, but a number of E-Boxes have been identified, as outlined in Section 5.1. In analogy with delayed $Per1$ and Calcium, we also define

$$Cry1(\tau_{cr}) \quad = \quad \int_{-\infty}^{t} Cry1(s)K_{Cry1}(t-s)ds.$$

In contrast to the modelling approach provided in Chapter 1, we do not consider here the possible effects of cooperativity between Calcium and $Per1$ in the binding; the reason for this is that the proposed model contains already several approximations of the real process, most notably that Calcium and *Per*1 mRNA are actually not directly binding the promoter. The available information, and the subsequent modelling, seems therefore not the most appropriate to investigate this aspect.

Finally, we can now analyse further the aggregation assumption mentioned in Section 5.2.1. We note in Chapter 1 that it is possible to aggregate different containers - cells in our case - containing reactions up to the second order, only if at least one of the reactants participating in a second order reaction can be assumed to be at approximately the same level in all containers. In the current scenario, the second order reactions would be the binding of 'Calcium' and PER/CRY proteins to the promoter (see Chapter 1 for analytical derivation of the transcription function from a full set of reactions describing binding and unbinding of the TFs to the promoter). We can assume Calcium to be at a similar level in cells close to each other: it is also present in the extra-cellular environment, hence cells are most likely not to be closed containers in its respect. As for PER/CRY proteins, this is not necessarily true. However, we focus our analysis on 2×2 pixel boxes, and a cell is believed to cover approximately squares of size 8×8 pixels. It is in theory possible for a square to be at the intersection of two or more cells. However, given the relatively large number of mRNA and protein molecules per cell involved, and the overall weak effect of aggregation observed in Chapter 1 (recall that we notice an effect of aggregation for an average molecule count of 15-20 proteins per cell, in one simulation scenario, and it does not seem to significantly affect inference), we believe that this is not likely to have a major impact on inference.

5.4 Simulation and mesoscopic approximation for systems with distributed delays

The main theoretical and technical complication of the new model in comparison to the model presented in Chapter 1 results from the introduction of the delay. This additional complexity acts at all levels, from the stochastic simulation, to the diffusion approximation and its subsequent linearisation, as well as the filtering process. We briefly introduce the methods available from the literature for this scenario, as well as those that we specifically develop. Here we focus on *Cry*1, as a first step in the direction of a full analysis which comprises also CRE, and, finally, *Per*1. The methods proposed can however be extended in a straightforward way to any reaction network comprising distributed delays.

5.4.1 Stochastic simulation

We simulate the simple auto-repressive feedback loop generated by $Cry1$. Recall from the set of reactions in Equation 5.7, that our model comprises two reactions: transcription and degradation of $Cry1$ mRNA. The hazard for the transcriptional reaction in Equation 5.7 is computed by evaluating the integral accounting for the delay up to a maximum delay time, and then, each selected reaction is assumed to take place immediately. $Cry1$ promoter is assumed to be at equilibrium, and the delay introduced accounts for $Cry1$ nuclear export, translation into protein, and nuclear import. We additionally assume that the reporter protein is a reasonable approximation of the underlying $Cry1$ mRNA.

The stochastic simulation algorithm (SSA) introduced in Chapter 1 can be used to simulate the approximate dynamics. We assume maximum delay time, $\tau_m = 30\,\mathrm{h}$, and an arbitrary initial condition for the time span of the maximum delay. In particular, the initial condition is given by the first 30 hours of data observations, properly rescaled, from one location of Figure 5.2. Values of simulated $Cry1$ are then stored at fixed time-intervals of duration 0.01 h, and, to mimic the real data, are summed over 0.5 hour, divided by their mean level, and corrupted with measurement error. Figure 5.5 shows the simulated time-series for 10 independent replications of the simulation algorithm, for two levels of signal to noise ratio, i.e. 20 and 100, approximately reproducing the levels observed in real data. Parameters are set in order to reproduce observed dynamics and are within the range of those estimated in the inferential process described in Chapter 6.

One clarification is now required with respect to the simulation methodology. There are different extensions of the SSA/Gillespie algorithm for scenarios which assume random delays *of the reactions* (see e.g. Galla, 2009). In some situations, it is in fact sensible to assume that the actual product of a reaction is not available for a time-interval of non-negligible length, which can be modelled with a delay (Anderson and Kurtz, 2011, for example, motivate the introduction of a fixed delay). However, in our case, the mRNA is assumed to be immediately available after transcription and the gamma delay distribution arises from the integration of the intermediate translation and translocation processes that the produced mRNA undergoes, as outlined in the previous section. Hence we assume, in some sense, a ‘delayed mRNA’ effect rather than a ‘delayed transcription’.

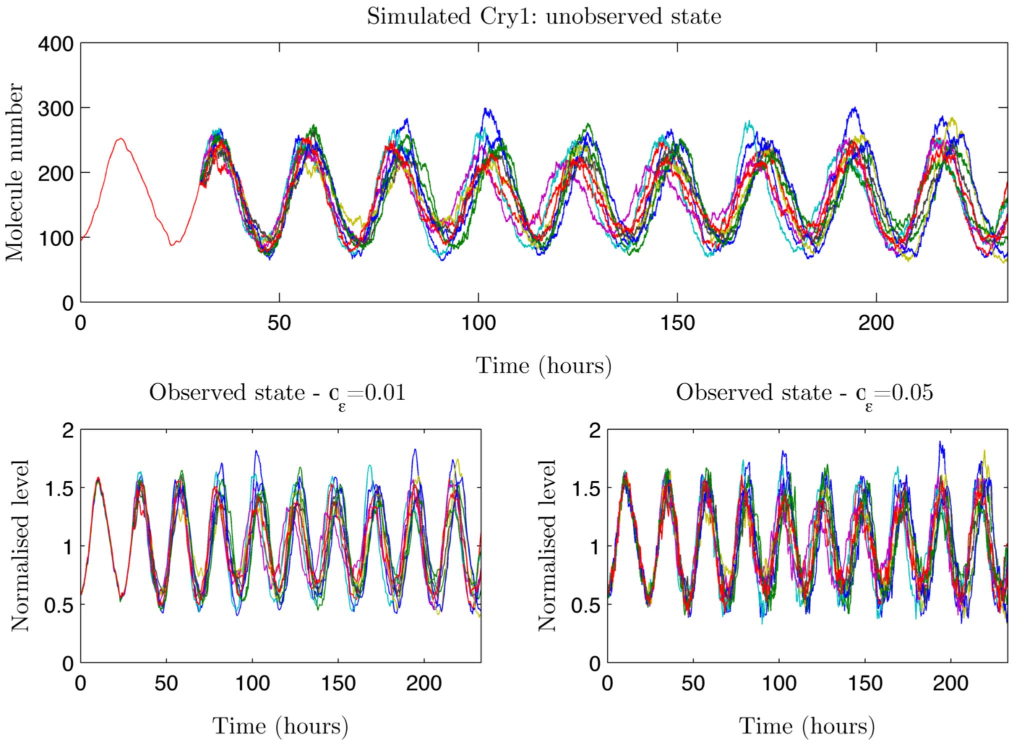


Figure 5.5: SSA simulations for the Set of reactions 5.7. Simulations of unobserved $Cry1$ mRNA (top), and observed $Cry1$ mRNA, rescaled by its mean level, integrated over 0.5 hour and corrupted with measurement error (bottom). Two levels of signal to noise ratio, i.e. 100 (bottom left) and 20 (bottom right). Each plot contains 10 independent replications, with parameters $R_0 = 90\,\text{molecules/h}$, $K_{pc} = 1.5 \times 10^2\,\text{molecules}$, $\mu_{M_g} = 0.25\,\text{h}^{-1}$, $E[\tau_{cr}] = 9\,\text{h}$, $SD[\tau_{cr}] = \sqrt{15}\,\text{h}$. The initial condition is given by the first 30 hours of the $Cry1$-luc time series, rescaled, from one location of Figure 5.2.

5.4.2 Diffusion approximation for systems with distributed delays

Here we introduce the diffusion approximation for our system, and more generally for models comprising delayed species. We also report the result of Brett and Galla (2013), for models comprising delayed reactions. Depending on case-specific modelling assumptions, one or the other approximation arises as a mesoscopic scale model for the underlying stochastic dynamics of the child mRNA.

Diffusion approximation - delayed species

Following the notation of Chapter 1, define a reaction network with p species and r reactions, with $p \times r$ stoichiometry matrix S, and vector of hazards $h(X) = [h_1(X), \ldots, h_r(X)]^T$, where we drop the dependence on Ω and c for ease of notation.

Divide the reactions in two groups: the set of z reactions not involving delayed species, with stoichiometry S_{nd} and hazard vector h_{nd}, and the set of w reactions comprising the random delays, with stoichiometry S_d and hazard vector h_d. The matrices S_{nd} and S_d are simply sub-matrices of S, i.e. we have $S = [S_d, S_{nd}]$. With respect to the hazards, $h_{nd}(X(t)) = [h_{w+1}(X(t)), \dots, h_{w+z}(X(t))]$, while

$$h_d\left(\int_{-\infty}^{t} X(s) \cdot K(t-s)\, ds\right) = \begin{bmatrix} h_1\left(\int_{-\infty}^{t} X(s) \cdot K(t-s) ds\right) \\ \vdots \\ h_w\left(\int_{-\infty}^{t} X(s) \cdot K(t-s)\right) \end{bmatrix},$$

The diffusion approximation for reaction networks comprising distributed delays which arise by elimination of intermediate species, is indeed straightforward. The delays are in fact not introducing additional stochasticity in the hazards, and the hazards themselves are 'deterministically' defined, given the knowledge of the state of the system up until the time of maximum delay. A more formal explanation of this statement is provided, along with a proof of the linear noise approximation for this scenario, in Section 5.5.2. The diffusion approximation arising from the reduced reaction network is, therefore

$$\begin{aligned} dX(t) &= \left[S_{nd} h_{nd}(X(t)) + S_d h_d\left(\int_{-\infty}^{t} X(s) \cdot K(t-s)\right)\right] dt \qquad (5.9) \\ &\quad + \sqrt{S_{nd}\,\mathrm{diag}[h_{nd}(X(t))] S_{nd}^T + S_d\,\mathrm{diag}\left[h_d\left(\int_{-\infty}^{t} X(s) \cdot K(t-s) ds\right)\right] S_d^T}\, dB(t), \end{aligned}$$

where $dB(t)$ is a p-dimensional Wiener process, and $K(\cdot)$ is the vector of the density functions associated with the delay of each species, and we have $K(\cdot) = 0$ if $s \geq t$.

Equation 5.9 provides a continuous approximation for the dynamics of the unobserved molecule counts of $Cry1$. However, we have to take into account that

- Observations are light intensities assumed to be proportional to the number of molecules: this is modelled by introducing a scaling factor κ.
- Signals are integrated over 0.5 hour.

Note that we do not take into account the fact that we observe LUC reporter proteins. This is possibly the major simplification of our model. We assume that dynamics of Luc mRNA are similar to those of $Cry1$, which, given the fact that they share the promoter region, means assuming a similar degradation. Moreover, the measurement equation is also assuming the observed LUC protein levels, possibly rescaled, to be a reasonable proxy for its mRNA levels. In practice, we expect the factor κ to lie between $Cry1$ mRNA and LUC protein levels, partially compensating for model mismatch.

We then write as in Chapter 1, a state-space form of the model at this stage of approximation, where we slightly rearrange Equation 5.9 to introduce a general methodology for the extended Kalman-Bucy filter with distributed delays,

$$\begin{aligned} Y_t &= \kappa \int_{t-\Delta_t}^{t} X(s)\ ds + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2) \qquad (5.10) \\ dX(t) &= \left(g(X(t)) + f\left(\int_{t-\tau_m}^{t} X(s) \cdot K(t-s)\ ds \right) \right) dt \\ &\quad + \left[\sqrt{l(X(t)) + q\left(\int_{t-\tau_m}^{t} X(s) \cdot K(t-s)\ ds \right)} \right] dB(t), \end{aligned}$$

where Δ_t represents the time-interval of signal integration and

$$\begin{aligned} g(X(t)) &= S_{nd} h_{nd}(X(t)) \\ l(X(t)) &= S_{nd}\,\mathrm{diag}[h_{nd}(X(t))] S_{nd}^T \\ f\left(\int_{t-\tau_m}^{t} X(s) \cdot K(t-s)\ ds \right) &= S_d h_d \left(\int_{t-\tau_m}^{t} X(s) \cdot K(t-s)\ ds \right) \\ q\left(\int_{t-\tau_m}^{t} X(s) \cdot K(t-s)\ ds \right) &= S_d\,\mathrm{diag}\left[h_d \left(\int_{t-\tau_m}^{t} X(s) \cdot K(t-s)\ ds \right) \right] S_d^T \\ h_d\left(\int_{t-\tau_m}^{t} X(s) \cdot K(t-s)\ ds \right) &= \nu \left(\int_{t-\tau_m}^{t} X(s) \cdot K(t-s)\ ds \right). \end{aligned}$$

Note also that we have introduced the maximum delay time τ_m in the integral of the delayed reactions, to account for truncation.

Diffusion approximation - delayed reactions

Brett and Galla (2013) provide a diffusion approximation for systems incorporating delayed reactions. Here we report only the main result, in the sub case where a reaction can only have a delayed or an immediate effect, along with an intuitive explanation, and refer to the paper for extensions and technical details.

Assume that a reaction can have either a delayed *or* an immediate effect, and so divide once again the reactions in two groups: the set of z non-delayed reactions, with stoichiometry S_{nd} and hazard vector $h_{nd}(X(t))$, and the set of w delayed reactions, with stoichiometry S_d and hazard vector $h_d(X(s))$. The matrices

S_{nd} and S_d, and the vector $h_{nd}(X(t))$, are defined as above, while

$$h_d(X(s)) \;=\; \Big[h_1(X(s))K_1(t-s) \quad \dots \quad h_w(X(s))K_w(t-s)\Big]^T ,$$

where $K_1(\cdot), ..., K_w(\cdot)$ are the density functions associated with each delay, and we have $K(\cdot) = 0$ if $s \geq t$. In this sub-case, the diffusion approximation, according to Brett and Galla (2013), is

$$\begin{aligned} dX(t) \;&=\; \left[S_{nd}h_{nd}(X(t)) + \int_{-\infty}^{t} S_d h_d(X(s))ds\right] dt \qquad (5.11) \\ &+ \sqrt{S_{nd}\,\mathrm{diag}[h_{nd}(X(t))]S_{nd}^T + \int_{-\infty}^{t} S_d\,\mathrm{diag}[h_d(X(s))]S_d^T ds}\; dB(t), \end{aligned}$$

where $dB(t)$ is a p-dimensional Wiener process.

The proof of Brett and Galla (2013) follows from time-discretisation of the underlying Markov process, and the definition of the generating function associated with the distribution of the number of reactions firing at time t, assumed to be Poisson. By drawing the continuous limit, it is shown to provide normal dynamics for the number of molecules $X(t)$. Intuitively, this formulation poses the delay as a characteristic of the reactions, rather than of the species involved. Therefore, the contribution of the k-th delayed reaction to the continuous approximation of the number of molecules X at time t, can be seen as a weighted average of all the past possible paths, i.e. the corresponding hazards evaluated at all the past times, and weighted according to the assumed distribution of the delay.

This formulation leads also to a straightforward linearisation of the system, as presented, once again, in Brett and Galla (2013). We return to this point in Section 5.5.

5.5 The extended Kalman-Bucy filter for systems with distributed delays

In this section we develop the extended Kalman-Bucy filter for systems with distributed delays for both the diffusion of Equation 5.11 and the diffusion of Equation 5.9. Linearisation of Equation 5.11 is provided by Brett and Galla (2013), while we propose a different linearisation approach for Equation 5.9. Recall from Chapter 2, that linearisation is particularly useful in an inferential framework, as it allows to obtain normal transition densities for dynamics of the unobserved states, and therefore a closed form for the likelihood. The likelihood can be then employed to perform inference about the unknown kinetic parameters, in both a Bayesian and a frequentist framework.

One drawback of linearisation is that it results, for our system, in mean and variance mismatch with respect to the CLE. For predictions far away in time, this mismatch can have in both cases an impact on the quality of predictions, and therefore, indirectly, on the likelihood and parameter estimation. To improve the quality of predictions, the system linearisation can be 'restarted' at every iteration, setting the initial point of the deterministic process and the noise at their optimal values - here, we mean optimal in the filtering sense, i.e. they are conditional on the available past observations. The non-restarted linearisation, also known as non-restarted LNA (see Chapters 1 and 2 for further details), for the CLE in Equation 5.11 is provided by Brett and Galla (2013). A restarted LNA for non-delayed systems is provided in Fearnhead et al. (2014), and leads to predictive densities analogous to those of the extended Kalman-Bucy filter.

Delays have been successfully incorporated in the extended Kalman filter in the literature, although not in a distributed form or for a large number of iterations in the past, according to Gopalakrishnan et al. (2011), which provides a review of the topic.

Here we provide a restarted version of the LNA of Brett and Galla (2013), which we denote extended Kalman-Bucy filter or restarted LNA for systems comprising delayed reactions. For the diffusion of Equation 5.9, we also provide a novel linearisation approach, leading to the extended Kalman-Bucy filter or restarted LNA for systems comprising delayed species. In addition, our formulation includes an approximate, but explicit, modelling of signal integration in the measurement process.

5.5.1 EKBF - delayed species

Following the same steps of Chapter 2 for the EKBF, we start from a time-discretised state-space model based on Model 5.10 of the type

$$\begin{aligned} Y_t &= F^T X_t + \epsilon_t \qquad (5.12)\\ X_t &= X_{t-\delta_t} + \delta_t g(X_{t-\delta_t}) + \delta_t f\left(\sum_{s=t-\delta_t-\tau_m}^{t-\delta_t} X_s \cdot K_{t-s}\right) \\ &\quad + \sqrt{\delta_t} d\left(X_{t-\delta_t}, \sum_{s=t-\delta_t-\tau_m}^{t-\delta_t} X_s \cdot K_{t-s})\right) Z_t \\ Z_t &\sim MVN(0, I), \end{aligned}$$

where $g(\cdot)$, $f(\cdot)$ and $d(\cdot)$ are general nonlinear functions, F is a $p \times q$ matrix (where q is the dimension of Y and p the dimension of X), $K_{t-s} = [K_{1,t-s}, ..., K_{p,t-s}]^T$ is a vector of weights, one for each X, evaluated at time s, and τ_m the maximum delay time. Note that we start from a model not comprising signal integration, for ease of notation and clarity. However, we address this generalisation of the model at the end of this section.

Suppose that an optimal estimate for the initial condition is available, i.e. $\pi(x_{0:\tau_m}|y_{0:\tau_m})$ distributed as a $\mathrm{MVN}(\rho_{0:\tau_m}, P_{x_{0:\tau_m}})$. In practice, this initial condition is generally not available and requires further *ad hoc* modelling. We present the model specification for the initial condition for our system in Section 5.6.

Now suppose that we wish to obtain an estimate of $\rho_{\tau_m+\delta_t} = \mathrm{E}[X_{\tau_m+\delta_t}|y_{0:\tau_m}]$, $P_{\tau_m+\delta_t} = \mathrm{Var}[X_{\tau_m+\delta_t}|y_{0:\tau_m}]$ and $P_{\tau_m,\tau_m+\delta_t} = \mathrm{Cov}[X_{\tau_m}, X_{\tau_m+\delta_t}|y_{0:\tau_m}]$.
Write

$$\begin{aligned} \mathrm{E}\left[X_{\tau_m+\delta_t}|y_{0:\tau_m}\right] &= \mathrm{E}\left[X_{\tau_m}|y_{0:\tau_m}\right] + \delta_t\, \mathrm{E}\left[g(X_{\tau_m})|y_{0:\tau_m}\right] \\ &\quad +\delta_t\, \mathrm{E}\left[f\left(\sum_s X_s \cdot K_{t-s}\right)|y_{0:\tau_m}\right], \end{aligned} \tag{5.13}$$

and Taylor-expand the nonlinear function $g(\cdot)$ about ρ_{τ_m}, and $f(\cdot)$ about $\sum_s \rho_s \cdot K_{t-s}$, $s \in [0, \tau_m]$, up to the first order

$$\begin{aligned} g(X_{\tau_m}) &\approx g(\rho_{\tau_m}) + J_g(\rho_{\tau_m})(X_{\tau_m} - \rho_{\tau_m}) \\ f\left(\sum_s X_s \cdot K_{t-s}\right) &\approx f\left(\sum_s \rho_s \cdot K_{t-s}\right) \\ &\quad +J_f\left(\sum_s \rho_s \cdot K_{t-s}\right) \\ &\quad \left(\sum_s X_s \cdot K_{t-s} - \sum_s \rho_s \cdot K_{t-s}\right) \\ &= f\left(\sum_s \rho_s \cdot K_{t-s}\right) \\ &\quad +J_f\left(\sum_s \rho_s \cdot K_{t-s}\right)\sum_s (X_s - \rho_s) \cdot K_{t-s}. \end{aligned}$$

To obtain $\rho_{\tau_m+\delta_t}$, plug the results into the expectation of Equation 5.13, i.e.

$$\begin{aligned} \mathrm{E}[X_{\tau_m+\delta_t}|y_{0:\tau_m}] &\approx \mathrm{E}[X_{\tau_m}|y_{0:\tau_m}] + \delta_t g(\rho_{\tau_m}) \\ &\quad +\delta_t J_g(\rho_{\tau_m})\, \mathrm{E}[X_{\tau_m} - \rho_{\tau_m}|y_{0:\tau_m}] \end{aligned}$$

$$
\begin{aligned}
& +\delta_t f\left(\sum_s \rho_s \cdot K_{t-s}\right)\\
& +\delta_t J_f\left(\sum_s \rho_s \cdot K_{t-s}\right)\sum_s \mathrm{E}\left[(X_s-\rho_s)\cdot K_{t-s}|y_{0:\tau_m}\right]\\
= \;& \rho_{\tau_m}+\delta_t g(\rho_{\tau_m})+\delta_t f\left(\sum_s \rho_s \cdot K_{t-s}\right).
\end{aligned}
\tag{5.14}
$$

The variance $P_{\tau_m+\delta_t}$ follows from

$$
\begin{aligned}
\mathrm{Var}[X_{\tau_m+\delta_t}|y_{0:\tau_m}] = \;& \mathrm{Var}[X_{\tau_m}|y_{0:\tau_m}]+\delta_t^2\,\mathrm{Var}[g(X_{\tau_m})|y_{0:\tau_m}]\\
& +\delta_t^2\,\mathrm{Var}\left[f\left(\sum_s X_s\cdot K_{t-s}\right)|y_{0:\tau_m}\right]\\
& +\delta_t\,\mathrm{Cov}\left[x_{\tau_m},g(X_{\tau_m})|y_{0:\tau_m}\right]+\delta_t\,\mathrm{Cov}\left[g(X_{\tau_m}),X_{\tau_m}|y_{0:\tau_m}\right]\\
& +\delta_t\,\mathrm{Cov}\left[X_{\tau_m},f\left(\sum_s X_s\cdot K_{t-s}\right)|y_{0:\tau_m}\right]\\
& +\delta_t\,\mathrm{Cov}\left[f\left(\sum_s X_s\cdot K_{t-s}\right),X_{\tau_m}|y_{0:\tau_m}\right]\\
& +\delta_t^2\,\mathrm{Cov}\left[g(X_{\tau_m}),f\left(\sum_s X_s\cdot K_{t-s}\right)|y_{0:\tau_m}\right]\\
& +\delta_t^2\,\mathrm{Cov}\left[f\left(\sum_s X_s\cdot K_{t-s}\right),g(X_{\tau_m})|y_{0:\tau_m}\right]\\
& +\delta_t\,\mathrm{E}\left[d\left(X_{\tau_m},\sum_s X_s\cdot K_{t-s}\right)d\left(X_{\tau_m},\sum_s X_s\cdot K_{t-s}\right)^T|y_{0:\tau_m}\right].
\end{aligned}
$$

Dropping the terms of order δ_t^2, and plugging in the Taylor expansion we obtain

$$
\begin{aligned}
\mathrm{Var}[X_{\tau_m+\delta_t}|y_{0:\tau_m}] \approx \;& \mathrm{Var}[X_{\tau_m}|y_{0:\tau_m}]\\
& +\delta_t\,\mathrm{Cov}\left[J_g(\rho_{\tau_m})(X_{\tau_m}-\rho_{\tau_m}),X_{\tau_m}|y_{0:\tau_m}\right]\\
& +\delta_t\,\mathrm{Cov}\left[X_{\tau_m},J_g(\rho_{\tau_m})(X_{\tau_m}-\rho_{\tau_m})|y_{0:\tau_m}\right]\\
& +\delta_t\,\mathrm{Cov}\left[X_{\tau_m},J_f\left(\sum_s \rho_s\cdot K_{t-s}\right)\sum_s(X_s-\rho_s)\cdot K_{t-s}|y_{0:\tau_m}\right]\\
& +\delta_t\,\mathrm{Cov}\left[J_f\left(\sum_s \rho_s\cdot K_{t-s}\right)\sum_s(X_s-\rho_s)\cdot K_{t-s},X_{\tau_m}|y_{0:\tau_m}\right]\\
& +\delta_t\,\mathrm{E}\left[D\left(X_{\tau_m},\sum_s X_s\cdot K_{t-s}\right)|y_{0:\tau_m}\right],
\end{aligned}
$$

where $D\left(X_{\tau_m}, \sum_s X_s \cdot K_{t-s}\right) = d\left(X_{\tau_m}, \sum_s X_s \cdot K_{t-s}\right) d\left(X_{\tau_m}, \sum_s X_s \cdot K_{t-s}\right)^T$, in analogy with the notation introduced in Chapter 2. In particular, assume $D\left(X_{\tau_m}, \sum_s X_s \cdot K_{t-s}\right)$ to be of the type

$$D\left(X_{\tau_m}, \sum_s X_s \cdot K_{t-s}\right) = l(X_{\tau_m}) + q\left(\sum_s X_s \cdot K_{t-s}\right),$$

where $l(\cdot)$ and $q(\cdot)$ are again general nonlinear functions, accounting for the effect of non-delayed and delayed species, respectively. Taylor expand $l(\cdot)$ and $q(\cdot)$ about ρ_{τ_m} and $\sum_s \rho_s \cdot K_{t-s}$, respectively, to obtain

$$\mathrm{E}\left[D\left(X_{\tau_m}, \sum_s X_s \cdot K_{t-s}\right) | y_{0:\tau_m}\right] = D\left(\rho_{\tau_m}, \sum_s \rho_s \cdot K_{t-s}\right).$$

Hence we can approximate the variance by

$$\begin{aligned}
\mathrm{Var}[X_{\tau_m+\delta_t}|y_{0:\tau_m}] \approx\; & P_{\tau_m} + \delta_t\left[J_g(\rho_{\tau_m})P_{\tau_m} + P_{\tau_m}^T J_g(\rho_{\tau_m})^T\right] \\
& +\delta_t\left[J_f\left(\sum_s \rho_s \cdot K_{t-s}\right)\left(\sum_s P_{s,\tau_m} \cdot K_{t-s}\right)\right] \\
& +\delta_t\left[\left(\sum_s P_{\tau_m,s} \cdot K_{t-s}\right) J_f\left(\sum_s \rho_s \cdot K_{t-s}\right)^T\right] \\
& +\delta_t D\left(\rho_{\tau_m}, \sum_s \rho_s \cdot K_{t-s}\right). \qquad (5.15)
\end{aligned}$$

Finally, the covariance is given by

$$\begin{aligned}
\mathrm{Cov}[X_{\tau_m+\delta_t}, X_{\tau_m}|y_{0:\tau_m}] =\; & \mathrm{Cov}\left[X_{\tau_m} + \delta_t g(X_{\tau_m}) + \delta_t f\left(\sum_s X_s \cdot K_{t-s}\right)\right. \\
& \left. +\sqrt{\delta_t} d(X_{\tau_m}, X_s) Z_t, X_{\tau_m}|y_{0:\tau}\right] \\
=\; & \mathrm{Cov}[X_{\tau_m}, X_{\tau_m}|y_{0:\tau_m}] + \delta_t\,\mathrm{Cov}[g(X_{\tau_m}), X_{\tau_m}|y_{0:\tau}] \\
& +\delta_t\,\mathrm{Cov}\left[f\left(\sum_s X_s \cdot K_{t-s}\right), X_{\tau_m}|y_{0:\tau_m}\right] \\
=\; & P_{\tau_m} + \delta_t J_g(\rho_{\tau_m})P_{\tau_m} \\
& +\delta_t J_f\left(\sum_s \rho_s \cdot K_{t-s}\right)\sum_s P_{s,\tau_m} \cdot K_{t-s}. \qquad (5.16)
\end{aligned}$$

Moreover, with analogous steps, for $i \in [1, \tau_m/\delta_t]$,

$$\begin{aligned}\mathrm{Cov}[X_{\tau_m+\delta_t}, X_{\tau_m-i\delta_t}|y_{0:\tau_m}] &= P_{\tau_m,\tau_m-i\delta_t} + \delta_t J_g(\rho_{\tau_m}) P_{\tau_m,\tau_m-i\delta_t} \\ &\quad +\delta_t J_f\left(\sum_s \rho_s \cdot K_{t-s}\right) \cdot \\ &\quad \sum_s P_{s,\tau_m-i\delta_t} \cdot K_{t-s}. \qquad (5.17)\end{aligned}$$

Since the system has been linearised, all the distributions involved are normal, and therefore in particular we have that $\pi(x_{\tau_m:\tau_m+\delta_t}|y_{0:\tau_m})$ is $\mathcal{N}(\rho_{\tau_m:\tau_m+\delta_t}, P_{\tau_m:\tau_m+\delta_t})$, where the mean and covariance matrix are given by Equations 5.14 to 5.17.
More generally, suppose we want to obtain the likelihood

$$L(y|\Psi) = \pi(y_{0:T}|\Psi),$$

where we start by considering Ψ as a given set of parameters. The likelihood can be factorised as

$$\begin{aligned}\pi(y_{0:T}|\Psi) &= \pi(y_{\tau_m+mM\Delta_t+\Delta_t:T}|y_{0:\tau_m+mM\Delta_t}, \Psi) \\ &\quad \cdots \pi(y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}|y_{0:\tau_m}, \Psi)\pi(y_{0:\tau_m}|\Psi),\end{aligned}$$

where m is the number of observations that we wish to predict before performing an update of the unobserved states mean and variance and M is the total number of blocks of observations, minus the block for the initial condition. Note that we use Δ_t, and not δ_t, for the time step of the observations y. The two quantities need not be the same, and generally they are not: the time step of the available data is indeed often too large to obtain a good approximation of the underlying unobserved continuous process, i.e. $\Delta_t >> \delta_t$

It is straightforward to obtain the density $\pi(y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}|y_{0:\tau_m}, \Psi)$, once $\pi(x_{\tau_m+\Delta_t:\tau_m+m\Delta_t}|y_{0:\tau_m}, \Psi)$ is available. By linearisation, the latter follows a multivariate normal distribution, whose mean and variance can be obtained by iterating the steps in Equations 5.14 to 5.17, up to time $\tau_m + m\Delta_t$. In particular, let $\rho_t = \mathrm{E}[X_t|y_{0:\tau_m}]$, $P_t = \mathrm{Var}[X_t|y_{0:\tau_m}]$ and $P_{t,s} = \mathrm{Cov}[X_t, X_s|y_{0:\tau_m}]$. We then have

$$\begin{aligned}\mathrm{E}[X_{t+\delta_t}|y_{0:\tau_m}] &\approx \rho_t + \delta_t g(\rho_t) + \delta_t f\left(\sum_s \rho_s \cdot K_{t-s}\right) \qquad (5.18) \\ \mathrm{Var}[X_{t+\delta_t}|y_{0:\tau_m}] &\approx P_t + \delta_t\left[J_g(\rho_t)P_t + P_t^T J_g(\rho_t)^T\right]\end{aligned}$$

$$
\begin{aligned}
& +\delta_t \left[J_f \left(\sum_s \rho_s \cdot K_{t-s} \right) \left(\sum_s P_{s,t} \cdot K_{t-s} \right) \right] \\
& +\delta_t \left[\left(\sum_s P_{t,s} \cdot K_{t-s} \right) J_f \left(\sum_s \rho_s \cdot K_{t-s} \right)^T \right] \\
& +\delta_t D \left(\rho_t, \sum_s \rho_s \cdot K_{t-s} \right) \qquad (5.19)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cov}[X_{t+\delta_t}, X_t | y_{0:\tau_m}] \approx\ & P_t + \delta_t J_g(\rho_t) P_t \\
& +\delta_t J_f \left(\sum_s \rho_s \cdot K_{t-s} \right) \sum_s P_{s,t} \cdot K_{t-s}. \qquad (5.20)
\end{aligned}
$$

For an arbitrary lag, given $i > j$, the covariance is

$$
\begin{aligned}
\mathrm{Cov}[X_i, X_j | y_{0:\tau_m}] =\ & P_{i-\delta_t,j} + \delta_t J_g(\rho_{i-\delta_t}) P_{i-\delta_t,j} \\
& +\delta_t J_f \left(\sum_s \rho_s \cdot K_{t-s} \right) \left(\sum_s P_{s,j} \cdot K_{t-s} \right). \qquad (5.21)
\end{aligned}
$$

We can also draw the continuous limit for the dynamics of the unobserved states. We have, as in Chapter 2,

$$
\begin{aligned}
\frac{\rho_{t+\delta_t} - \rho_t}{\delta_t} \approx\ & g(\rho_t) + f \left(\sum_s \rho_s \cdot K_{t-s} \right) \\
\frac{P_{t+\delta_t} - P_t}{\delta_t} \approx\ & \left[J_g(\rho_t) P_t + P_t^T J_g(\rho_t)^T \right] \\
& + \left[J_f \left(\sum_s \rho_s \cdot K_{t-s} \right) \left(\sum_s P_{s,t} \cdot K_{t-s} \right) \right] \\
& + \left[\left(\sum_s P_{t,s} \cdot K_{t-s} \right) J_f \left(\sum_s \rho_s \cdot K_{t-s} \right)^T \right] \\
& + D \left(\rho_t, \sum_s \rho_s \cdot K_{t-s} \right).
\end{aligned}
$$

As $\delta_t \to 0$,

$$
\begin{aligned}
\frac{d\rho(t)}{dt} \approx\ & g(\rho(t)) + f \left(\int_{t-\tau_m}^{t} \rho(s) \cdot K(t-s) ds \right) \qquad (5.22) \\
\frac{dP(t)}{dt} \approx\ & \left[J_g(\rho(t)) P(t) + P(t)^T J_g(\rho(t))^T \right]
\end{aligned}
$$

$$\begin{aligned}
&+\left[J_f\left(\int_{t-\tau_m}^{t} \rho(s)\cdot K(t-s)ds\right)\left(\int_{t-\tau_m}^{t} P(s,t)\cdot K(t-s)ds\right)\right]\\
&+\left[\left(\int_{t-\tau_m}^{t} P(t,s)\cdot K(t-s)ds\right)J_f\left(\int_{t-\tau_m}^{t} \rho(s)\cdot K(t-s)ds\right)^T\right]\\
&+D\left(\rho(t),\int_{t-\tau_m}^{t} \rho(s)\cdot K(t-s)ds\right).
\end{aligned}\tag{5.23}$$

We provide an equivalent derivation of the latter two equations in Section 5.5.2.

Returning to our problem, we then use the fact that

$$\begin{aligned}
Y_{\tau_m+\Delta_t:\tau_m+m\Delta_t} &= F^T X_{\tau_m+\Delta_t:\tau_m+m\Delta_t}+\epsilon_{\tau_m+\Delta_t:\tau_m+m\Delta_t},\\
\epsilon_{\tau_m+\Delta_t:\tau_m+m\Delta_t} &\sim \mathcal{N}(0,\sigma_\epsilon^2 I),
\end{aligned}$$

It follows that $\pi(y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}|y_{0:\tau_m})$ follows a multivariate normal distribution with mean $F^T\rho_{\tau_m+\Delta_t:\tau_m+m\Delta_t}$, and variance/covariance matrix $F^T(P_{\tau_m+\Delta_t:\tau_m+m\Delta_t}+\sigma_\epsilon^2 I)F$.

We now wish to update the unobserved states, to obtain an optimal estimate of the mean and variance of $\pi(x_{m\Delta_t:\tau_m+m\Delta_t}|y_{0:\tau_m+m\Delta_t})$. From the previous steps, the joint mean and variance of $\pi(x_{m\Delta_t:\tau_m+m\Delta_t}, y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}|y_{0:\tau_m})$ is

$$\lambda=\begin{pmatrix}\lambda_{X_{m\Delta_t:\tau_m+m\Delta_t}}\\ \lambda_{Y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}}\end{pmatrix}=[\rho_{m\Delta_t:\tau_m+m\Delta_t}, F^T\rho_{\tau_m+\Delta_t:\tau_m+m\Delta_t}]^T,$$

and

$$\begin{aligned}
\Lambda &= \begin{pmatrix}\Lambda_{X_{m\Delta_t:\tau_m+m\Delta_t}} & \Lambda_{X_{m\Delta_t:\tau_m+m\Delta_t},Y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}}\\ \Lambda_{Y_{\tau_m+\Delta_t:\tau_m+m\Delta_t},X_{m\Delta_t:\tau_m+m\Delta_t}} & \Lambda_{Y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}}\end{pmatrix}\\
&= \begin{pmatrix}P_{m\Delta_t:\tau_m+m\Delta_t} & P_{m\Delta_t:\tau_m+m\Delta_t,\tau_m+\Delta_t:\tau_m+m\Delta_t}F\\ F^TP_{\tau_m+\Delta_t:\tau_m+m\Delta_t,m\Delta_t:\tau_m+m\Delta_t} & F^T\left(P_{\tau_m+\Delta_t:\tau_m+m\Delta_t}+\sigma_\epsilon^2 I\right)F\end{pmatrix}.
\end{aligned}$$

By conditioning on $y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}$, we obtain $\pi(x_{m\Delta_t:\tau_m+m\Delta_t}|y_{0:\tau_m+m\Delta_t})$ as a MVN $(\rho^*_{m\Delta_t:\tau_m+m\Delta_t}, P^*_{m\Delta_t:\tau_m+m\Delta_t})$, where

$$\begin{aligned}
\rho^*_{m\Delta_t:\tau_m+m\Delta_t} &= \rho_{m\Delta_t:\tau_m+m\Delta_t}+K(Y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}-\lambda_{Y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}})\\
P^*_{m\Delta_t:\tau_m+m\Delta_t} &= P_{m\Delta_t:\tau_m+m\Delta_t}-K\Lambda_{Y_{\tau_m+\Delta_t:\tau_m+m\Delta_t},X_{\tau_m+\Delta_t:\tau_m+m\Delta_t}}\\
K &= \Lambda_{X_{m\Delta_t:\tau_m+m\Delta_t},Y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}}\Lambda^{-1}_{Y_{\tau_m+\Delta_t:\tau_m+m\Delta_t}}.
\end{aligned}\tag{5.24}$$

This procedure provides optimal estimates of the mean and variance conditional on all the observations from time 0 to $\tau_m+m\Delta_t$. The values of $\rho_{m\Delta_t:\tau_m+m\Delta_t}$

and $P_{m\Delta_t:\tau_m+m\Delta_t}$ are then set equal to their optimal estimates $\rho^*_{m\Delta_t:\tau_m+m\Delta_t}$ and $P^*_{m\Delta_t:\tau_m+m\Delta_t}$, and the same steps are repeated until the last time-point, T.

Note that integration of the observed signal can be straightforwardly performed by substituting F^T with $\widetilde{F}$, a block diagonal matrix, with diagonal elements F^T, and considering in the observation equation $X_{\tau_m+\delta_t:\tau_m+m\Delta_t}$.

It is important to stress the main difference introduced by the presence of the delay in the filtering process. Irrespectively of the choice of m, i.e. irrespectively of the number of observations we wish to predict before updating our estimates of the mean and variance, every time we perform filtering we need to update *all the estimates in the past*, until the time of maximum delay from the current time-point. We can see this process as, indeed, a 'partial smoothing': at the end of the algorithm, the conditioning on the future observations is limited to the maximum delay time τ_m. On the other hand, the advantage is that a partially smoothed estimate of the unobserved states is obtained 'for free', together with the predictions and the likelihood. This can be enough for practical purposes, as it is generally unlikely that observations which are very distant in time, have a significant impact on present and future states.

Clearly, the additional complexity introduced by the delay, comes at higher computational cost. When the dimension of X and, more crucially, the number of unobserved states included in the maximum delay time is high, the algorithm can be significantly slower than in non-delayed cases. The cost comes mainly from the evaluation of the covariances, at every iteration and for all the time-points back until the maximum delay. We provide exact running times for the system under study in Section 5.6.

5.5.2 LNA derivation - delayed species

We now follow the same steps outlined in Appendix A.2 for the derivation of the LNA of non-delayed systems, as provided by Anderson and Kurtz (2011), and derive the LNA for systems comprising delayed species. We hence show that normal transition densities analogous to those obtained for the EKBF for distributed delays of the species, can also directly arise from linearisation of the reduced Markov process which assumes integration of the delayed species in the hazards. In other words, normal and linear transition densities can be derived without first resorting to the diffusion approximation (this is again analogous to the non-delayed case, see Anderson and Kurtz, 2011).

Formally, recall that $Z(t) = X(t)/\Omega$, where $X(t)$ is the vector of molecules counts of the stochastic system, $z(t) = x(t)/\Omega$, where $x(t)$ is the deterministic limit of the system, and Ω is the system size. Let S_{nd}, S_d, h_d and h_{nd} be defined as in Section 5.4.2, recalling that

$$h_d\left(\int_{-\infty}^{t} X(s)\cdot K(t-s)\ ds\right) \quad = \quad \begin{bmatrix} h_1\left(\int_{-\infty}^{t} X(s)\cdot K(t-s)ds\right) \\ \vdots \\ h_w\left(\int_{-\infty}^{t} X(s)\cdot K(t-s)\right) \end{bmatrix},$$

is the vector of hazards for the reactions comprising delayed species. The hazards definition is the core assumption of the following approximation. Following steps analogous to Anderson and Kurtz (2011), first define the quantity

$$P^{\Omega}(t) = \sqrt{\Omega}(Z(t) - z(t)). \tag{5.25}$$

Note that we have in this case

$$\begin{aligned} Z(t) \quad &\approx \quad Z(0) + \frac{1}{\Omega}S_{nd}Y\left(\Omega\int_0^t \tilde{h}_{nd}(Z(s),c)ds\right) \\ &\quad +\frac{1}{\Omega}S_dY\left(\Omega\int_0^t \tilde{h}_d\left(\int_{-\infty}^{s} Z(u)\cdot K(t-u)du, c\right)ds\right), \end{aligned}$$

and recall that $\tilde{h}_{nd}$ and $\tilde{h}_d$ are the hazards arising from the law of mass action. We then have

$$\begin{aligned} P^{\Omega}(t) \quad &\approx \quad P^{\Omega}(0) + \sqrt{\Omega}\left[\frac{1}{\Omega}S_{nd}Y\left(\Omega\int_0^t \tilde{h}_{nd}\left(Z(s),c\right)ds\right)\right. \\ &\quad \left. - \int_0^t S_{nd}\tilde{h}_{nd}\left(z(s),c\right)ds\right] \\ &\quad +\sqrt{\Omega}\left[\frac{1}{\Omega}S_dY\left(\Omega\int_0^t \tilde{h}_d\left(\int_{-\infty}^{s} Z(u)\cdot K(t-u)du, c\right)ds\right)\right. \\ &\quad \left. - \int_0^t S_d\tilde{h}_d\left(\int_{-\infty}^{s} z(u)\cdot K(t-u)du, c\right)ds\right] \\ &= \quad P^{\Omega}(0) + \frac{1}{\sqrt{\Omega}}S_{nd}\tilde{Y}\left(\Omega\int_0^t \tilde{h}_{nd}\left(Z(s),c\right)ds\right) \\ &\quad + \int_0^t \sqrt{\Omega}S_{nd}\tilde{h}_{nd}\left(Z(s),c\right)ds - \int_0^t \sqrt{\Omega}S_{nd}\tilde{h}_{nd}\left(z(s),c\right)ds. \\ &\quad +\frac{1}{\sqrt{\Omega}}S_d\tilde{Y}\left(\Omega\int_0^t \tilde{h}_d\left(\int_{-\infty}^{s} Z(u)\cdot K(t-u)du, c\right)ds\right) \\ &\quad + \int_0^t \sqrt{\Omega}S_d\tilde{h}_d\left(\int_{-\infty}^{s} Z(u)\cdot K(t-u)du, c\right)ds \end{aligned}$$

$$-\int_0^t \sqrt{\Omega} S_d \tilde{h}_d \left(\int_{-\infty}^s z(u) \cdot K(t-u) du, c \right) ds.$$

The normal approximation to the Poisson process (see again Anderson and Kurtz, 2011) gives

$$\begin{aligned} &\frac{1}{\sqrt{\Omega}} S_d \tilde{Y} \left(\Omega \int_0^t \tilde{h}_d \left(\int_{-\infty}^s Z(u) \cdot K(t-u) du, c \right) ds \right) \approx \\ &\approx S_d B \left(\text{diag} \left[\int_0^t \tilde{h}_d \left(\int_{-\infty}^s Z(u) \cdot K(t-u) du, c \right) ds \right] \right), \end{aligned}$$

and

$$\frac{1}{\sqrt{\Omega}} S_{nd} \tilde{Y} \left(\Omega \int_0^t \tilde{h}_{nd} \left(Z(s), c \right) ds \right) \approx S_{nd} B \left(\text{diag}[\int_0^t \tilde{h}_{nd} \left(Z(s), c \right) ds] \right),$$

where B is a Wiener process.

The first order Taylor expansion of $h_{nd}\left(Z(s), c\right)$ is performed about $z(s)$, while $h_d \left(\int_{-\infty}^s Z(u) \cdot K(t-u) du, c \right)$ is expanded about $\int_{-\infty}^s z(u) \cdot K(t-u) du$. We then have

$$\tilde{h}_{nd} \left(Z(s), c \right) \approx \tilde{h}_{nd} \left(z(s), c \right) + J_{\tilde{h}_{nd}} \left(z(s) \right) \left(Z(s) - z(s) \right)$$

and

$$\begin{aligned} \tilde{h}_d \left(\int_{-\infty}^s Z(u) \cdot K(t-u) du, c \right) \approx\ & \tilde{h}_d \left(\int_{-\infty}^s z(u) \cdot K(t-u) du, c \right) \\ & + J_{\tilde{h}_d} \left(\int_{-\infty}^s z(u) \cdot K(t-u) du \right) \\ & \left(\int_{-\infty}^s Z(u) \cdot K(t-u) du \right. \\ & \left. - \int_{-\infty}^s z(u) \cdot K(t-u) du \right). \end{aligned}$$

Combining the two results, we obtain

$$\begin{aligned} P^\Omega(t) \approx\ & P^\Omega(0) + S_{nd} B \left(\text{diag}[\int_0^t \tilde{h}_{nd} \left(Z(s), c \right) ds] \right) \\ & + S_{nd} \int_0^t J_{\tilde{h}_{nd}} \left(z(s) \right) P^\Omega(s) ds \\ & + S_d B \left(\text{diag} \left[\int_0^t \tilde{h}_d \left(\int_{-\infty}^s Z(u) \cdot K(t-u) du, c \right) ds \right] \right) \\ & + \int_0^t S_d J_{\tilde{h}_d} \left(\int_{-\infty}^s z(u) \cdot K(t-u) du \right) \int_{-\infty}^s P^\Omega(u) K(t-u) du\ ds. \end{aligned}$$

As $\Omega \to \infty$, we have

$$\begin{aligned} P(t) \quad \approx \quad & P(0) + S_{nd}B\left(\operatorname{diag}\left[\int_0^t \tilde{h}_{nd}\left(z(s), c\right) ds\right]\right) \\ & + \int_0^t S_{nd}J_{\tilde{h}_{nd}}\left(z(s)\right) P(s)ds \\ & + S_dB\left(\operatorname{diag}\left[\int_0^t \tilde{h}_d\left(\int_{-\infty}^s z(u)\cdot K(t-u)du, c\right) ds\right]\right) \\ & + \int_0^t S_dJ_{\tilde{h}_d}\left(\int_{-\infty}^s z(u)\cdot K(t-u)du\right)\int_{-\infty}^s P(u)K(t-u)du\ ds. \end{aligned}$$

Finally, it follows from Equation 5.25 that

$$Z(t) \approx Z_{LNA}(t) = z(t) + \frac{P(t)}{\sqrt{\Omega}},$$

and

$$X(t) \approx X_{LNA}(t) = x(t) + \sqrt{\Omega}P(t).$$

In analogy with the LNA for non-delayed systems, linearity implies that $X(t)$ follows a normal distribution at any arbitrary time t. Moreover, the time-evolution of its mean and variance are described by Equations 5.22 and 5.23, respectively.

5.5.3 EKBF - delayed reactions

In this section we outline a form of the filter which deals with systems incorporating delayed reactions. Proof of the diffusion approximation and its linearisation is provided in Brett and Galla (2013); here we extend the methodology by adding a Kalman updating step. We start in this case from a time-discretised state-space model, where the evolution of the unobserved states is given by the discretisation of Equation 5.11, i.e.

$$\begin{aligned} Y_t &= F^TX_t + \epsilon_t \\ X_t &= X_{t-\delta_t} + \delta_t g(X_{t-\delta_t}) + \delta_t\left(\sum_{s=t-\delta_t-\tau_m}^{t-\delta_t} W_{t-s}f(X_s)\right) \\ & \quad + \sqrt{\delta_t}d\left(X_{t-\delta_t}, \sum_{s=t-\delta_t-\tau_m}^{t-\delta_t} W_{t-s}f(X_s)\right) Z_t \\ Z_t &\sim MVN(0, I), \end{aligned}$$

where the quantities involved are defined as in the delayed species case, apart from the matrix of weights W_{t-s}, which has dimension $p \times w$; p is the dimension of X and w is the total number of delayed reactions (or, more generally delayed nonlinear transformations of the inputs X). Suppose again that an optimal estimate for the initial condition is available, i.e. $\pi(x_{0:\tau_m}|y_{0:\tau_m})$ follows a $\mathrm{MVN}(\rho_{0:\tau_m}, P_{x_{0:\tau_m}})$.

Taylor expansion of the nonlinear function $g(\cdot)$ can again be performed about ρ_{τ_m}, while $f(\cdot)$ is expanded about ρ_s, $s \in [0, \tau_m]$, up to the first order:

$$\begin{aligned} g(X_{\tau_m}) &\approx g(\rho_{\tau_m}) + J_g(\rho_{\tau_m})(X_{\tau_m} - \rho_{\tau_m}) \\ f(X_s) &\approx f(\rho_s) + J_f(\rho_s)(X_s - \rho_s). \end{aligned}$$

Following steps analogous to those of Section 5.5.1., we can obtain the equations for the evolution of mean and variance of X

$$\begin{aligned} \mathrm{E}[X_{t+\delta_t}|y_{0:\tau_m}] &\approx \rho_t + \delta_t g(\rho_t) + \delta_t \sum_s W_{t-s} f(\rho_s). \\ \mathrm{Var}[X_{t+\delta_t}|y_{0:\tau_m}] &\approx P_t + \delta_t \left[J_g(\rho_t)P_t + P_t^T J_g(\rho_t)^T\right] \\ &\quad +\delta_t \sum_s \left[W_{t-s} J_f(\rho_s) P_{s,t} + P_{t,s} J_f(\rho_s)^T W_{t-s}^T\right] \\ &\quad +\delta_t D\left(\rho_t, \sum_s W_{t-s} f(\rho_s)\right) \\ \mathrm{Cov}[X_{t+\delta_t}, X_t|y_{0:\tau_m}] &\approx P_t + \delta_t J_g(\rho_t) P_t + \delta_t \sum_s W_{t-s} J_f(\rho_s) P_{s,t}. \end{aligned}$$

For an arbitrary lag, given $i > j$, the covariance is

$$\mathrm{Cov}[X_i, X_j|y_{0:\tau_m}] = P_{i-\delta_t,j} + \delta_t J_g(\rho_{i-\delta_t}) P_{i-\delta_t,j} + \delta_t \sum_s W_{t-s} J_f(\rho_s) P_{s,j}.$$

We can once again draw the continuous limit, and obtain the LNA of Brett and Galla (2013) for the dynamics of the unobserved states. We indeed have, as $\delta_t \to 0$,

$$\begin{aligned} \frac{d\rho(t)}{dt} &\approx g(\rho(t)) + \int_{t-\tau_m}^{t} W(t-s) f(\rho(s)) ds. \\ \frac{dP(t)}{dt} &\approx \left[J_g(\rho(t))P(t) + P(t)^T J_g(\rho(t))^T\right] \\ &\quad + \int_{t-\tau_m}^{t} W(t-s) J_f(\rho(s)) P(s,t) + P(t,s) J_f(\rho(s))^T W(t-s)^T ds \\ &\quad + D\left(\rho(t), \int_{t-\tau_m}^{t} W(t-s) f(\rho(s)) ds\right). \end{aligned}$$

The Kalman update, and the computation of the likelihood can then be performed following exactly the same steps as in Section 5.5.1.

A last note is about the choice of m. Kalman filtering methodologies are typically aimed at online prediction, i.e. they predict future observations as new information becomes available. In this context, it is appropriate to correct at every iteration the estimates of the mean and variance, and restart the system. At the other extreme, the LNA has had a wider application in the inferential framework, where all the observations of interest are already available, and therefore in its non-restarted version, m is more naturally chosen to be equal to the total number of observations. The disadvantage of this choice, however, has already been outlined, although several possibilities are available between these two extremes. We follow in this work the 'traditional' filtering framework, and we therefore assume $m = 1$. Further investigation of the effect of m on the performance of the methodology can be of interest, but it is beyond the scope of the present work.

5.6 Application of the EKBF for delayed species to Cry1 model

Although a continuous expression for the time evolution of the mean and variance of the unobserved states can be easily defined, computations require in practice a discretisation of the system. We therefore focus, from now on, on a time-discretised version of the model. Model 5.12, applied for clarity to the $Cry1$ case, is

$$\begin{aligned} Y_t &= \kappa F^T \begin{bmatrix} X_t \\ X_{t-\delta_t} \\ \vdots \\ X_{t-\Delta_t+\delta_t} \end{bmatrix} + \epsilon_t \qquad (5.26) \\ X_t &= X_{t-\delta_t} + \delta_t \left[\nu \left(\sum_{s=t-\delta_t-\tau_m}^{t-\delta_t} X_s \cdot K_{t-s} \right) - \mu_{M_g} X_{t-\delta_t} \right] \\ &\quad + \sqrt{\delta_t} \sqrt{\nu \left(\sum_{s=t-\delta_t-\tau_m}^{t-\delta_t} X_s \cdot K_{t-s} \right) + \mu_{M_g} X_{t-\delta_t}} Z_t \\ \epsilon_t &\sim \mathcal{N}(0, \sigma_\epsilon^2) \\ Z_t &\sim \mathcal{N}(0, 1), \end{aligned}$$

where
$$\nu\left(\sum_{s=t-\delta_t-\tau_m}^{t-\delta_t} X_s\cdot K_{t-s}\right)=\frac{R_0}{1+\left(\frac{\sum_s X_s\cdot K_{t-s}}{K_{pc}}\right)^{n_{cr}}},$$

and F is a vector of ones of length (Δ_t/δ_t); furthermore, we assume Δ_t to be a multiple of δ_t. Note that this definition of F implements integration of the measurement process in a discretised form.

As mentioned in Section 5.5, a first issue is the specification of the initial condition, i.e. in the mean and variance of $\pi(x_{0:\tau_m}|y_{0:\tau_m})$. This a relatively crucial issue, as a good estimate of the initial condition, represents a 'reliable' input for the first states after τ_m, and therefore helps in achieving reliable parameter inference. On the other hand, a bad estimate may result in low quality predictions under the true parameters values, and therefore lead to biased parameters estimates.

Different strategies can be adopted to obtain the distribution of the initial condition. A first, relatively easy, solution is to propose a sufficiently flexible and general model for the unobserved states receiving the delayed inputs, in our application the gene mRNA. Depending on the number of parameters that we are willing to introduce, more or less sophisticated models can be adopted. We try to keep the dimensionality of the parameter space to a minimum, and so we assume a step function for the transcription rate ν, thus eliminating the dependence on past $Cry1$ (see Hey et al., 2015; Jenkins et al. 2013). We assume a single switch point, located at the peak-time of the observed $Cry1$-*luc*. This would add in total only two parameters: the two transcription rates, before and after the switch. However, there is one drawback with this approach, namely that it can result in a too restrictive model for the variance. The transcription and degradation rate in fact fully define the intrinsic variance, and a 'rough' choice for the transcription rate - due to the fact that we adopt a very restrictive approach on the number of switches, setting it to one - may underestimate the actual uncertainty about the signal. This leads to a poor partial smoothing, and therefore poor input estimates. To overcome this issue, we introduce a third parameter, denoted by β, which multiplies the diffusion term of our unobserved state equation. We then have that the time-evolution of the unobserved $Cry1$ is given by

$$\begin{aligned}X_t &= X_{t-\delta_t}+\delta_t\left(\mathrm{I}_{\{0:t_{ch}\}}\nu_1+\mathrm{I}_{\{t_{ch}:\tau_m\}}\nu_2-\mu_{M_g,in}X_{t-\delta_t}\right)\\ &\quad+\sqrt{\delta_t\beta}\,\mathrm{diag}\left(\sqrt{\mathrm{I}_{\{0:t_{ch}\}}\nu_1+\mathrm{I}_{\{t_{ch}:\tau_m\}}\nu_2+\mu_{M_g,in}X_{t-\delta_t}}\right)Z_t,\end{aligned}\tag{5.27}$$

where we adopt the usual notation for the quantities involved, and, in addition, I is the indicator function, t_{ch} is the switch time, and $\mu_{M_g,in}$, is the degradation rate of the initial condition, additionally introduced in order to avoid a possible source of bias in the estimation process, due to model mismatch of the initial condition. Once the optimal estimate for the initial condition has been obtained, we can proceed as in the general case outlined in the previous section. The mean and covariance of the unobserved state are propagated according to Equations 5.18 to 5.21. At each observed time-point, we update the mean and variance estimates of the unobserved state according to Equation 5.24.

It is possible to verify the performance of the methodology by setting the parameters to their true values used in simulations. Figures 5.6 (a) and (b) show the one step-ahead prediction densities and the partial smoothing of the unobserved state, as well as the one step-ahed prediction density of the observed state, along with their true trajectories, for one sample simulation in Figure 5.5, and for the two simulation scenarios. The time-interval δ_t is set equal to 0.1 h. We can observe an overall good coverage, as well as the shrinking of the variability intervals in the partial smoothing density.

To compare quantitatively the behaviour of the filter for different values of δ_t, we have computed the empirical coverage of both the predictive and the partial smoothing densities at level 95%. Results are shown in Table 5.2. We observe, as expected, a general improvement in terms of coverage as δ_t becomes closer to its true simulation value of 0.01 h. However, the major improvement is seen in the partial smoothing density, and in the low measurement error scenario. Moreover, there is clearly a minor improvement when lowering δ_t below 0.1 h, in terms of matching nominal and empirical coverage, for both the predictive and partial smoothing densities, and for both measurement error scenarios.

	One step-ahead prediction		Partial smoothing	
	$\sigma_\epsilon = 0.01$	$\sigma_\epsilon = 0.05$	$\sigma_\epsilon = 0.01$	$\sigma_\epsilon = 0.05$
$\delta_t = 0.5\,\mathrm{h}$	0.97 (1.79×10^{-2})	0.95 (2.60×10^{-2})	0.45 (5.18×10^{-2})	0.88 (4.05×10^{-2})
$\delta_t = 0.25\,\mathrm{h}$	0.98 (1.32×10^{-2})	0.96 (1.50×10^{-2})	0.86 (3.23×10^{-2})	0.94 (2.42×10^{-2})
$\delta_t = 0.1\,\mathrm{h}$	0.97 (1.23×10^{-2})	0.96 (1.36×10^{-2})	0.94 (1.39×10^{-2})	0.95 (1.67×10^{-2})
$\delta_t = 0.05\,\mathrm{h}$	0.96 (1.18×10^{-2})	0.95 (1.28×10^{-2})	0.95 (1.10×10^{-2})	0.95 (1.60×10^{-2})
$\delta_t = 0.01\,\mathrm{h}$	0.96 (1.25×10^{-2})	0.95 (1.53×10^{-2})	0.95 (8.83×10^{-3})	0.95 (1.37×10^{-2})

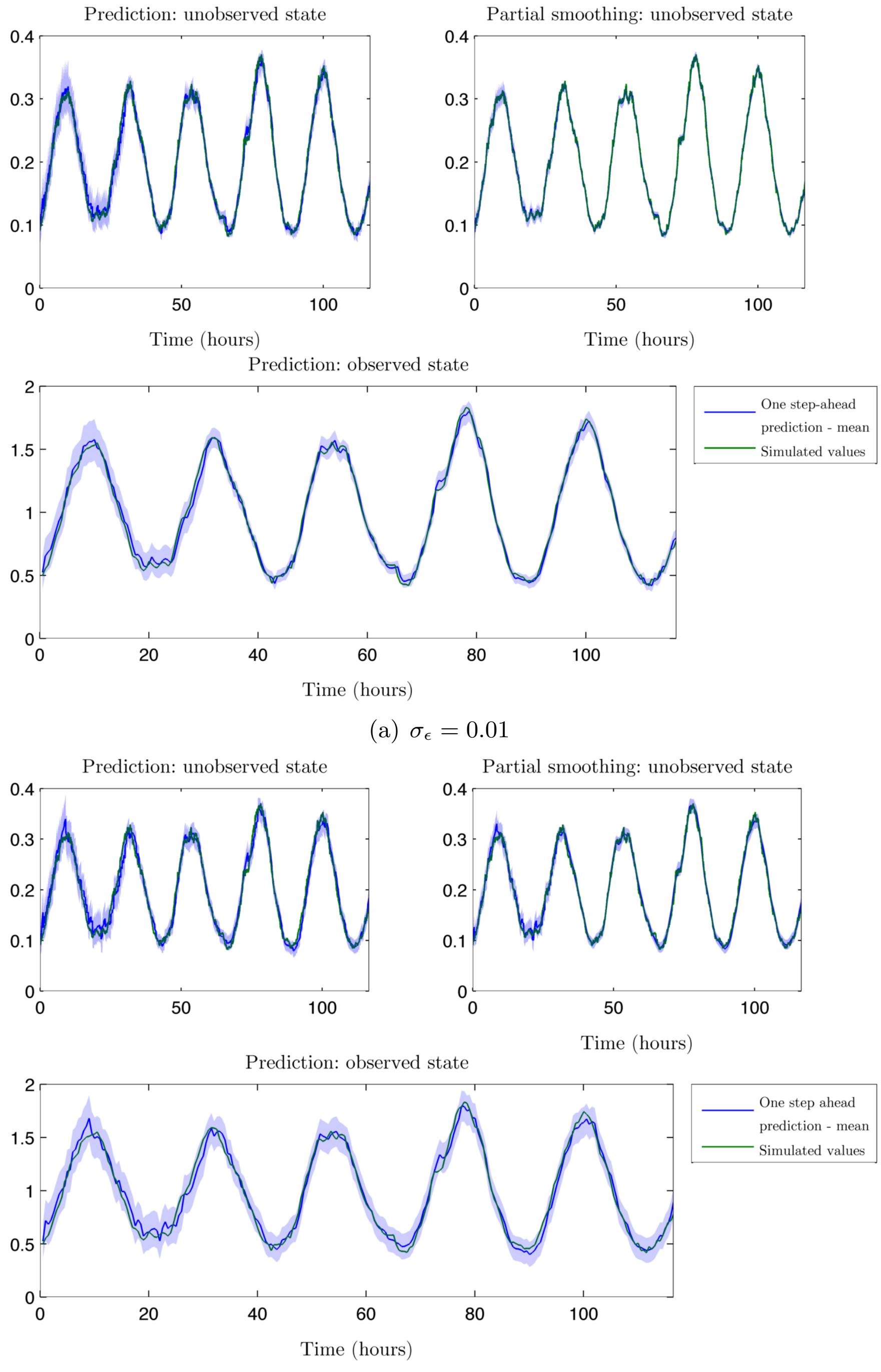
Table 5.2: EKBF with distributed delays of the species: empirical coverages at level 95%, for different unobserved states time-grids, predictive and partial smoothing density, for two levels of measurement error variance. Mean and standard deviation (in brackets) of 10 independent runs of the filter for Model 5.26, with unobserved state initial condition as in Equation 5.27, applied to the simulated data of Figure 5.5.

Table 5.3 provides average running times of the filter and selected sub-functions. We observe that running times increase more than linearly as the unobserved state time-grid is refined, ranging from 4.70×10^{-2} s for $\delta_t = 0.5$ h to up to 3.46×10^{2} s for $\delta_t = 0.01$ h. Note also that the computation of the unobserved states covariance accounts for a significant proportion of the total running time, with an increasing computational cost as the time-grid is thinned.

	Filter (full)	Unobserved states	Kalman update
		prediction (covariance)	(covariance)
$\delta_t = 0.5$ h	4.70×10^{-2} s (1.17×10^{-2})	3.99×10^{-3} s (3.15×10^{-3})	6.49×10^{-3} s (4.09×10^{-3})
$\delta_t = 0.25$ h	0.11 s (1.54×10^{-2})	3.17×10^{-2} s (1.06×10^{-2})	1.64×10^{-2} s (6.71×10^{-3})
$\delta_t = 0.1$ h	0.35 s (1.27×10^{-2})	0.15 s (1.34×10^{-2})	7.55×10^{-2} s (1.71×10^{-2})
$\delta_t = 0.05$ h	3.64 s (6.17×10^{-2})	2.20 s (5.57×10^{-2})	0.64 s (2.10×10^{-2})
$\delta_t = 0.01$ h	3.46×10^{2} s (1.95)	3.08×10^{2} s (1.73)	17.6 s (0.10)

Table 5.3: EKBF with distributed delays of the species: running times of the full filter and of selected sub-functions, for different unobserved states time-grids. Mean and standard deviation (in brackets) of 10 independent runs of the filter for Model 5.26, with unobserved state initial condition as in Equation 5.27, applied to the simulated data of Figure 5.5, $m = 1$ and $\sigma_\epsilon = 0.01$. Running times are based on simulations run on a RM One 310 computer, Core i7 3400 MHz processor, and 16 GB of RAM.

We have so far introduced a model and a filtering methodology for the *Cry*1-*luc* data. The ultimate goal of our analysis is to perform inference on the model parameters. The filtering methodology is crucial for this purpose, as it provides normal transition densities for the dynamics of the *Cry*1-*luc* and hence a tractable likelihood. Given the significant increase in the computational cost of the filtering methodology as the time-grid is refined, it is therefore sensible to investigate whether such a refinement has a significant impact on parameter likelihood. This aspect is studied in Section 6.1.1, before moving to inference validation on simulated data, and, finally, the full data analysis of the *Cry*1-*luc* observed data.



(a) $\sigma_\epsilon = 0.01$

(b) $\sigma_\epsilon = 0.05$

Figure 5.6: EKBF with distributed delays of the species, mean (blue) and ±2 SD (shaded blue): one step-ahead prediction of the unobserved state (top left); partial smoothing (top right); one step-ahead prediction of the observed state (bottom). True simulated time-series superimposed in green. Model 5.26, with unobserved state initial condition as in Equation 5.27, applied to the simulated data of Figure 5.5, for two levels of measurement error standard deviation. $\delta_t = 0.1\,\mathrm{h}$, $m = 1$.

Chapter 6

Inference for spatio-temporal Cry1-luc data from the SCN

In this chapter we present the results from the parameter estimation process on both simulated data and *Cry*1-*luc* observed data. The latter consist of three experimental replicates, two of these spanning over a time-interval of 144 hours, i.e. six cycles, while one experiment has been running for 116.5 hours, i.e. approximately five cycles. We perform the simulation study assuming the shorter duration, as additional observations are only likely to improve inference. Inference is then performed on the three data-sets, and results unified by applying a two-stage Bayesian hierarchical model (Lunn et al., 2013). Finally, the spatial structure of the hierarchical parameter estimates is investigated by means of an exploratory spatial analysis.

Inference is performed in a Bayesian framework, the likelihood being provided by the filtering methodology introduced in Chapter 5.

6.1 Inference validation on the simulated data

The performance of the estimation algorithm for the parameters of Model 5.26, having an unobserved state initial condition as in Equation 5.27, is studied on the set of simulated data of Figure 5.5. In particular, we perform the simulation study on 10 i.i.d. replications of two simulation scenarios, the first which assumes measurement error standard deviation $\sigma_\epsilon = 0.01$, and the second $\sigma_\epsilon = 0.05$. We first compare the effect of the time-grid for the evolution of the unobserved states, δ_t, by computing univariate log-likelihoods. We then design and run an MCMC algorithm, to re-estimate the parameters used for simulation.

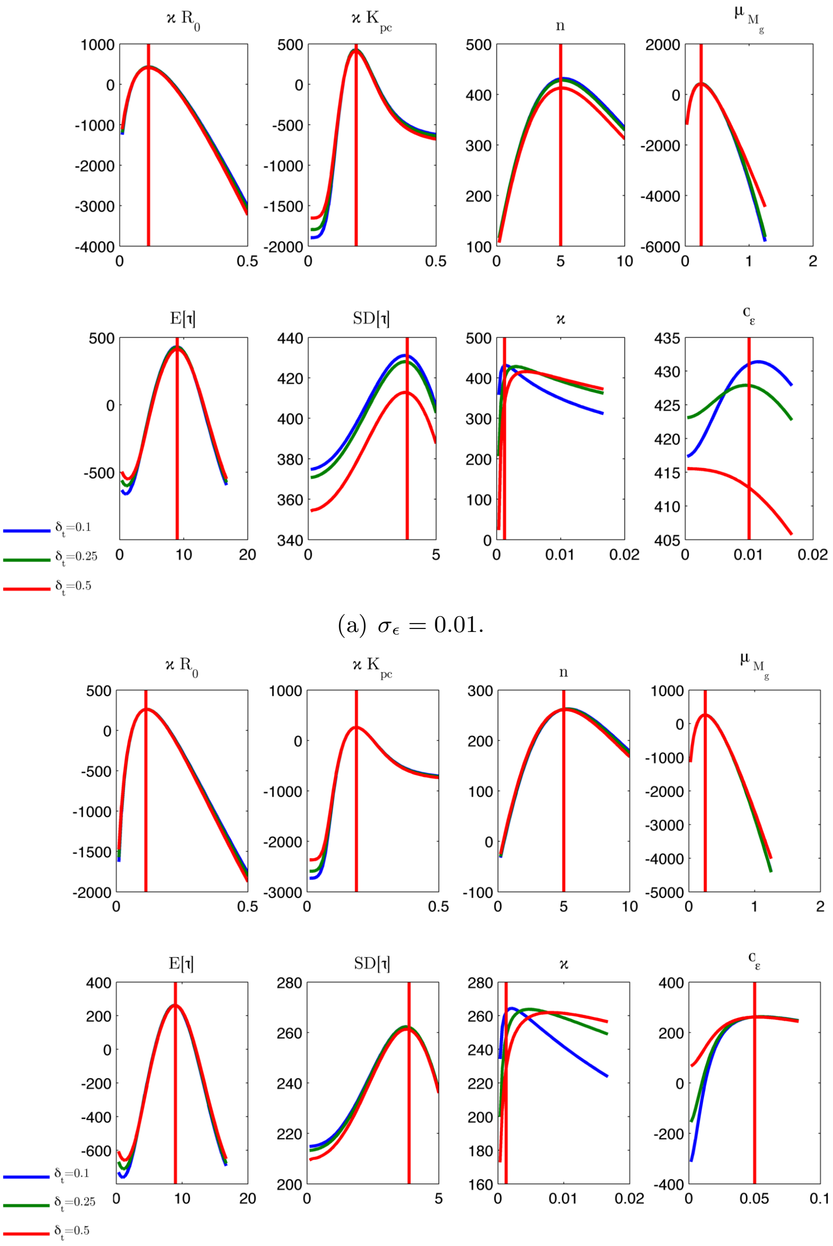
6.1.1 Parameter likelihood for the simulated data

The computational cost of the algorithm represents an important point, which becomes crucial when considering the inferential application of the algorithm in an MCMC framework. We have provided in Section 5.6 the computational cost and coverage properties of the proposed filtering methodology for different choices of δ_t. Examined time-grids which assume δ_t lower than 0.1 h seem to be prohibitively expensive when considering the application of the filter in an MCMC context; in fact, 2×10^5 iterations of the filter when assuming $\delta_t = 0.01$ h would require approximately 800 days, compared to approximately 8 days of the $\delta_t = 0.05$ h case, and approximately 19 hours for $\delta_t = 0.1$ h. Note also that multiple likelihood evaluations are required for each MCMC iteration if a Metropolis-within-Gibbs scheme is adopted. Moreover, the increased computational cost seems to be not well balanced by significant advantages in terms of coverage of the filter predictive density. We therefore focus on the time-grids having δ_t equal to 0.1 h, 0.25 h and 0.5 h. Among the selected time-grids, it is sensible to investigate whether a higher refinement provides a significant advantage also for inferential purposes, e.g by providing a maximum of the log-likelihood closer to the true simulation value. We therefore compute univariate log-likelihoods for each parameter, with the remaining parameters set at their true values. The computation is performed on one simulation replicate from each simulation scenario, and the results are shown in Figures 6.1 (a) and (b).

We notice a minor effect of the time-grid on most of the parameters involved. However, the choice of a fine time-grid seems more relevant for the parameters κ and σ_ϵ, where we observe a log-likelihood peak closer to the true simulation value as δ_t decreases. This result motivates the choice of a delayed acceptance MCMC algorithm (Christen and Fox, 2005; Golightly et al., 2015), which allows to exploit the fast likelihood computation provided by the filter for $\delta_t = 0.5$ h to explore the parameter space, but finally accepts values according to the likelihood provided by the filter for $\delta_t = 0.1$ h. The designed MCMC algorithm and the simulation study results are presented in the following section.

6.1.2 Inference for the simulated data

We sample all the parameters involved on the logarithmic scale, except the mean and standard deviation of the distributed delay. As usual, we move κ from the observation equation, to the unobserved state equation; we refer again to Section 2.5.1 for the effect of this change on parameter estimates. Priors are set to be $\mathcal{N}(0, 10^2)$ for $\log(\kappa R_0)$, $\log(\kappa K_{pc})$, $\log(\kappa \nu_1)$ and $\log(\kappa \nu_2)$, and $\mathcal{N}(\log(1.5), 5^2)$ for



(a) $\sigma_\epsilon = 0.01$.

(b) $\sigma_\epsilon = 0.05$.

Figure 6.1: Univariate log-likelihood plots for a sub-set of parameters (initial condition not included) of Model 5.26, with unobserved state initial condition as in Equation 5.27, with the remaining parameters set at their true simulation values (initial condition adjusted according to visual inspection). Red lines set at the true simulation values (for $\delta_t = 0.1\,\mathrm{h}$). Model applied to the simulated data of Figure 5.5, for two simulation scenarios assuming $\sigma_\epsilon = 0.01$ (top) and $\sigma_\epsilon = 0.05$ (bottom), $m = 1$.

$\log(n)$, given their biological interpretation. We employ $\mathcal{N}(0, 20^2)$ for $\log(\kappa V[X_0])$, $\log(\kappa E[X_0])$, $\log(\sigma_\epsilon)$ and $\log(\kappa)$, given the approximate nature of the initial condition, and to allow for very low measurement error or high molecules numbers. We adopt a $\mathcal{N}(\log(0.58), 0.25^2)$ prior for the degradation rates (initial condition and following cycles) of *Cry*1-*luc*, following Yamaguchi et al. (2003). The authors report the mean half-life in the text, while we have assumed the standard deviation from visual inspection of their Figure S1. In particular, we adopt a value of 0.25, which seems to be not over-restrictive. Finally, the prior for the mean of the distributed delay is set to be $U(0, 23)$ and the prior for the standard deviation $U(0, 20)$. The latter two priors can be justified by the circadian rhythmicity of the data: it is sensible to assume that the cellular product of the previous cycle is mainly responsible for the dynamics of the consecutive one. Initial conditions for the parameter chains are randomly drawn from the prior densities.

The MCMC algorithm adopted has a pilot run of 3×10^3 iterations, in which a mixture of single-components adaptive proposal variance schemes, and adaptive block proposal covariance matrix schemes (Roberts and Rosenthal, 2009), are employed, in order to optimise exploration of the posterior density and computational efficiency. The unobserved states time-grid here is set to $\delta_t = 0.5\,\text{h}$, to obtain a fast first exploration of the posterior density.

From iteration 3×10^3, we define three blocks of parameters, one for the transcription function parameters plus degradation, one for the initial condition parameters, and one for scale and measurement error standard deviation, and we adopt a random walk scheme in which the variance of the proposals is proportional to the variance of the previously accepted values. Moreover, a delayed acceptance component is introduced (Christen and Fox, 2005; Golightly et al., 2015; Sherlock et al., 2016; Sherlock et al., 2015). In a delayed acceptance scheme, samples are first proposed according to a 'fast' likelihood evaluation, which means in our case adopting $\delta_t = 0.5\,\text{h}$; samples are then accepted or rejected according to the usual acceptance ratio of Equation 2.14. If the proposed samples are accepted, a slower and more precise evaluation of the likelihood is performed, meaning in our scenario that δ_t is set to $0.1\,\text{h}$. The acceptance ratio of the nested step is, then, (Christen and Fox, 2005; Golightly et al., 2015)

$$\alpha_{DA} = \min\left\{1, \frac{\pi_{0.1}(y_{0:T}|\Psi^*)}{\pi_{0.1}(y_{0:T}|\Psi)} \frac{\pi_{0.5}(y_{0:T}|\Psi)}{\pi_{0.5}(y_{0:T}|\Psi^*)}\right\},$$

where we recall that Ψ^* is the proposed sample of parameters, accepted in the fast likelihood evaluation step, and Ψ the accepted sample from the previous iteration of the MCMC algorithm. Moreover, we denote by $\pi_{0.1}$ the likelihood under the $\delta_t = 0.1\,\mathrm{h}$ case, and by $\pi_{0.5}$ the likelihood under the $\delta_t = 0.5\,\mathrm{h}$ one. The choice of a delayed acceptance scheme is motivated by the fact that the proposed filter becomes computationally time-demanding as the chosen δ_t decreases. Indeed, we have a computational cost which is approximately 10 times higher under the $\delta_t = 0.1\,\mathrm{h}$ case than in the $\delta_t = 0.5\,\mathrm{h}$ one (see Table 5.3); in the example of Sherlock et al. (2015) for a random walk Metropolis scheme, an increase in computational speed by a factor of 10 is sufficient to achieve an increase in efficiency of the algorithm by a factor of three with respect to a non-delayed scheme, given an optimal jump size for the proposals. The gain in efficiency is also influenced by the curvature and position of the stage one proposal with respect to the target density. In our case, the univariate log-likelihoods seem of comparable curvature or slightly flatter for most of the parameters under $\delta_t = 0.5\,\mathrm{h}$, and their maxima are relatively close under the two time-intervals. Following the results of Sherlock et al. (2015), this suggests that the 'cheap' approximation is already a relatively good approximation, while being significantly faster, and the delayed acceptance scheme can therefore result in a significant gain in efficiency. Optimisation of the scheme according to the jump size of the proposals may require further investigation, although the proportionality factors for the covariance matrices of the proposals are chosen in order to improve the overall performance of the chains, based on visual investigation of the trace-plots (some variability between different runs in the quality of the mixing is also present).

After the pilot run, we discard 10^5 iterations as a burn-in, and thin the chains by retaining one sample every 100 iterations. Figure 6.2 (a) and (b) show the posterior densities for the transcription function parameters, $\log(\sigma_\epsilon)$ and $\log(\kappa)$, for the two simulation scenarios. We report bias in the estimation of $\log(\mu_{M_g})$, due to the fact that the assumed prior is concentrated away from the assumed simulation value. The simulation value is set in order to approximately reproduce the behaviour of real data, and thus it allows to investigate the possible effect of the degradation rate bias on the remaining parameters in the estimation process. We observe, indeed, that $\log(\kappa R_0)$, as well as $E[\tau]$ tend to be slightly overestimated, while $\log(n)$ show a downwards bias. A weaker underestimation of κ is also observed. The remaining parameters seem to be not significantly affected.

An additional run of the estimation algorithm, which assumes a degradation prior density centred close to the true simulation value, allows to confirm that the observed bias is only due to the mismatch of the degradation rate with respect to its prior. We adopt in this case a $\mathcal{N}(\log(0.3), 0.35^2)$ prior, which allows for relatively high dispersion and a minor mean mismatch. As we can observe in Figure 6.3 (a) and (b), the intervals in this case generally contain the true simulation value of the parameters, we only report a relatively poorer coverage for $\log(n)$ in the simulation scenario having $\sigma_\epsilon = 0.01$, where 3 out 10 values fall outside the 95% HPDIs, which however falls to only one case in the $\sigma_\epsilon = 0.05$ case.

Mixing of the chains is also generally poorer in the $\sigma_\epsilon = 0.01$ case, we believe due to the highly challenging correlation structure and more peaked likelihood.

6.2 Cry1-luc data analysis

In this section we present inferential results for the three available experimental replicates of *Cry*1-*luc* spatio-temporal data in the SCN.

6.2.1 Exploratory analysis of the Cry1-luc data

As we ideally want to investigate spatial variation of the model parameters, a first sensible step is to investigate spatial variation of the available *Cry*1-*luc* time-series data across the SCN. Relevant aspects of circadian time-series are the amplitude, period and phase. To compute these quantities, we follow the same procedure outlined in 4.2.1 for the Nanostring data, which we recall resorts to spectral analysis. The exploratory analysis is performed on the de-trended *Cry*1-*luc* data for experiment 1, while the slope of a linear regression fitted on the peak light-intensities versus time has not resulted significant in experiment 2 and 3 (0 is included in the 95 % level HPDIs), hence we perform for these experiments the exploratory analysis on the raw data.

We first identify the most relevant frequency of the observed data as the frequency at which the periodogram reaches its maximum. As expected, we find the periodogram mode at the circadian frequency for almost all the observed locations in all the three experiments. Deviations are observed for some locations at the very top, top-left, of the images, which also generally show very noisy behaviour. The estimated period is thus 23.3 hours for experiment 1, and 24 hours for experiments 2 and 3 in almost all non-extreme locations. We then compute amplitude and phase of the harmonic corresponding to this leading frequency. Figure 6.4 shows the spatial variation of the amplitude across the right half of the SCN, for the three experiments. We can see that in all cases it tends to be higher in the central region, and decreases as we move towards more peripheral locations.

A second relevant aspect is given by the phase of the observed time-series. Figure 6.5 shows a comparison, again across the three experimental replicates, for the leading circadian frequency. Note that the phase time is computed by assuming that each experiment starts at $t = 0$. If we observe the time-distribution of the phases, we observe that the last experiment has earlier phases and thus seems to start later than experiment 1 and 2. We also observe an interesting pattern which seems to be conserved across the three replicates: the peripheral regions show a phase advance of approximately 1-2 hours with respect to the more central area; the central area spans approximately from the bottom left part of the section, up to the centre-bottom right extreme, in a bow-like shape. A spatial arrangement of *Cry*1-*luc* phases in the SCN which sees earlier phases in the upper-left dorsal area, and later phases as we move towards the lower-right ventral area, is indeed also observed in Maywood et al. (2013).

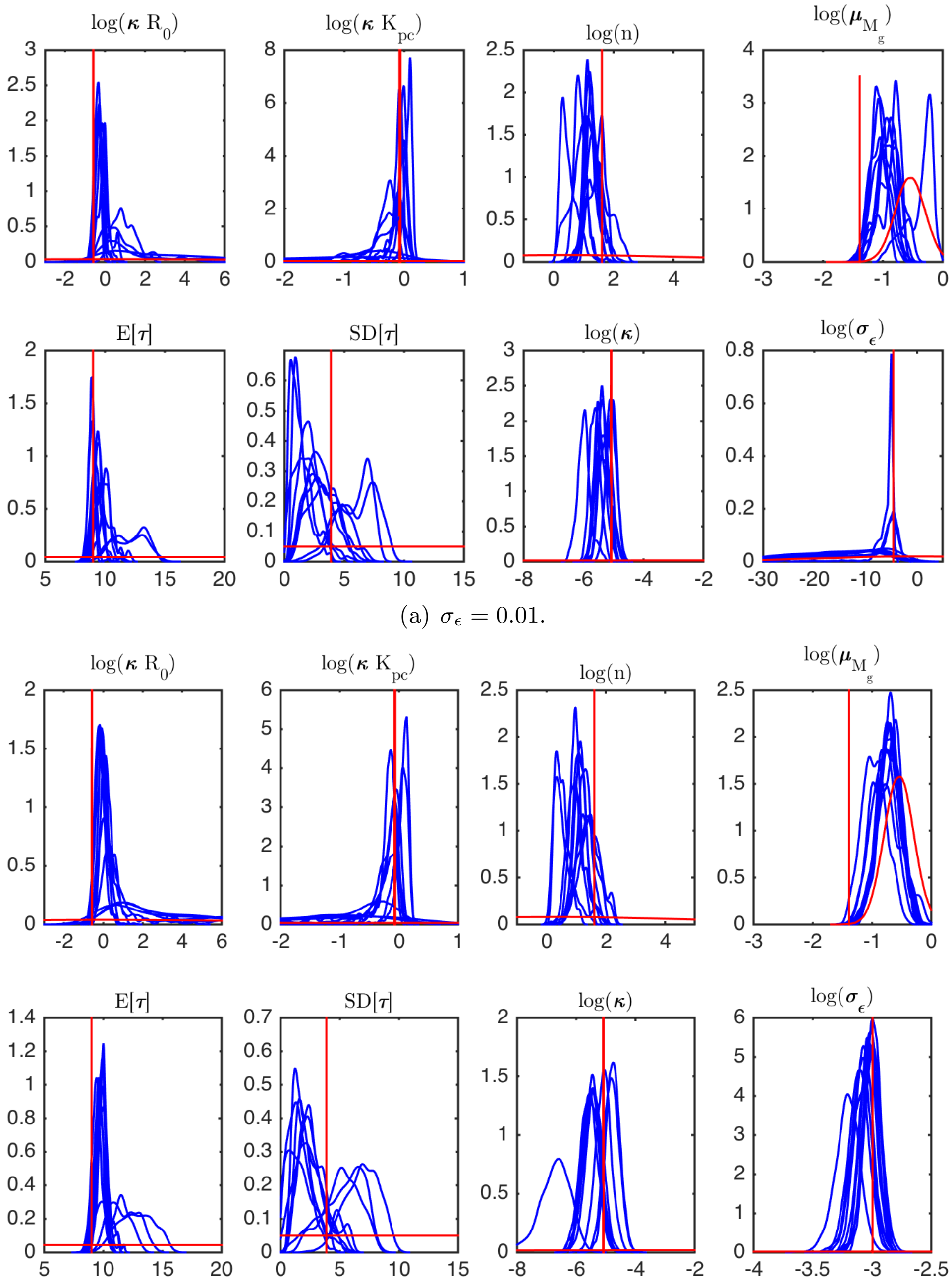
Note that we select slightly more than 100 locations for each experiment, equally spaced across the SCN, and falling within a boundary defined by a threshold light intensity equal to 0.1; the threshold is applied to light intensities of roughly the same time of the first circadian cycle in the three experiments, estimated by comparing the medians of the phases estimated within each experiment. Assuming experiment 3 as baseline time-reference, we thus take $t = 0\,\text{h}$ for experiment 3, $t = 5.5\,\text{h}$ experiment 1, and $t = 4.5\,\text{h}$ for experiment 2. The background light intensity in each figure is *Cry*1-*luc* at these selected times. We remark that, in all the three cases, hour 0 defines the beginning of the experiment, i.e. there is no relationship with circadian time.

6.2.2 Inference for single experiments

We present the results of the estimation process by plotting the posterior median estimates of the parameters of interest, along with a measure of their variability, in a spatial fashion across the SCN. We ideally want to both visualise patterns of spatial variation, and have an estimate of their reliability. At coordinates corresponding to the analysed 2×2 pixel boxes, we therefore plot a dot with colour scale proportional to posterior density median of each parameter, and size inversely proportional to the corresponding coefficient of quartile variation (see e.g. Bonett, 2006). Note that proportionalities of both colour and size can be different for different parameters, in order to obtain visually interpretable plots. In analogy with Figures 6.4 and 6.5, the background is given by *Cry*1-*luc* light intensity at comparable times of the first circadian cycle across the three experiments.

We note in Figure 6.6 high values of n, with respect to the known number of binding sites, as well as low values of μ_{M_g} with respect to the prior of Yamaguchi et al. (2003). These parameters are indeed believed to compensate for model approximation, as our model represents only a simplified representation of the more complex process outlined in Chapter 5. We also observe a low degree of spatial variation for κR_0, κK_{pc}, if we ignore a proportion of extreme estimates characterised by a higher dispersion, which we again attribute to model approximation. The remaining parameters seem to exhibit a more evident spatial structure, with similar values clustering together, as well as exhibiting decreasing or increasing trends as we move from central to more peripheral locations. In Figure 6.7 we first observe an increase in the median estimates of σ_ϵ from central to more external locations, suggesting that peripheral locations are characterised by a lower signal to (measurement) noise ratio. Central locations tend also to show higher responsiveness of the promoter, as indicated by a high estimate of n, as well as a higher mean and standard deviation of the delay, and scale κ. These estimates point in the direction of a system comprising a higher intrinsic noise in the central area of the SCN, with longer and noisier delays in the feedback cycles. This result is particularly interesting if considering that the core area of the SCN, which covers approximately the lower half of the SCN, is the recipient of light impulses coming from the retina; we may put forward the hypothesis that intrinsic noise contributes to the higher responsiveness of the upper core SCN to external inputs, although we note that such a result would also imply different mechanisms to be taking place between lower and upper core SCN regions. The link between intrinsic noise and responsiveness to external signals, is supported by mathematical studies on the effect of noise on oscillatory systems, and in particular Steuer et al. (2003) show that, for a given amplitude of the input, noisy systems may exhibit higher amplitudes in the outputs than deterministic ones. Moreover, studies have found that intrinsic noise has a key role in generating circadian oscillations as a consequence of extra-cellular signalling, when individual cell clocks are disrupted by mutation of BMAL1 (Ko et al., 2010).

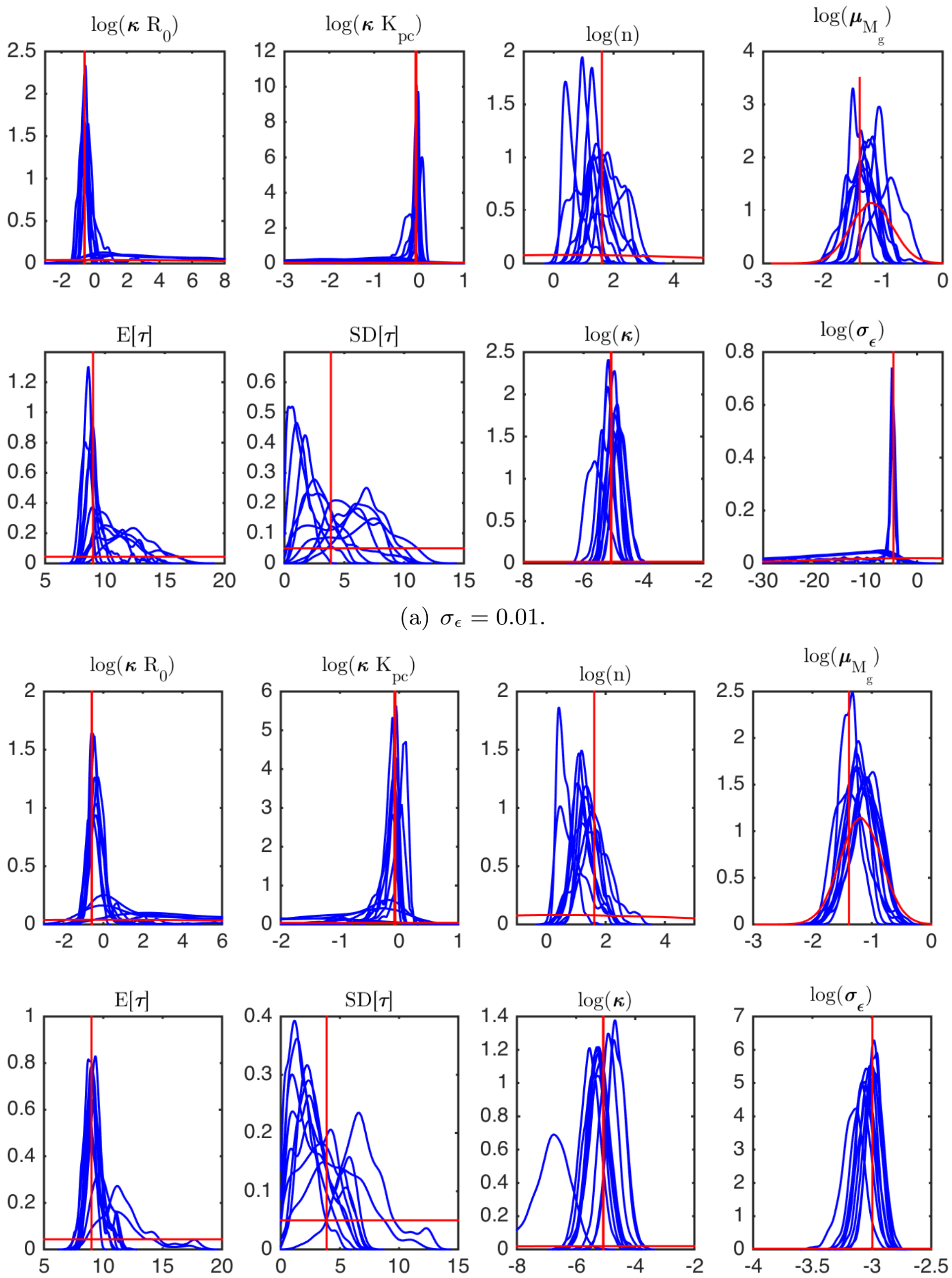
Finally, comparison among the three experiments show similarities in parameter estimates. In order to obtain a clearer and more robust picture of the inferred spatial dynamics, a natural step is to perform a meta-analytic study, which we formulate via a Bayesian hierarchical model. This is presented in Section 6.3.



(a) $\sigma_\epsilon = 0.01$.

(b) $\sigma_\epsilon = 0.05$.

Figure 6.2: Kernel densities estimates of the model parameters posterior densities, excluding the parameters of the initial condition. Model 5.26, with unobserved state initial condition as in Equation 5.27, applied to the simulated data of Figure 5.5, for the two simulation scenarios assuming $\sigma_\epsilon = 0.01$ (top) and $\sigma_\epsilon = 0.05$ (bottom), $m = 1$. MCMC samples for 10 independent replications. The red vertical line is at the true value, and the prior density is also superimposed in red. Prior for the degradation rate from Yamaguchi et al. (2003).



(a) $\sigma_\epsilon = 0.01$.

(b) $\sigma_\epsilon = 0.05$.

Figure 6.3: Kernel densities estimates of the model parameters posterior densities, excluding the parameters of the initial condition. Model 5.26, with unobserved state initial condition as in Equation 5.27, applied to the simulated data of Figure 5.5, for the two simulation scenarios assuming $\sigma_\epsilon = 0.01$ (top) and $\sigma_\epsilon = 0.05$ (bottom), $m = 1$. MCMC samples for 10 independent replications. The red vertical line is at the true value, and the prior density is also superimposed in red. Prior for the degradation rate centred close to the true simulation value.

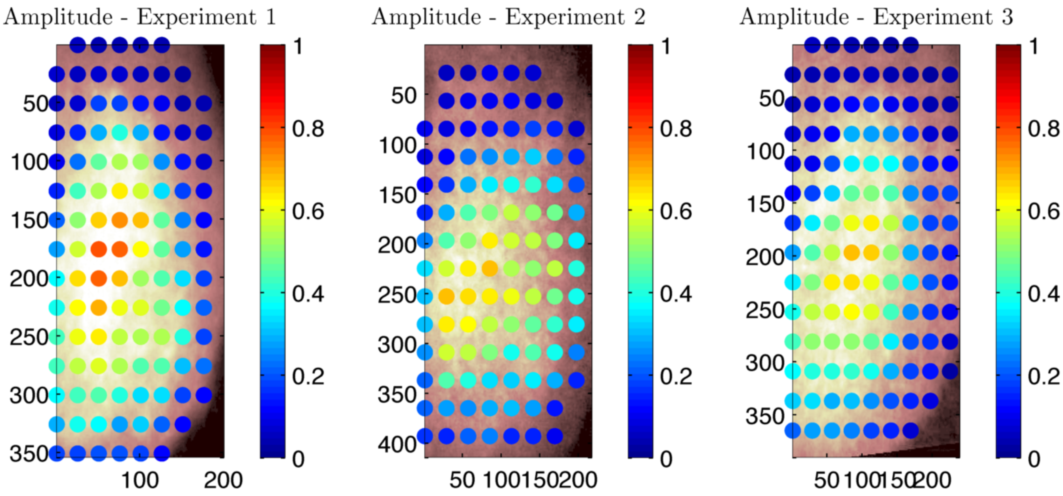


Figure 6.4: Amplitude of the harmonic corresponding to the leading frequency of the observed *Cry*1-*luc* light intensities, for selected locations in the right half of the SCN, and averaged over 2 × 2 pixels blocks. Points at the selected locations. Experiment 1 (left), 2 (middle) and 3 (right). Data from Hastings lab. at MRC, Cambridge.

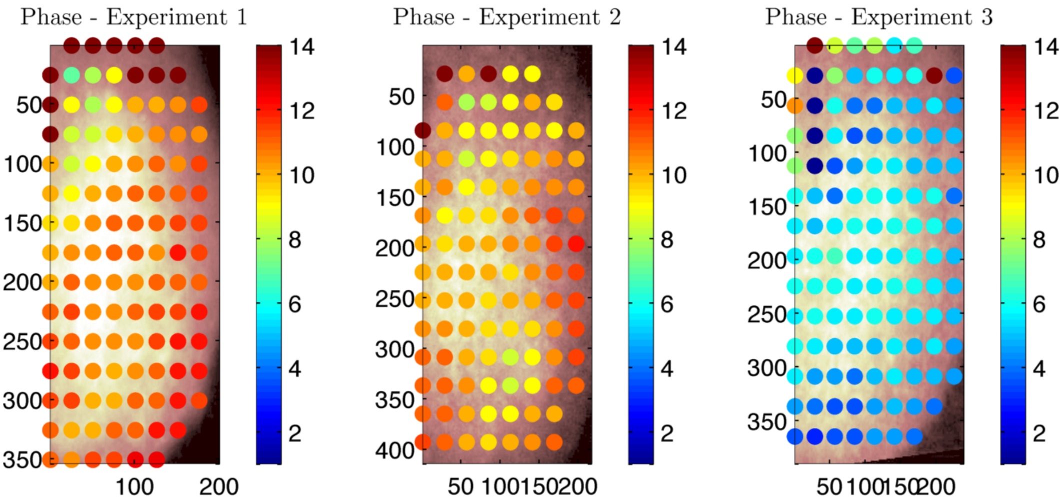


Figure 6.5: Phase of the harmonic corresponding to the leading frequency of the observed *Cry*1-*luc* light intensities, for selected locations in the right half of the SCN, and averaged over 2 × 2 pixels blocks. Points at the selected locations. Experiment 1 (left), 2 (middle) and 3 (right). Data from Hastings lab. at MRC, Cambridge.

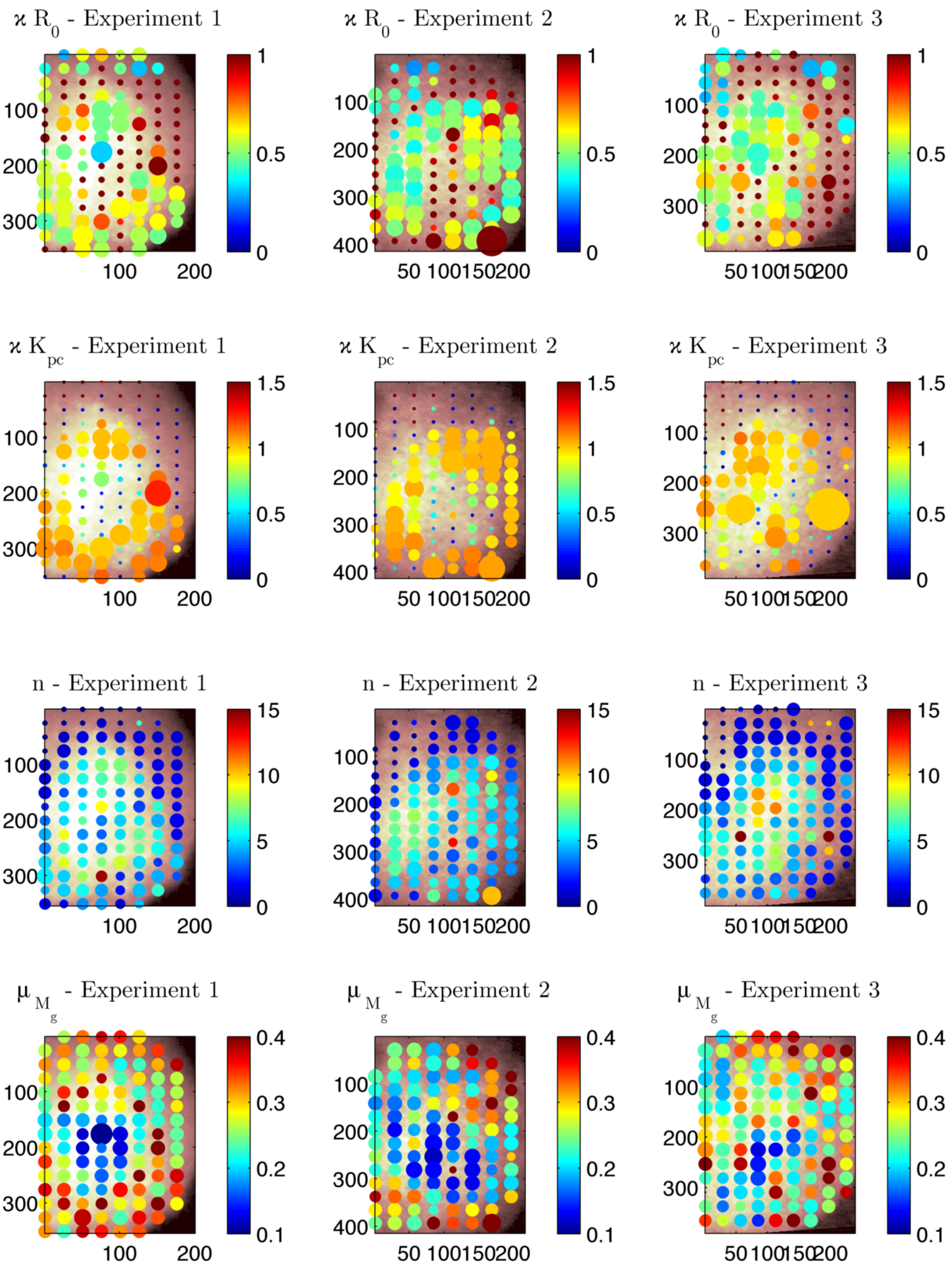


Figure 6.6: Posterior estimates for κR_0, κK_{pc}, n and μ_{M_g}; colour scale proportional to the median, size inversely proportional to the coefficient of quartile variation. Samples are transformed in the original scale to compute the interquartile coefficient of variation. The size scale may vary between different parameters for plotting purposes, but are equal for the same parameter across the three experiments. The MCMC estimation algorithm for Model 5.26, with unobserved state initial condition as in Equation 5.27, is run on *Cry*1-*luc* intensities averaged over 2 × 2 pixel boxes, at the plotted points locations. The chains of two locations in experiment 3 show very poor mixing, and their estimates are therefore not included for plotting purposes (very low coefficient of quartile variation). Data from Hastings lab. at MRC, Cambridge.

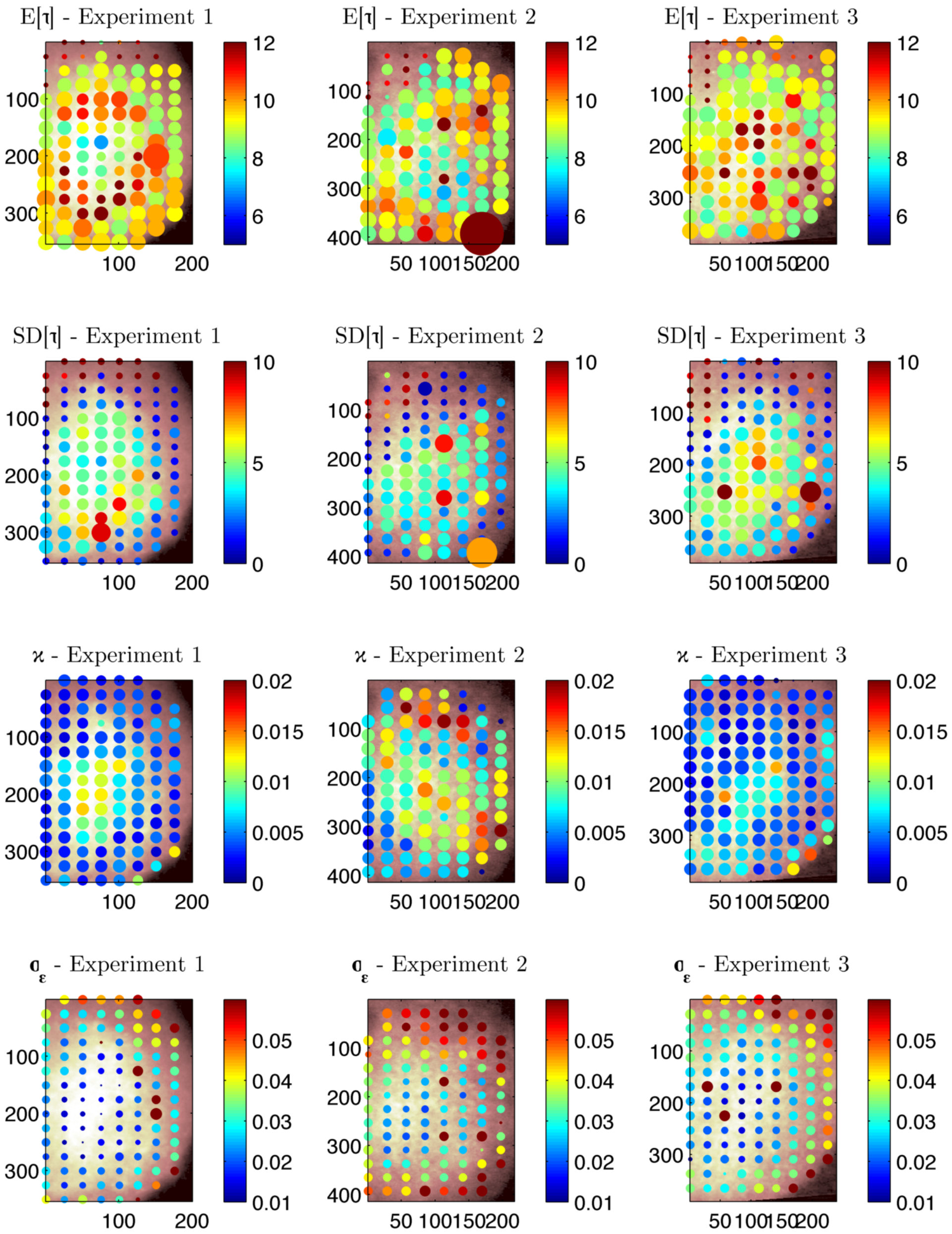


Figure 6.7: Posterior estimates for $E[\tau]$, $SD[\tau]$, σ_ϵ and κ, colour scale proportional to the median, size inversely proportional to the coefficient of quartile variation. Samples are transformed in the original scale to compute the interquartile coefficient of variation. The size scale may vary between different parameters for plotting purposes, but are equal for the same parameter across the three experiments. The MCMC estimation algorithm for Model 5.26, with unobserved state initial condition as in Equation 5.27, is run on *Cry*1-*luc* intensities averaged over 2 × 2 pixel boxes, at the plotted points locations. The chains of two locations in experiment 3 show very poor mixing, and their estimates are therefore not included for plotting purposes (very low coefficient of quartile variation). Data from Hastings lab. at MRC, Cambridge.

6.2.3 Diagnostics of Cry1-luc model fit

Before moving to the meta-analytic study, we perform a check of the model fit through inspection of the standardised residuals. In particular we investigate residual periodicity and normality. If a good fit is achieved by the model, then the residuals should be approximately normal and should not exhibit any more circadian, or other relevant, cyclicity. In analogy with the *Arabidopsis Thaliana* analysis, we obtain samples of residuals for a thinned set of parameters samples from the MCMC algorithm. The procedure is performed for each location and for each experiment. To account for the delay, we focus on the residuals obtained from fitting the model to the data after the first 30 hours. The initial condition is in fact only instrumental into obtaining reasonable starting estimates of the mean and variance of the process, and its fit is not of particular interest for our purposes. Finally, in order to obtain a periodogram estimate at approximately 24 and 12 hour periods, we focus on residual periodicity only of the last four cycles for experiment 1 and the last five cycles for experiment 2 and 3. This is required by the discrete temporal nature of the observations, and consequently of the residuals.

Figure 6.8 shows the median estimates of the Shapiro-Wilk test, as well as the 12 hour and 24 hour normalised periodogram estimates across the SCN, following the same procedure of Section 4.2.2. Recall that the Shapiro-Wilk test is computed with the *swtest* function in MATLAB (Saida, 2007), which performs either the Shapiro-Wilk or the Shapiro-Francia test based on the sample kurtosis. A simulation study performed on 5×10^4 vectors of i.i.d samples drawn from a $\mathcal{N}(0, 1)$ and having the same length of the available residuals, provides a median value for the test, under the null hypothesis of normality, which is equal to 0.992 for experiment 1, and 0.994 for experiment 2 and 3. We do not notice in Figure 6.8 (top) any evident spatial pattern, and the normality assumption seems generally adequate.

With respect to the residual periodicity, we notice in Figure 6.8 (bottom) some residual 24 hour periodicity, in particular in the first two experiments and in the central region of the SCN. The 12 hour residual periodicity in Figure 6.8 (centre), on the other hand, seems to be more wide-spread, being consistently present in the three experiments, and in a wider area than the 24 hour one. The interpretation is not straightforward, but we can assume that residual periodicity is caused by processes which are not explicitly included in the model. We may postulate an influence of the choice of a multiplicative scaling factor κ to account for the fact that we observe light intensities, and not actual molecules numbers or concentrations. This assumption may be too simplistic, as light emission comes indeed from an enzymatic reaction, possibly better described by a Hill functional form. It is also

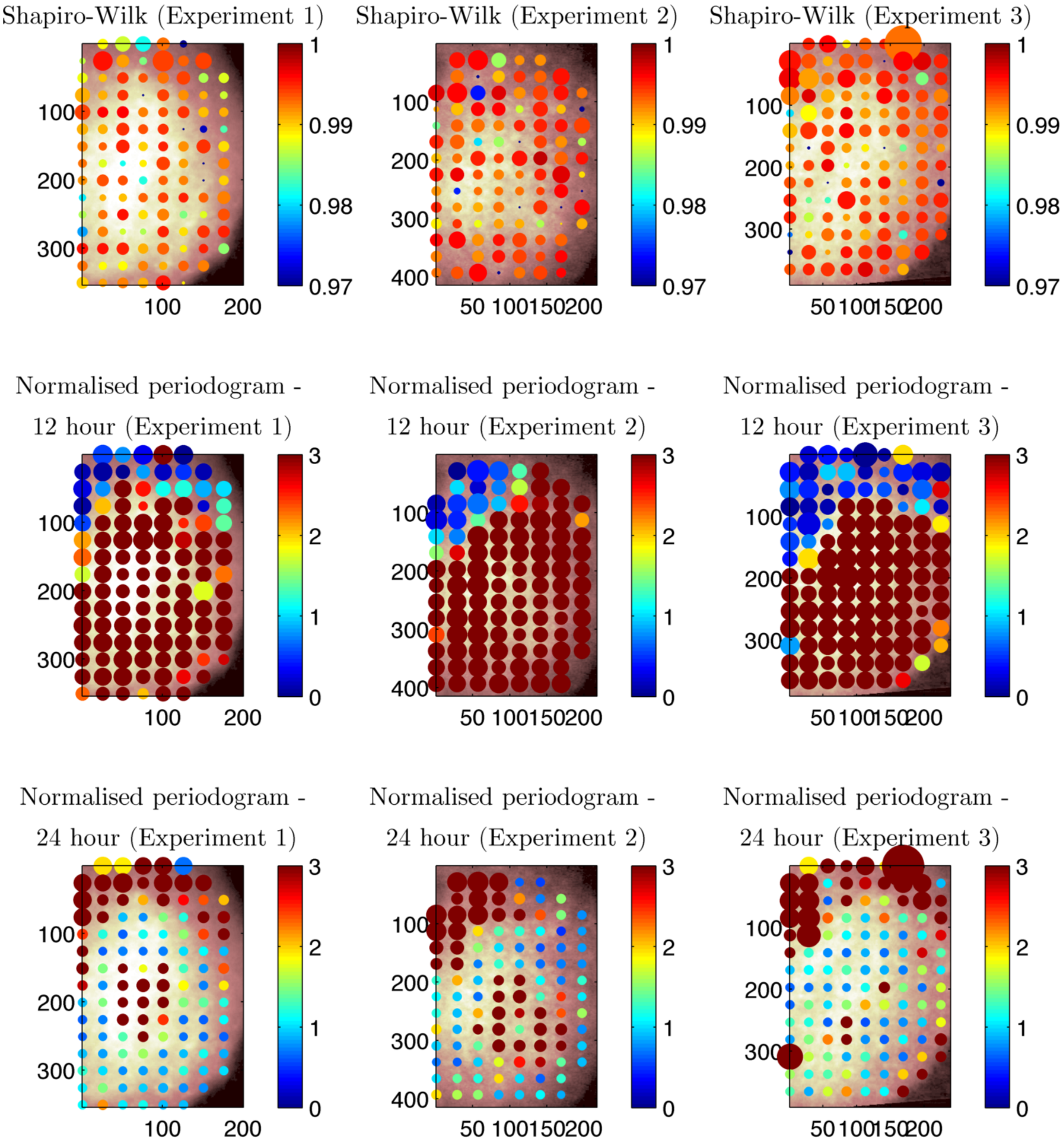


Figure 6.8: Diagnostics plots for model fit: Shapiro-Wilk test (top), normalised periodogram estimate for the 12 hour period (centre), and normalised periodogram estimate for the 24 hour period (bottom), computed on the standardised residuals for Model 5.26, with unobserved state initial condition as in Equation 5.27. Colour scale is proportional to the median, size inversely proportional to the coefficient of quartile variation. Proportionalities may vary between different measures for plotting purposes, but are equal for the same measure across the three experiments. The MCMC estimation algorithm is run on *Cry*1-*luc* intensities averaged over 2×2 pixel boxes, at the plotted points locations. Data from Hastings lab. at MRC, Cambridge.

interesting to note that the 24 hour residual periodicity is concentrated in the more central locations. Other processes may be taking place, e.g. influence of intercellular signalling on *Per* transcription, and therefore on PER protein; recall in fact that the is the PER/CRY complex which effectively represses *Cry*1 transcription, while our model only accounts for *Cry*1 auto-repression.

Although the residual periodicity, especially for the 12 hour period, provides an indication that a more detailed model may be required, for relatively wide regions of the SCN most of the density mass of 24 hour period residual rhythmicity is below the 95% threshold level. The diagnostics seem to suggest overall care in the interpretation of the parameters, but it also importantly provides a spatial indication on where additional processes that are not accounted for by our model, may be taking place.

6.3 Hierarchical Bayesian meta-analysis

The results of the single experiments provide a first useful insight on the transcription function parameters values across the SCN, but a unified estimate across the three experiments may provide a more powerful and interpretable summary of the results. In a Bayesian framework, this objective translates into assuming the parameters of each experiment as generated by a common prior distribution, whose parameters are called hyperparameters. Each hyperparameter is assigned in turn a prior distribution, called hyperprior.

Following Lunn et al. (2013), we define a hierarchical model in which the lower layer of the hierarchy is represented by the *Cry*1-*luc* time-series for each SCN location and for each experiment, i.e. the vectors $y_{z,l}$, where $z = 1, 2, 3$ defines the experiment and $l = 1, ..., L$ defines the location. The likelihood is computed via the EKBF for systems with delayed species introduced in 5.5.1, conditional on the location and the experiment specific parameters $\Psi_{z,l}$, whose estimates are shown in Section 6.2.2.

The vectors of the experiment and the location specific parameters $\Psi_{z,l}$ define the second layer of the hierarchy. We restrict the hierarchical analysis to the parameters of the transcription function, the scale parameter κ and the measurement error standard deviation σ_ϵ, as our parameters of interest. Each parameter is assigned a normal prior distribution, i.e. $\psi_{z,l,u} \sim N(\alpha_{l,u}, \beta_{l,u})$, and we assume $\mathcal{N}(0, 10^4)$ conjugate independent hyperpriors on each hypermean $\alpha_{l,u}$, and $IGa(0.001, 0.001)$ conjugate independent hyperpriors on each hypervariance $\beta_{l,u}$, $l = 1, ..., L$ and $u = 1, ..., 8$, where u denotes the individual parameter component. Note that we adopt the logarithmic parametrisation also for the mean and standard deviation of the delay.

Following Lunn et al. (2013), the analysis is divided into two stages; the first computationally expensive stage is reported in Section 6.2.2 and is aimed at obtaining samples from the posterior distribution

$$\pi(\Psi_{z,l}|y_{z,l}) \propto \pi(y_{z,l}|\Psi_{z,l})\pi(\Psi_{z,l}),$$

which are obtained independently for each experiment z and each location l.

In the second stage, the full target posterior density is given by (Lunn et al., 2013)

$$\pi(\Psi_l, A_l, B_l|y_l) \propto \pi(A_l)\pi(B_l)\prod_{z=1}^{Z}\pi(y_{z,l}|\Psi_{z,l})\pi(\Psi_{z,l}|A_l, B_l).$$

Samples for the parameters and the hyperparameters are obtained by sampling sequentially from the distributions (Lunn et al., 2013)

$$\begin{aligned}
\pi(A_l|\Psi_l, B_l, y_l) &\propto \pi(A_l)\prod_{z=1}^{Z}\pi(\Psi_{z,l}|A_l, B_l)\\
\pi(B_l|\Psi_l, A_l, y_l) &\propto \pi(B_l)\prod_{z=1}^{Z}\pi(\Psi_{z,l}|A_l, B_l)\\
\pi(\Psi_l|A_l, B_l, y_l) &\propto \prod_{z=1}^{Z}\pi(y_{z,l}|\Psi_{z,l})\pi(\Psi_{z,l}|A_l, B_l). \qquad (6.1)
\end{aligned}$$

Lunn et al. (2013) propose to employ the first stage samples of $\Psi_{z,l}$ as proposals in a Metropolis-Hastings scheme to sample from each $\pi(\Psi_{z,l}|, A_l, B_l, y_l)$ in Equation 6.1. This proposal has the major advantage that computationally expensive evaluations of the likelihood are no longer required. Indeed, the acceptance ratio for the Metropolis-Hastings step simplifies into

$$\alpha_{HI} = \min\left\{1, \frac{\pi(\Psi^*_{z,l}|A_l, B_l)}{\pi(\Psi_{z,l}|A_l, B_l)}\frac{\pi(\Psi_{z,l})}{\pi(\Psi^*_{z,l})}\right\}, \qquad (6.2)$$

where $\Psi^*_{z,l}$ is the proposed sample, drawn uniformly at random from the stage one samples. Note that, since we have transformed the samples for the mean and standard deviation of the delay by applying the logarithm, a density transformation has to be performed on the prior densities of stage 1 in Equation 6.2.

In order to define clusters of 'equivalent' locations from the three experiments, a first difficulty is represented by the fact that images have different sizes in each experiment. To make locations in the SCN comparable across the three

experiments, we first perform a change of coordinates, by setting the origin of the axes at the centre of mass, i.e. light intensity, of each experiment, at a selected time. In particular, we employ the times selected in 6.2.1, corresponding to approximately the end of the first quarter of the first circadian cycle.
We then obtain clusters of locations which satisfy the following criteria:

1. each cluster contains exactly one location from each experiment;

2. each cluster contains locations having the minimum euclidean distance between each other, among the available locations;

3. if the same location of a given experiment is selected more than once, the cluster with the overall minimum euclidean distance is retained.

The procedure identifies 93 clusters in total. There is a minimal loss in information, given by the fact that not all the analysed locations in the first stage of the analysis, are retained in the second stage. However, the locations lost correspond mostly to peripheral locations, with low observed circadian rhythmicity, and relatively uninformative results in the first stage of the analysis. To avoid this loss, one could alternatively define a grid on the image having the biggest size, and then form clusters according to the locations belonging to each box. If there is significant 'between experiment' variability, however, the inferred hypermeans and hypervariances may be influenced mostly by the experiment participating with the highest number of locations.

Figure 6.9 shows medians and dispersion of the transcription function parameters, κ and σ_ϵ. The background half-SCN image is obtained by superimposition of the half-SCN background images from the three experiments, and dots are plotted at the centre of each cluster of locations. Values are obtained by drawing samples from the parameter hierarchical prior distributions, for a thinned set of MCMC samples of the corresponding hypermeans and hypervariances. In practice, we draw samples from the hierarchical distributions of each location, which should summarise location-specific distributions by properly merging the results from the three experimental replicates. The usual dot size and colour interpretation is adopted. We notice that a smoother picture of the parameter variation across the SCN is obtained. In particular, the overall picture seems to suggest that while κR_0, κK_{pc}, κ tend to follow the variation of amplitudes shown in Figure 6.4, the remaining parameter may account for the variation of phases, and in particular we recover the bow-like shape of Figure 6.5, particularly in the variation of the mean and standard deviation of the delay.

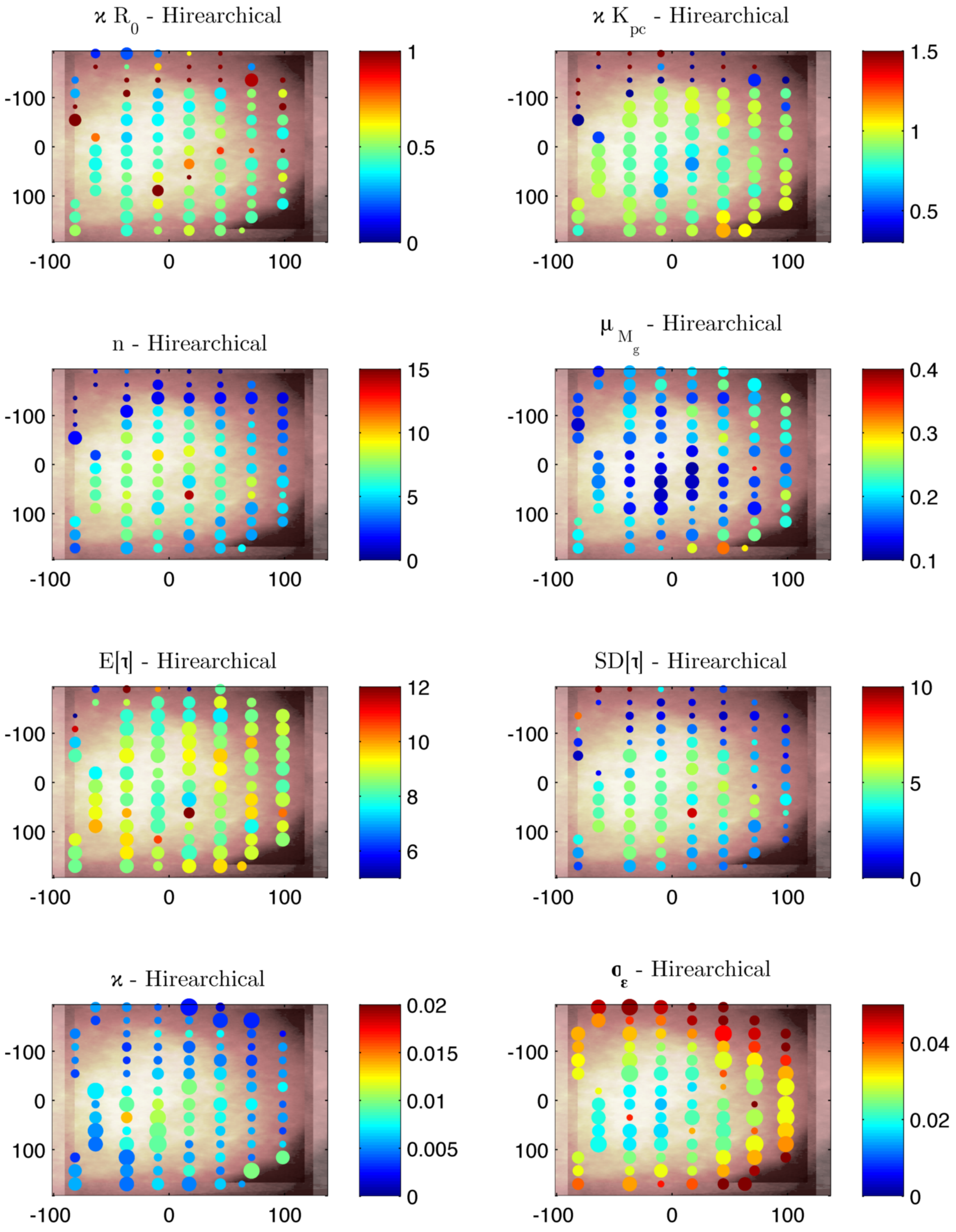


Figure 6.9: Hierachical medians of κR_0, κK_{pc}, n, μ_{M_g}, $E[\tau]$, $SD[\tau]$, κ and σ_ϵ. Samples are transformed in the original scale to compute the interquartile coefficient of variation. The colour scale proportional to the median of the transformed parameters samples, the size inversely proportional to their coefficient of quartile variation. Dots are plotted at the centre of each cluster of locations from the three experiments. Scales of colour and size may vary between different parameters for plotting purposes. Data from Hastings lab. at MRC, Cambridge.

One can also consider a third layer, represented by the overall SCN. This raises however two issues: first, the choice of an appropriate hierarchy, and second, closely linked to the first, the relevance of the results for the aim of our analysis. The first problem can be seen, in some broad sense, as the 'egg-or-chicken' problem, as one could also argue that the overall SCN represents the lower layer of the hierarchy, while the experiment represents the higher one. On the other hand, recall that our aim is to investigate the spatial distribution of the parameters across the SCN; in this sense, a hypermean and hypervariance for each location across the three experiments, provides a more informative quantity than a hypermean and hypervariance for each experiment across the overall SCN. This consideration brings us to the second issue, as, once location specific hypermeans and hypervariances are obtained, we still wish to investigate the degree and type of spatial variation of these quantities across the SCN. Wikle et al. (1998) advise to resort to 'proper' spatial modelling of space-varying parameters when a strong spatial structure is present. We therefore investigate the degree of spatial variation and correlation by means of a preliminary spatial analysis of the location-specific meta-analytic parameters distributions. This is the focus of Section 6.4.

6.4 Towards a spatial model?

The final aim of the present work is to investigate the hypothesis of spatial variation of the transcription function parameters across the SCN. To define spatial variation we refer to the first law of geography, as stated by Tobler (1970): *'everything is related to everything else, but near things are more related than distant things'*. We resort for this purpose to an exploratory analysis, in which we first aim at identifying a possible trend in the mean of the parameters across the SCN, and, secondly, possible residual spatial correlation in the de-trended data.

Our observations consist of draws from the hierarchical prior distribution of the parameters of each location, i.e. we draw samples for each parameter u, $u = 1, ..., 8$, and each location l, $l = 1, ..., L$, from a $\mathcal{N}(a_{u,l}, b_{u,l})$, where $a_{u,l}$ and $b_{u,l}$ represent a thinned set of MCMC samples from the meta-analysis of Section 6.3.

For each parameter u, we assume that $\Psi_u = (\psi_{1,u}, ..., \psi_{L,u})$ is a spatial process at locations $\{1, ..., L\}$. We drop the dependence on u from now on, for ease of notation. Most of the available spatial exploratory and modelling approaches assume some form of stationarity. In particular, a process Ψ defined on a space Q is second-order stationary if the mean of the process is constant and the correlation between any two locations depends only on their distance and is invariant to translation, i.e. $\forall q \in \text{Q E}[\Psi_q] = \mu$ and $\forall q, t, h \in Q : c(q + h, t + h) = \text{Cov}(\Psi_{q+h}, \Psi_{t+h}) = C(q - t)$ (Gaetan and Guoyon, 2010).

Geostatistical approaches generally assume second-order stationarity, aim at modelling spatial dependence by fitting a suitable functional form to the correlation function $\chi(h) = C(h)/C(0)$, and are defined on a continuous space (Gaetan and Guoyon, 2010); $\chi(h)$ is usually dependent on unknown parameters, which are to be estimated (see e.g. Diggle, Moyeed and Tawn, 1998). On the other hand, if a discrete set of locations is assumed, alternative approaches include autoregressive models, e.g. spatial autoregressive (SAR), or, more broadly, Markov random fields (Gaetan and Guyon, 2010). Weaker stationarity assumptions are more frequently made in autoregressive and Markov random field models, i.e. only mean stationarity may be required.

We first focus on exploring the presence of a mean trend, i.e. the condition for first-order stationarity. We assess the presence of a trend with respect to the spatial coordinates, by regressing our parameter samples against the parametric equation of an ellipse, to investigate a trend which moves from the central area of the SCN towards more peripheral locations, possibly in a specific angular direction. A regression in which the spatial coordinates act as covariates is known as trend surface analysis (see e.g. Agterberg, 1984).

We then explore the regression residuals by means of Moran's I (Moran, 1950) and Geary's c (Geary, 1954). These two quantities are more related to local autoregressive models than geostatistical approaches, as they resemble, respectively, the autocorrelation coefficient and the Durbin-Watson statistics found in the time-series literature (Waller and Gotway, 2004). We note that, however, these methods are only adopted here as an exploratory approach to investigate possible residual spatial correlation, and do not necessarily imply a modelling choice in favour of an autoregressive model.

The fitted regression model is the parametric equation of an ellipse, where we define x_l and y_l as the x and y coordinates of the pixels locations at which the hierarchical model is estimated, assuming the origin of the axes to be at the centre of mass of the SCN. Namely, the model is

$$\psi_l = \zeta_0 + \zeta_1 x_l^2 + \zeta_2 y_l^2 + \zeta_3 x_l + \zeta_4 y_l + \zeta_5 x_l y_l + \epsilon_l \quad \epsilon_l \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{6.3}$$

and is fitted to the sample of parameter values. In order to avoid the influence of extreme values and outliers, we resort to robust regression implemented in the *MASS* package in R (Ripley et al., 2013), which down-weights very extreme observations by so-called Huber weights. Median coefficient estimates and HPDIs are provided in Table 6.1.

We can see that the parameters $\log(n)$, $\log(SD[\tau])$, $\log(\mu_{M_g})$, $\log(\kappa)$ and $\log(\sigma_\epsilon)$ all show evidence of a spatial trend, according to the assumed model. The result seems to suggest that the SCN location has a significant influence on these parameters and, in particular, a significant decreasing spatial trend from central to peripheral locations is seen for $\log(n)$, while an increasing spatial trend in the same directions is seen for $\log(\sigma_\epsilon)$, as the significance and sign of both ζ_1 and ζ_2 indicates. Moreover, significance of either ζ_3 or ζ_4 indicates a shift in the origin of the spatial trend with respect to the centre of mass, along the x or the y axis, respectively. Such shift is observed for $\log(\mu_{M_g})$, and $\log(\sigma_\epsilon)$ along the x axis, and for $\log(n)$ and $\log(SD[\tau])$ along the y axis. Finally, significance of ζ_2, and not of ζ_1, indicates a spatial trend which moves from from the origin of the spatial trend towards more external regions only along the y axis. We observe this behaviour for $\log(\mu_{M_g})$, $\log(\kappa)$ and $\log(SD[\tau])$.

We then apply Moran's I and Geary's c on the regression residuals. Let W be a weight matrix with elements $w_{i,j}$, $i = 1, ..., L$ and $j = 1, ..., L$, which define the hypothesised range and strength of spatial connection between any two locations i and j. We return on the definition of W later in this section. Moran's I is defined, in our notation, as

$$I = \frac{L}{\sum_{i=1}^{L}\sum_{j=1}^{L} w_{i,j}} \frac{\sum_{i=1}^{L}\sum_{j=1}^{L} w_{i,j}(\epsilon_i - \bar{\epsilon})(\epsilon_j - \bar{\epsilon})}{\sum_{i=1}^{L}(\epsilon_i - \bar{\epsilon})^2},$$

while Geary's c has the form

$$c = \frac{L-1}{2\sum_{i=1}^{L}\sum_{j=1}^{L} w_{i,j}} \frac{\sum_{i=1}^{L}\sum_{j=1}^{L} w_{i,j}[(\epsilon_i - \bar{\epsilon}) - (\epsilon_j - \bar{\epsilon})]^2}{\sum_{i=1}^{L}(\epsilon_i - \bar{\epsilon})^2}.$$

Under the null hypothesis of no spatial association, Moran's I has expected value equal to $-1/(n-1)$, while Geary's c has expected value equal to 1 (Cliff and Ord, 1973). Departures from these values indicate the presence of positive spatial association (for $m > -1(n-1)$ and $c < 1$), or negative spatial association (for $m < -1(n-1)$ and $c > 1$). Both measures assess the degree of similarity between close units, although Geary's c is more influenced by local autocorrelation (Viton, 2010)

We now discuss the choice of the weight matrix W. Each element $w_{i,j}$ of W represents the strength of spatial connection between locations i and j. According to Getis (2009), W was originally defined based on a neighbourhood criterion: in the geographical context, for example, its elements can be assumed equal to 1 when two units (countries) share a border, and zero otherwise. The diagonal is set to zero

	$\log(\kappa R_0)$	$\log(\kappa K_{pc})$
ζ_0	$-0.7[-0.86, -0.51]$	$-0.23[-0.43, -3.49 \times 10^{-2}]$
ζ_1	$-5.44 \times 10^{-6}[-4.37 \times 10^{-5}, 4.03 \times 10^{-5}]$	$3.33 \times 10^{-7}[-7.53 \times 10^{-5}, 7.99 \times 10^{-5}]$
ζ_2	$7.88 \times 10^{-7}[-1.67 \times 10^{-5}, 1.72 \times 10^{-5}]$	$5.09 \times 10^{-6}[-3.47 \times 10^{-5}, 3.76 \times 10^{-5}]$
ζ_3	$1.66 \times 10^{-3}[-9.72 \times 10^{-4}, 4.20 \times 10^{-3}]$	$-2.00 \times 10^{-5}[-8.15 \times 10^{-3}, 5.85 \times 10^{-3}]$
ζ_4	$-2.08 \times 10^{-4}[-1.94 \times 10^{-3}, 1.32 \times 10^{-3}]$	$4.22 \times 10^{-4}[-3.34 \times 10^{-3}, 5.90 \times 10^{-3}]$
ζ_5	$-1.97 \times 10^{-5}[-4.58 \times 10^{-5}, 5.67 \times 10^{-6}]$	$1.03 \times 10^{-5}[-4.98 \times 10^{-5}, 8.49 \times 10^{-5}]$

	$\log(n)$	$\log(\mu_{M_g})$
ζ_0	$2.08[1.91, 2.26]$	$-1.93[-2.04, -1.79]$
ζ_1	$-8.89 \times 10^{-5}[-1.38 \times 10^{-4}, -3.95 \times 10^{-5}]$	$2.10 \times 10^{-5}[-3.87 \times 10^{-6}, 4.42 \times 10^{-5}]$
ζ_2	$-4.28 \times 10^{-5}[-5.99 \times 10^{-5}, -2.56 \times 10^{-5}]$	$1.12 \times 10^{-5}[3.10 \times 10^{-6}, 1.75 \times 10^{-5}]$
ζ_3	$2.88 \times 10^{-3}[-7.71 \times 10^{-4}, 6.50 \times 10^{-3}]$	$1.25 \times 10^{-3}[6.26 \times 10^{-5}, 2.75 \times 10^{-3}]$
ζ_4	$4.54 \times 10^{-3}[2.58 \times 10^{-3}, 6.48 \times 10^{-3}]$	$2.62 \times 10^{-4}[-4.10 \times 10^{-4}, 1.08 \times 10^{-3}]$
ζ_5	$-2.94 \times 10^{-5}[-7.08 \times 10^{-5}, 2.15 \times 10^{-6}]$	$-5.18 \times 10^{-6}[-1.92 \times 10^{-5}, 6.89 \times 10^{-6}]$

	$\log(\mathrm{E}[\tau])$	$\log(SD[\tau])$
ζ_0	$2.17[2.08, 2.27]$	$1.48[1.28, 1.67]$
ζ_1	$-3.50 \times 10^{-6}[-2.33 \times 10^{-5}, 1.61 \times 10^{-5}]$	$-4.68 \times 10^{-5}[-1.03 \times 10^{-4}, 1.91 \times 10^{-5}]$
ζ_2	$-6.90 \times 10^{-7}[-6.67 \times 10^{-6}, 6.46 \times 10^{-6}]$	$-2.77 \times 10^{-5}[-4.45 \times 10^{-5}, -1.07 \times 10^{-5}]$
ζ_3	$3.94 \times 10^{-4}[-8.78 \times 10^{-4}, 1.57 \times 10^{-3}]$	$-9.93 \times 10^{-4}[-4.12 \times 10^{-3}, 2.47 \times 10^{-3}]$
ζ_4	$2.78 \times 10^{-4}[-3.95 \times 10^{-4}, 9.65 \times 10^{-4}]$	$2.90 \times 10^{-3}[9.76 \times 10^{-4}, 4.39 \times 10^{-3}]$
ζ_5	$-2.44 \times 10^{-6}[-1.54 \times 10^{-5}, 9.77 \times 10^{-6}]$	$3.91 \times 10^{-6}[-3.19 \times 10^{-5}, 4.14 \times 10^{-5}]$

	$\log(\kappa)$	$\log(\sigma_\epsilon)$
ζ_0	$-4.88[-5.15, -4.59]$	$-3.97[-4.14, -3.80]$
ζ_1	$-3.51 \times 10^{-5}[-1.13 \times 10^{-4}, 2.70 \times 10^{-5}]$	$3.69 \times 10^{-5}[3.18 \times 10^{-6}, 6.40 \times 10^{-5}]$
ζ_2	$-1.54 \times 10^{-5}[-3.09 \times 10^{-5}, -5.82 \times 10^{-7}]$	$2.69 \times 10^{-5}[1.90 \times 10^{-5}, 3.64 \times 10^{-5}]$
ζ_3	$2.08 \times 10^{-4}[-2.72 \times 10^{-3}, 4.09 \times 10^{-3}]$	$2.20 \times 10^{-3}[6.40 \times 10^{-4}, 4.08 \times 10^{-3}]$
ζ_4	$1.24 \times 10^{-3}[-3.67 \times 10^{-4}, 2.83 \times 10^{-3}]$	$-5.93 \times 10^{-4}[-1.48 \times 10^{-3}, 1.69 \times 10^{-4}]$
ζ_5	$1.50 \times 10^{-5}[-1.65 \times 10^{-5}, 4.71 \times 10^{-5}]$	$8.36 \times 10^{-7}[-1.50 \times 10^{-5}, 1.73 \times 10^{-5}]$

Table 6.1: Estimates of the coefficients of the regression model 6.3, applied to the samples from the hierarchical prior distribution of the parameters $\log(\kappa R_0)$, $\log(\kappa K_{pc})$, $\log(n)$, $\log(\mu_{M_g})$, $\log(E[\tau])$, $\log(SD[\tau])$, $\log(\kappa)$ and $\log(\sigma_\epsilon)$. Median estimates, with 95 % HPDIs in brackets. Robust regression with Huber weights, *rlm* function in the R package MASS. We highlight in grey parameters whose 95 % HPDIs do not include 0.

by definition. This approach was generalised in Cliff and Ord (1960), by providing the general form of Moran's I for any choice of the weight matrix W. Here we assume the weights to be inversely proportional to the euclidean distance between the pixels locations, for different maximum distances, after which the weight is set to zero, partially following Cliff and Ord (1960), later reproduced in Bivand (2015). We perform the calculation of I and c using the *spdep* R package (Bivand, 2015). Tables 6.2 and 6.3 provide the median estimates of I and c, together with 95%

HPDIs, for the transcription function, measurement error standard deviation and scale parameters. Note that for interpretability purposes we report the standardised coefficient (with sign reversed for Geary's c as in the R function *geary.test*).

	$\log(\kappa R_0)$	$\log(\kappa K_{pc})$	$\log(n)$
$Dist_1$	$0.43[-2.20, 3.21]$	$-7.08 \times 10^{-3}[-4.32, 4.47]$	$0.33[-5.33, 5.37]$
$Dist_2$	$0.54[-2.69, 2.84]$	$-0.16[-4.78, 3.93]$	$7.95 \times 10^{-2}[-4.94, 4.25]$
$Dist_3$	$0.39[-2.08, 3.07]$	$-2.12 \times 10^{-2}[-3.95, 4.16]$	$-0.3[-3.82, 3.67]$

	$\log(\mu_{M_g})$	$\log(\mathrm{E}[\tau])$	$\log(SD[\tau])$
$Dist_1$	$8.29 \times 10^{-2}[-1.97, 1.90]$	$-0.34[-5.20, 4.15]$	$0.46[-1.62, 2.77]$
$Dist_2$	$-0.23[-2.04, 1.91]$	$-0.60[-3.64, 2.99]$	$0.24[-1.85, 2.54]$
$Dist_3$	$-0.37[-1.84, 1.69]$	$-0.79[-3.39, 2.54]$	$-7.07 \times 10^{-3}[-2.01, 1.92]$

	$\log(\kappa)$	$\log(\sigma_\epsilon)$	
$Dist_1$	$-0.13[-2.45, 1.82]$	$-0.18[-2.02, 1.42]$	
$Dist_2$	$-0.36[-2.18, 1.49]$	$-0.38[-1.85, 1.24]$	
$Dist_1$	$-0.54[-2.11, 1.15]$	$-0.52[-1.69, 1.19]$	

Table 6.2: Moran's I statistics (Moran's I standard deviate) applied to the samples from the hierarchical prior distribution of the parameters $\log(\kappa R_0)$, $\log(\kappa K_{pc})$, $\log(n)$, $\log(\mu_{M_g})$, $\log(E[\tau])$, $\log(SD[\tau])$, $\log(\kappa)$ and $\log(\sigma_\epsilon)$, after removal of the mean trend and for maximum distance $Dist_1 = 27$ pixels, $Dist_2 = 54$ pixels and $Dist_3 = 81$ pixels. Weights inversely proportional to distance. Median estimates, with 95 % HPDIs in brackets.

Neither Moran's I nor Geary's c provide any evidence against the null hypothesis of no spatial correlation. It has to be noted, however, that the absence of correlation does not imply independence, as nonlinear effects may be taking place.

The overall results of this preliminary analysis seem therefore to suggest the presence of a mean trend, with no residual spatial correlation. The mean trend is observed especially for $\log(n)$, $\log(SD[\tau])$, $\log(\mu_{M_g})$, $\log(\kappa)$ and $\log(\sigma)$, and confirms the visual impression of a central region with higher responsiveness of the promoter, higher intrinsic noise and variability in the delays, and lower degradation rate, and a peripheral region with a larger measurement error standard deviation.

Given the predominant role of the mean trend, the implementation of a hierarchical spatial model represents a possible extension of this analysis: a third layer can indeed be defined, e.g. by defining a mean hyper-hyperparameter of the form fitted by the linear regression, or by exploring alternative and more detailed models. Such an extension is however beyond the scope of the present work.

	$\log(\kappa R_0)$	$\log(\kappa K_{pc})$	$\log(n)$
$Dist_1$	$0.27[-1.69, 2.15]$	$-7.48 \times 10^{-2}[-3.04, 3.45]$	$0.25[-2.97, 4.50]$
$Dist_2$	$1.20 \times 10^{-2}[-1.86, 2.38]$	$0.53[-2.62, 3.88]$	$0.81[-2.71, 4.43]$
$Dist_3$	$5.26 \times 10^{-2}[-1.47, 2.00]$	$0.97[-1.54, 3.56]$	$1.04[-1.27, 3.35]$

	$\log(\mu_{M_g})$	$\log(\mathrm{E}[\tau])$	$\log(SD[\tau])$
$Dist_1$	$1.66 \times 10^{-2}[-1.36, 2.42]$	$0.24[-2.64, 3.71]$	$0.7[-1.57, 2.92]$
$Dist_2$	$-0.25[-2.10, 1.77]$	$0.66[-2.18, 2.93]$	$0.75[-1.46, 2.77]$
$Dist_3$	$-0.38[-1.87, 1.51]$	$0.58[-1.29, 2.23]$	$0.8[-1.30, 2.54]$

	$\log(\kappa)$	$\log(\sigma_\epsilon)$	
$Dist_1$	$0.34[-2.05, 1.95]$	$-0.29[-1.83, 1.41]$	
$Dist_2$	$0.2[-1.92, 2.30]$	$-0.48[-1.94, 1.11]$	
$Dist_3$	$0.27[-1.65, 2.13]$	$-0.61[-1.91, 0.97]$	

Table 6.3: Geary's c statistics (Geary's c standard deviate, with sign reversed as in the R function *geary.test*, so that is has the same direction as Moran's I) applied to the samples from the hierarchical prior distribution of the parameters $\log(\kappa R_0)$, $\log(\kappa K_{pc})$, $\log(n)$, $\log(\mu_{M_g})$, $\log(E[\tau])$, $\log(SD[\tau])$, $\log(\kappa)$ and $\log(\sigma_\epsilon)$, after removal of the mean trend and for maximum distance $Dist_1 = 27$ pixels, $Dist_2 = 54$ pixels and $Dist_3 = 81$ pixels. Weights inversely proportional to distance. Median estimates, with 95 % HPDIs in brackets.

6.5 Discussion

In the last part of this work, we have applied a negative feedback loop model to $Cry1$-*luc* observed spatio-temporal data in mice SCN. The feedback loop is modelled by means of a nonlinear stochastic model for the dynamical evolution of $Cry1$ mRNA, and comprises a distributed delay. Inference is performed on the parameters of the model by linearising the nonlinear functions involved, with a methodology that extends the extended Kalman-Bucy filter to be applicable to systems incorporating distributed delays.

Parameter inference has been performed at both a single-experiment level, and by pooling the results of three independent experimental replicates by adopting a hierarchical Bayesian meta-analytic approach. The results have revealed a significant mean spatial trend of the parameter estimates, and in particular suggest the presence of a central SCN region with higher promoter responsiveness, higher intrinsic noise, lower degradation rate and a higher variability in the distribution of the delay. On the other hand, some care is required in the interpretation of the results, as model fit diagnostics reveal a significant residual correlation at period 12 hour across the SCN, and a 24 hour residual correlation particularly in the central region. Moreover, the degradation estimates tend to assume much lower mean values than those measured in Yamaguchi et al. (2003). Model approximation is the most likely source of both the residual correlation and the low estimate of the degradation rate, and in particular we may be missing additional processes acting on the PER/CRY protein complex, e.g. inter-cellular signalling, as the 24 hour residual correlation in the central SCN can suggest. An additional approximation of the underlying dynamics is performed by assuming the reporter protein light intensities as proxies for $Cry1$ mRNA. An extension of the model, or alternative models of transcriptional regulation can be explored, e.g the model of Kim et al. (2014) does not resort on Hill-type input-output functions, but rather derives the transcription function by assuming that repression is achieved through sequestration of the activator CLOCK/BMAL complex by the PER/CRY complex. The authors also point out the relevance of this alternative formulation of the transcription function in order to achieve synchrony of Per periods across cells, as a consequence of extra-cellular signalling, as well as in order to describe the underlying dynamics possibly more realistically. The model of Ananthasubramaniam et al. (2014) reviewed in section 5.2.2, on the other hand, also implements cell synchronisation by means of extra-cellular VIP, through Hill-type equations. A comparison between our results and those obtained by assuming a transcription function of the form proposed by Kim et al. (2014), can definitely be of interest.

To conclude, this study has investigated spatial differences of the transcriptional dynamics of $Cry1$, as reflected by the inferred model parameter estimates. It suggests that the central SCN region exhibits of a higher responsiveness of the promoter, a lower degradation rate, as well as a higher intrinsic noise and variability in the distribution accounting for the feedback loop delay. This result is of particular interest, as it highlights the importance of intrinsic noise, and therefore of stochastic modelling of such transcriptional processes, and may have a biological explanation in observing that the core SCN, which partially overlaps with our 'central' locations, receives the light impulses coming the retina; Steuer et al. (2003) shows, indeed, that, in an input-output relationship, responsiveness of the output amplitudes to the input ones, is higher in stochastic than in deterministic systems.

Conclusions

In this work we have studied different scenarios of stochastic transcriptional regulation, arising from the binding of transcription factor proteins to the promoter of a putative regulated child gene. Our work extends previous modelling in this area, and in particular Tkačik and Walczak (2011), and Nachman et al. (2004), by assuming the presence of two transcription factors as regulators of the child gene, and by developing our modelling approach from an exact stochastic description of the system, which also incorporates the binding and unbinding reactions of the two TFs to the promoter, as well as their binding cooperativity. In order to ensure practical parameter identifiability, we have applied available approximation approaches to the system under study, and we have checked their accuracy through simulation. Additionally, we have investigated the effect of data aggregation across different cells, for our given set of reactions and parameter values, and outlined the condition required for achieving meaningful parameter inference in this scenario. In particular, while zero-th and first order reactions do not pose significant problems in this respect, due to the linearity of the hazards, if second order reactions are included in the system care should be taken. In order to be able to 'safely' perform inference on the aggregated samples, at least one of the two reactants should be at approximately the same level in each cell. Such condition is more easily satisfied when, assuming independent and identically distributed processes for the time-evolution of the species in each cell, high molecule numbers can be assumed for one of the two species involved in a second order reaction. In our application, the transcription factor proteins participate in a second order reaction when binding the promoter, and the aggregation assumption implies assuming similar protein, and consequently mRNA, levels for each TF, but the same assumption is not required for the promoter state (which we treat as chemical species). We have found, in our specific application, that inference on aggregated values may be safely performed when average molecule numbers for the TFs are as low as 15-20 in each cell, although a minimal mismatch in the mean and variability of the child mRNA simulated trajectory has been observed in one simulation scenario.

We have then outlined an approximate model for transcriptional regulation, by additionally assuming promoter equilibrium. Effectively, the approximation results in a birth and death process for the child mRNA, which is then formulated as a state-space model. Traditional filtering methodologies, adopted to estimate the likelihood in this modelling framework, generally assume that the same unit is observed over time. However, biological experiments often imply destructive sampling. This characteristic has profound consequences on the correlation structure of the samples, and specifically it implies independence (Stathopoulos and Girolami, 2013). As a consequence, we have shown that in this framework performing smoothing is equivalent to performing filtering at each time point, as other observations, close or distant in time, do not add any information about the underlying process. Note that smoothing is performed conditionally on the parameters, which are therefore assumed to be known or set to a fixed value, and which fully determine the mean and variance evolution of the unobserved states.

By performing inference for this regulatory framework, we have observed that it is possible to retrieve all the parameters involved in an approximate model which assumes that the two TFs are observed. The model is fitted to the original SSA simulations of the full network. However, minor biases in the parameter estimates are observed, we believe due to the strong correlation structure between the parameters themselves. We have also observed that successful inference requires that both the TFs are dynamically influencing the child gene transcription, as well as that observations are available at a high frequency, and exhibit a high signal to noise ratio.

The availability of experimental data concerning circadian genes in both plant and mammals has motivated variations of our first proposed approach for transcriptional regulation. For the plant circadian data, we have considered in particular a scenario which assumes that one important transcription factor for the putative child gene has not been observed, and we have fitted such a model to the available rhythmic mRNA expression levels (Carré lab.). The data analysis so performed has allowed to investigate in more depth the relevance of a known important circadian regulator in the *Arabidopsis Thaliana*, namely LHY. We have observed that, indeed, the inferred unobserved TFs profiles of genes belonging to the circadian clock, are correlated with the observed time-series of LHY protein itself. Moreover, by clustering the inferred unobserved mRNA profiles of the child genes, we have observed a possible association between cluster of expression and presence of known LHY binding sites in their promoters, but not between cluster of expression and LHY transcriptional effect, as measured in an additional induction experiment (Carré lab.). Our analysis has exploited the capabilities of mechanistic state-space models for transcriptional regulation, as its application has resulted not only in the inference of the model parameters, but also in the estimation of a whole distribution for the unobserved TF and mRNA profiles. Such a distribution of profiles has also allowed to easily quantify uncertainty about the output of the posterior correlation and clustering analyses. The main limitation of this first data application is likely to be in the relatively simple model that we have assumed for the transcriptional dynamics, and in the approximate handling of the unobserved TF by means of a truncated Fourier series representation. These choices are motivated by the scarcity of available observations and prior information for each gene. Nevertheless, we have gained some biological insight on the regulatory role of LHY, and in particular our results point in the direction of a complex form of regulation of the child genes, in which LHY is an important factor, but not sufficient to explain the variety of observed phases and profiles. The association between the presence of binding sites and cluster of expression, also suggests that factors binding to the same binding sites as LHY may represent an important piece of the current picture, and further research in this direction may be of interest, e.g. by comparing the inferred unobserved TF profiles with the time-series of additional transcription factor candidates.

Finally, we have investigated a form of transcriptional regulation which arises from an auto-regulatory feedback loop. Feedback loops can be effectively described by introducing a delay in the model. However, the presence of the delay has several implications on the state-space modelling framework that we have considered in the first two parts of this work; most notably, Markovianity can be assumed only over a long time-interval, which 'goes back' up until the assumed maximum delay time. When sampling is not destructive, such as in our available mammalian clock data, the consequence is that filtering requires the update of all the unobserved state estimates in the past having significant weight in the distribution of the delay (a fixed delay simply implies a distribution with all its mass concentrated at one time-point). We have developed a novel filtering methodology for systems which comprise distributed delays, based on extension of the extended Kalman-Bucy filter. We have shown that the computational speed of the procedure makes it applicable in the context of an MCMC algorithm for parameter inference, and checked its empirical coverage and induced univariate parameter likelihood. Moreover, we have taken into account the role of *temporal* aggregation of the samples, and incorporated the aggregation process in the measurement equation. We have observed that explicit modelling of aggregation across time has not a strong effect on most of the model parameters; however, a more significant influence is seen for the measurement error and scale parameters, for which the log-likelihood peak gets visibly closer to the true parameter values as a finer and finer aggregation time-grid is considered. A thinner time grid implies, however, a higher computational cost, as a higher number of unobserved states needs to be estimated. We have therefore performed inference on available $Cry1$-*luc* spatio-temporal imaging data recorded in mice SCN (Hastings lab.), by adopting a delayed-acceptance MCMC algorithm, which takes advantage of a fast likelihood evaluation by assuming a 'rough' time-grid ($\delta_t = 0.5\,\mathrm{h}$), to perform a first selection of the proposed MCMC samples, and finally accepts them according to a slow a more precise likelihood evaluation under the assumption of a thinner time-grid for the unobserved states ($\delta_t = 0.1\,\mathrm{h}$).

Independent runs of the algorithm on three experimental replicates have shown similarities in the parameter estimates, which we have then merged with two-stage a hierarchical meta-analytic Bayesian approach (Lunn et al., 2013). Finally we have investigated the spatial distribution of the hierarchical parameter samples. The spatial analysis has revealed a mean trend which moves approximately from the central area of the SCN, towards more peripheral locations, for the parameters $\log(n)$, $\log(SD[\tau])$, $\log(\mu_{M_g})$, $\log(\kappa)$ and $\log(\sigma)$. No residual spatial association has been observed, as assessed by means of Moran's I and Geary's c. The results highlight the importance of both stochastic modelling, as implied by the spatial distribution of κ, and of accounting for a distributed, rather than for a fixed, delay, as indicated by the significant spatial trend of $\log(SD[\tau])$. Moreover, the parameter estimates may partially reflect known functional differences between the core and dorsal areas of the SCN; as the core area identifies approximately the lower half of the SCN, and partially overlaps with out 'central' locations, we may thus hypothesise that upper core regions are characterised by more stochastic form of regulation, having more distributed delays, and a higher responsiveness of the promoter. However, model fit also suggests care in the parameter interpretation, as a significant 12 hour periodicity is observed in the residuals across the whole SCN section, and a 24 hour periodicity is observed particularly in the central area. Such residual periodicity is possibly induced by processes not explicitly taken into account by our model, such as extra-cellular signalling, or the reporter process. Alternative formulations of the transcription function can also be investigated, such as those proposed in Kim et al. (2014), and a comparison of the results may provide additional insights on the underlying process. Moreover, our analysis of $Cry1$-*luc* data is only a first step into a model which should take advantage of additional CRE-*luc* and, eventually, $Per1$-*luc* and Calcium levels, measured in an analogous spatio-temporal fashion across the SCN (Hastings lab.). Due to the fact that Cry is auto-repressed by PER/CRY, CRE is only activated by the protein CREB, responsive to Calcium itself, and *Per*1 is both repressed by the PER/CRY complex, and activated by CREB/Calcium, the analysis of *Cry*1-*luc*, CRE-*luc* and *Per*1-*luc* would ideally allow to disentangle the contribution of PER/CRY, Calcium, and the possible interaction of PER/CRY and Calcium on transcription, respectively. Such an analysis may contribute to the current understanding of the role of extra-cellular signalling on gene expression in the SCN, by explicitly accounting for intrinsic noise. The explicit modelling of intrinsic stochasticity can be of particular interest, as it has been shown in different studies to play an important role, broadly, in entrainment of signals to an external input (Steuer et al., 2003), and, specifically, in generating circadian cyclicity as a consequence of extra-cellular signalling in the SCN (Ko et al., 2010).

Appendix A

Modelling stochastic transcriptional regulation: additional review material and further details

A.1 Exact transition density for monomolecular reactions systems

A closed form for the transition density of systems involving only zero-th and first order reactions has been derived in Jahnke and Huisinga (2007).
Following the authors, define with $c_{j,i}$ the reaction rate of conversion of one molecule of the j-th species into one molecule of the i-th species, and with $A(t)$ the matrix with elements $a_{j,i}(t)$, where $a_{j,i}(t) = c_{j,i}(t)$ for $i \neq j \geq 1$, while $a_{i,i}(t) = -\sum_{j=0}^{p} c_{j,i}(t)$. Finally, define the vector of birth rates $b(t) = (c_{0,1}(t), ..., c_{0,p}(t))^T$.

> **Theorem 1** (Jahnke and Huisinga, 2007, §3.3) *For a monomolecular system with initial distribution $p(0, \cdot) = \delta_{x_0}(\cdot)$, for some $x_0 \in \mathbb{N}^p$, the probability distribution at time $t > 0$ is given by*
>
> $$p(t, \cdot) = Poi(\cdot, \lambda(t)) * M(\cdot, x_1(0), \pi^{(1)}(t)) * ... * M(\cdot, x_p(0), \pi^{(p)}(t)).$$
>
> *The vectors $\pi^{(i)}(t) \in [0, 1]^p$ and $\lambda(t) \in \mathbb{R}^p$ are the solutions of the reaction rate equations*
>
> $$\dot{\pi}^{(i)}(t) \quad = \quad A(t)\pi^{(i)}(t),$$

$$\dot{\lambda}(t) \quad = \quad A(t)\lambda(t) + b(t),$$

with $\pi^{(i)}(0) = \epsilon_i$, $\lambda(0) = 0$.

The term ϵ_i is equal to 1 if all the molecules at $t = 0$ belong to species i.

A.2 Derivation of the LNA

There are different possible derivations of the LNA (see e.g. Wilkinson, 2012; Komorowski et al., 2009; Stathopoulos and Girolami, 2013; Wallace, 2010; Anderson and Kurtz, 2011), here we follow Anderson and Kurtz (2011). The authors start by defining the quantity

$$P^{n\Omega}(t) = \sqrt{n\Omega}(Z(t) - z(t)),$$

which gives

$$Z(t) = z(t) + \frac{P^{n\Omega}(t)}{\sqrt{n\Omega}}.$$

The aim is to derive an approximation for $P^{n\Omega}(t)$, which, for large $n\Omega$, is normal for all t.

From 1.4 , it follows that (Anderson and Kurtz, 2011)

$$\begin{aligned} P^{n\Omega}(t) &\approx P^{n\Omega}(0) + \sqrt{n\Omega}\left[\frac{1}{n\Omega}SY\left(n\Omega\int_0^t \tilde{h}(Z(s),c)ds\right) - \int_0^t S\tilde{h}(z(s),c)ds\right] \\ &= P^{n\Omega}(0) + \frac{1}{\sqrt{n\Omega}}S\tilde{Y}\left(n\Omega\int_0^t \tilde{h}(Z(s),c)ds\right) \\ &\quad + \int_0^t \sqrt{n\Omega}\left(S\tilde{h}(Z(s),c) - S\tilde{h}_k(z(s),c)\right)ds. \end{aligned}$$

The authors then apply two approximations. The first one is the normal approximation to the Poisson process, i.e.

$$S\frac{1}{\sqrt{n\Omega}}\tilde{Y}\left(n\Omega\int_0^t \tilde{h}(Z(s),c)ds\right) \approx SB\left(\text{diag}\left[\int_0^t \tilde{h}(Z(s),c)ds\right]\right),$$

where B is an r-dimensional Wiener process. The second is the first order Taylor expansion of $\tilde{h}(Z(s),c)$ about the deterministic limit $z(t)$

$$\tilde{h}(Z(s),c) \approx \tilde{h}(z(s),c) + J_{\tilde{h}}(z(s))(Z(s) - z(s)),$$

which effectively eliminates nonlinearities. The process becomes

$$P^{n\Omega}(t) \approx P^{n\Omega}(0) + SB\left(\operatorname{diag}\left[\int_0^t \tilde{h}(Z(s),c)ds\right]\right) + \int_0^t SJ_{\tilde{h}}(z(s))P^{n\Omega}(s)ds,$$

and, as $n\Omega \to \infty$, it follows that (Anderson and Kurtz, 2011)

$$P(t) \approx P(0) + SB\left(\operatorname{diag}\left[\int_0^t \tilde{h}(z(s),c)ds\right]\right) + \int_0^t SJ_{\tilde{h}}(z(s))P(s)ds,$$

which gives the desired result.

A.3 First derivatives of the transcription function

The first partial derivatives of the transcription function of Equation 1.5 have the form

$$\begin{aligned}\frac{d\nu(x_{P_A},x_{P_B})}{dx_{P_A}} \quad = \quad & \frac{\frac{1}{K_A}(R'_A - R'_0) + \frac{1}{K_A}\frac{1}{K_B}x_{P_B}(R'_A - \frac{1}{K_c}R'_0) + \frac{1}{K_A}\frac{1}{K_B^2}\frac{1}{K_c}x_{P_B}^2(R'_{A,B} - R'_B)}{\left(1 + \frac{x_{P_A}}{k_A} + \frac{x_{P_B}}{K_B} + \frac{1}{K_c}\frac{x_{P_A}}{K_A}\frac{x_{P_B}}{K_B}\right)^2} \\ & + \frac{\frac{1}{K_A}\frac{1}{K_B}x_{P_B}(\frac{1}{K_c}R'_{A,B} - R'_B)}{\left(1 + \frac{x_{P_A}}{k_A} + \frac{x_{P_B}}{K_B} + \frac{1}{K_c}\frac{x_{P_A}}{K_A}\frac{x_{P_B}}{K_B}\right)^2},\end{aligned}$$

and

$$\begin{aligned}\frac{d\nu(x_{P_A},x_{P_B})}{dx_{P_B}} \quad = \quad & \frac{\frac{1}{K_B}(R'_B - R'_0) + \frac{1}{K_A}\frac{1}{K_B}x_{P_A}(R'_B - \frac{1}{K_c}R'_0) + \frac{1}{K_A^2}\frac{1}{K_B}\frac{1}{K_c}x_{P_A}^2(R'_{A,B} - R'_A)}{\left(1 + \frac{x_{P_A}}{k_A} + \frac{x_{P_B}}{K_B} + \frac{1}{K_c}\frac{x_{P_A}}{K_A}\frac{x_{P_B}}{K_B}\right)^2} \\ & + \frac{\frac{1}{K_A}\frac{1}{K_B}x_{P_A}(\frac{1}{K_c}R'_{A,B} - R'_A)}{\left(1 + \frac{x_{P_A}}{k_A} + \frac{x_{P_B}}{K_B} + \frac{1}{K_c}\frac{x_{P_A}}{K_A}\frac{x_{P_B}}{K_B}\right)^2}.\end{aligned}$$

Setting $K_c = 1$ gives

$$\frac{d\nu(x_{P_A},x_{P_B})}{dx_{P_A}} = \frac{(\frac{1}{K_A} + \frac{1}{K_A}\frac{x_{P_B}}{K_B})((R'_A - R'_0) + (R'_{A,B} - R'_B)\frac{x_{P_B}}{K_B})}{(1 + \frac{x_{P_A}}{k_A} + \frac{x_{P_B}}{K_B} + \frac{x_{P_A}}{K_A}\frac{x_{P_B}}{K_B})^2},$$

and

$$\frac{d\nu(x_{P_A},x_{P_B})}{dx_{P_B}} = \frac{(\frac{1}{K_B} + \frac{1}{K_B}\frac{x_{P_A}}{K_A})((R'_B - R'_0) + (R'_{A,B} - R'_A)\frac{x_{P_A}}{K_A})}{(1 + \frac{x_{P_A}}{k_A} + \frac{x_{P_B}}{K_B} + \frac{x_{P_A}}{K_A}\frac{x_{P_B}}{K_B})^2}.$$

Appendix B

Arabidopsis thaliana modelling: additional information and modelling tools

B.1 Prior information on the dissociation coefficients: additional details

The average concentration of LHY per cell is computed as follows. The average level of LHY protein is of about 2.14×10^4 molecules per cell, from additional information provided by personal communication with I. Carré. According to Wang (2013), the nuclear area of wild type rosette leaves in the *Arabidopsis thaliana* is about $70\,\mu\mathrm{m}^3$. Therefore, the nuclear volume results about $4.41 \times 10^2\,\mu\mathrm{m}^3$. Following Price et al. (1973), the ratio between the *Arabidopsis thaliana* nuclear and cellular volume is 0.16, leading to a cellular volume of approximately $2753.54\,\mu\mathrm{m}^3$.

Conversion from molecules to moles, leads to $8.06 \times 10^{-23}\,\mathrm{M} \cdot \mu\mathrm{m}^{-3}$ for the average LHY protein level. Finally, given that $1\,\mathrm{l}$ of solution occupies $10^{15}\,\mu\mathrm{m}^3$, it follows that average LHY protein concentration per cell is approximately $1.29 \times 10^{-8}\,\mathrm{M}$.

As a final remark, one main critical point of the dissociation coefficients estimates provided, is that the available values are only based on one experiment, therefore there is no information about their variability.

B.2 Switch tool

The switch tool presented in Jenkins et al. (2013) allows estimating the times at which a gene changes transcription, as well as the transcription rates (in units of the real data) and degradation rate of its mRNA, given its mRNA time series. An underlying dynamical model of the form

$$\frac{dM(t)}{dt} = \nu_i - \mu M(t) \tag{B.1}$$

is assumed, where $M(t)$ denotes the mRNA at time t, μ is the degradation rate, and ν_i, $i = 0, ..., w$ denotes the transcription rate for t between time s_i and s_{i+1}, called switch time. The residuals between the observed data and the solution of the differential equation (B.1) are assumed to be normal with mean 0 and level-dependent standard deviation.

The degradation rate is assumed to be constant, while a switch is denoted by a change in the transcription rate, which could be classified as either 'On', if the transcription rate is higher after the switch, or 'Off', if the transcription rate becomes lower.

A reversible jump MCMC is employed to estimate the switch points. The measurement error variance is estimated with a Gibbs step, the degradation rate with a Metropolis-Hastings step. The transcription rates are obtained via weighted least squares, given all the other parameters in the model. Although the latter does not belong to the traditional framework of Bayesian regression, it has been verified through simulation to provide no significant mismatch in estimation, while highly improving the computational speed of the algorithm.

B.3 Fourier series representation

Here we report Theorem 8.3 in West (1999, Chapter 8), which defines the Fourier series representation of any observed time-series. In particular, the theorem states that

> **Theorem 2** *(West, 1999, Chapter 8) Any sequence of n real numbers $\beta_1, ...\beta_n$ can be written as*
>
> $$\beta_j = a_0 + \sum_{q=1}^{h} H_q(j),$$
>
> *where h is the largest integer not exceeding $n/2$ and the coefficients a_0,*

a_q, b_q, $a_{n/2}$ and $b_{n/2}$ are given by

$$\begin{aligned}
a_0 &= \frac{1}{n}\sum_{j=0}^{n-1}\beta_j, \quad a_{n/2} = \frac{1}{n}\sum_{j=0}^{n-1}(-1)^j\beta_j, \quad b_{n/2} = 0 \\
a_q &= \frac{2}{n}\sum_{j=0}^{n-1}\beta_j\cos(\alpha qj), \quad b_q = \frac{2}{n}\sum_{j=0}^{n-1}\beta_j\sin(\alpha qj) \;\; 1 \le q < n/2 \qquad \text{(B.2)}
\end{aligned}$$

The function

$$H_q(j) = a_q\cos(\alpha qj) + b_q\sin(\alpha qj) = A_q\cos(\alpha qj + \gamma_q),$$

is the so-called q-th harmonic, evaluated at j. The coefficients $A_q = \sqrt{a_q^2 + b_q^2}$ and $\gamma_q = \arctan(-b_q/a_q)$, represent, respectively, its half-amplitude and phase. Finally, the quantity $\alpha = 2\pi/n$ is the frequency of the first harmonic. Essentially, the original series is decomposed in h sinusoidal functions, the q-th sinusoidal having frequency $\alpha q = 2\pi q/n$.

Appendix C

Arabidopsis Thaliana simulation study: additional plots

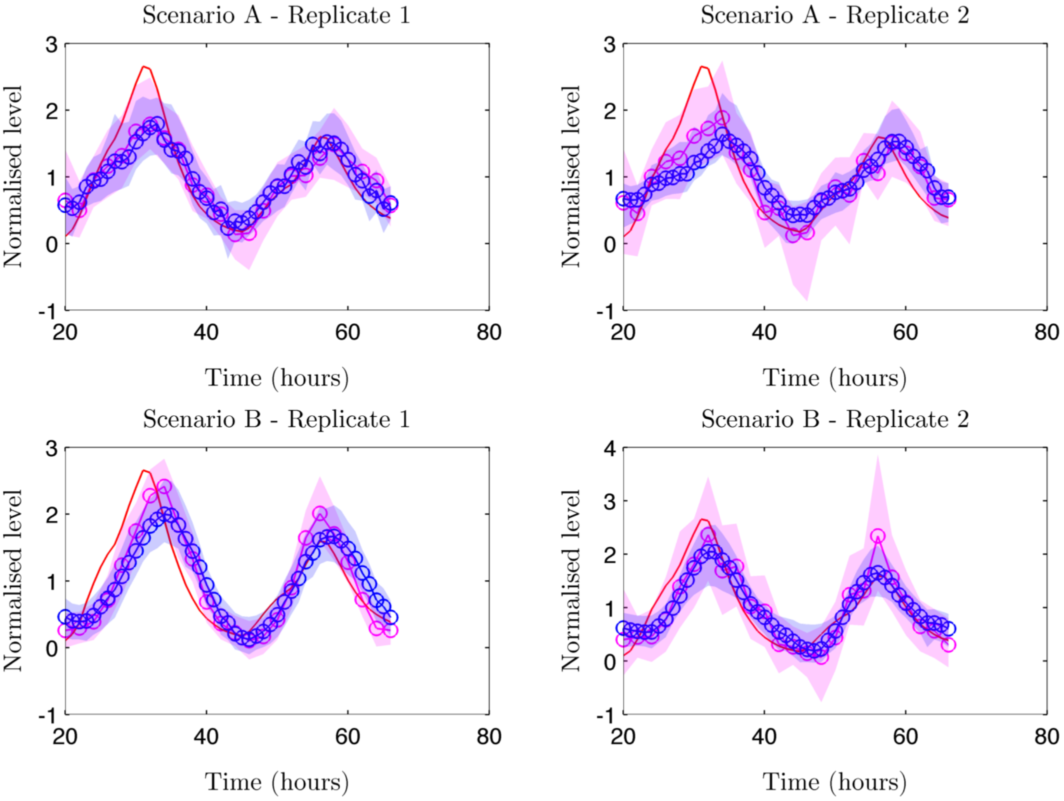


Figure C.1: Unobserved TF inference (smoothing): posterior median and 95 % HPDIs. Estimates obtained for $\Delta_t = 1\,\mathrm{h}$ (blue) and $\Delta_t = 2\,\mathrm{h}$ (magenta). True simulated TF B (from the diffusion approximation) superimposed in red. Model 3.6, as applied to diffusion approximation data simulated according to the parameters of Figure 3.5. MCMC samples for two replicates of scenario A (top), and two replicates of scenario B (bottom).

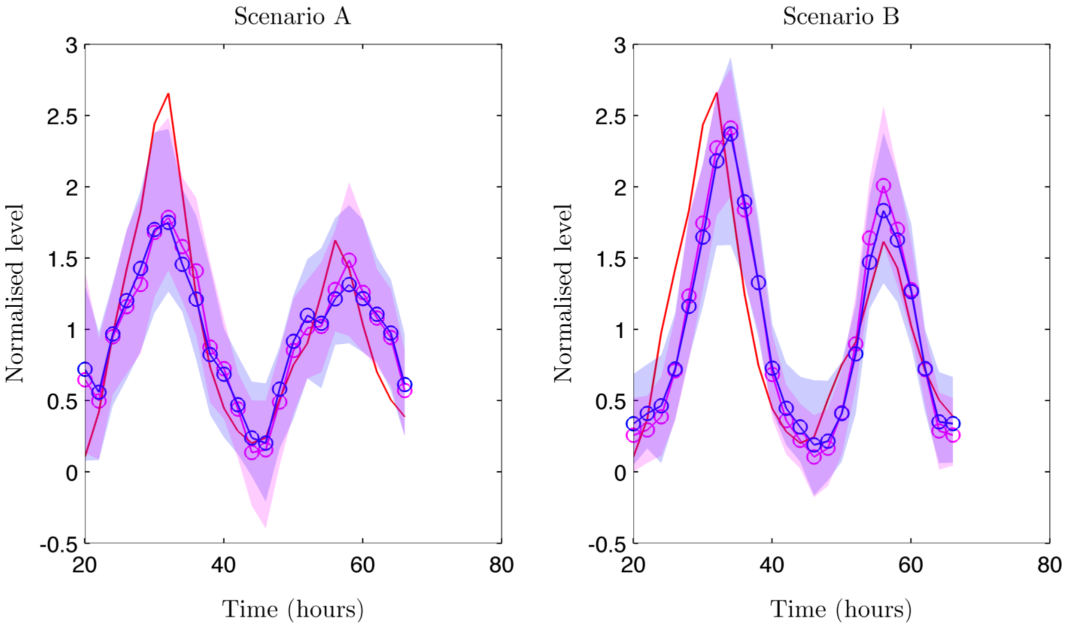
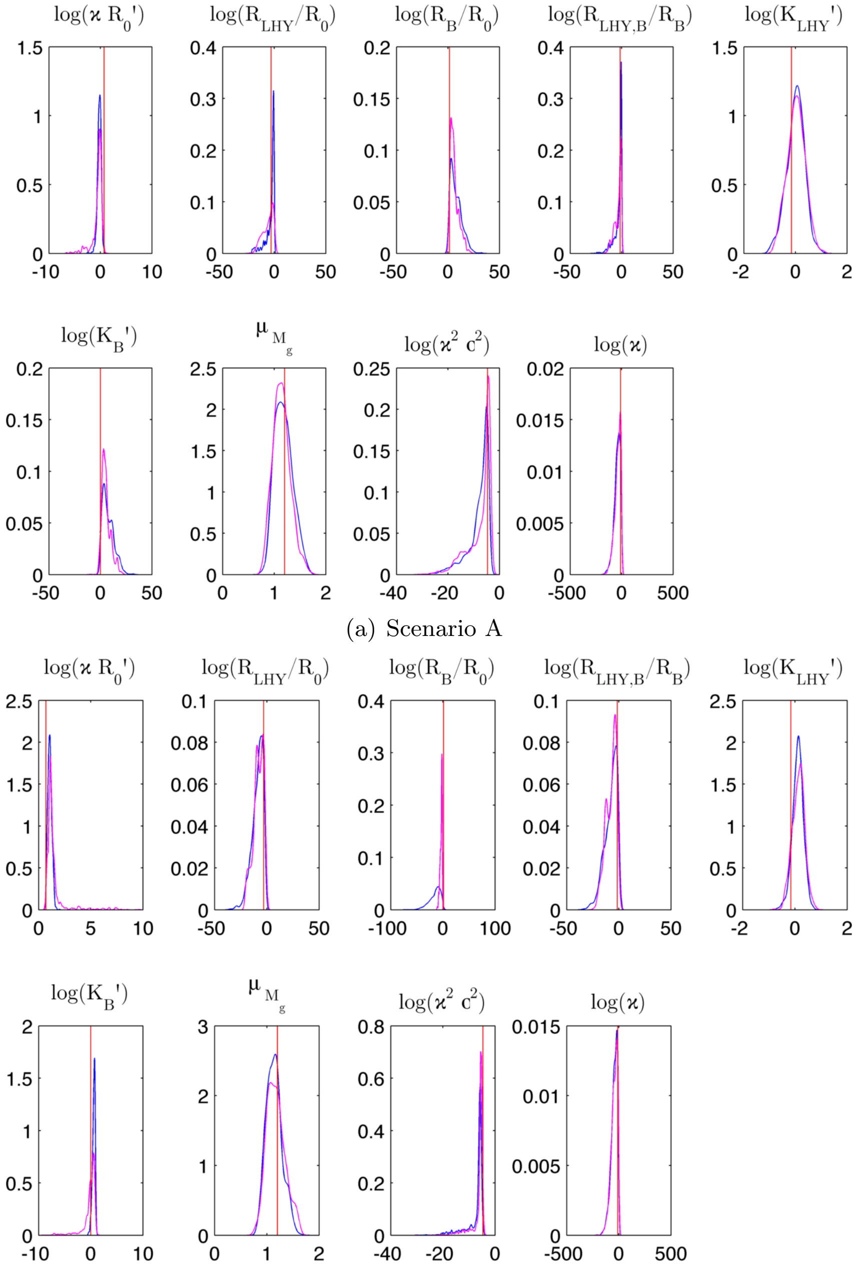


Figure C.2: Unobserved TF inference (smoothing): posterior median and 95 % HPDIs. Estimates obtained for diffusion approximation simulated data (blue) and SSA simulated data (magenta). True simulated TF B (from the diffusion approximation) superimposed in red. Model 3.6, as applied to SSA simulated data from the model in Table 1.1 and parameter values as in the scenarios of Figure 4.11, and diffusion approximation data simulated according to the scenarios of Figure 3.5. MCMC samples for one replicate scenario A (top) and B (bottom).



(a) Scenario A

(b) Scenario B

Figure C.3: Comparison of the kernel density estimates of the transcription function, noise, scale and degradation parameters of Model 3.6: diffusion approximation simulated data (blue) and SSA simulated data (magenta); the red line is at the true value. MCMC samples for one SSA simulation with parameters as in Figure 4.11 and one diffusion approximation simulation with parameters as in Figure 3.5, scenario A (top) and B (bottom). The red vertical line is at the true value.

Appendix D

Arabidopsis Thaliana data analysis: additional plots

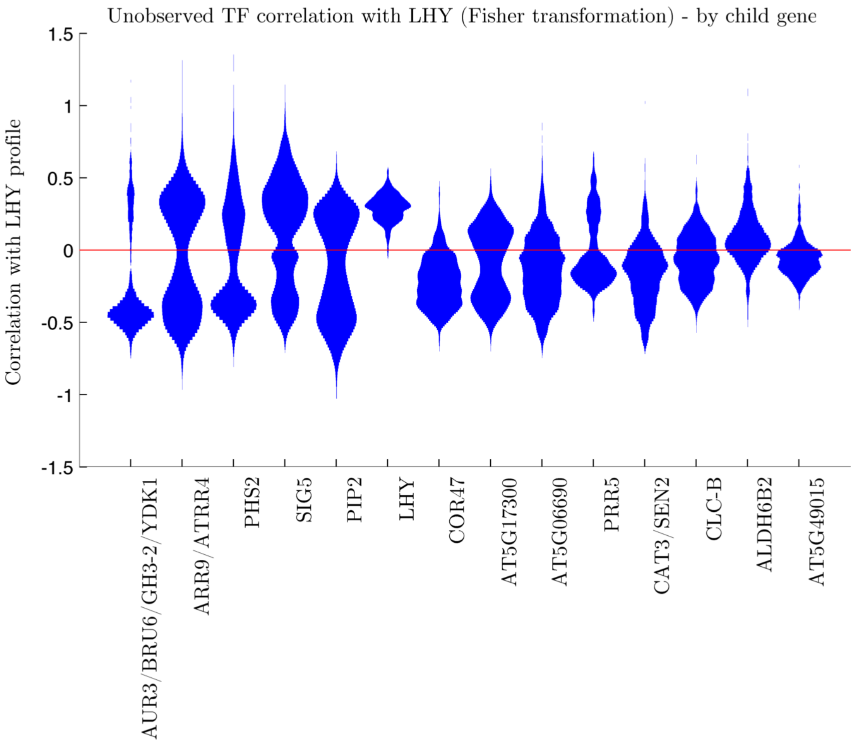
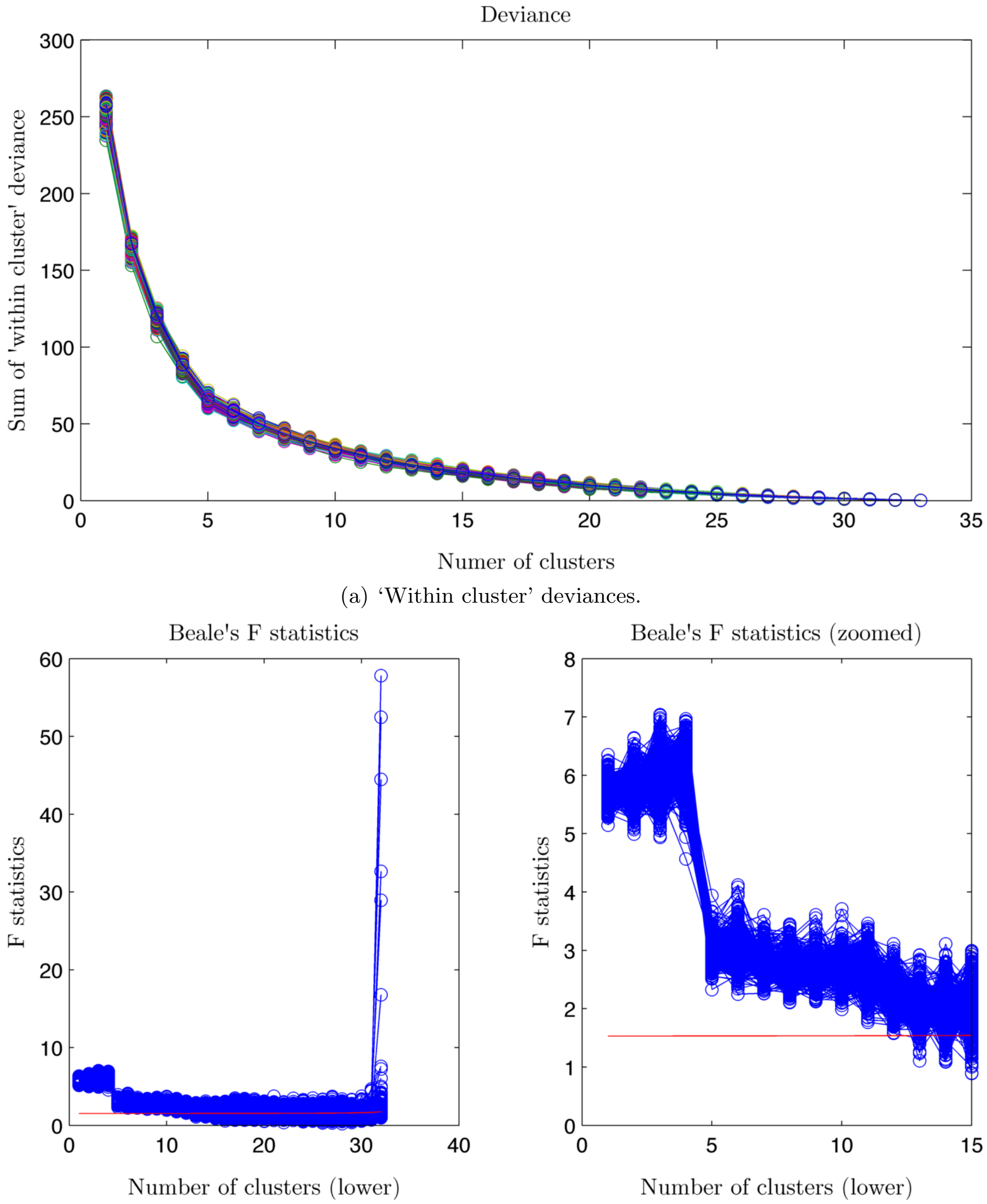


Figure D.1: Violin plots for the correlation of the posterior unobserved TFs profiles and smoothed observed LHY. Fisher transformation of the correlation coefficient. Genes between positions 21 and 34, in order of median posterior correlation, in absolute value. Note that for a limited number of genes only one mode satisfies the model fit requirements. Rhythmic Nanostring genes, with satisfactory explained circadian rhythmicity, Carré lab. Plot code from Dorn (2009).



(a) ‘Within cluster’ deviances.

(b) Beale’s F statistics. Beale’s F statistics test the hypothesis that an increase from k to $k + 1$ clusters, $k = 1, ..., K$, provides a significant decrease in the ’within cluster’ deviance; here we plot k on the x axis. The red line is the threshold for significance at level $\alpha = 5\%$.

Figure D.2: ‘Within cluster’ deviances and Beale’s F statistics for the cluster partitions identified by the k-means algorithm, as applied to the samples posterior profiles of the child mRNA, and for an increasing number of clusters. Each line corresponds to one sample matrix of posterior profiles of the child mRNA, where each row corresponds to a gene. Rhythmic Nanostring genes, with satisfactory explained circadian rhythmicity. Carré lab. at Warwick.

Bibliography

Adams, S., Manfield, I., Stockley, P., & Carré, I. A. (2015). Revised morning loops of the Arabidopsis circadian clock based on analyses of direct regulatory interactions. *PloS one, 10*(12), e0143943.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical science*, 131-153.

Agterberg, F. P. (1984). Trend surface analysis. In *Spatial statistics and models* (pp. 147-171). Springer Netherlands.

Alabadí, D., Oyama, T., Yanovsky, M. J., Harmon, F. G., Más, P., & Kay, S. A. (2001). Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science, 293*(5531), 880-883.

Alon, U. (2006). *An introduction to systems biology: design principles of biological circuits.* CRC press.

An, S., Tsai, C., Ronecker, J., Bayly, A., Herzog, E. D. (2012). Spatiotemporal distribution of vasoactive intestinal polypeptide receptor 2 in mouse suprachiasmatic nucleus. *Journal of Comparative Neurology, 520*(12), 2730-2741.

Ananthasubramaniam, B., Herzog, E. D., & Herzel, H. (2014). Timing of neuropeptide coupling determines synchrony and entrainment in the mammalian circadian clock. *PLoS computational biology, 10*(4), e1003565.

Anderson, D. F., & Kurtz, T. G. (2011). Continuous time Markov Chain models for chemical reaction networks. In *Design and analysis of biomolecular circuits* (pp. 3-42). Springer New York.

Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research, 34*(suppl 2), W369-W373.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceed-*

ings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 268-282.

Beale, E. M. L. (1969). Euclidean cluster analysis. Scientific Control Systems Limited.

Bialek, W., & Setayeshgar, S. (2005). Physical limits to biochemical signaling. *Proceedings of the National Academy of Sciences of the United States of America, 102*(29), 10040-10045.

Bonett, D. G. (2006). Confidence interval for a coefficient of quartile variation. *Computational Statistics Data Analysis, 50*(11), 2953-2957.

Brancaccio, M., Maywood, E. S., Chesham, J. E., Loudon, A. S., & Hastings, M. H. (2013). A Gq-Ca 2+ axis controls circuit-level encoding of circadian time in the suprachiasmatic nucleus. *Neuron, 78*(4), 714-728.

Brett, T., & Galla, T. (2013). Stochastic processes with distributed delays: chemical langevin equation and linear-noise approximation. *Physical review letters, 110*(25), 250601.

Chattopadhyay, S., Puente, P., Deng, X. W., & Wei, N. (1998). Combinatorial interaction of light-responsive elements plays a critical role in determining the response characteristics of light-regulates promoter in Arabidopsis. *The Plant Journal, 15*(1), 69-77.

Christen, J. A., & Fox, C. (2012). Markov chain Monte Carlo using an approximation. *Journal of computational and graphical statistics.*

Collingwood, T. N., Urnov, F. D., & Wolffe, A. P. (1999). Nuclear receptors: coactivators, corepressors and chromatin remodeling in the control of transcription. *Journal of molecular endocrinology, 23*(3), 255-275.

Colwell, C. S. (2011). Linking neural activity and molecular oscillations in the SCN. *Nature Reviews Neuroscience, 12*(10), 553-569.

Covington, M. F., Maloof, J. N., Straume, M., Kay, S. A., & Harmer, S. L. (2008). Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and developement. *Genome biology, 9*(8), 1.

Crick, F. H. (1958, January). On protein synthesis. In *Symp Soc Exp Biol* (Vol. 12, No. 138-63, p. 8).

Crick, F. (1970). Central dogma of molecular biology. *Nature 227*(5258), 561-563.

DeWoskin, D., Myung, J., Belle, M. D., Piggins, H. D., Takumi, T., & Forger, D. B. (2015). Distinct roles for GABA across multiple timescales in mammalian circadian timekeeping. *Proceedings of the National Academy of Sciences, 112*(29), E3911-E3919.

Dibner, C., & Schibler, U. (2015). Circadian timing of metabolism in animal models and humans. *Journal of internal medicine, 277*(5), 513-527.

Dibner, C., Schibler, U., & Albrecht, U. (2010). The mammalian circadian timing system: organization and coordination of central and peripheral clocks. *Annual review of physiology, 72*, 517-549.

Dodd, A. N., Salathia, N., Hall, A., Kèvei, E., Tóth, R., Nagy, F., ... & Webb, A. A. (2005). Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science, 309*(5734), 630-633.

Doob, J. L. (1945). Markoff chains–denumerable case. *Transactions of the American Mathematical Society, 58*(3), 455-473.

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering, 12*, 656-704.

El Cheikh, R., Lepoutre, T., & Bernard, S. (2012). Modeling biological rhythms in cell populations. *Mathematical Modelling of Natural Phenomena, 7*(6), 107-125.

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic Gene Expression in a Single Cell. *Science Signaling, 297*(5584) 1183.

Evans, J. A., Leise, T. L., Castanon-Cervantes, O., & Davidson, A. J. (2013). Dynamic interactions mediated by nonredundant signaling mechanisms couple circadian clock neurons. *Neuron, 80*(4), 973-983.

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis (¿ Wiley Series in Probability and Statistics).

Fabbris, L. (1997). *Statistica multivariata: analisi esplorativa dei dati.* McGraw-Hill Libri Italia.

Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nature Reviews Genetics, 10*(9), 605-616.

Fearnhead, P., Giagos, V., & Sherlock, C. (2014) Inference for reaction networks using the Linear Noise Approximation. *Biometrics, 70*(2), 457-466.

Ferm, L., Lötstedt, P., & Hellander, A. (2008). A hierarchy of approximations of the master equation scaled by a size parameter. *Journal of Scientific Computing, 34*(2), 127-151.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society, 85*(1), 87-94.

Forger, D. B., & Peskin, C. S. (2005). Stochastic simulation of the mammalian circadian clock. *Proceedings of the National Academy of Sciences of the United States of America, 102*(2), 321-324.

Fuhr, L., Abreu, M., Pett, P., & Relógio, A. (2015). Circadian systems biology: When time matters. *Computational and structural biotechnology journal, 13*, 417-426.

Fustin, J. M., O'Neill, J. S., Hastings, M. H., Hazlerigg, D. G., & Dardente, H. (2009). Cry1 circadian phase in vitro: wrapped up with an E-box. *Journal of biological rhythms, 24*(1), 16-24.

Gaetan, C., & Guyon, X. (2010). *Spatial statistics and modeling* (Vol. 81). New York: Springer.

Galla, T. (2009). Intrinsic fluctuations in stochastic delay systems: Theoretical description and application to a simple model of gene regulation. *Physical Review E, 80*(2), 021909

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician, 5*(3), 115-146.

Getis, A. (2009). Spatial weights matrices. *Geographical Analysis, 41*(4), 404-410.

Gillespie, C. S., & Golightly, A. (2016). Diagnostics for assessing the linear noise and moment closure approximations. *Statistical Applications in Genetics and Molecular Biology, 15*(5), 363-379.

Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry, 81*(25), 2340-2361.

Gillespie, D.T. (1992). A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications, 188*(1), 404-425.

Gillespie, D.T. (2000). The chemical Langevin equation. *The Journal of Chemical Physics, 113*(1), 297-306.

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in medicine, 33*(11), 1946-1978.

Golightly, A., Henderson, D. A., & Sherlock, C. (2015). Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statistics and Computing, 25*(5), 1039-1055.

Golightly, A., & Wilkinson, D. J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics, 61*(3), 781-788.

Gonze, D., Bernard, S., Waltermann, C., Kramer, A., & Herzel, H. (2005). Spontaneous synchronization of coupled circadian oscillators. *Biophysical journal*, 89(1), 120-129.

Goodwin, B. C. (1965). Oscillatory behavior in enzymatic control processes. *Advances in enzyme regulation, 3*, 425-437.

Gopalakrishnan, A., Kaisare, N. S., & Narasimhan, S. (2011). Incorporating delayed and infrequent measurements in Extended Kalman Filter based nonlinear state estimation. *Journal of Process Control, 21*(1), 119-129.

Goutelle, S., Maurin, M., Rougier, F., Barbaut, X., Bourguignon, L., Ducher, M., & Maire, P. (2008). The Hill equation: a review of its capabilities in pharmacological modelling. *Fundamental and Clinical Pharmacology, 22*(6), 633-648.

Grima, R. (2012). A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics, 136*(15), 154105.

Harmer, S. L. (2009). The circadian system in higher plants. *Annual review of plant biology, 60*, 357-377.

Harmer, S. L., Hogenesch, J. B., Straume, M., Chang, H. S., Han, B., Zhu, T., ... & Kay, S. A. (2000). Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science, 290*(5499), 2110-2113.

Harmer, S. L., & Kay, S. A. (2005). Positive and negative factors confer phase-specific circadian regulation of transcription in Arabidopsis. *The Plant Cell, 17*(7), 1926-1940.

Hastings, M. H., Brancaccio, M., & Maywood, E. S. (2014). Circadian pacemaking in cells and circuits of the suprachiasmatic nucleus. *Journal of neuroendocrinology, 26*(1), 2-10.

Hastings, M. H., Maywood, E. S., & O'Neill, J. S. (2008). Cellular circadian pacemaking and the role of cytosolic rhythms. *Current Biology, 18*(17), R805-R815.

Hey, K. L., Momiji, H., Featherstone, K., Davis, J. R., White, M. R., Rand, D. A., & Finkenstädt, B.(2015). A stochastic transcriptional switch model for single cell imaging data. *Biostatistics*, kxv010.

Horne, J. H., & Baliunas, S. L. (1986). A prescription for period analysis of unevenly sampled time series. *The Astrophysical Journal, 302*, 757-763.

Huang, W., Prez-Garca, P., Pokhilko, A., Millar, A. J., Antoshechkin, I., Riechmann, J. L., & Mas, P. (2012). Mapping the core of the Arabidopsis circadian clock defines the network structure of the oscillator. *Science, 336*(6077), 75-79.

Jahnke, T., & Huisinga, W. (2007). Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of mathematical biology, 54*(1), 1-26.

Jazwinski, A. H. (2007). *Stochastic processes and filtering theory.* Courier Corporation.

Jenkins, D.J., Finkenstädt, B., & Rand, D. A. (2013). A temporal switch model for estimating transcriptional activity in gene expression. *Bioinformatics, 29*(9), 1158-1165.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering, 82*(1), 35-45.

Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Fluids Engineering, 83*(1), 95-108.

Kim, J. K., & Forger, D. B. (2012). A mechanism for robust circadian timekeeping via stoichiometric balance. *Molecular systems biology, 8*(1), 630.

Kim, J. K., Kilpatrick, Z. P., Bennett, M. R., Josić, K. (2014). Molecular mechanisms that regulate the coupled period of the mammalian circadian clock. *Biophysical journal, 106*(9), 2071-2081.

Ko, C. H., Yamada, Y. R., Welsh, D. K., Buhr, E. D., Liu, A. C., Zhang, E. E., ... & Takahashi, J. S. (2010). Emergence of noise-induced oscillations in the central circadian pacemaker. *PLoS Biol, 8*(10), e1000513.

Komorowski, M., Finkenstädt, B., Harper, C. V., & Rand, D. A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC bioinformatics, 10*(1), 1.

Korenčič, A., Bordyugov, G., Rozman, D., Goličnik, M., & Herzel, H. (2012). The interplay of cis-regulatory elements rules circadian rhythms in mouse liver. *PLoS One, 7*(11), e46835.

Kulikov, G. Y., & Kulikova, M. V. (2014). Accurate numerical implementation of the continuous-discrete extended Kalman filter. *IEEE Transactions on Automatic Control, 59*(1), 273-279.

Kurtz, T. G. (1972). The relationship between stochastic and deterministic models for chemical reactions. *The Journal of Chemical Physics, 57*(7), 2976-2978.

Lachowicz, M. (2011). Microscopic, mesoscopic and macroscopic descriptions of complex systems. *Probabilistic Engineering Mechanics, 26*(1), 54-60.

Latchman, D. (2007). *Gene regulation*, Taylor & Francis.

Lee, C., Etchegaray, J. P., Cagampang, F. R., Loudon, A. S., & Reppert, S. M. (2001). Posttranslational mechanisms regulate the mammalian circadian clock. *Cell, 107*(7), 855-867.

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell, 152*(6), 1237-1251.

Locke, J. C., Kozma-Bognár, L., Gould, P. D., Fehér, B., Kevei, E., Nagy, F., ... & Millar, A. J. (2006). Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana. *Molecular systems biology, 2*(1), 59.

Lunn, D., Barrett, J., Sweeting, M., & Thompson, S. (2013). Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 62*(4), 551-572.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Mann, R. S., & Carrol, S. B. (2002). Molecular mechanisms of selector gene function and evolution. *Current opinion in genetics & development, 12*(5), 592-600.

Marinari, E., & Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters), 19*(6), 451.

Maywood, E. S., Drynan, L., Chesham, J. E., Edwards, M. D., Dardente, H., Fustin, J. M., ... & Hastings, M. H. (2013). Analysis of core circadian feedback loop in suprachiasmatic nucleus of mCry1-luc transgenic reporter mouse. *Proceedings of the National Academy of Sciences, 110*(23), 9547-9552.

Maywood, E. S., Reddy, A. B., Wong, G. K., O'Neill, J. S., O'Brien, J. A., McMahon, D. G., ... & Hastings, M. H. (2006). Synchronization and maintenance of timekeeping in suprachiasmatic circadian clock cells by neuropeptidergic signaling. *Current Biology, 16*(6), 599-605.

McClung, C. R. (2006). Plant circadian rhythms. *The Plant Cell, 18*(4), 792-803.

McQuarrie, D. A. (1967). Stochastic approach to chemical kinetics. *Journal of applied probability, 4*(3), 413-478.

Menkens, A. E., Schindler, U., & Cashmore, A. R.(1995). The G-box: a ubiquitous regulatory DNA element in plants bound by the GBF family of bZIP proteins. *Trends in biochemical sciences, 20*(12), 506-510.

Michael, T. P., & McClung, C. R. (2002). Phase-specific circadian clock regulatory elements in Arabidopsis. *Plant Physiology, 130*(2), 627-638.

Michael, T. P., Mockler, T. C., Breton, G., McEntee, C., Byer, A., Trout, J. D., ... & Givan, S. A. (2008). Network discovery pipeline elucidates conserved time-of-day specific cis-regulatory modules. *PLoS Genet, 4*(2), e14.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1/2), 17-23.

Myung, J., Hong, S., DeWoskin, D., De Schutter, E., Forger, D. B., & Takumi, T. (2015). GABA-mediated repulsive coupling between circadian clock neurons in the SCN encodes seasonal time. *Proceedings of the National Academy of Sciences, 112*(29), E3920-E3929.

Nachman, I., Regev, A., & Friedman, N. (2004). Inferring quantitative models of regulatory networks from expression data. *Bioinformatics, 20*(suppl 1), i248-i256.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and computing, 6*(4), 353-366.

Oates, C. J., & Mukherjee, S. (2012). Network inference and biological dynamics. *The annals of applied statistics, 6*(3), 1209.

Pace, L., & Salvan, A. (1997). *Principles of statistical inference: from a Neo-Fisherian perspective* (Vol. 4). World scientific.

Park, P. J. (2009). Chip-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics, 10*(10), 669-680.

Petris, G., Petrone, S., & Campagnoli, P. (2009). *Dynamic linear models with R.* Springer.

Prado, R., & West, M. (2010). *Time series: modeling, computation, and inference.* CRC Press.

Price, H. J., Sparrow, A. H., & Nauman, A. F. (1973). Correlations between nuclear volume, cell volume and DNA content in meristematic cells of herbaceous angiosperms. *Experientia, 29*(8), 1028-1029.

Quiroga, R. Q. (2009). Bivariable and multivariable analysis of EEG signals.

Rao, C. V., & Arkin, A. P. (2003). Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. *The Journal of chemical physics, 118*(11), 4999-5010.

Relógio, A., Westermark, P. O., Wallach, T., Schellenberg, K., Kramer, A., & Herzel, H. (2011). Tuning the mammalian circadian clock: robust synergy of two loops. *PLoS Comput Biol, 7*(12), e1002309-e1002309.

Ribeiro, A., Zhu, R., & Kauffman, S. A. (2006). A general modeling strategy for gene regulatory networks with stochastic dynamics. *Journal of Computational Biology, 13*(9), 1630-1639.

Roberts, G. O., & Rosenthal, J. S. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics, 18*(2), 349-367.

Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science, 16*(4), 351-367.

Salome, P. A., & McClung, C. R. (2004). The Arabidopsis thaliana clock. *Journal of Biological Rhythms, 19*(5), 425-435.

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature, 489*(7414), 109-113.

Särkkä, S. (2007). On unscented Kalman filtering for state estimation of continuous-time nonlinear systems. *Automatic Control, IEEE Transactions on, 52*(9), 1631-1641.

Särkkä, S. (2013). *Bayesian filtering and smoothing (Vol. 3).* Cambridge University Press.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin, 2*(6), 110-114.

Scargle, J. D. (1982). Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal, 263*, 835-853.

Schindler, U., Beckmann, H., & Cashmore, A. R. (1992). TGA1 and G-box binding factors: two distinct classes of Arabidopsis leucine zipper proteins compete for the G-box-like element TGACGTGG. *The Plant Cell, 4*(10), 1309-1319.

Shaphiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika, 52*(3), 591-611.

Sherlock, C., Golightly, A., & Henderson, D. A. (2016). Adaptive, delayed-acceptance MCMC for targets with expensive likelihoods. *arXiv preprint arXiv:1509.00172.*

Sherlock, C., Thiery, A., & Golightly, A. (2015). Efficiency of delayed-acceptance random walk Metropolis algorithms. *arXiv preprint arXiv:1506.08155.*

Singer, H. (2002). Parameter estimation of nonlinear stochastic differential equations: simulated maximum likelihood versus extended Kalman filter and It-Taylor expansion. *Journal of Computational and Graphical Statistics, 11*(4), 972-995.

Singer, H. (2006). *Continuous-discrete unscented Kalman filtering.* Fernuniversität.

Smith, H. (2010). *An introduction to delay differential equations with applications to the life sciences* (Vol. 57). Springer Science Business Media.

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics, 13*(9), 613-626.

Stathopoulos, V., & Girolami, M. A. (2013). Markov Chain Monte Carlo inference for Markov jump processes via the linear noise approximation. *Philosophical Transactions of the Royal Statistical Society A: Mathematical, Physical and Engineering Sciences, 371*(1984), 20110541.

Steuer, R., Zhou, C., & Kurths, J. (2003). Constructive effects of fluctuations in genetic and biochemical regulatory systems. *Biosystems, 72*(3), 241-251.

Stewart, T., Strijbosch, L. W. G., Moors, H., & Batenburg, P. V. (2007). A simple approximation to the convolution of gamma distributions.

Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., & Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science, 332*(6028), 472-474.

Takahashi, N., Hirata, Y., Aihara, K., & Mas, P. (2015). A hierarchical multi-oscillator network orchestrates the Arabidopsis circadian system. *Cell, 163*(1), 148-159.

Terejanu, G. A. (2003). Extended kalman filter tutorial. Online. Disponible: http://users. ices. utexas. edu/ terejanu/files/tutorialEKF. pdf.

Thompson, W. H., & Fransson, P. (2016). On stabilizing the variance of dynamic functional brain connectivity time series. *arXiv preprint arXiv:1603.00201.*

Tian, L., Hires, S. A., Mao, T., Huber, D., Chiappe, M. E., Chalasani, S. H., ... & Looger, L. L. (2009). Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nature methods, 6*(12), 875-881.

Tkačik, G., & Walczac, A. M. (2011). Information transmission in genetic regulatory networks: a review. *Journal of Physics: condensed matter, 23*(15), 153102.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46*(sup1), 234-240.

Travnickova-Bendova, Z., Cermakian, N., Reppert, S. M., & Sassone-Corsi, P. (2002). Bimodal regulation of mPeriod promoters by CREB-dependent signaling and CLOCK/BMAL1 activity. *Proceedings of the National Academy of Sciences, 99*(11), 7728-7733.

Van Der Merwe, R., Doucet, A., De Freitas, N., & Wan, E. (2000, August). The unscented particle filter. In *NIPS* (Vol. 2000, pp. 584-590).

Van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry* (Vol. 1). Elsevier.

Viton, P. A. (2010). Notes on spatial econometric models. *City and regional planning, 870*(03), 9-10.

Voss, T. C., & Hager, G. L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. Nature Reviews Genetics, 15(2), 69-81.

Wallace, E. W. (2010). A simplified derivation of the linear noise approximation. *arXiv preprint arXiv:1004.4280.*

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data* (Vol. 368). John Wiley Sons.

Wang, H., Dittmer, T. A., & Richards, E. J. (2013). Arabidopsis CROWDED NUCLEI (CRWN) proteins are required for nuclear size control and heterochromatin organization. *BMC plant biology, 13*(1), 1.

Welch, B. L. (1947). The generalization ofstudent's' problem when several different population variances are involved. *Biometrika, 34*(1/2), 28-35.

West, M. (1999). *Bayesian forecasting.* John Wiley Sons, Inc.

Wikle, C. K., Berliner, L. M., & Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics, 5*(2), 117-154.

Wilkinson, D. J. (2012). *Stochastic Modelling for Systems Biology.* Chapman & Hall.

Yamaguchi, S., Isejima, H., Matsuo, T., Okura, R., Yagita, K., Kobayashi, M., & Okamura, H. (2003). Synchronization of cellular clocks in the suprachiasmatic nucleus. *Science, 302*(5649), 1408-1412.

Source code

Altman, Y. (2009). export_fig. MATLAB Central File Exchange.

Bivand, R. (2015). Spatial dependence: weighting schemes, statistics and models.

Dorn, J. (2009). Violin Plots for plotting multiple distributions (distributionPlot.m). MATLAB Central File Exchange.

Greene, C. (2014a). histf. MATLAB Central File Exchange.

Greene, C. (2014b). legalpha. MATLAB Central File Exchange.

Henson, R. (2005). Flow Cytometry Data Reader and Visualization. MATLAB Central File Exchange.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package MASS. CRAN Repository. http://cran. r-project. org/web/packages/MASS/MASS. pdf.

Saida, A. B. (2007). Shapiro-Wilk and Shapiro-Francia normality tests. MATLAB Central [online], 15.

Vehtari, A. (2001). GPstuff. MATLAB Central File Exchange.