Modelling, Simulation, and Inferring Regulatory Networks

Maksat Ashyraliyev

Modelling, Simulation, and Inferring Regulatory Networks

Maksat Ashyraliyev

Modelling, Simulation, and Inferring Regulatory Networks

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 20 oktober 2009, te 12:00 uur

door

Maksat Ashyraliyev

geboren te Ashgabat, Toerkmenistan

Promotiecommissie:

Promotor:	prof. dr. J.G. Verwer
Overige leden:	prof. dr. P.W. Hemker
	prof. dr. C.A.J. Klaassen
	prof. dr. M.A. Peletier
	prof. dr. R.P. Stevenson
	dr. J.A. Kaandorp
	dr. J. Jaeger

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research described in this thesis has been supported by the Netherlands Organisation for Scientific Research (NWO) under project number 635.100.010. It was carried out at the Centrum Wiskunde & Informatica (CWI), the Dutch national research institute for mathematics and computer science, in Amsterdam.

Thomas Stieltjes Institute
for Mathematics

To my parents,
Alleberen and Govherjan.

Contents

Chapter 1

Introduction

Mathematical modelling of real-life systems is widely used in different applied fields of science. Models are often based on certain known laws (rules) for the processes under consideration. Depending on those rules, a deterministic model can be a set of (non)linear differential equations, such as Ordinary Differential Equations (ODEs), Partial Differential Equations (PDEs), Delay Differential Equations (DDEs), etc., or combinations of these. Additionally, algebraic relations can be coupled to differential equations resulting in a system of Differential Algebraic Equations (DAEs). The analytical solution of the model can be found very rarely and therefore, one often needs to compute a numerical solution by using various numerical techniques.

A mathematical model, to be meaningful, is expected to describe the phenomenon in a sufficiently accurate way. In addition to that, it is desirable for a model to be an *analytical* and *predictive* tool. Analytical means that the model can be used for a better understanding of the underlying mechanism of a system. So, understanding a model leads to understanding the real system. Predictive implies that the model is able to simulate experiments before actually performing them or to predict the system behaviour in the situations that cannot be dealt with experimentally.

Mathematical models usually have a number of parameters. Some parameters are known from the literature or can be directly obtained from experiments. However, in many cases parameters are not known and cannot be measured. On the other hand, other quantities (observables) involved in the model, such as state variables or combinations thereof, can be quantified in experiments. Then, the unknown parameters can be estimated by fitting model outputs to the data.

Parameter estimation or data fitting typically starts with a guess about parameter values and then changes those values to minimize the discrepancy between model and data using a particular metric which is called a cost function (or fitness function). This inverse problem has a number of pitfalls. The first question is whether the parameters for the mathematical model can be identified uniquely

assuming that for all observables continuous and error-free data are available. This is the subject of *a priori* identifiability or structural identifiability analysis of the mathematical model [1].

When the mathematical models are nonlinear in the parameters, the fitness function may have many local minima that do not correspond to its lowest value. Although *local search methods*, like gradient-based algorithms, are efficient optimizers [2] for nonlinear least-squares problems once a sufficiently good initial guess for the parameter values is available, they can get trapped in one of the local minima rather than finding the global minimum. On the contrary, *global search methods*, like Simulated Annealing (SA) [3] explore the whole search space but have slow convergence. The latter is very crucial for large-scale problems. Sequential application by using a global search method followed by a local gradient-based technique allows to find the global minimum precisely and fast [4].

Finally, given a particular set of experimental data, and one particular acceptable model parameterization obtained by a parameter estimation procedure, does not mean that all obtained parameters can be trusted. After the minimum has been found, an *a posteriori* or practical identifiability study can show how well the parameter vector has been determined given a data set that is possibly sparse and noisy. That this part of model fitting should not be underestimated is shown by Gutenkunst et al. [5]. For all 17 systems biology models that they considered, the obtained parameters are sloppy, meaning not well-defined. It is shown that the sloppiness is an intrinsic property of such models and cannot be removed by using more comprehensive and more accurate data. On the other hand, one could argue that often the precise value of a parameter is not required to draw qualitative conclusions [6].

The main application field for the mathematical models studied in this dissertation lies in developmental biology. Genetic regulation plays a fundamental role in the developmental processes, such as the body plan formation of an organism. The insect body plan consists of repeating units called segments. Segment determination happens at the early developmental stages. The main focus here is on the fruit fly *Drosophila melanogaster* for which the boundaries of segments are determined during the blastoderm stage of development [7]. Segment determination is controlled by a relatively small set of segmentation genes. The whole segmentation gene network has a hierarchical structure whose levels consist of gap, pair-rule, and segment-polarity genes [8, 9]. In such a regulatory cascade, genes from the upstream layer are involved in the regulation of the genes from the downstream layer; additionally, genes from the same level are regulating each other as well. The gap gene system is of a particular interest in this hierarchy. It constitutes the first zygotic step in the segmentation gene network. The expression of gap gene domains are initially established by maternal input. These spatial domains further change dynamically due to the influence from the terminal maternal system and most importantly due to the regulatory interactions between gap genes. In turn, gap genes are involved in the regulation of pair-rule and segment-polarity genes from downstream layers. The latter establish a segmental pre-pattern of

gene expression by the onset of gastrulation. Although the gap gene system has been studied extensively using genetic and molecular approaches [10] (and references therein), still understanding of the underlying mechanism of formation of gap gene expression patterns remains incomplete. This is not an unusual situation; not all questions can be answered experimentally. In some cases, genetic evidences result in ambiguities for interactions between the genes in a regulatory network. These remaining blanks can be filled by using mathematical models of regulatory networks.

Mathematical modelling has been proven to be a powerful tool to study regulatory networks. Among many different approaches, we restrict ourselves to the case where a model is given by a system of nonlinear ODEs (lattice-differential equations), a so-called 'connectionist' model of development proposed in [11]. The time evolution of gene products is described with a sigmoid function to model genetic regulation, a decay term, and a simple difference formula to model diffusion. A connection matrix consisting of numbers ('weights') is used to model the regulatory influences in the input of the sigmoid regulation function. Extensive research has been done using this model to study the gap gene system in the early development of the *Drosophila* [10, 12–15]. Some analytical results on the behaviour of this model are presented in [16]. This model is also studied in the continuous version, as a reaction-diffusion PDE system [17]. With the connectionist model and available quantitative expression data for all relevant genes involved in the gap gene network [18, 19], regulatory interactions between gap genes have been successfully inferred using different optimization methods. Obtained results have given significant insight into the functioning of the gap gene system, i.e. the underlying mechanism of spatio-temporal pattern formation of gap gene products. However, conclusions in all previous works have been based on the values of estimated parameters only and the important link, the identifiability analysis of inferred regulatory parameters, is missing. This has a serious implication as all conclusions can be unreliable.

We note that the derivation of the connectionist model [11] for modelling regulatory networks is based on a number of assumptions (simplifications) and therefore, such a model has limitations because it cannot capture many important details of development. In order to overcome this, the model has to be extended. Undoubtedly, there will be a price to pay. More detailed models will have larger complexity and their numerical integration as well as the parameter estimation procedure may give rise to real numerical challenges.

An important limitation of the connectionist model is that cells are modelled as static discrete lattice sites on a grid. In a developmental process gene expression patterns in space and time are leading to various biomechanical responses of the cells, like cell migration, adhesion, growth, death, etc. The responses, in turn, can influence the gene expression pattern. For example, a migrating cell may change the spatial gene expression pattern or in cell division the components of the mother cell may be unequally distributed over the daughter cells which can also affect the gene expression pattern. In many cases a simple connectionist

model will not be sufficient to explain the development of a spatio-temporal gene expression pattern in an aggregate of migrating cells and a more elaborate, cell-based, model is required to capture the coupling between gene expression patterns and moving cells.

The protein synthesis in the connectionist model is assumed to be instantaneous. This is clearly a simplification of the biological system as the transcription of genes and the translation of mRNA into proteins take some time. Neglecting these details has resulted in an artificially high level of gene products in the model outputs at early stages of time integration when the standard connectionist model has been applied for the gap gene system in *Drosophila* [10, 12]. Time delays in the production of proteins have to be incorporated in order to have a more realistic model to correctly describe early gap gene expression and regulation. This will change the connectionist model into a system of DDEs. For parameter estimation and the practical identifiability analysis in the resulting DDE models a detailed investigation is needed as the model solution can be nonsmooth in the model parameters [20].

1.1 Parameter Estimation

In this section[1)] we shall briefly present different aspects of parameter estimation which will be used throughout the thesis (for a detailed description see Chapter 2). The methodology is illustrated on the basis of a simple enzymatic reaction.

1.1.1 Problem definition

For ease of presentation, we consider a model given by the system of ODEs of the general form [2)]:

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{y}(\mathbf{t},\theta)}{\mathrm{d}t} = \mathbf{f}(t,\mathbf{y}(t,\theta),\mathbf{u}(t),\theta), & 0 < t \le T, \\ \mathbf{y}(t,\theta) = \mathbf{y}_0(\theta), & t = 0, \end{cases} \tag{1.1.1}$$

where t denotes time, the m-dimensional vector θ contains all unknown parameters, $\mathbf{y}$ is an n-dimensional vector containing the state variables (e.g. concentration values), $\mathbf{u}$ are the externally input signals, and $\mathbf{f}$ is a given vector function, differentiable with respect to t, $\mathbf{y}$ and θ. When components of the initial state vector $\mathbf{y}_0$ are not known, they are considered as unknown parameters, so $\mathbf{y}_0$ may depend on θ. Assume that the parameter vector θ should satisfy (non)linear constraints:

$$\mathbf{c}(t,\mathbf{y}(t,\theta),\mathbf{u}(t),\theta) \ge 0, \qquad 0 < t \le T. \tag{1.1.2}$$

[1)] Part of this section is based on the journal publication [21].

[2)] The presented material is also applicable to systems of DAEs, discretized PDEs and DDEs.

Let us assume that N measurements for the state variables [3] in (1.1.1) are available. Each measurement, which we denote by $\tilde{y}_i$, is specified by the time t_i when the c_i-th component of the vector $\mathbf{y}$ is measured. The corresponding model value for a specific parameter vector θ is denoted by $y_{c_i}(t_i, \theta)$. We assume that a true solution θ^* exists for which (1.1.1) is a sufficiently accurate mathematical description approximating reality. In this case, the difference $|y_{c_i}(t_i, \theta^*) - \tilde{y}_i|$ is solely due to experimental errors. The vector of weighted discrepancies between the model values and the experimental values is denoted by $\mathrm{Y}(\theta)$. The m-dimensional optimization problem is given by the task to minimize some measure, $S(\theta)$, for the weighted discrepancy $\mathrm{Y}(\theta)$. By far the most used measure is the Euclidean norm or the sum of weighted squares:

$$S(\theta) = \sum_{i=1}^{N} \frac{(y_{c_i}(t_i, \theta) - \tilde{y}_i)^2}{\sigma_i^2} = \mathrm{Y}^T(\theta)\mathrm{Y}(\theta), \tag{1.1.3}$$

see [22]. This measure results from the maximum likelihood estimator (MLE) theory under the assumption that the experimental errors are independent and normally distributed with standard deviation σ_i and zero mean. When these assumptions do not hold, other measures might be used like the sum of the absolute values. The MLE theory then does not apply and the statistical analysis in the Section 1.1.2 does not hold.

1.1.2 Identifiability and Determinability

Whether the parameters for the mathematical model can be found is dependent on (a) the mathematical model; (b) the significance of the data; and (c) the experimental errors. In the following, we assume that the model is properly scaled such that both the parameter values and the state variables are of the same order of magnitude. Otherwise, a proper scaling should be applied to the model.

A priori identifiability

A parameter is globally identifiable if it can be uniquely determined given the input profile u(t) and assuming continuous and error-free data for the observables of the model. If there is a countable number of solutions the parameter is locally identifiable; it is unidentifiable if there exist uncountable many solutions. A model is structurally globally/locally identifiable if all its parameters are globally/locally identifiable.

It is advisable to always perform an *a priori* analysis to determine *a priori* global identifiability of the model. However, for realistic situations (i.e. nonlinear models of a large size) it is very difficult to obtain any results with known techniques.

[3] The problem can be formulated in the same way when other quantities, such as the combinations of state variables in (1.1.1), are measured.

A posteriori identifiability

Practical or *a posteriori* identifiability analysis studies whether the parameters can be determined with the available, noisy, experimental data. The difficulty in estimating the parameters in a quantitative mathematical model is not so much how to compute them, but more how to assess the quality of the obtained parameters because this not only depends on how well the model describes the phenomenon studied and on the existence of a unique set of parameters, but also on whether the experimental data are sufficient in number, sufficiently significant and sufficiently accurate. With respect to the first two requirements, a sufficient and significant amount of data, it is clear that, whatever method one uses to fit a model with experimental data: to estimate m unknown parameters, one needs at least m experimental values. On the other hand, it is not necessary to have experimental data for all state variables involved in the model at all possible time points, often only a few measurements for the right observable at significant times are needed. The last question, sufficiently accurate data, is related to the fact that measurement errors imply that we do not have precise data points to fit our model with, but that each point represents a whole cloud of possible data values, implying also that the inferred parameters are not point-values but are contained in a cloud. Depending on the model, the cloud of possible parameter values varies in size and shape and can be much larger than the original uncertainty in the data.

The most applied method [23, 24] to study the uncertainty in the parameters is to compute the sensitivity matrix $J(\hat{\theta}) = \frac{\partial \mathrm{Y}(\hat{\theta})}{\partial \theta}$, where $\hat{\theta}$ is the parameter estimate obtained by minimizing (1.1.3). This can be done either by finite differencing or by solving the variational equations [4)]. How close the estimate $\hat{\theta}$ is to the true parameter vector θ^* is expressed by the $(1-\alpha)$-confidence region for θ^*, given by:

$$(\theta^* - \hat{\theta})^T \left(J^T(\hat{\theta}) J(\hat{\theta}) \right) (\theta^* - \hat{\theta}) \leq r_\sigma, \tag{1.1.4}$$

with $r_\sigma = \frac{m}{N-m} S(\hat{\theta}) F_\alpha(m, N-m)$, where $F_\alpha(m, N-m)$ is the upper α part of Fishers distribution with m and $N-m$ degrees of freedom. The $(1-\alpha)$-confidence region implies that there is a probability of $1-\alpha$ that the true parameter vector θ^* lies in this ellipsoid that is centered at $\hat{\theta}$ and has its principal axes directed along the eigenvectors of $J^T(\hat{\theta})J(\hat{\theta})$. The length of the principal axes is proportional to the reciprocal of the corresponding singular values of $J^T(\hat{\theta})J(\hat{\theta})$. Note that this is a linear analysis, and local both with respect to θ and to the given data points.

Now, the practical identifiability of parameters can be assessed in different ways. First, if the model has only two or three parameters it can be done by visual inspection of the ellipsoidal region. For high-dimensional problems this is not possible. But inspecting the eigenvectors and the singular values of the Fisher information matrix $J^T(\hat{\theta})J(\hat{\theta})$ can reveal the identifiability or unidentifiability of

[4)] Variational equations are obtained by taking the derivative of the system (1.1.1) with respect to the parameters. This results in m ODE systems in the variables $\frac{\partial \mathbf{y}(t,\theta)}{\partial \theta_i}$, $i = 1, \ldots, m$.

the parameters or linear combinations of them (for a detailed description see Chapter 2). Finally, there are a number of easy-to-compute indicators. Assuming that all other parameters are exact, a confidence interval for a specific parameter is the intersection of the ellipsoidal region with the parameter axis. This is the dependent confidence interval:

$$\triangle^D \theta_i = \frac{r_\sigma}{\sqrt{\left(J^T(\hat{\theta})J(\hat{\theta})\right)_{ii}}}. \tag{1.1.5}$$

The independent confidence interval is given by the projection of the ellipsoidal region onto the parameter axis:

$$\triangle^I \theta_i = r_\sigma \sqrt{\left(\left(J^T(\hat{\theta})J(\hat{\theta})\right)^{-1}\right)_{ii}}. \tag{1.1.6}$$

If dependent and independent confidence intervals are similar and small, $\hat{\theta}_i$ is well-determined. In case of a strong correlation between parameters, the dependent confidence intervals underestimate the confidence region, whereas the independent confidence intervals overestimate it. Another way to obtain information about the correlations between parameters is to look at the covariance matrix $\mathrm{cov} = (J^T J)^{-1}$. The correlation coefficient of the i-th and j-th parameter is given by:

$$\rho_{ij} = \frac{\mathrm{cov}_{ij}}{\sqrt{\mathrm{cov}_{ii}\mathrm{cov}_{jj}}}. \tag{1.1.7}$$

Determinability

Practical unidentifiability, i.e. parameter sloppiness, indicated for instance by large confidence intervals, is a serious problem when the models are used to extract certain information from the solution of the inverse problem and the precise values of parameter estimates are important for that. However, in the models of gene networks, which will be discussed in the next section, the regulatory influences are represented by regulatory weights. Although these parameters are in principle allowed to take any real value, the sign of the weight has a specific biological implication. It indicates how a certain gene regulates another gene in the network. Therefore, the precise values of those parameters are not always important as long as they have certain characteristics, like being positive or negative.

If *a posteriori* identifiability analysis results in a parameter uncertainty range which lies in the characteristic range we call this parameter (qualitatively) *determinable*. Note that for those parameters which have to be determined quantitatively, i.e. having no specific characteristics, determinability refers to *a posteriori* identifiability.

1.1.3 Example

On the basis of a very simple artificial example [25] we show the influence of the experimental data on the parameter determinability. Consider the mathematical model for the simple enzymatic reaction:

$$\begin{cases} \dfrac{d[S]}{dt} = -k_1[E][S] + k_2[C] \\ \dfrac{d[C]}{dt} = k_1[E][S] - k_2[C] - k_3[C] \\ [E] + [C] = [E_0] + [C_0] \end{cases} \tag{1.1.8}$$

with as state variables the concentrations of the substrate [S], the enzyme [E], and complex [C]. Suppose the initial concentration of the state variables, $[S_0]$, $[E_0]$ and $[C_0]$ is known, and the concentration of [C] is measured rather precisely at regular time points $t = 1, \ldots, 20$. For this example, the measurements are generated artificially by adding an independent, normally distributed perturbance with zero expectation and a fixed variance to the model results (red +-marks in Figure 1.1). The initial parameter values are $\theta = (k_1, k_2, k_3) = (6, 0.8, 1.2)$. With these parameter values, the model results are given by the solid lines in the left plot in Figure 1.1. Fitting the model to these measurements with the Levenberg-Marquardt method [26] results in the parameter vector $\hat{\theta} = (0.683, 0.312, 0.212)$ (for the model results, see Figure 1.1, right).

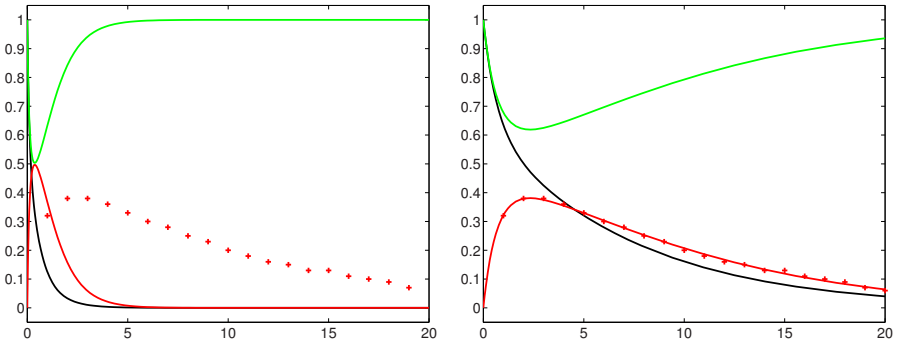


Figure 1.1: Model results for initial (left) and final (right) parameter vector, black: [S], red: [C], green: [E]; and measurements of [C]: red +.

We define the discrepancy of the model with respect to the data:

$$e(\theta) = (c(t_i, \theta) - \tilde{c}_i)_{i=1,\ldots,N} \tag{1.1.9}$$

the vector of the differences between the ith data value, $\tilde{c}_i$, which is the measured concentration of [C] at time t_i, and the corresponding value from the model,

$c(t_i, \theta)$. In the present example, the sensitivity matrix J is an $N \times 3$ matrix, with $N = 20$. For this simple three-parameter problem, one can easily visualize the confidence region (1.1.4) and we can see from the left plot in Figure 1.2 that the true parameter vector lies in a small disc around $\hat{\theta}$, implying that we can estimate all three parameters with a reasonable accuracy by measuring only the complex (or any of the two other concentrations in this case). With 95% confidence, all parameters can be estimated with one digit accuracy and k_3 even with two digits. Using only three well-chosen time-points for measuring ($t = 1, 2, 20$), the axes-length of the ellipsoid increases with a factor of about 4, but still all parameters can be determined reasonably well. Suppose now that it is not possible to measure before time $t = 6$ but that we take 20 samples of the complex at regular times from $t = 6, \ldots, 20$. Suppose also that the same parameter vector $\hat{\theta}$ results from minimizing the least squares error $e^T e$. In this case, the confidence region gives much more reason for distrusting the result. As can be seen in Figure 1.2 (right), the true parameter vector now lies in a long elongated cigar and especially for k_1 and k_2 we can not even trust the order of magnitude. It is clear that it is not easy to *a priori* give an indication whether experimental data are sufficient in number and sufficiently significant. With three lucky data points, one can estimate three parameters, but 20 data points in a region where nothing happens are not sufficient.

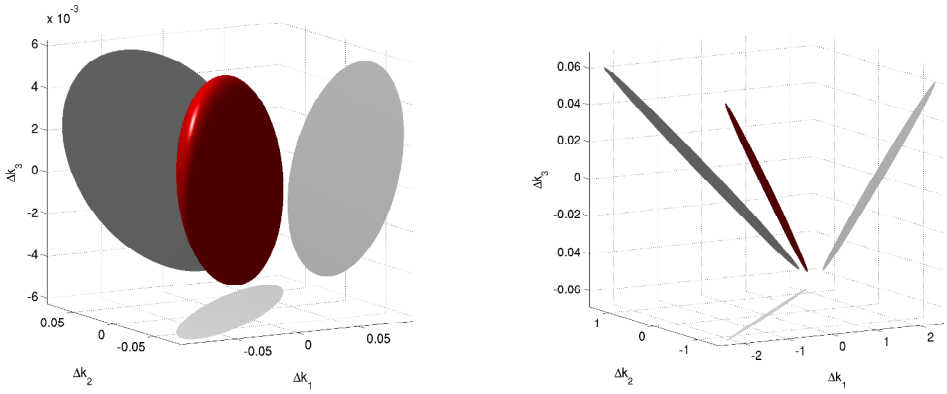


Figure 1.2: Confidence region $\triangle k$ (cf. (1.1.4)) in parameter space around computed parameter vector (origin in the plots) and its projection on the parameter planes. The region contains the true parameter vector with a 95% probability. Left: 20 measurements at $t = 1, \ldots, 20$; right: 20 measurements at time points distributed uniformly over [6,20].

Next, we examine the influence of experimental noise, (i.e. whether the experimental data are sufficiently accurate). Because r_σ is proportional to the variance of the measurement error distribution, the principal axes of the ellipsoidal con-

fidence region are proportional to the standard deviation. Roughly speaking: reducing the (standard deviation of the) error by a factor of two, implies that a parameter, or combination of parameters, can be determined more accurately by a factor of two. This means that to shrink the ellipsoidal confidence region for the $t > 6$ experiment (Figure 1.2, right) such that it results in the same accuracy as the experiment with measurements between 1 and 20, one has to reduce the variance of the experimental error beyond reasonable experimental accuracy. Finally, if we just look at the computable information from the Fisher matrix we get for the confidence intervals (1.1.5)-(1.1.6):

Exp.	$\triangle^D(k_1)$	$\triangle^D(k_2)$	$\triangle^D(k_3)$	$\triangle^I(k_1)$	$\triangle^I(k_2)$	$\triangle^I(k_3)$
[1,20]	0.033	0.028	0.005	0.076	0.067	0.005
[6,20]	0.074	0.047	0.004	2.217	1.267	0.060

The correlation matrices for the two cases are:

$$R_{20} = \begin{pmatrix} 1 & 0.9 & -0.37 \\ -0.9 & 1 & -0.45 \\ -0.37 & -0.45 & 1 \end{pmatrix} \qquad R_6 = \begin{pmatrix} 1 & 0.999 & -0.997 \\ -0.999 & 1 & -0.996 \\ -0.997 & -0.996 & 1 \end{pmatrix} \tag{1.1.10}$$

This simple to compute information shows that, for the second case, the parameters are strongly correlated and the model is not identifiable.

1.2 Modelling of Developmental Regulatory Networks

Different formalisms have been used to describe genetic regulatory networks. Among them are directed graphs, Bayesian networks, ODEs, PDEs, stochastic equations, qualitative equations, Boolean networks, etc. For a review of various models and simulation aspects we refer to [27]. Here we restrict ourselves to some models using differential equations.

1.2.1 Connectionist model of development

Assume that N_c cells are arranged in a row equally distanced from each other. Consider the gene network consisting of N_g genes which regulate each other. Assume that these genes are additionally regulated by N_e external inputs, like maternal protein. Denote the time-varying concentration of the product of gene a and the concentration of the external protein e in cell i by g_i^a and g_i^e, respectively. The connectionist model describes the change in concentrations of each

gene product in each cell over time by the following system of ODEs [11]:

$$\begin{aligned} \frac{dg_i^a}{dt} &= R_a \Phi_a \left(\sum_{b=1}^{N_g} W_a^b g_i^b + \sum_{e=1}^{N_e} E_a^e g_i^e + h_a \right) && \text{(genetic regulation)} \\ &\quad -\lambda_a g_i^a && \text{(decay)} \\ &\quad +D_a \left(g_{i+1}^a - 2g_i^a + g_{i-1}^a \right) && \text{(diffusion)} \end{aligned} \tag{1.2.1}$$

As can be seen from the model, the change in gene product concentration is supposed to depend on three factors.

The first term describes the genetic regulation. W and E are the matrices of genetic regulatory coefficients whose elements characterize the influence of gene product b and external protein e on gene product a, respectively. Regulatory parameters represent repression (if < 0), activation (if > 0) or no interaction (if ≈ 0). It is important to note that the regulatory parameters are independent of the cell number i, so the "machinery" for the genetic regulation is the same in every individual cell. h_a summarizes the effect of general transcription factors on gene a, and R_a is the maximum rate of synthesis from gene a. Φ is the sigmoid function

$$\Phi(x) = \frac{1}{2} \left(\frac{x}{\sqrt{x^2 + 1}} + 1 \right). \tag{1.2.2}$$

Its aim is to prohibit negative influence by inhibitors and to provide saturation for activators. The second term describes the decay of gene products, and the third term the exchange of gene products between neighboring cells (diffusion).

The connectionist model (1.2.1) has been used to simulate the gap gene network in the early development of the *Drosophila* [10, 12–15]. For this system also the continuous version of (1.2.1), given by the reaction-diffusion PDE system

$$\begin{aligned} \frac{\partial c^a(x,t)}{\partial t} &= R_a \Phi_a \left(\sum_{b=1}^{N_g} W_a^b c^b(x,t) + \sum_{e=1}^{N_e} E_a^e g^e(x,t) + h_a \right) \\ &\quad -\lambda_a c^a(x,t) + D_a \frac{\partial^2 c^a(x,t)}{\partial x^2}, \end{aligned} \tag{1.2.3}$$

has been used [17].

Model (1.2.1), despite of its simplicity, has been proven to be a suitable tool to study the gap gene system. This is apparently due to an exceptional property of the *Drosophila* blastoderm. At the early developmental stages when the segment determination occurs the blastoderm is a syncytium, so that nuclei are not yet surrounded by membranes and therefore, the cell-cell signaling can be neglected [12]. Cellularization of the syncytial embryo occurs at later stages when the segments are determined. Therefore, the simple connectionist model (1.2.1) is sufficient to describe such a biological system in a consistent way.

Finally, we note that the model (1.2.1) is formulated for the case when cells are arranged in a row, i.e. when the gene expression pattern is a function of position along one spatial axis. However, this is not a severe restriction. The model can be easily extended to higher-dimensional spatial domains.

1.2.2 Cell-based model

Here, we propose a cell-based model for simulating regulatory networks that is capable of quantitatively reproducing spatial and temporal gene expression patterns in developmental processes. This model is a generalization of (1.2.1). It distinguishes between the genetic regulation which takes place inside the cells (species g) and the diffusion of gene products (species c) through the organism. Mathematically speaking this amounts to a continuum-discrete hybrid model where discrete objects exchange species with the surrounding environment modelled as a continuum. Inside the cells one has genetic regulation and decay (biochemical reactions). We assume that the concentration inside these cells is homogeneous (no diffusion). Outside the cells species only diffuse and decay. The exchange of gene products between the cells and the extracellular matrix is described by a secretion/absorption function. The complete model reads:

Intracellular For the gene products g in all cells $i = 1, \ldots, N_c$

$$\begin{aligned} \frac{dg_i^a(t)}{dt} &= R_a \Phi_a \left(\sum_{b=1}^{N_g} W_a^b g_i^b + \sum_{e=1}^{N_e} E_a^e g_i^e + h_a \right) && \text{(genetic regulation)} \\ &\quad -\lambda_a g_i^a && \text{(decay)} \\ &\quad -\frac{1}{|\partial V_i|} \int_{\partial V_i} S_a \left(g_i^a, c^a(\mathbf{x}(S), t) \right) dS, && \text{(secretion/absorption)} \end{aligned} \tag{1.2.4}$$

where ∂V_i denotes the surface of the cell i.

Extracellular matrix For the gene products c in the domain Ω

$$\begin{aligned} \frac{\partial c^a(\mathbf{x}, t)}{\partial t} &= \nabla \cdot (D_a \nabla c^a) && \text{(diffusion)} \\ &\quad -\lambda_a' c^a && \text{(decay)} \\ &\quad +\sum_{i=1}^{N_c} \frac{V_i}{|\partial V_i|} \delta(\partial V_i, \mathbf{x}) S_a(g_i^a, c^a(\mathbf{x}, t)) && \text{(secretion/absorption)} \end{aligned} \tag{1.2.5}$$

where V_i denotes the volume of the cell i, $\delta(\Gamma, \mathbf{x})$ is a delta function with support on Γ:

$$\int_{\Re^d} \delta(\Gamma, \mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int_{\Gamma} f(\mathbf{X}(S)) dS, \qquad \mathbf{X}(S) \in \Gamma.$$

The secretion/absorption function S_a models membrane processes. It can take a simple linear form, such as $S_a(g_i^a, c^a(\mathbf{x},t) = \alpha\,(g_i^a - c^a(\mathbf{x},t))$ with α being a constant, or it can be represented in some nonlinear form.

This model additionally can be coupled with a biomechanical model of cell aggregates, so that spatial and temporal morphogen gradients stemming from the genetic regulation simulation can influence the states of the cells within the cell aggregate, and/or lead to: cell migration, cell-layer-contraction, adhesion, growth, secretion of skeletal elements, and programmed cell death.

The numerical approximation of model (1.2.4)-(1.2.5) is not trivial. A substantial difficulty for the numerical integration of the coupled system lies in the computation of the solution of the PDE system (1.2.5). This is caused by the singular reaction source terms, singular in the sense that within the spatial domain the source is defined by a Dirac delta function expression on a lower dimensional surface.

1.2.3 Model with delays

In the derivation of model (1.2.1) it is assumed that protein synthesis occurs instantaneously. However, it is known that it takes time from the start of the transcription to the final moment when the protein is produced. In order to have a more realistic description of this procedure, time delay in protein production has to be incorporated in (1.2.1).

The following DDE model has been proposed in [28]:

$$\begin{aligned}\frac{dg_i^a}{dt} &= R_a\;\theta(t-\tau_a)\;\Phi_a\left(\sum_{b=1}^{N_g} W_a^b g_i^b(t-\tau_a) + \sum_{e=1}^{N_e} E_a^e g_i^e(t-\tau_a) + h_a\right)\\ &\quad -\lambda_a g_i^a(t) + D_a\left(g_{i+1}^a(t) - 2g_i^a(t) + g_{i-1}^a(t)\right),\end{aligned} \tag{1.2.6}$$

where τ_a is the time lag representing delay in the protein production corresponding to gene a, $\theta(t) = 0$ during mitosis (no synthesis) and $\theta(t) = 1$ otherwise. Using this model for the gap gene network in the *Drosophila* embryo, fitting the gap gene data has not given satisfactory results. It is not clear, whether that failure is due to the model itself which is still missing important steps in the protein production process or due to identifiability issues.

A new model has been developed [29], which distinguishes between time delays for transcription and translation. Additionally, this model includes the mRNA concentrations as state variables. The change in mRNA concentrations g_i^a is

prescribed by the DDEs:

$$\frac{dg_i^a}{dt} = R_a\ \theta(t-\tau_a^\alpha)\ \Phi_a\left(\sum_{b=1}^{N_g} W_a^b G_i^b(t-\tau_a^\alpha) + \sum_{e=1}^{N_e} E_a^e G_i^e(t-\tau_a^\alpha) + h_a\right) - \lambda_a g_i^a(t), \tag{1.2.7}$$

where G_i^a denotes the concentration of the product of gene a in nucleus i and τ_a^α accounts for transcriptional delay. The change in protein concentrations is given by:

$$\frac{dG_i^a}{dt} = \alpha_a g_i^a(t-\tau_a^\beta) - \tilde{\lambda}_a G_i^a(t) + D_a\left(G_{i+1}^a(t) - 2G_i^a(t) + G_{i-1}^a(t)\right), \tag{1.2.8}$$

where τ_a^β accounts for translational delay.

Both delay models (1.2.6) and (1.2.7)-(1.2.8) have a discontinuous right-hand side. As a consequence, their solutions are nonsmooth in parameters and therefore they need special attention, especially regarding the validity of the determinability analysis.

1.3 Outline of the thesis

This thesis is the compilation of four articles. These articles are self-contained and the corresponding chapters can be read separately. As we already emphasized, modelling and inferring regulatory networks is the central study subject in this thesis.

Chapter 2: *Parameter estimation and determinability analysis applied to Drosophila gap gene circuits*, M. Ashyraliyev, J. Jaeger, J.G. Blom, BMC Systems Biology 2:83 (2008).

In this chapter, we present the methodology for parameter estimation and the determinability analysis. We apply that to study determinability of regulatory interactions in the gap gene network in early *Drosophila* embryos using a standard connectionist model (see Section 1.2.1 of this chapter).

Chapter 3: *Parameter estimation for a model of gap gene circuits with time-variable external inputs in Drosophila*, M. Ashyraliyev, CWI Report, MAS-E0904 (2009). This report contains background for the paper *Gene circuit analysis of the terminal gap gene huckebein* by M. Ashyraliyev, K. Siggens, H. Janssens, J.G. Blom, M. Akam, J. Jaeger accepted for publication in PLoS Computational Biology.

In this chapter, we further investigate the gap gene system by reformulating the network. In contrast to the gap gene network considered in Chapter 2, we implement the products of upstream genes as time-variable external inputs, including a protein Huckebein which has been missing in the network in all previous studies of the gap gene system. We show that with the reformulated network the

determinability of the regulatory parameters significantly improves. Additionally, we illustrate that for data fitting the Weighted Least Squares (WLS) sum is a more suitable measure than the Ordinary Least Squares (OLS) sum which has been used in all previous studies.

Chapter 4: *On the numerical solution of diffusion-reaction equations with singular source terms*, M. Ashyraliyev, J.G. Blom, J.G. Verwer, J. Comp. Appl. Math 216, pp. 20-38 (2008).

In this chapter, we present a numerical study for reaction-diffusion problems having singular reaction source terms. These type of problems arise when the cell-based model is used for simulating regulatory networks (see Section 1.2.2 of this chapter). We emphasize that the focus of Chapter 4 is entirely numerical and that in this chapter no genuine biological model is used.

Chapter 5: *On parameter estimation for delay models with discontinuous right-hand sides*, M. Ashyraliyev, CWI Report, MAS-E0908 (2009).

In the final Chapter 5, we study delay models with discontinuous right-hand side. These type of problems arise when the model of regulatory networks incorporates a delay in the protein production (see Section 1.2.3 of this chapter).

Chapter 2

Parameter estimation and determinability analysis applied to *Drosophila* gap gene circuits

Abstract

Background: Mathematical modeling of real-life processes often requires the estimation of unknown parameters. Once the parameters are found by means of optimization, it is important to assess the quality of the parameter estimates, especially if parameter values are used to draw biological conclusions from the model.

Results: In this paper we describe how the quality of parameter estimates can be analyzed. We apply our methodology to assess parameter determinability for gene circuit models of the gap gene network in early *Drosophila* embryos.

Conclusions: Our analysis shows that none of the parameters of the considered model can be determined individually with reasonable accuracy due to correlations between parameters. Therefore, the model cannot be used as a tool to infer quantitative regulatory weights. On the other hand, our results show that it is still possible to draw reliable qualitative conclusions on the regulatory topology of the gene network. Moreover, it improves previous analyses of the same model by allowing us to identify those interactions for which qualitative conclusions are reliable, and those for which they are ambiguous.

2.1 Background

Many real-life processes can be modeled by non-linear Ordinary Differential Equations (ODEs) or Partial Differential Equations (PDEs). In developmental biology, for instance, systems of reaction-diffusion equations are used to model spatio-temporal patterns of gene expression [30]. A common difficulty is that the model equations usually have a large number of unknown parameters, such as weights for regulatory interactions, diffusion coefficients, decay and reaction rates, etc. Sometimes, it is feasible to determine the missing parameters experimentally, but in most cases this is difficult or even impossible. However, one can usually measure other quantities involved in the model. For instance, experimentalists can quantify mRNA or protein concentrations using microarrays, quantitative PCR, in situ hybridization or immunofluorescence. Unknown model parameters can then be found by parameter estimation techniques based on fitting the model solution to the measured gene expression data.

Once the parameter estimates have been computed, it is very important to know how reliable they are. For this, confidence regions can be determined, which allow us to assess the quality of the parameter estimates. Whether the parameters for the mathematical model can be found assuming that sufficient and error-free data is available is the subject of *a priori identification* or *structural identifiability* analysis. Once the parameter estimates have been computed, an *a posteriori* or *practical identifiability* study can show how well the parameters have been determined given a data set that is possibly sparse and noisy. For the subject of structural and practical identifiability we refer to [1, 21, 23] and references therein. Ideally, one would wish to determine all parameters accurately enough. In practice, however, this is usually not possible and one has to face an uncertainty in the parameter values. This can be due to several reasons: First, the model could be 'wrong'. In this paper, we do not focus on this possibility assuming that the 'right' model is available (i.e. a model which represents the underlying mechanism of the modeled process accurately and correctly). Second, the data used for fitting could be insufficient or too noisy. Finally, a recent study by Gutenkunst et al. [5] revealed that even if a correct model is used with a comprehensive set of data, many models used in systems biology still exhibit parameter 'sloppiness'. This means that some model parameters can be determined with great certainty ('stiff' parameters), while estimates of other ('sloppy') parameters can vary by orders of magnitude without significantly influencing the quality of the fit. Identifiability is a mathematical notion. For biological implications the precise values of parameters are not always important as long as they have certain characteristics, like being (roughly) positive, negative or zero. If *a posteriori* analysis results in a parameter uncertainty range which lies in the characteristic range we call this parameter *determinable*. Note that for those parameters which have to be determined quantitatively, i.e. having no specific characteristics, determinability refers to *a posteriori* identifiability.

Parameter sloppiness implies that very different sets of estimated parameters

can lead to accurate model predictions. Therefore, it is not a serious problem if the main purpose of a model is to predict the dynamical behavior of the system, and little significance is attributed to parameter values. This is the case for all models considered by Gutenkunst et al. [5]. Parameter sloppiness becomes much more problematic, however, when models are used explicitly to extract biological information from estimated parameter values. In particular, this affects attempts at reverse engineering gene regulatory networks underlying cellular or developmental processes, where models are used to infer regulatory interactions—and hence regulatory network topology—from quantitative gene expression data.

As a case study, we consider the gap gene system of the vinegar fly *Drosophila melanogaster*. Gap genes constitute the first step in a regulatory cascade that leads to the determination of body segment positions along the major (or anterior-posterior, A–P) body axis during early *Drosophila* development [8]. The biological function of the gap gene system is to interpret long-range protein gradients implemented by the products of the maternal co-ordinate genes (e.g. *bicoid (bcd)*, *hunchback (hb)* and *caudal (cad)*; see [10, 12, 31] and references therein). Zygotic gap genes, such as *hb*, *Krüppel (Kr)*, *knirps (kni)* and *giant (gt)*, are activated or repressed by these maternal gradients, which establishes their expression in broad, overlapping regions of the embryo. These spatial domains of gap gene expression are stabilized and refined by gap-gap cross-repression. In turn, gap genes are involved in regulation of pair-rule and segment-polarity genes, the latter of which establish a segmental pre-pattern of gene expression by the onset of gastrulation.

The gap gene system is one of the best characterized developmental gene networks available today. It has been studied extensively using genetic and molecular approaches (see [10] and references therein). More importantly for our purposes, quantitative expression data are available for all relevant maternal co-ordinate and gap genes [18, 19], and those data have been used to infer regulatory interactions between gap genes using different global and local optimization strategies [10, 12–14]. In this study, we use parameter values from these earlier studies as starting points for local optimization to obtain a large set of parameter estimates. We then apply analysis of confidence regions to those parameter sets to establish how well these estimates can be determined based on the available experimental data. We discuss the implications our results have for modeling of the gap gene system and for the biological interpretation of estimated parameter values. Finally, we note that the analysis can easily be adapted to other systems, and we strongly recommend its use to systems biology models in which large emphasis is put on the biological interpretation of estimated parameter values.

2.2 Methods

We consider a model given by the system of ODEs of the general form:

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{f}(t, \mathbf{y}, \theta), & 0 < t \leq T, \\ \mathbf{y}(t, \theta) = \mathbf{y}_0(\theta), & t = 0. \end{cases} \tag{2.2.1}$$

Here the m-dimensional vector θ contains all unknown parameters, $\mathbf{y}$ is an n-dimensional state vector, and $\mathbf{f}$ is a given vector function, differentiable with respect to t, $\mathbf{y}$ and θ. When components of the initial state vector $\mathbf{y}_0$ are not known, they are considered as unknown parameters. Thus, $\mathbf{y}_0$ may depend on θ.

As mentioned above, we assume that (2.2.1) is the 'right' model for the problem we are interested in, implying that (2.2.1) is a sufficiently accurate mathematical description approximating reality. This means that all relevant knowledge about the modeled processes is incorporated correctly in the vector function $\mathbf{f}$. Thus, the only uncertainty in (2.2.1) is the vector of unknown parameters θ. Furthermore, it means that there exists a 'true' value θ^* for the parameters θ such that (2.2.1) represents reality. Therefore, in principle, all unknown parameters can be determined if sufficient and accurate enough data are available.

Quantities that can be experimentally measured are called *observables*. The theory of identifiability holds in general for observables being a combination of state variables. However, for the sake of simplicity we consider here a particular case when only the components of state vector are measured. Let us assume that for fitting (2.2.1) there are N measurements available. Each measurement, which we denote by $\tilde{y}_i$, is specified by the time t_i when the c_i-th component of the state vector $\mathbf{y}$ is measured. The corresponding model value obtained from (2.2.1) is denoted by $y_{c_i}(t_i, \theta)$. The assumptions outlined above imply that the difference $|\tilde{y}_i - y_{c_i}(t_i, \theta^*)|$ is solely due to experimental error. We denote the vector of discrepancies between the theoretical values and the measured values by $\mathrm{Y}(\theta)$. Then the least squares estimate $\hat{\theta}$ of the parameters is the value of θ that minimizes the sum of squares [22, 32]

$$S(\theta) = \sum_{i=1}^{N} (y_{c_i}(t_i, \theta) - \tilde{y}_i)^2 = \mathrm{Y}^T(\theta)\mathrm{Y}(\theta). \tag{2.2.2}$$

We note that (2.2.2) is an appropriate measure under certain assumptions only, which we will discuss below. Other measures might be used when these assumptions do not hold.

2.2.1 Parameter Estimation by the Levenberg-Marquardt Method

There exist a number of different optimization techniques for parameter estimation. The choice of technique usually depends on the type of model equations

(deterministic or stochastic), on the number of unknown parameters (moderate or large), as well as on the dependence of model solutions on parameters (linear or nonlinear, continuous or discontinuous). A model (2.2.1), being nonlinear in θ, leads in general to a least squares problem (2.2.2) that has several minima. Firstly, because the problem has more than one solution. And secondly, because the fitness function (2.2.2) can have several minima that are not corresponding to the lowest value of the fitness landscape, so-called local minima. *Local search methods*, like Levenberg-Marquardt (LM), can get easily trapped in one of the local minima rather than finding a global minimum. To explore the whole search space one needs *global search methods*, like the Evolutionary algorithm used in [14]. Unfortunately, these methods converge very slow once near a minimum. For nonlinear least-squares problems gradient-based methods are efficient optimizers [2] once a sufficiently good initial guess for the parameter values is available. In this paper we use as initial guesses the solutions from the global search in [14] and the LM method [26] for the local optimization. In this way, we believe that the chance of missing the global minimum is small and the determination of all the minima is precise and fast. For a survey on optimization methods in biochemical models we refer to [21, 33].

In general, any gradient-based optimization procedure seeks a correction $\delta\theta$ for the parameter vector, such that $S(\theta+\delta\theta) \leq S(\theta)$ holds. The LM method [26] determines the correction as the solution of the equations

$$\left(J^T(\theta)J(\theta) + \lambda I_m\right)\delta\theta = -J^T(\theta)\mathrm{Y}(\theta), \tag{2.2.3}$$

where $\lambda \geq 0$ is a control parameter (see below), I_m is the identity matrix of size m and the Jacobian $J(\theta) = \frac{\partial \mathrm{Y}(\theta)}{\partial \theta}$ is the so-called 'sensitivity' matrix of size $N \times m$. The entry $J_{i,j}$ in $J(\theta)$ shows how sensitive the model response is at the i-th data point for a change in the j-th parameter. The LM method can be seen as the combination of two gradient-based approaches: Gauss-Newton and steepest descent [2]. If $\lambda = 0$ in (2.2.3), it coincides with the Gauss-Newton method. However, when the matrix $J^T(\theta)J(\theta)$ is (almost) singular, to solve (2.2.3), λ has to be positive and for large λ the LM method approaches the steepest descent method. During the optimization λ is adapted such that the algorithm strives to exploit the fast convergence of the Gauss-Newton method whenever this is possible [26, 34].

In order to solve (2.2.3), the singular value decomposition (SVD) [35] of the matrix $J(\theta)$ can be used, i.e.

$$J(\theta) = U(\theta)\ \Sigma(\theta)\ V^T(\theta), \tag{2.2.4}$$

where $U(\theta)$ is an orthogonal matrix of size $N \times m$, such that $U^T(\theta)U(\theta) = I_m$, $V(\theta)$ is an orthogonal matrix of size $m \times m$, such that $V^T(\theta)V(\theta) = V(\theta)V^T(\theta) = I_m$, and $\Sigma(\theta)$ is a diagonal matrix of size $m \times m$ which contains all singular values σ_i in non-increasing order. Then the correction $\delta\theta$ can be found as

$$\delta\theta = -V(\theta)\ \left(\Sigma^2(\theta) + \lambda I_m\right)^{-1}\Sigma(\theta)\ U^T(\theta)\ \mathrm{Y}(\theta). \tag{2.2.5}$$

Later, when we study the reliability of the parameters computed, the SVD will play an important role again.

In order to execute an LM optimization step, the vector of discrepancies $\mathrm{Y}(\theta)$, the matrix $J(\theta)$ and its SVD have to be evaluated for each new estimate of θ. For this purpose, for Y and the entries of J one needs to resolve (2.2.1) and the additional system of variational equations ($i = 1, 2, \ldots, m$)

$$\begin{cases} \dfrac{\partial}{\partial t}\dfrac{\partial \mathbf{y}}{\partial \theta_i} = \dfrac{\partial \mathbf{f}}{\partial \theta_i} + \dfrac{\partial \mathbf{f}}{\partial \mathbf{y}}\dfrac{\partial \mathbf{y}}{\partial \theta_i}, & 0 < t \le T, \\ \dfrac{\partial \mathbf{y}(t,\theta)}{\partial \theta_i} = \dfrac{\partial \mathbf{y}_0(\theta)}{\partial \theta_i}, & t = 0, \end{cases} \tag{2.2.6}$$

respectively. We note that the costs for performing the SVD and computing the correction (2.2.5) are negligible in comparison with the computational costs for solving (2.2.1) and (2.2.6).

Thus, a single LM step requires the numerical solution of $m + 1$ coupled systems, each one consisting of n ODEs. Fortunately, these systems are coupled in a special way, namely, for each $i = 1, 2, \ldots, m$, system (2.2.6) is a system of linear ODEs for $\frac{\partial \mathbf{y}}{\partial \theta_i}$, coupled only with (2.2.1). The system of equations (2.2.6) has the same stiffness as (2.2.1), so for numerical stability the same step size can be used for the time integration of (2.2.1) and (2.2.6) (note that ODE stiffness is determined by the eigenvalues of the Jacobian matrix $\frac{\partial \mathbf{f}}{\partial \mathbf{y}}$ and is not related to parameter stiffness as described above). Therefore, the one-way coupling can be used to solve (2.2.1) and (2.2.6) efficiently.

Numerical integration of (2.2.1) and (2.2.6) requires a fast and reliable ODE solver. Search in the parameter space may lead to some values of θ such that the systems of ODEs become stiff [36]. It is well known that for stiff ODE systems explicit schemes can give rise to numerical instability or alternatively require extremely small time steps. Therefore, an implicit scheme is the best choice for time integration for stability reasons. Using an implicit scheme allows us to exploit the specific coupling between (2.2.1) and (2.2.6) in an efficient way. At each time step τ integrating first (2.2.1) provides the solution vector $\mathbf{y}$. This requires the LU decomposition of the Jacobian matrix $I_m - \tau\frac{\partial \mathbf{f}}{\partial \mathbf{y}}$. Using this LU decomposition the calculation of $\frac{\partial \mathbf{y}}{\partial \theta_i}$ from (2.2.6) reduces to a simple forward substitution and backsubstitution. In our simulations we used tailor-made code [25] based on the implicit multistep Backward Differentiation Formulas (BDF) [37].

When the unknown parameters have to obey certain constraints—linear or non-linear—some additional work is needed. If the correction $\delta\theta$ found by (2.2.5) leads to violation of some constraints, then by the introduction of Lagrange multipliers a modified correction can be found, which fits all constraints. For the constrained minimization problem we refer the reader to [25].

For additional modeling and numerical aspects of this method we refer the reader to Appendix.

2.2.2 Statistical Analysis of Parameter Estimates

Above we used θ^* to denote the 'true' parameter vector, for which (2.2.1) describes reality with sufficient accuracy, and by $\hat{\theta}$ we denote the parameter vector which minimizes (2.2.2). Even having a 'right' model and an estimate $\hat{\theta}$ for the parameter vector which fits the data well, does not mean that the whole modeling problem is resolved successfully. It is important to know how reliable the obtained estimate is. This is the subject of *a posteriori* identifiability analysis [1, 23, 24]. One way to look at this is inspecting the fitness landscape $S(\theta)$ in the neighbourhood of $\hat{\theta}$. If it is a sharp trough then the true parameter vector θ^* and the obtained minimum $\hat{\theta}$ are close. If it is flat in one or more directions, like the surface for a 2-parameter case in Figure 2.1(a), then the minimum found can be far apart from the true parameter vector. Near the minimum, where the gradient of $S(\theta)$ vanishes, this surface is approximated by the second derivative or Hessian of $S(\theta)$. If the model is linear in the parameters the Hessian is equal to $J^T J$. This linearity assumption and some statistical analysis underlie the following [22, 32, 36].

We assume that the measurement errors in $\tilde{y}_i$ are independent of each other and normally distributed and that the error distributions have zero mean and constant standard deviation σ. Then, $\hat{\theta}$ is a maximum likelihood estimate [22, 32]. By assumption the model with the 'true' solution θ^* describes reality, thus

$$\tilde{y}_i \approx y_{c_i}(t_i, \theta^*) + \epsilon_i, \quad i = 1, 2, \ldots, N,$$

where ϵ_i are the measurement errors, for which

$$\hat{\theta} - \theta^* \sim N_m \left(0, \sigma^2 \left(J^T(\hat{\theta}) J(\hat{\theta})\right)^{-1}\right) \tag{2.2.7}$$

holds approximately [22]. Here $N_m(\cdot, \cdot)$ denotes the m-dimensional multivariate normal distribution. Notice that (2.2.7) holds exactly when $\mathbf{y}$ is linear in θ. Next we can define a region around $\hat{\theta}$ in which the 'true' parameter vector θ^* lies with a certain probability $1 - \alpha$. This $(1 - \alpha)$-*confidence region* is determined by the inequality

$$(\theta^* - \hat{\theta})^T \left(J^T(\hat{\theta}) J(\hat{\theta})\right) (\theta^* - \hat{\theta}) \leq \frac{m}{N - m} S(\hat{\theta}) F_\alpha(m, N - m), \tag{2.2.8}$$

where $F_\alpha(m, N - m)$ is the upper α part of Fisher's distribution with m and $N - m$ degrees of freedom. Geometrically these confidence regions are given by the contours of $S(\hat{\theta})$ (for linear models), cf. Figure 2.1(a).

The ellipsoid defined by (2.2.8), is centered at $\hat{\theta}$ and has its principal axes directed along the eigenvectors of $J^T(\hat{\theta}) J(\hat{\theta})$. Using the SVD (2.2.4) for $J(\hat{\theta})$, we get

$$J^T(\hat{\theta}) J(\hat{\theta}) = V(\hat{\theta}) \Sigma^2(\hat{\theta}) V^T(\hat{\theta}),$$

and the eigenvectors of $J^T(\hat{\theta}) J(\hat{\theta})$ are the columns of the matrix $V(\hat{\theta})$. So, the ellipsoid has its principal axes directed along the column vectors of the matrix

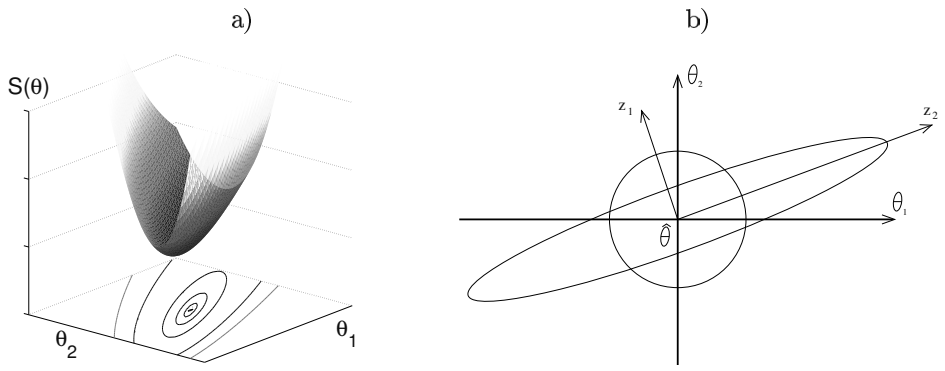


Figure 2.1: A graphical representation of the confidence region in the 2-dimensional case. a) Example of a fitness landscape $S(\theta)$ for a linear model and the contours corresponding to different confidence regions. b) Example of an ellipsoidal confidence region and an accuracy sphere, where principal axes of the ellipsoid, z_1 and z_2, define the new coordinate system which is a rotation of actual parameter space $(\theta_1;\theta_2)$. Clearly, z_1 is well-determined, while z_2 is not.

$V(\hat{\theta})$. Moreover, the radii along these principal axes are inversely proportional to the corresponding singular values σ_i, the diagonal elements of $\Sigma(\hat{\theta})$. This all can be seen by using the following transformation (rotation)

$$\mathbf{z} = V^T(\hat{\theta})(\theta^* - \hat{\theta}), \tag{2.2.9}$$

yielding

$$(\theta^* - \hat{\theta})^T \left(V(\hat{\theta})\Sigma^2(\hat{\theta})V^T(\hat{\theta})\right)(\theta^* - \hat{\theta}) = \mathbf{z}^T\Sigma^2(\hat{\theta})\mathbf{z} = \sum_{i=1}^{m} \sigma_i^2 z_i^2. \tag{2.2.10}$$

On the other hand, since $S(\hat{\theta})/(N-m)$ is an unbiased estimator of σ^2, the equation for the ellipsoid can be rewritten as

$$\sum_{i=1}^{m} \sigma_i^2 z_i^2 = r_\sigma^2, \tag{2.2.11}$$

where $r_\sigma^2 \approx m\sigma^2 F_\alpha(m, N-m)$ is proportional to the variance in the measurement errors. This form is more convenient to deal with because $\mathbf{z}$ can be considered as a set of uncorrelated variables, and once the conclusion has been drawn for the determinability of $\mathbf{z}$, the problem can be transformed back, revealing us the quality of $\hat{\theta}$.

Now, we assume that the model (2.2.1) is properly scaled, such that all parameter values are of the same order of magnitudes, and that we are interested

only in the first few digits of the parameter values. Let us introduce the sphere given by

$$\sum_{i=1}^{m} z_i^2 = r_\epsilon^2,$$

where r_ϵ defines the level of accuracy one desires for the parameter estimates. For instance, if the parameters are of order $O(1)$ and one is interested only in the first two digits to the right of the decimal point, then $r_\epsilon = 0.01$. In order to be able to determine z_i accurately enough, the radius along the ellipsoid's i-th principal axis shouldn't exceed the radius of the sphere, which leads us to the following inequality

$$\sigma_i \geq \frac{r_\sigma}{r_\epsilon}. \tag{2.2.12}$$

A graphical representation of the ellipsoid and the sphere for the 2-dimensional case is given in Figure 2.1(b).

If only the first k largest singular values satisfy (2.2.12), then only the first k entries of $\mathbf{z}$ are estimated with the required accuracy and no sufficient information is available for the remaining components of $\mathbf{z}$. Each of the first k entries of $\mathbf{z}$ defines a parameter or a linear combination of parameters which is well-determined. If a principal axis of the ellipsoid makes a significant angle with the axes in parameter space (i.e., there exists more than one significant entry in the eigenvector), this implies correlation between parameters in $\hat{\theta}$.

To summarize, the level of noise in the data in combination with the accuracy requirement for the parameter estimates, defines the threshold for significant singular values in the matrix Σ. The number of singular values exceeding this threshold determines the number of parameter relations that can be derived from the experiment. How these relations relate to the individual parameters is described by the corresponding columns in the matrix V. The largest entries in these columns indicate the well-determined parameters. This method is illustrated on the basis of a simple enzymatic reaction in [21].

Finally, (2.2.11) indicates that having, for instance, two times more accurate data so that the standard deviation σ is halved, will decrease the radii along the ellipsoid's principal axis by a factor of 2. Therefore, in case of very small singular values σ_i (i.e. strongly elongated ellipsoids) more accurate data obtained by the experimentalist will not improve the quality of the corresponding parameter estimates by much. In such a case, one certainly needs additional measurements of a different type (e.g. different components, different time points, or in the case of PDEs different spatial points).

Another way of assessing the information from the confidence region is by looking at confidence intervals of the parameter estimates $\hat{\theta}_i$ $(i = 1, 2, \ldots, m)$. From (2.2.8) one can derive dependent and independent confidence intervals. The *dependent confidence interval* is the intersection of the ellipsoidal region with the

i-th parameter axis

$$\left\{ \theta_i : |\theta_i - \hat{\theta}_i| \leq \frac{r_\sigma}{\sqrt{\left(V(\hat{\theta})\Sigma^2(\hat{\theta})V^T(\hat{\theta}) \right)_{ii}}} \right\}, \tag{2.2.13}$$

i.e. one assumes that all other parameters are exactly determined. The *independent confidence interval* is the projection of the ellipsoidal region onto the i-th parameter axis

$$\left\{ \theta_i : |\theta_i - \hat{\theta}_i| \leq r_\sigma \sqrt{\left(V(\hat{\theta})\Sigma^{-2}(\hat{\theta})V^T(\hat{\theta}) \right)_{ii}} \right\}. \tag{2.2.14}$$

Clearly, small independent confidence intervals for $\hat{\theta}_i$ indicate that it is well-determined. However, in some cases considering only individual confidence intervals can be misleading. For instance, in the presence of strong correlations between parameters, the dependent confidence intervals underestimate the confidence region while the independent confidence intervals overestimate it.

From (2.2.7), the covariance matrix of $\hat{\theta}$ is given by

$$\sigma^2 \left(J^T(\hat{\theta})J(\hat{\theta}) \right)^{-1} = \sigma^2 V(\hat{\theta})\Sigma^{-2}(\hat{\theta})V^T(\hat{\theta}). \tag{2.2.15}$$

Then, by denoting $B(\hat{\theta}) = V(\hat{\theta})\Sigma^{-2}(\hat{\theta})V^T(\hat{\theta})$, the *correlation coefficient* between $\hat{\theta}_i$ and $\hat{\theta}_j$ can be computed by

$$\rho_{ij} = \frac{B_{ij}}{\sqrt{B_{ii}B_{jj}}}. \tag{2.2.16}$$

We note that by computing individual confidence intervals and correlations between parameters, one is not able to assess the determinability of linear combinations of parameters. This can be seen only by using the first approach, i.e. by inspection of the V and Σ matrix.

2.2.3 The Biological Test Problem: Gap Gene Circuits

We apply the methodology described above to assess parameter determinability of gene circuit models for the gap gene network in early *Drosophila* development. Here, we provide a brief outline of gap gene circuit models. More detailed information can be found in [10, 12, 38].

Segment determination occurs during the blastoderm stage of *Drosophila* development, between 1.5 and 3 hours after egg laying [7]. During this stage, the embryo consists of a syncytium; there are no cell membranes between nuclei. These nuclei constitute the basic objects of the model. They are arranged in a

row along the A–P axis. Nuclei divide rapidly and synchronously [39]. Periods between mitotic divisions are called cleavage cycles, where cycle n occurs between mitoses $n-1$ and n. The models considered here run from early cycle 13 ($t = 0.0$ min) to the onset of gastrulation at the end of cycle 14A ($t = 71.1$ min). Mitosis occurs at the end of cycle 13, between $t = 16.0$ min and $t = 21.1$ min [39].

Gene circuit models describe the change in concentrations of each gap gene product in each nucleus over time by the following system of ODEs

$$\frac{dg_i^a}{dt} = R_a \Phi \left(\sum_{b=1}^{N_g} W_a^b g_i^b + m_a g_i^{\mathrm{Bcd}} + h_a \right) - \lambda_a g_i^a + D_a \left(g_{i+1}^a - 2g_i^a + g_{i-1}^a \right). \tag{2.2.17}$$

a and b denote regulated genes and regulators, respectively. a and b are integer indices representing *cad*, *hb*, *Kr*, *kni*, *gt* as well as the terminal gap gene *tailless (tll)*. g_i^a denotes the concentration of the product of gene a in nucleus i. The Bcd gradient remains constant over time, and is not regulated by the other genes in the model. g_i^{Bcd} denotes the concentration of Bcd protein in nucleus i. $N_g = 6$ is the number of genes in the model (excluding Bcd), and the function

$$\Phi(x) = \frac{1}{2} \left(\frac{x}{\sqrt{x^2 + 1}} + 1 \right) \tag{2.2.18}$$

is a sigmoid regulation-expression function.

During mitosis, protein production is shut down. Nuclei divide instantaneously at the end of mitosis and the distance between them is halved. Gap gene circuits cover the region from 35% to 92% A–P position, which includes 30 (cycle 13) and 58 (14A) nuclei. Therefore, system (2.2.17) consists of 180 and 348 ODEs during cycles 13 and 14A, respectively. Initial conditions are prescribed by maternal gradients of Bcd, Cad and Hb, and zero levels for all other gene products. We use no-flux boundary conditions at $i = 0$ and $i = i_{\max}$.

In system (2.2.17) there are $m = 66$ unknown parameters. These include the genetic interconnection or regulatory weight matrix W of size $N_g \times N_g$ where the matrix elements W_a^b represent the regulation of gene a by gene b, while maternal coefficients m_a represent the regulatory effect of Bcd on gene a. Regulatory parameters represent repression (if < 0), activation (if > 0) or no interaction (if ≈ 0). Other parameters include promoter thresholds h_a, promoter strengths R_a, diffusion coefficients D_a, and decay rates λ_a. Estimates for these parameters have been obtained in previous studies by fits to quantitative expression data [19] using global search methods such as parallelized Lam Simulated Annealing [10, 12] or the Stochastic Ranking Evolution Strategy (followed by downhill simplex direct search) [14] and using a first-improvement local search method with randomized order of examination [13]. In the latter the initial parameter estimates are obtained by using a splitting strategy: parameters λ_a and D_a are estimated by assuming that the protein production is constant within certain spatio-temporal domains which reduces (2.2.17) to a system of linear equations uncoupled for each

gene (the boundaries of production domains are obtained from data); parameters in the nonlinear part of the model are estimated by fitting the production term in (2.2.17) with the data given as input, as closely as possible, to the quadrilateral production regions.

The data set used for model fitting consists of $N = 2702$ measurements of protein concentrations at nuclear resolution (using multi-channel immunofluorescent antibody assays; available online [19]). Measurements were taken at one time point during cycle 13 (T_0), and eight time points T_i ($1 \leq i \leq 8$) during cycle 14A (Figure 2.2a). Measurements for the concentrations of all gene products represented in the model at all time points are available, except for Cad at T_7 and T_8, and Tll before T_3. The level of measurement error in the data is less than 5%, see [40]. Each data point represents concentration values which have been averaged across 9–62 embryos. Therefore, from the Central Limit Theorem (CLT) we assume that the experimental errors are approximately normally distributed.

The quality of the parameter estimates is measured by the root mean square (RMS) of the discrepancy vector

$$RMS(\theta) = \sqrt{\frac{1}{N}\sum_{a=1}^{N_g}\sum_{j=0}^{N_t}\sum_{i=1}^{N_c}\alpha_j^a\left(g_i^a(T_j,\theta)_{model} - g_i^a(T_j)_{data}\right)^2}, \tag{2.2.19}$$

where $N_t = 8$ is the number of time classes, N_c is the number of nuclei and α_j^a is equal to zero for Tll at $j = 0, 1, 2$ and for Cad at $j = 7, 8$, and is equal to one otherwise. A solution is considered to be 'good' if $RMS < 12.0$ and if there are no visible pattern defects in the model response [10, 12–14]. It is important to note that the RMS only shows the quality of the fit of the model to the data but does not give any information about the quality of the parameter estimates. Our aim is to find the parameter estimates that give a good fit and to apply statistical analysis in order to investigate how reliable these estimates are.

The search space for parameters is defined by the linear constraints

$$10.0 \leq R_a \leq 30.0, \quad 0.0 < D_a \leq 0.3, \quad 5.0 \leq \frac{ln(2)}{\lambda_a} \leq 20.0, \quad a = 1, \ldots, N_g, \tag{2.2.20}$$

and by the nonlinear constraints

$$\sum_{b=1}^{N_g}\left(W_a^b g_{max}^b\right)^2 + \left(m_a g_{max}^{\mathrm{Bcd}}\right)^2 + (h_a)^2 \leq 10^4, \quad a = 1, \ldots, N_g, \tag{2.2.21}$$

where g_{max}^b and g_{max}^{Bcd} are the maximum values in the data set for proteins b and Bcd, respectively. Note that in [10, 12, 13] threshold parameters h_a for genes Kr, Kni, gt, and hb are fixed to negative values representing a constitutively repressed state for the corresponding genes [10, 12]. Fixing some parameters to specific values may severely restrict the search space leaving some solutions out of

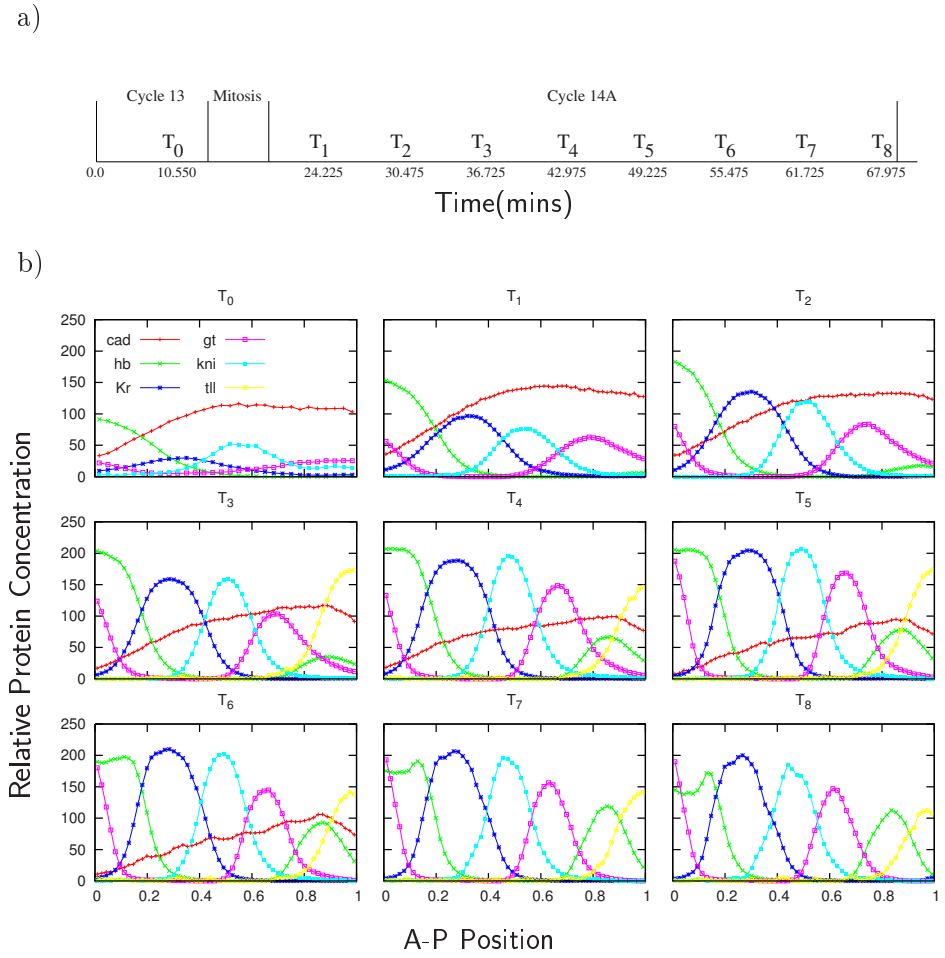


Figure 2.2: Data. a) Time axis and the points when measurements were taken: one in cycle 13 and eight in cycle 14A; the duration of mitosis is also indicated. b) Quantitative gene expression data at different time points. Graphs show relative protein concentration (with a range from 0 to 255 fluorescence units) plotted against position on the A–P axis (the trunk region of the embryo, from 35% to 92% A–P position is scaled to relative co-ordinates $[0, 1]$).

consideration. Contrary to their approach, we include threshold parameters for these genes in the search by putting the constraints $-10.0 \leq h_a \leq 0.0$.

In order to make the analysis of parameter estimation easier, we scale all parameters used in (2.2.17) in the following way:

$$\tilde{R}_a = 0.1R_a,\ \tilde{D}_a = 10D_a,\ \tilde{\lambda}_a = 10\lambda_a,\ \tilde{W}_a^b = 10^2 W_a^b,\ \tilde{m}_a = 10^2 m_a,\ \tilde{h}_a = h_a,$$

for all genes a and b. Note that the choice of the scaling factors for R_a, D_a, and λ_a is based on the search ranges of the corresponding parameters. The choice of the scaling factors for regulatory weights W_a^b and maternal coefficients m_a is based on the fact that the maximum level of protein concentration for all genes in the data set is of order $O(10^2)$. Thus, all scaled parameters are of order $O(1)$.

2.3 Results and Discussion

We use 80 different parameter sets, obtained by global search [14], as initial guess for the parameter values and apply the LM method to estimate all 66 unknown parameters of the gap gene circuit model (2.2.17), such that the state variables fit the given data (see Figure 2.2b), subject to (non)linear constraints (2.2.20)-(2.2.21). Once the parameters are estimated we apply our statistical analysis to asses the quality of the parameter estimates.

2.3.1 Optimization Results

Least squares estimation of the 66 parameters of the gap gene circuit model (full search case) using the LM method yields a significant decrease of the RMS (2.2.19) in all simulations (see Table 2.1). There are only 5 (out of 80) initial parameter sets with $RMS < 10.0$ (best fit: $RMS = 9.56$). After using the LM method there are 71 final parameter sets with $RMS < 10.0$, among which there are 64 with their RMS evenly distributed between 8.37 and 9.43. None of these low-scoring parameter sets show any visible patterning defects (see Figure 2.1 in Suppl. Mat. [6]), while most solutions with larger RMS do. As it is difficult to make a distinction between these 64 parameter sets based on RMS values and expression patterns only, we take all of them into account for our analysis. We note that there is no guarantee that a better solution might not have been missed by our parameter estimation procedure. However, since the initial points for LM search were found by a global search method [14], it is likely that the search space for unknown parameters is explored sufficiently enough.

Parameter estimates found by the LM method significantly improve solution fits found in previous studies (see Figure 2.3) [10, 12–14]. However, there are two problems, mentioned in [10, 12], that remain unsolved with the new parameter estimates. The first one concerns the artificially high level of gap gene expression during early cycle 13. The model responses are much larger than the data values yielding large positive discrepancies (Figure 2.1 in Suppl. Mat. [6]). This is

	RMS<10.0	10.0≤RMS<12.0	12.0≤RMS<14.0	RMS≥14.0
θ^{in}	5	36	21	18
$\hat{\theta}_{full}$	71	3	1	5
$\hat{\theta}_{fixed}$	63	7	2	8

Table 2.1: RMS distribution for parameter estimates. Entries in the table show the number of parameter estimates with corresponding ranges for RMS, where θ^{in} correspond to initial parameter estimate; $\hat{\theta}_{full}$ and $\hat{\theta}_{fixed}$ correspond to the parameter estimates after using the LM method in the full search case and the case of fixed promoter thresholds, respectively.

probably due to the lack of protein production delays in the model [10]. The second one concerns the incorrect shift of the posterior Hb domain (data not shown), which is due to the absence of the terminal gap gene *huckebein (hkb)* from our current models [10, 12].

Many parameters have a broad range of possible values, meaning that they are not uniquely determined (Figures 2.2-2.3 in Suppl. Mat. [6]). Classification of all parameter estimates for regulatory weights into 'activating', 'repressing' or 'no interaction' categories is shown in Figure 2.4(a). The resulting network topology is in very good agreement with the results obtained in [10, 12–14]. Specifically,

(A1) Cad and Bcd activate gap genes *hb*, *Kr*, *kni*, and *gt*;

(A2) gap genes *hb*, *Kr*, *kni*, and *gt* show auto-activation;

(A3) Tll represses gap genes *Kr*, *kni*, and *gt*;

(A4) gap genes with mutually exclusive expression domains strongly repress each other; these correspond to weights W_{gt}^{Kr}, W_{Kr}^{gt}, W_{hb}^{kni}, and W_{kni}^{hb}.

Previous results also suggested that pairs of overlapping gap genes (*hb* and *gt*, *gt* and *kni*, *kni* and *Kr*, as well as *Kr* and *hb*) either show no or weak repressive interactions among each other. Note that some of these weights differ slightly from earlier analyses [10, 14]. In all of these cases the difference is extremely slight and depends on the threshold chosen to categorize an interaction as 'very weak repression' or 'no interaction' (for example W_{kni}^{Kr} or W_{Kr}^{kni} in Figure 2.4(a); see also scatter plots in Figure 2.2 in Suppl. Mat. [6]). It is therefore unlikely that such differences are biologically significant. The only activation between overlapping gap genes is predicted for the effect of Gt on *hb*. In addition, we find that Kni activates *gt* in a majority of solutions. In both of these cases, the significance of the interactions does not lie in their weak activating effect (which has no discernible biological function), but rather in the absence of repression [10, 12].

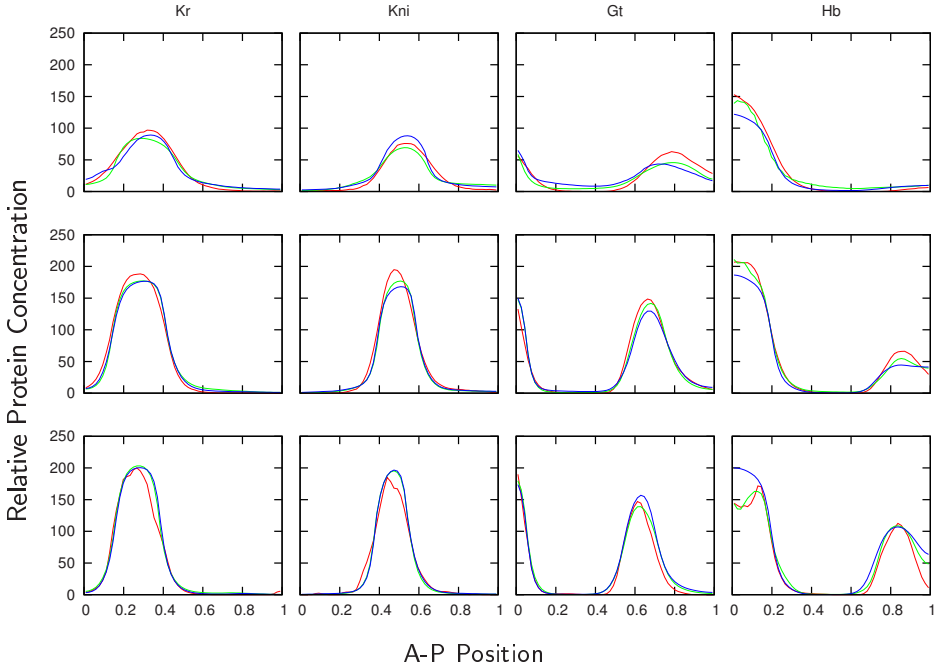


Figure 2.3: Model responses vs Data. Comparison between data (red lines), patterns obtained by a parameter set from [10] (blue lines) and patterns with a parameter set yielded from the LM search (green lines) for the expression of gap genes Kr, Kni, gt, and hb at early ($t = 24.225$ min, T_1, first row) mid- ($t = 42.975$ min, T_4, second row) and late ($t = 67.975$ min, T_8, last row) cycle 14A. Axes are as in Figure 2.2.

a)

	bcd	*cad*	*hb*	*Kr*	*gt*	*kni*	*tll*
cad	56/ 5/ 3	64/ 0/ 0	64/ 0/ 0	64/ 0/ 0	64/ 0/ 0	64/ 0/ 0	64/ 0/ 0
hb	0/ 0/64	0/ 1/63	0/ 0/64	0/56/ 8	1/26/37	64/ 0/ 0	3/37/24
Kr	0/ 0/64	0/ 0/64	20/44/ 0	0/ 4/60	64/ 0/ 0	38/26/ 0	64/ 0/ 0
gt	0/ 0/64	0/ 0/64	7/43/14	64/ 0/ 0	0/ 0/64	0/10/54	41/23/ 0
kni	5/ 4/55	0/ 6/58	64/ 0/ 0	33/31/ 0	46/18/ 0	0/ 0/64	64/ 0/ 0
tll	39/ 7/18	10/ 6/48	38/15/11	60/ 2/ 2	57/ 3/ 4	62/ 2/ 0	0/10/54

b)

	bcd	*cad*	*hb*	*Kr*	*gt*	*kni*	*tll*
cad	47/ 8/ 5	60/ 0/ 0	60/ 0/ 0	60/ 0/ 0	60/ 0/ 0	60/ 0/ 0	59/ 1/ 0
hb	0/ 0/60	0/ 0/60	0/ 0/60	0/60/ 0	2/23/35	60/ 0/ 0	3/57/ 0
Kr	0/ 0/60	0/ 0/60	18/42/ 0	0/ 1/59	60/ 0/ 0	28/32/ 0	60/ 0/ 0
gt	0/ 0/60	0/ 0/60	26/32/ 2	60/ 0/ 0	0/ 0/60	0/38/22	35/25/ 0
kni	3/ 1/56	0/ 0/60	60/ 0/ 0	23/33/ 4	33/27/ 0	0/ 0/60	58/ 2/ 0
tll	40/ 4/16	5/ 3/52	29/ 8/23	55/ 1/ 4	52/ 4/ 4	60/ 0/ 0	0/11/49

Figure 2.4: Regulatory topology of the gap gene network. Maternal coefficients and regulatory weight matrix for the gap gene system based on parameter sets found by the LM method: a) 64 solutions in the full search case; b) 60 sets in the case of fixed promoter thresholds. Triplets show the number of parameter sets in which a regulatory weight falls into one of the following categories: repression (values ≤ -0.005)/ no interaction (values between -0.005 and 0.005)/ activation (values ≥ 0.005). Based on the highest value in the triplets, the table is coloured such that the background colours represent activation (green), no interaction (light-blue), or repression (pink).

2.3.2 Parameter Determinability

We applied the statistical analysis introduced in the Methods section to the 64 parameter sets obtained by the LM method to asses the quality of our estimates. Ellipsoidal confidence regions corresponding to parameter estimates are given by (2.2.10). None of the parameter estimates lies in the ellipsoidal confidence regions of all other parameter sets. This does not necessarily imply that there is no unique 'true' solution for the parameter vector, since the ellipsoidal confidence regions—or at least some of them—may still have a non-empty intersection.

For each parameter set $\hat{\theta}$, the SVD (2.2.4) of the Jacobian $J(\hat{\theta})$ yields the matrices $V(\hat{\theta})$ and $\Sigma(\hat{\theta})$. In order to find the number of singular values in $\Sigma(\hat{\theta})$ satisfying the accuracy inequality (2.2.12), i.e. to determine how many (combinations of) parameters are determinable, we need to quantify r_ϵ and r_σ. We are interested only in the first decimal digit of the scaled parameters, and therefore we take $r_\epsilon = 0.1$. For $\alpha = 0.05$ we obtain $r_\sigma \approx 9.4\ RMS(\hat{\theta})$ (the choice of α does not make much difference here due to the large value of N).

Investigation of all parameter sets shows that on average, 15 singular values satisfy (2.2.12) meaning that at most 15 parameters or linear combinations of them can be determined with one digit accuracy. There is a set of parameters which have significant entries in the first 15 columns of all V matrices. It includes regulatory weights W_{Kr}^{cad}, W_{gt}^{cad}, W_{kni}^{cad}, W_{tll}^{cad}, W_{Kr}^{hb}, promoter thresholds h_{Kr}, h_{gt}, h_{tll}, decay rate λ_{cad}, and promoter strength R_{Kr}. However, inspection of the first 15 columns of the V matrices shows that there is not a single parameter which can be determined individually with the chosen accuracy. Thus, each column has a number of significant entries implying that the principal axis of the confidence ellipsoid is at an angle with the corresponding axes in parameter space. This indicates the presence of correlations between parameters.

Dependent and independent confidence intervals for each parameter set can be computed by (2.2.13) and (2.2.14), respectively. We check if the corresponding confidence intervals for regulatory weights fall entirely into the 'repression', 'no interaction', or 'activation' categories. Results in Figure 2.4(a) do not change when only dependent confidence intervals are taken into account. However, including independent confidence intervals one can no longer make correct qualitative conclusions about many of the entries in the regulatory weight matrix.

The sizes of the independent confidence intervals give an indication about the determinability of the corresponding regulatory weights. There is a set of eight regulatory weights which have relatively small confidence intervals for all 64 parameter sets (Figure 2.4 in Suppl. Mat. [6]). It includes W_{hb}^{Kr}, W_{hb}^{tll}, W_{Kr}^{hb}, W_{Kr}^{Kr}, W_{Kr}^{kni}, W_{gt}^{hb}, W_{gt}^{gt}, and W_{gt}^{kni}. For instance, Figure 2.5(a) shows the confidence intervals for W_{hb}^{Kr}. This regulatory weight is well determined qualitatively, i.e. the independent confidence intervals fall entirely into one category and therefore the type of the regulation can be concluded. The model predicts that Kr does not regulate *hb*. Note that the confidence intervals for these eight parameters in the scaled case are of order $O(10^{-1})$ and therefore they are not determinable with the

chosen accuracy level $r_\epsilon = 0.1$. In fact, they are determinable only if we choose $r_\epsilon = 1.0$.

The remaining regulatory weights have larger independent confidence intervals (Figure 2.4 in Suppl. Mat. [6]) and therefore they are not determined quantitatively. Among them are some regulatory weights for which qualitative conclusions can be deduced from the results. For example, panels (d) and (e) of Figure 2.5 show the confidence intervals for regulatory weights W^{hb}_{kni} and W^{Kr}_{gt}, respectively. Although these two regulatory weights are not determinable quantitatively, there is a qualitative difference between them. The independent confidence intervals for W^{Kr}_{gt} do not extend significantly into the positive part of the plane. Therefore, one can make a qualitative conclusion for this weight: the model predicts that Kr does not activate *gt*. Note that this is a weaker conclusion than predicting repression for this weight from Figure 2.4(a). In contrast, we cannot draw any qualitative conclusions about W^{hb}_{kni}. Thus, our analysis does not confirm the repression of *kni* by Hb inferred from Figure 2.4(a) (but does not contradict it either). To demonstrate that repression of *kni* by Hb is not strictly necessary to fit the data correctly, we fix this weight to zero while performing parameter estimation. The obtained parameter set has a $RMS = 9.24$ and produces patterns with no visible defects (Figure 2.7 in Suppl. Mat. [6]).

Based on the confidence intervals, we summarize the qualitative conclusions for the most important regulatory weights in the gap gene system:

(B1) Cad and Bcd do not repress gap genes *hb*, *Kr*, and *gt*; no conclusions can be made for regulation of *kni* by Cad and Bcd;

(B2) gap genes *hb*, *Kr*, *kni*, and *gt* do not show auto-repression;

(B3) Tll does not activate gap gene *gt*; no conclusions can be made for regulation of *Kr* and *kni* by Tll;

(B4) gap genes with mutually exclusive expression domains *gt* and *Kr* do not activate each other; no conclusions can be made for regulatory interactions between *hb* and *kni*.

Interactions between overlapping gap genes are mostly weakly repressive or absent, and are largely consistent with Figure 2.4(a): confidence intervals for W^{Kr}_{hb}, W^{hb}_{Kr}, W^{kni}_{Kr}, W^{hb}_{gt}, and W^{kni}_{gt} indicate no interaction, while confidence intervals for W^{gt}_{kni}, and W^{Kr}_{kni} suggest the absence of activation. Finally, confidence intervals for W^{gt}_{hb} indicate the absence of repression.

Obviously, our qualitative conclusions (B1)-(B4) are weaker than the conclusions (A1)-(A4) made from Figure 2.4(a) by considering only the values of parameter estimates. Note that for all genes, promoter thresholds h, promoter strengths R, diffusion coefficients D, and decay rates λ have extremely large independent confidence intervals (Figure 2.5 in Suppl. Mat. [6]) meaning that all these parameters are not determinable.

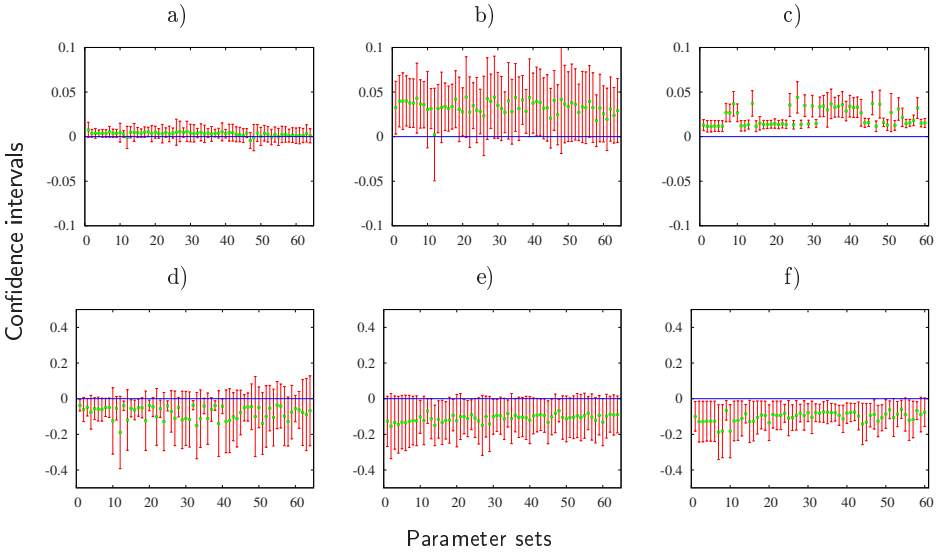


Figure 2.5: Confidence intervals for parameter estimates. Dependent (green lines) and independent (red lines) confidence intervals for regulatory weights W_{hb}^{Kr} (a), W_{hb}^{cad} (b), W_{kni}^{hb} (d), and W_{gt}^{Kr} (e) in the full search case and for regulatory weights W_{hb}^{cad} (c) and W_{gt}^{Kr} (f) in the case of fixed promoter thresholds. Confidence intervals are plotted along the vertical axis for all 64 parameter sets in the full search case and 60 parameter sets in the case of fixed promoter thresholds.

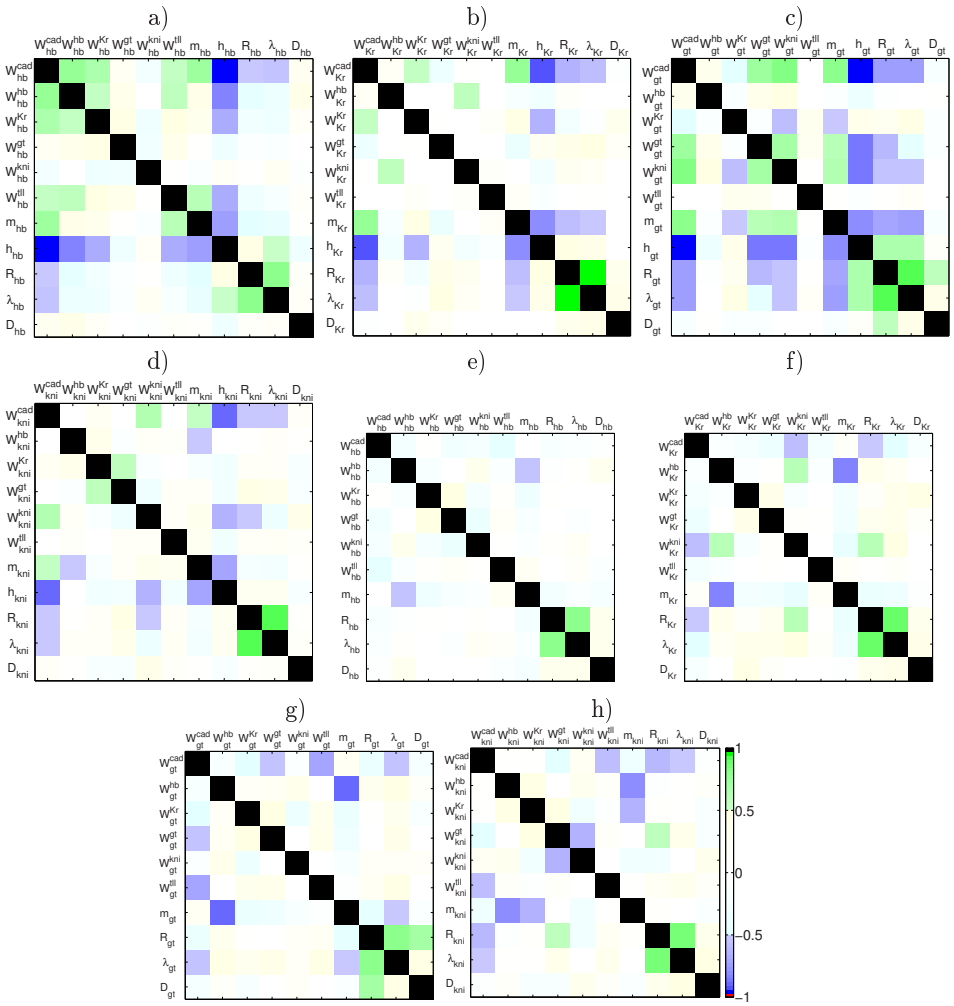


Figure 2.6: Correlations between parameters. Diagonal blocks corresponding to gap genes *hb* (a,e), *Kr* (b,f), *gt* (c,g), and *kni* (d,h) from the mean correlation matrix in the full search case (a,b,c,d) and the mean correlation matrix in the case of fixed promoter thresholds (e,f,g,h).

The large difference between dependent and independent confidence intervals indicates the presence of correlations between parameters. Individual confidence intervals are not informative for understanding the reason of poor determinability of parameters when their estimates are correlated. Using (2.2.16), we find the correlation matrix for each parameter set. To detect the most significant correlations between parameters present in all correlation matrices, we calculate an averaged matrix—which we call the mean correlation matrix—whose entries are the mean values of the corresponding correlation coefficients in the individual correlation matrices. The obtained mean correlation matrix has a block diagonal structure such that each block corresponds to a given gene and contains the correlation coefficients between parameters for the same gene (Figure 2.6 in Suppl. Mat. [6]). Panels (a),(b),(c), and (d) of Figure 2.6 show the blocks corresponding to gap genes *hb*, *Kr*, *gt*, and *kni*, respectively. Note that the correlations corresponding to the most significant entries in the mean correlation matrix (with absolute values greater than 0.5) are statistically present in all individual correlation matrices because corresponding standard deviations are relatively small (less than 0.2).

2.3.3 Parameter Estimation with Fixed Promoter Thresholds

The main insight from the mean correlation matrix is that we observe strong correlations of regulatory parameters with promoter thresholds. For instance, regulation of *hb*, *Kr*, *gt*, and *kni* by Bcd and Cad, and auto-regulation are all strongly correlated with their corresponding h_a (see panels (a),(b),(c), and (d) of Figure 2.6). This may explain the poor determinability for these interactions. We checked this hypothesis by fixing promoter thresholds h_a for gap genes *hb*, *Kr*, *gt*, and *kni* in (2.2.17) to a value of -3.5, similar to the approach used in [10, 12, 14]. We find that also in this case, least squares estimation using the LM method yields a significant decrease of the RMS (see Table 2.1). There are 63 parameter sets with $RMS < 10.0$ (best fit: $RMS = 8.66$). Among these, there are 60 parameter sets which have no visible patterning defects (Figure 3.1 in Suppl. Mat. [6]) and these were taken into account in the following analysis.

The resulting regulatory network topology (see Figure 2.4(b)) largely corresponds to that obtained without fixing threshold parameters (full search case) with a few minor exceptions. W_{kni}^{Kr}, W_{Kr}^{kni}, and W_{gt}^{kni} now all fall into the 'no interaction' category while the full search found mutual repression between *Kr* and *kni*, and activation of *gt* by Kni (compare panels (a) and (b) of Figure 2.4). As discussed above, these changes represent very small quantitative changes in the parameter values and depend on the (somewhat arbitrary) choice of cut-off between regulatory categories (compare scatter plots in Figures 2.2 and 3.2 in Suppl. Mat. [6]). Therefore, they are unlikely to be biologically significant, while all our main qualitative conclusions (A1)-(A4) on gap gene network topology are fully consistent with our results using fixed threshold parameters.

On the other hand, we observe significant improvement in determinability

of some regulatory weights when we compute dependent and independent confidence intervals for each parameter set by (2.2.13) and (2.2.14), respectively (see Figure 3.4 in Suppl. Mat. [6]). As an example, Figure 2.5(c) shows the confidence intervals for the regulatory weight W_{hb}^{cad} with fixed promoter thresholds. There is a quantitative improvement in the determinability of this parameter indicated by smaller independent confidence intervals in the case of fixed threshold parameters (compare panels (b) and (c) of Figure 2.5). But there are also qualitative changes. The model now predicts the activation of *hb* by Cad. Similarly, Figure 2.5(f) shows the confidence intervals for the regulatory weight W_{gt}^{Kr} with fixed promoter thresholds. Comparison of the panels (e) and (f) of Figure 2.5 shows that there is no quantitative difference between the two approaches for this weight. However, we see a qualitative improvement for the case of fixed threshold parameters. The independent confidence intervals in Figure 2.5(f) lie in the negative part of the plane for almost all parameter estimates and therefore, repression is now predicted for this weight while the plot in Figure 2.5(e) corresponding to full search case predicts only the absence of activation.

Based on the confidence intervals, we summarize the qualitative conclusions for the essential regulatory weights in the gap gene model in the case of fixed promoter thresholds:

(C1) Cad activates gap genes *hb*, *Kr*, *kni*, and *gt*;

(C2) Bcd does not repress gap genes *hb*, *Kr*, and *gt*; no conclusions can be made for regulation of *kni* by Bcd;

(C3) gap genes *hb*, *Kr*, and *gt* have auto-activation; gap gene *kni* does not have auto-repression;

(C4) Tll does not activate gap gene *gt*; no conclusions can be made for the regulation of *Kr* and *kni* by Tll;

(C5) mutually exclusive gap genes *gt* and *Kr* repress each other; no conclusions can be made for regulations between *hb* and *kni*.

For interaction among overlapping gap genes, the confidence intervals in the case of fixed promoter thresholds are fully consistent with those for the full search case, even though three of these interactions fall into different categories in the analysis based on parameter values only (compare panels (a) and (b) of Figure 2.4). This shows that confidence intervals can be used to check the significance of ambiguities in predicted interactions based on parameter classification alone. However, although conclusions (C1)-(C5) show qualitative improvement for some regulations in comparison with (B1)-(B4), they are still weaker than those drawn from classifying parameter values only (A1)-(A4).

Similar to the full search case, we compute the mean correlation matrix to detect the significant correlations between parameters (see Figure 3.6 in Suppl.

Mat. [6]). The obtained mean correlation matrix also has a block diagonal structure. However, there is a number of significant entries in off-diagonal blocks. Panels (e),(f),(g), and (h) of Figure 2.6 show the diagonal blocks corresponding to gap genes *hb*, *Kr*, *gt*, and *kni*, respectively. In the absence of dominating correlations between regulatory parameters and thresholds h_a we can now identify biologically significant parameter correlations. Here we restrict ourselves to describe some correlations which can be interpreted in biological terms with the emphasis on those for which at least one parameter is 'sloppy':

- Strong negative correlation is present between W_{kni}^{hb} and m_{kni}. That is, strong repression of *kni* by Hb needs to be overcome through increased activation by Bcd. Note that both parameters are poorly determined. In the circuit with W_{kni}^{hb} set to zero, Bcd actually represses *kni* (see Table 2.1 in Suppl. Mat. [6]). This contradicts genetic and molecular evidence indicating that both repression of *kni* by Hb and its activation by Bcd are present in the embryo[41, 42].

- There are complex correlations between the (very small, or absent) repressive effects of Hb on *Kr* and *gt*, and the activation of those two genes by Bcd. This confirms earlier results indicating that the balance between activation and repression from maternal genes is crucial for correct gap gene expression [43].

- The importance of the balance between activation and repression is highlighted by the following: repression of *kni* and *gt* by Tll can be compensated through increased activation by Cad, repression of *kni* by Kr can be compensated through increased activation by Bcd, while repression of *kni* by Gt can be overcome by increased *kni* auto-activation in the posterior of the embryo.

- Increased *hb* auto-activation is compensated through decreased activation of *hb* by Bcd indicating that broad maternal activation and auto-regulation are somewhat redundant.

- There is a strong positive correlation present between m_{Kr} and m_{gt}. This correlation is most likely indirect, due to repressive interaction between *gt* and *Kr*. Increased activation of *Kr* by Bcd must be balanced by increased activation of *gt* by Bcd to maintain balance of mutual repression between *Kr* and *gt*.

- There are correlations between activation of *Kr* and *gt* by Bcd and their respective promoter strengths and decay rates. Such correlations are to be expected as stronger expression or increased protein stability can compensate for weaker activation by Bcd.

We note that some of the 'sloppy' parameters, such as W_{gt}^{Kr}, W_{Kr}^{gt}, W_{hb}^{kni}, and W_{Kr}^{tll} are not (strongly) correlated to any of other parameters and their sloppiness

remains unclear. The last is completely uncorrelated parameter. Posteriorly Kr is strongly repressed by Gt and somewhat weaker by Hb and Kni. Apparently, due to these interactions repression of Kr by Tll is somewhat redundant in the model.

In summary, the above suggests that complex correlations between regulatory weights as well as correlations between those weights and promoter strength or protein decay rates are an unavoidable property of complex biological networks, as some interactions or changes in expression rate can always compensate for changes in others.

2.3.4 Parameter Correlations: Data vs Model

Poor determinability of most of the parameters in the gap gene model is due to correlations between parameters. Here we investigate whether these correlations are caused by shortcomings of the data or the model.

At first glance, it seems that insufficient accuracy of the data cannot be the reason for correlations. More accurate data would simply make the ellipsoid confidence region shrink but not rotate. Therefore, it cannot significantly improve the determinability of the parameters (see also [5]). We checked this by assuming that a larger data set was available: Say we had measurements for all gene products, in all nuclei, at 71 uniformly distributed time points (instead of 9). With these choices the total number of measurements would be $N = 21180$. Suppose that we have found that one of our parameter estimates $\hat{\theta}$ minimizes the sum of squares (2.2.2). Since the Jacobian depends only on the model responses and not on the values of the data, we can generate a new Jacobian $\tilde{J}(\hat{\theta})$ including all 'ghost' data points. From the SVD of the corresponding $\tilde{J}(\hat{\theta})$ we get the matrices $\tilde{V}(\hat{\theta})$ and $\tilde{\Sigma}(\hat{\theta})$ which define new ellipsoidal regions. The ellipsoids are slightly rotated in comparison with the initial ones but not enough to make the principal axes of the ellipsoid get closer to the parameter axes, i.e. the correlations between parameters are not removed.

Each data point is actually a sample mean, obtained by averaging gene concentrations from individual embryos. Therefore, measurement errors most likely have a normal distribution with zero mean. However, their standard deviations may vary for different data points. Assume that for the i-th data point K_i measurements from individual embryos are used and assume that the standard deviation of this sample s_i is known. Then the normal distribution of the sample mean has a standard deviation which can be estimated by $\sigma_i = \frac{s_i}{\sqrt{K_i}}$. For the dataset we used, both s_i and K_i are available from the FlyEx database [19]. Once all σ_i are found, we can use a weighted least squares estimation such that θ minimizes the sum

$$S(\theta) = \sum_{i=1}^{N} w_i^2 (y_{c_i}(t_i, \theta) - \tilde{y}_i)^2$$

instead of (2.2.2). We take the weights w_i inversely proportional to σ_i such that the weighted least squares yields the maximum likelihood estimate. Also in this case, we find that the obtained parameter estimates have the same type of correlations as those obtained with an ordinary least squares fit (data not shown).

Correlations between parameters can be due to hidden dependencies in the data set. To investigate whether this is the case, we conduct an inverse experiment. We choose one of the parameter sets obtained by the LM search, with an $RMS = 8.38$, and we denote it by θ^*. By integrating the model equations with θ^* we generate an exact data set at the same data points as the initial data set. To the exact data values we add errors drawn from the normal distribution with zero mean and standard deviation equal to 8.5. From the exact and the perturbed data set, we compute $RMS(\theta^*) = 8.17$. The perturbed dataset is used for the parameter estimation by means of the LM search. By constructing this inverse problem, we make sure that the assumption about the independence of the measurement errors is correct. With 40 different initial values of θ from [14] we obtain 34 parameter estimates having RMS between 7.95 and 8.25. Inspection of the corresponding V matrices shows that parameters are not determinable due to the correlations, similar to the original problem.

We conclude that the observed correlations between parameters are a property of the model, not of the data. Since an explicit form of the dependence of the state vector on the parameters is not known, the use of reparametrization techniques is not feasible. Note that the majority of parameters in (2.2.17) appear in the argument of the sigmoid regulation-expression function Φ. If the model (2.2.17) is used to obtain only qualitative information, such as the signs of regulatory weights, then the particular mathematical form of this function is of no importance [38]. However, it has to be studied if the choice of the sigmoid function affects the determinability of parameters.

2.4 Conclusions

In this paper we have applied the Levenberg-Marquardt (LM) optimization method to obtain a set of parameter estimates for gap gene circuit models. We then used statistical analysis to study the quality of these estimates, i.e. how well the parameters are determined with the available experimental data. Our analysis shows that none of the model parameters can be determined individually with reasonable accuracy due to correlations between parameters. Therefore, current gene circuit models cannot be used as a tool to infer quantitative regulatory weights for the gap gene network.

With this caveat in mind, however, it is still possible to draw qualitative conclusions on the regulatory topology of the gap gene network. These conclusions are weaker than, but entirely consistent with those made by only considering the values of parameter estimates [10, 12–14]. Therefore, they are also fully consistent with genetic and molecular evidence on gap gene regulation (see [10],

and references therein).

Our analysis allows us to determine exactly which interactions predicted by gene circuit analysis remain ambiguous. If considered in isolation, this ambiguity poses a serious problem for inferring regulatory interactions from expression data as it leaves important aspects of gap gene regulation unresolved. We show that more and better data will not necessarily improve parameter estimates. On the other hand, our results using fixed threshold parameters indicate that at least some of these ambiguous aspects can be resolved by reducing parameter correlations through fixing some parameters in the optimization. Others may disappear if more realistic models are used: for instance, models incorporating protein production delays, or reduced models incorporating *cad* and *tll* as time-variable external inputs as these genes are not regulated by gap genes themselves. Further research into parameter correlations within complex network models will be required to explore what kind of improved models or optimization constraints lead to better parameter determinability.

Still it remains doubtful whether an approach can be found which leads to complete parameter determinability. The study by Gutenkunst et al. [5] indicates that parameter sloppiness is a very common phenomenon among models used in systems biology. Our results corroborate this as it is difficult to see how, for example, correlations between regulatory weights could be eliminated from a network model.

The situation is not hopeless, however, as genetic evidence can help us clarify these remaining ambiguous interactions. Such evidence is itself ambiguous in many cases, as it is often difficult to interpret mutant phenotypes. But it is also complementary to and completely independent of the evidence gained by reverse engineering approaches such as the one used here [10]. This means that its ambiguities are often complementary to the ones described in this study. For instance, while cross-repressive feedback between *hb* and *kni* is not supported (but also not contradicted) by our current models, it is very strongly supported by genetic evidence [42]. Based on this, we conclude that systems biology approaches are most successful if they combine experimental and theoretical insights in a consistent and powerful manner.

Other biological interpretations of parameter sloppiness are possible. Our results on the interactions between *hb* and *kni* indicate that although present in the *Drosophila* embryo, they are not strictly necessary to maintain correct gap gene expression, and may be at least partially redundant with or replaceable by other regulatory interactions in the system. It is interesting to think about this from an evolutionary point of view, as such redundancy or replaceability allows the network to be re-wired while maintaining correct gap gene expression.

2.5 Appendix: Technical aspects

Here, we outline some general remarks one should be aware of to apply the methodology in practice.

Remark 2.5.1. If the model is given by a system of Partial Differential Equations (PDEs), then by applying a spatial discretization, it can be reduced to the system of Ordinary Differential Equations (ODEs) (2.2.1). However, in such a case one has to be careful with the choice of the grid size of the spatial discretization. On the one hand, the grid should be fine enough, so that the numerical errors introduced by spatial discretization are negligible in comparison with the level of noise in the data. On the other hand, requiring an extremely fine grid would increase the size of the system (2.2.1). The latter may be crucial in terms of computational complexity.

Remark 2.5.2. When the model includes algebraic equations, the systems of ODEs (2.2.1) and (2.2.6) change to Differential Algebraic Equations (DAEs). Since we use an implicit solver for the time integration, the method we have described is readily applicable for that type of models.

Remark 2.5.3. Given $\mathbf{f}$ and $\mathbf{y}_0$ in (2.2.1), the partial derivatives $\frac{\partial \mathbf{f}}{\partial \mathbf{y}}$, $\frac{\partial \mathbf{f}}{\partial \theta_i}$, $\frac{\partial \mathbf{y}_0}{\partial \theta_i}$ $(i = 1, \ldots, m)$ in (2.2.6) can be, in principle, found analytically. However, for large scale problems when $\mathbf{f}$ has a complicated nonlinear form, this can be a tedious work to do. In such cases, these derivative functions can be generated automatically by using a symbolic mathematics package, like *Maple* [44] or *Mathematica* (Wolfram Research, Inc).

Remark 2.5.4. Numerically solving (2.2.6) has limitations for large scale problems due to computational costs. Another approach to approximate the matrix $J(\theta)$ could be by means of divided differences. The j-th column of $J(\theta)$ is then given by

$$\frac{\partial \mathbf{Y}(\theta)}{\partial \theta_j} \approx \frac{\mathrm{Y}(\tilde{\theta}^j) - \mathrm{Y}(\theta)}{\delta \tilde{\theta}_j},$$

where the vector $\tilde{\theta}^j$ is obtained by a small perturbation $\delta\tilde{\theta}_j$ in the j-th entry of θ. In this case, for one Levenberg-Marquardt (LM) step system (2.2.1) has to be numerically integrated $m+1$ times. With regard to the computational costs, when $\mathbf{f}$ is non-linear, it is more expensive than the approach where the linear systems of variational equations are solved. Another drawback of the divided difference method is that the numerical approximations introduce additional errors.

Remark 2.5.5. For large scale problems computation on a single processor can become unfeasible and one needs to use a parallel machine. Parallelization of the computational work when (2.2.1) and (2.2.6) are solved numerically is only possible at the level of the time step of the integrator. Therefore, it will be

inefficient due to heavy communication. The advantage of the divided difference approach is that in this case (2.2.1) is solved for $m+1$ different values of θ independently of each other. Therefore, parallelization of the computational work by divided difference method is trivial and can be very efficient.

Remark 2.5.6. Variational equations (2.2.6) coupled with (2.2.1) result in a system consisting of $m+1$ subsystems of the size n. The Jacobian of the coupled system has the following form:

$$\begin{pmatrix} \dfrac{\partial \mathbf{f}}{\partial \mathbf{y}} & 0 & \dots & 0 \\ \dfrac{\partial^2 \mathbf{f}}{\partial \theta_1 \partial \mathbf{y}} + \dfrac{\partial^2 \mathbf{f}}{\partial \mathbf{y}^2}\dfrac{\partial \mathbf{y}}{\partial \theta_1} & \dfrac{\partial \mathbf{f}}{\partial \mathbf{y}} & 0.. & 0 \\ \vdots & 0 & \ddots & 0 \\ \dfrac{\partial^2 \mathbf{f}}{\partial \theta_m \partial \mathbf{y}} + \dfrac{\partial^2 \mathbf{f}}{\partial \mathbf{y}^2}\dfrac{\partial \mathbf{y}}{\partial \theta_m} & 0 & ..0 & \dfrac{\partial \mathbf{f}}{\partial \mathbf{y}} \end{pmatrix}$$

The structure of the Jacobian reveals the one-way coupling of the system. Using an implicit scheme we can exploit this specific coupling between (2.2.1) and (2.2.6) in an efficient way. At each time step τ integrating first (2.2.1) provides the solution vector $\mathbf{y}$. This requires the LU decomposition of $I_m - \tau \frac{\partial \mathbf{f}}{\partial \mathbf{y}}$. Using this LU decomposition the calculation of $\frac{\partial \mathbf{y}}{\partial \theta_i}$ from (2.2.6) reduces to a simple forward substitution and backsubstitution.

Chapter 3

Parameter estimation for a Model of Gap Gene Circuits with Time-Variable External Inputs in *Drosophila*

Abstract

We study a model for spatio-temporal pattern formation of gap gene products in the early development of *Drosophila*. In contrast to previous studies of gap gene circuits, our model incorporates a number of proteins as time-variable external inputs, including a protein Huckebein which is necessary for setting up the correct posterior domain boundary and its shift in time for the gap gene *hunchback*. Unknown model parameters are inferred by fitting the model outputs to the gap gene data and statistical analysis is applied to investigate the quality of the parameter estimates.

Our results, while being consistent with previous findings, at the same time provide a number of improvements. Firstly, it takes into account correct regulation of *hunchback* at the posterior part of the embryo. Secondly, confidence interval analysis shows that the regulatory topology of the gene network in our model which consists of parameters representing the regulation between genes is more consistent with experimental evidence.

Our results also reveal that for data fitting the Weighted Least Squares sum is a more suitable measure than the Ordinary Least Squares sum which has been used in all previous studies. This is confirmed by a better fit of the boundaries of the gap gene expression domains and an absence of patterning defects in the model outputs, as well as by a correct prediction of mutant phenotypes.

3.1 Introduction

Gap genes constitute the first step in a regulatory cascade that leads to the determination of body segment positions along the major (or anterior-posterior, A–P) body axis during early *Drosophila* development [8]. They are involved in the regulation of pair-rule and segment-polarity genes, the latter of which establish a segmental pre-pattern of gene expression by the onset of gastrulation.

The gap gene system in the early *Drosophila melanogaster* is a well studied developmental gene network (see [10] and references therein). Initially the system is set up by spatial gradients of maternal proteins Bicoid (Bcd), Hunchback (Hb), and Caudal (Cad). Zygotic gap genes, such as *hunchback* (*hb*), *Krüppel* (*Kr*), *knirps* (*kni*), and *giant* (*gt*), are regulated by these maternal gradients, which establishes their expression in broad, overlapping regions of the embryo. These spatial domains of gap gene expression are stabilized and refined by gap-gap cross-repression and regulation by zygotic terminal gap genes *tailless* (*tll*) and *huckebein* (*hkb*).

The gap gene system has been studied extensively using a model of genetic regulatory networks described by a system of reaction-diffusion equations [10]. Quantitative expression data available for all relevant maternal coordinate and gap genes [18, 19] (except for *hkb*) have been used to infer regulatory interactions between gap genes using different global and local optimization strategies [6, 10, 12–14]. The gap gene system has been modeled by a 6-gene network, including *hb*, *Kr*, *kni*, *gt*, *tll*, and *caudal* (*cad*). The maternal protein Bcd has been incorporated as an external input constant in time. Although the obtained results have given significant insight into the underlying mechanism of the gap gene system, further investigation is needed for some important issues.

Results for the 6-gene model revealed a major patterning defect for the expression of gap gene *hb*. The posterior boundary of the posterior *hb* domain was not established correctly. Moreover, anterior shift of this boundary as well as the shift of the domain peak found in data was not reproduced by model outputs. This was explained by the absence of the terminal gap gene *hkb* in the 6-gene model. Huckebein (Hkb) is the main repressor of *hb* in that region of the embryo [45]. The missing *hkb* gene was also predicted to have an influence on the regulatory topology inferred from data. The model wrongly predicted that *hb* is not regulated by Tailless (Tll), while it is known that *hb* is activated by Tll [45]. This contradiction was explained by the ambiguous role of Tll in the regulation of *hb*. On the one hand, in the absence of the repressor Hkb, Tll has to take over its repressing function. On the other hand, Tll is an activator of *hb*. This dual role yields a cancellation effect and the model predicts that *hb* is not regulated by Tll.

Results of a parameter determinability analysis for the 6-gene model in [6] show that the parameter estimates corresponding to the regulation of *cad* and *tll* by gap genes are highly unreliable. The observed uncertainty was explained by the fact that the products of maternal genes (such as *cad*) and terminal gap genes (such as

tll) regulate gap genes, but not vice versa. Despite the reasonable fit obtained for the expression of *cad* and *tll* in the 6-gene model, the unrealistic assumption that their dynamics is prescribed by the regulation by gap genes increases the level of uncertainty in the gap gene model. Due to the correlations between parameters in the model, this influences the determinability of other, biologically relevant, regulatory weights.

In this work we consider a reduced 4-gene model, including only gap genes *hb*, *Kr*, *gt*, and *kni*. In contrast to the 6-gene model, we now incorporate *cad* and *tll* as time-variable external inputs. Thereby, in our model the expression of *cad* and *tll* is obtained directly from data rather than being computed as state variables. A second important change is that data for gene *hkb* have become available [46]. Similar to *tll*, the terminal gap gene *hkb* is not regulated by other gap genes and therefore, it is also included in the model as time-variable external input. Finally, we incorporate the maternal gradient Bcd in our new model as external input, similar to previous studies. However, the data suggest that the protein Bcd varies with time rather than being constant. Therefore, contrary to the 6-gene model, we allow Bcd to be time-variable input.

Thus, we replace the previously studied 6-gene network by a more realistic, reduced 4-gene network with four external time-variable inputs. This significantly decreases the size of the problem, both with regard to the number of equations in the model to be solved and the number of unknown parameters to be estimated. We will infer the regulatory topology and we will investigate the parameter determinability for the reduced model. Most importantly, we will show that despite the simplifications we made, the reduced model not only gives comparable results as the 6-gene model but also overcomes the above mentioned shortcomings. Note that the reduced gap gene network has also been recently investigated in [15], but in that study Bcd has been kept constant in time and *hkb* has not been used in the model.

Inference of the parameters is done by fitting model outputs to experimental data, i.e., by minimizing a cost function which measures the difference between them. The choice of the cost function is important for obtaining unbiased estimates and the computation of statistical quantities for parameter estimates (such as confidence intervals and correlations coefficients). It greatly depends on the nature of errors in the data. In all previous works [6, 10, 12–15], the Ordinary Least Squares (OLS) measure has been used for the optimization and the statistical analysis. It is well known that OLS is suitable if the measurement errors are independent of each other and normally distributed with a constant standard deviation. However, the data for the gap genes suggest that the level of noise in the measurements varies both in space and in time. In such case, the Maximum Likelihood Estimates (MLE) can be obtained only if the Weighted Least Squares (WLS) sum is used as a distance measure with the weights chosen to be inversely proportional to standard deviations [22]. Since the standard deviations are available from [19], there is no additional computational work needed when the WLS sum is minimized in comparison with the OLS case. In this work we will obtain

parameter estimates and study the parameter determinability using both the OLS and the WLS measures, and we will provide a detailed comparison between both results. We will demonstrate that for the problem under consideration, WLS gives indeed a more suitable measure than OLS.

The chapter is organized as follows. In Section 3.2 we describe the necessary materials and methods that are used. In Section 3.3 we give the results of our simulations. We conclude this chapter with a discussion in Section 3.4. In the Appendix we include all additional plots.

3.2 Materials and Methods

3.2.1 Gap Gene Circuits

Segment determination occurs during the blastoderm stage of *Drosophila* development, between 1.5 and 3 hours after egg laying [7]. During this stage, the embryo consists of a syncytium: there are no cell membranes between the nuclei. These nuclei constitute the basic objects of the model. They are arranged in a row along the A–P axis. Nuclei divide rapidly and synchronously [39]. Periods between mitotic divisions are called cleavage cycles, where cycle n occurs between mitoses $n-1$ and n. The models considered here run from early cycle 13 ($t = 0.0$ min) to the onset of gastrulation at the end of cycle 14A ($t = 71.1$ min). Mitosis occurs at the end of cycle 13, between $t = 16.0$ min and $t = 21.1$ min [39].

Gene circuit models describe the change in concentrations of each gap gene product in each nucleus over time by the following system of ODEs

$$\frac{dg_i^a}{dt} = R_a \Phi\left(\sum_{b=1}^{N_g} W_a^b g_i^b + \sum_{e=1}^{N_e} E_a^e g_i^e + h_a\right) - \lambda_a g_i^a + D_a\left(g_{i+1}^a - 2g_i^a + g_{i-1}^a\right), \tag{3.2.1}$$

where a and b refer to regulated gap genes and regulators, respectively, and e refers to external regulators. Here, a and b are integer indices representing *hb*, *Kr*, *kni*, and *gt*; e is an integer representing the regulators Bcd, Cad, Tll, and Hkb. The independent variable g_i^a denotes the concentration of the product of gene a in nucleus i; the input variable g_i^e denotes the concentration of the external protein e in nucleus i. $N_g = 4$ is the number of gap genes and $N_e = 4$ is the number of external proteins in the model. The function

$$\Phi(x) = \frac{1}{2}\left(\frac{x}{\sqrt{x^2+1}} + 1\right) \tag{3.2.2}$$

is a sigmoid regulation-expression function. The first term in the right hand side of (3.2.1) models the protein synthesis, while the second and third terms correspond to protein decay and protein diffusion, respectively.

During mitosis, protein synthesis is shut down. Nuclei divide instantaneously at the end of mitosis and the protein concentrations from each mother nucleus

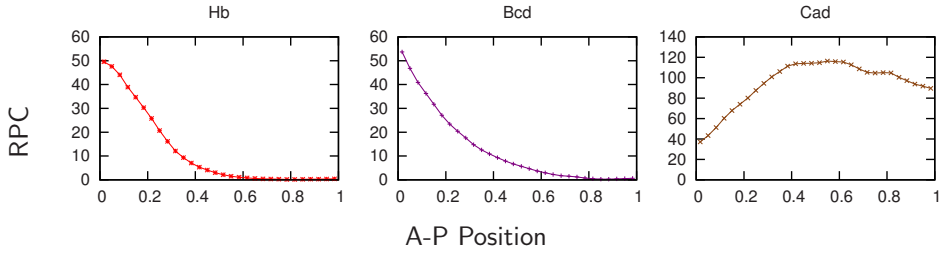


Figure 3.1: Quantitative gene expression for Hb, Bcd, and Cad at $t = 0$. Lines show the relative protein concentration (RPC) plotted against the position on the A–P axis (the trunk region of the embryo, from 35% to 92% A–P position is scaled to relative co-ordinates $[0, 1]$).

are copied to its daughter nuclei. The distance between nuclei is halved which is implemented in the model by reducing the diffusion coefficients D_a by the factor of 4. Gap gene circuits cover the region from 35% to 92% of the A–P axis, which includes $N_c = 30$ and $N_c = 58$ nuclei at cycles 13 and 14A, respectively. Therefore, system (3.2.1) consists of 120 and 232 ODEs during cycles 13 and 14A, respectively. At the boundary points $i = 1$ and $i = N_c$ we replace the diffusion term in right hand side of (3.2.1) by $D_a(g^a_{i+1} - g^a_i)$ and $D_a(g^a_{i-1} - g^a_i)$, respectively. This way we mimic the homogeneous Neumann (no flux) boundary conditions.

Gap genes *Kr*, *kni*, and *gt* are not expressed in the embryo before cycle 13. Therefore, zero initial conditions are taken for these. The initial condition for *hb* is prescribed by the maternal gradient of Hb shown in Figure 3.1. It is obtained by averaging the measurements from 18 individual embryos at cycle 12 ($t = -6.2$) and then using linear interpolation between this averaged pattern and *hb* data at cycle 13 ($t = 10.55$). Measurements for *hb* from individual embryos at cycle 12 (data without background) and the averaged *hb* pattern at cycle 13 are all available from [19].

In system (3.2.1) there are $m = 48$ unknown parameters. These include the genetic interconnection or regulatory weight matrices W and E of size $N_g \times N_g$ and $N_g \times N_e$, respectively. The matrix elements W^b_a and E^e_a represent the regulation of gap gene a by gene b and gene e, respectively. Regulatory parameters represent repression (if < 0), activation (if > 0) or no interaction (if ≈ 0). The other parameters are promoter thresholds h_a, promoter strengths R_a, diffusion coefficients D_a, and decay rates λ_a.

Data The data set used for model fitting consists of $N = 1976$ measurements of protein concentrations (available from [19]). Measurements were taken at one time point during cycle 13 (T_0), and eight time points T_i ($1 \le i \le 8$) during cycle 14A (Figure 3.2). Measurements for the concentrations of all gap gene products represented in the model at all time points are available. Each data point

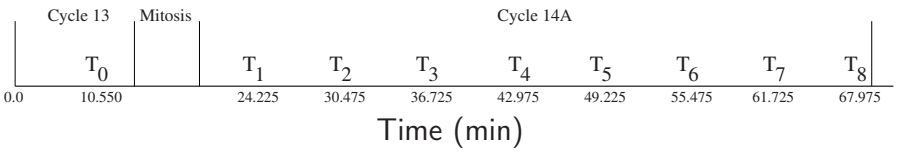


Figure 3.2: Time axis and the points when measurements were taken: one in cycle 13 and eight in cycle 14A; the duration of mitosis is also indicated.

represents concentration values which have been averaged over the bin (volume) from the measurements taken in individual embryos [18]. The number of embryos varies from 9 to 62 for different genes and different time points (with exception for *kni* at T_0 where only measurements from 4 embryos are available). Since from each embryo a few values per bin are available, the number of individual measurements used in the computation of the averaged value (sample mean) is much larger than the number of embryos. Using the Central Limit Theorem (CLT) we may assume that the experimental errors are approximately normally distributed [47]. Figure 3.3 shows the gap gene data at all time points (solid lines) and the standard deviations of the experimental errors (shaded areas).

To our knowledge, the presence of any hidden dependencies in the available dataset has not been investigated in literature. Although the measurements from different embryos are most likely to be uncorrelated (assuming that there was no systematic error in experiments), it is not known whether the gene expression data in the same embryo are correlated. In this work, we assume that the experimental errors are independent of each other.

External Inputs To solve (3.2.1), one needs the level of gene expression for external inputs at all $t \in [0, T]$, where $T = 71.1$. Measurements for Bcd, Cad, and Tll at all time points T_i ($0 \leq i \leq 8$) are available from [19] except for Bcd and Cad at T_7 and T_8. We obtain the patterns for Cad at those missing time points by integrating measurements from individual embryos (from [19]), 13 at T_7 and 12 at T_8. A similar procedure for Bcd however leads to an artificially high level of gene expression for Bcd at T_7 and T_8 and therefore they are not used here. Data for Hkb at all time points T_i are obtained from [46]. Figure 3.4 shows the relative protein concentration of external genes at all time points.

Genes Tll and Hkb are not expressed before cycle 13 and therefore we use a zero level for them at $t = 0$. Bcd and Cad at $t = 0$ have initial maternal gradients shown in Figure 3.1. We obtain them in the same way as the initial data for *hb*, i.e., by averaging the data from individual embryos at cycle 12 and then using linear interpolation between the patterns at cycles 12 and 13.

Now, the values of the external genes at any $t \in [0, T_8]$ can be linearly interpolated from the data at $t = 0, T_0, T_1, \ldots T_8$. The expression of Bcd for $t > T_6$ is linearly extrapolated from the values at T_5 and T_6, while the expression of other

external inputs for $t > T_8$ is linearly extrapolated from corresponding values at T_7 and T_8. If the extrapolated value is negative then we replace it with zero. Finally, we note that higher order interpolations give rise to artifacts from experimental noise [15] and therefore they are not used here.

Parameter inference We denote each measurement by $g_i^a(T_j)_{data}$, specified by the time T_j when the concentration of the gene product a in nucleus i is measured. The corresponding model value obtained from (3.2.1) is denoted by $g_i^a(T_j)_{model}$. The estimation of unknown parameters in (3.2.1) amounts to minimizing the cost function

$$CF = \sum_{a=1}^{N_g} \sum_{i=1}^{N_c} \sum_{j=0}^{N_t} v_{ij}^a \left(g_i^a(T_j)_{model} - g_i^a(T_j)_{data} \right)^2, \tag{3.2.3}$$

where v_{ij}^a are positive weights, $N_g = 4$ is the number of gap genes, N_c is the number of nuclei (30 and 58 during cycles 13 and 14A, respectively), and $N_t = 8$ is the number of time classes. When all weights are equal to one, (3.2.3) is the OLS sum. Note that previously in the studies of gap gene circuits, only OLS is used as cost function to minimize. The quality of the fit of the model to the data is measured by the root mean square (RMS) given by

$$RMS = \sqrt{\frac{1}{N} \sum_{a=1}^{N_g} \sum_{j=0}^{N_t} \sum_{i=1}^{N_c} \left(g_i^a(T_j)_{model} - g_i^a(T_j)_{data} \right)^2}, \tag{3.2.4}$$

where $N = 1976$ is the total number of all measurements. A solution is considered to be 'good' if $RMS < 12.0$ and if there are no visible pattern defects in the model response [10, 12–14].

We note that OLS is an appropriate measure under certain assumptions only. Namely, all measurement errors have to be independent of each other and be from a normal distribution with zero mean and constant standard deviation. The latter does not hold for our dataset [19]. The shaded areas in Figure 3.3 show how the standard deviation varies per gene and both in space and time. Note that the standard deviation (the level of noise) becomes smaller at late time points. Also important is that the standard deviation at the domain boundaries is relatively small and the level of noise in the non-expressing regions is almost negligible indicating that the stripe locations at the end of cycle 14A are determined with little variation [40].

When the weights v_{ij}^a in (3.2.3) are taken inversely proportional to the corresponding standard deviations, the cost function becomes the WLS distance. We emphasize here that this is a theoretically more justified measure than the OLS measure due to the variation in the experimental errors. Since the standard deviations are available in [19], minimization of the WLS sum has no additional computational expenses compared to the corresponding procedure for the OLS sum.

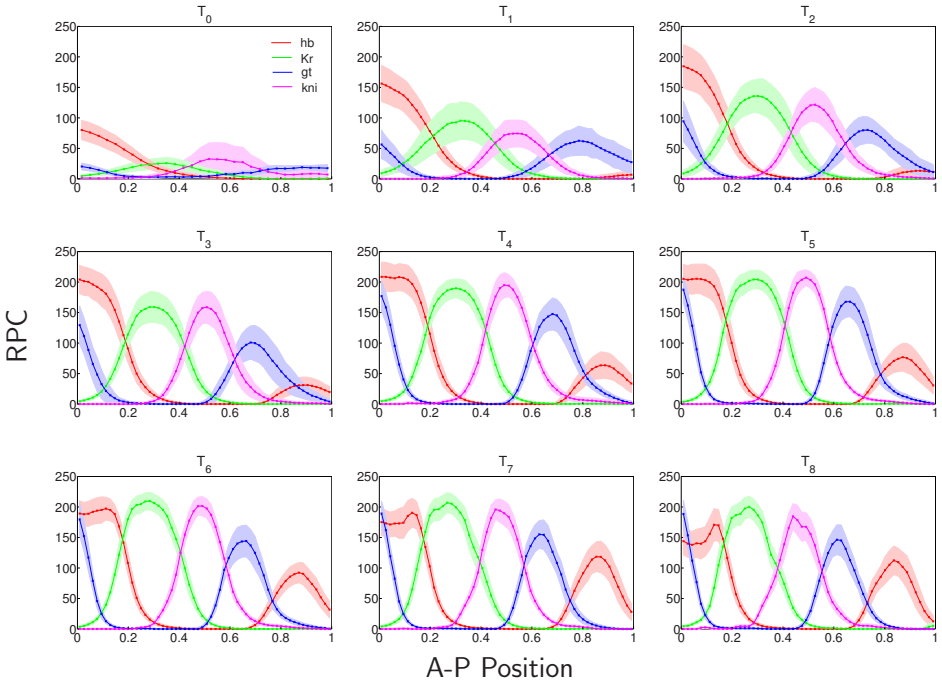


Figure 3.3: Quantitative gap gene expression data (solid lines) at the different time points. The shaded areas give the range of one standard deviation of the experimental error. Axes are as in Figure 3.1.

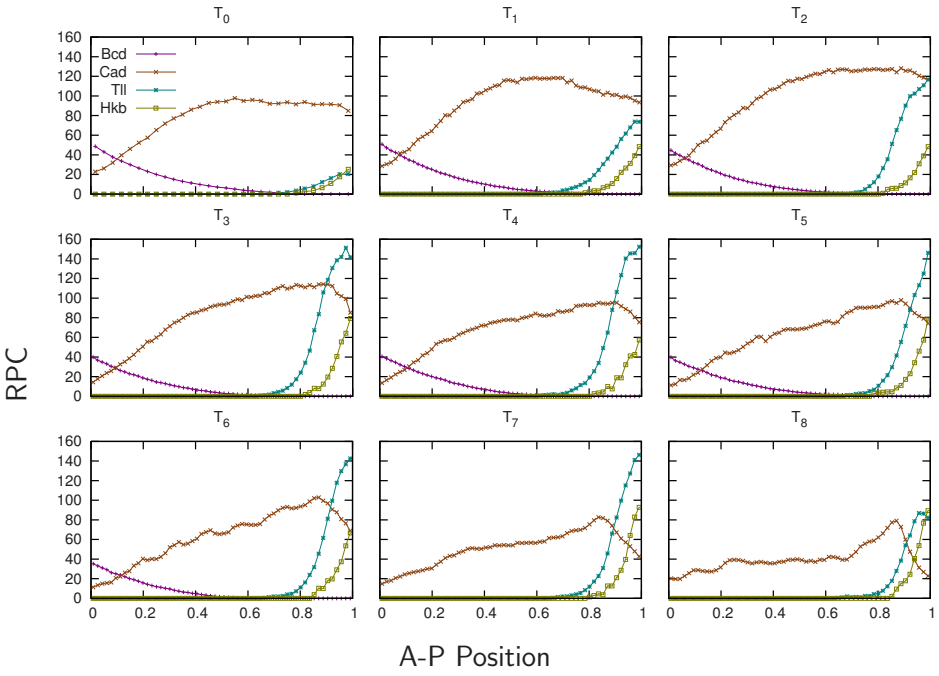


Figure 3.4: Quantitative data for external inputs at different time points. Note that Bcd at T_7 and T_8 is not available. Axes are as in Figure 3.1.

In this work we use both the OLS sum and the WLS sum as the cost function to minimize. Our aim is to demonstrate that WLS is a more suitable measure than OLS not only in theory but also in practice. Throughout this chapter we will use the notations *OLS search* and *WLS search* meaning that the OLS and WLS sums, respectively, are minimized. Similarly, *OLS results* and *WLS results* indicate the parameter estimates obtained by minimizing OLS and WLS sums, respectively.

For practical reasons it is better to constrain the parameter space, especially for the global search optimization methods. Similar to previous studies of the gap gene system [6, 10, 12–14], we define the search space for the parameters by the linear constraints

$$10.0 \leq R_a \leq 30.0, \quad 0.0 < D_a \leq 0.3, \quad 5.0 \leq \frac{ln(2)}{\lambda_a} \leq 20.0, \quad a = 1, \ldots, N_g, \tag{3.2.5}$$

and by the nonlinear constraints

$$\sum_{b=1}^{N_g} \left(W_a^b \; g_{max}^b\right)^2 + \sum_{e=1}^{N_e} \left(E_a^e \; g_{max}^e\right)^2 + (h_a)^2 \leq 10^4, \quad a = 1, \ldots, N_g, \tag{3.2.6}$$

where g_{max}^b and g_{max}^e are the maximum values in the data set for proteins b and e, respectively. Note that in [10, 12, 13] the threshold parameters h_a for genes Kr, Kni, gt, and hb are fixed to negative values representing a constitutively repressed state for the corresponding genes [10, 12]. In [6] it is shown that fixing promoter thresholds improves the parameter determinability in comparison to the case when they are estimated along with other parameters. Therefore, we take $h_a = -2.5, \; a = 1, \ldots, N_g$ in all simulations, which leaves us with 44 unknown parameters in (3.2.1) to be estimated.

Mutation analysis The regulation of gene b on gene a is studied experimentally in the following way, called mutation: gene b is knocked out in the embryo and from the change in the expression of gene a the possible type of regulation is deduced. If the expression of gene a decreases (increases), then it is assumed that b is activator (repressor). If the mutation does not affect the expression of gene a then it means that b does not regulate gene a. Experiments with double mutants (when two different genes are knocked out) are also widely used. Similarly, mutation can also be done by over-expressing a certain gene to study its effect on the expression of the other genes. Although the conclusions based on mutant analysis can be ambiguous in some cases, such as indirect influence, still this method is a commonly applied approach in genetics.

Once the regulatory weights in the gap gene model (3.2.1) are estimated based on wild type data, mutation analysis can be easily conducted in *silico* [48]. Namely, b mutants can be modelled simply by setting W_a^b (or E_a^b) for all gap genes a to zero and leaving all other parameter estimates unchanged. It is an important

issue whether the model with parameter estimates found using only wild type data can predict correct mutant phenotypes. Although the quantitative mutant data is not available, qualitative behaviour for mutant phenotypes of gap gene products in *Drosophila* is well studied. For instance, the posterior *hb* domain fails to retract from the posterior pole of the embryo in *hkb* mutant embryos [45], indicating that Hkb represses *hb*. The posterior *hb* domain is absent in *tll* mutant embryos [45], indicating that Tll activates *hb*.

3.2.2 Methods

We consider a model given by the system of ODEs of the general form:

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{f}(t, \mathbf{y}, \theta), & 0 < t \leq T, \\ \mathbf{y}(t, \theta) = \mathbf{y}_0, & t = 0. \end{cases} \tag{3.2.7}$$

Here the m-dimensional vector θ contains all unknown parameters, $\mathbf{y}$ is an n-dimensional state vector, and $\mathbf{f}$ is a given vector function, differentiable with respect to t, $\mathbf{y}$ and θ.

Let us assume that for fitting (3.2.7) there are N measurements available. Each measurement, which we denote by $\tilde{y}_i$, is specified by the time t_i when the c_i-th component of the state vector $\mathbf{y}$ is measured. The corresponding model value obtained from (3.2.7) is denoted by $y_{c_i}(t_i, \theta)$. We denote the vector of weighted discrepancies between the theoretical values and the measured values by $\mathrm{Y}(\theta)$. Then the least squares estimate $\hat{\theta}$ of the parameters is the value of θ that minimizes the sum of squares

$$S(\theta) = \sum_{i=1}^{N} w_i^2 (y_{c_i}(t_i, \theta) - \tilde{y}_i)^2 = \mathrm{Y}^T(\theta)\mathrm{Y}(\theta), \tag{3.2.8}$$

where w_i are positive weights. If the measurement errors in $\tilde{y}_i$ are independent of each other, normally distributed with standard deviations σ_i, and the weights w_i are proportional to $1/\sigma_i$, then $\hat{\theta}$ is a maximum likelihood estimate [22].

Parameter Estimation

In general, model (3.2.7)—being nonlinear in θ—leads to a least squares problem (3.2.8) that has several minima, first because the problem has more than one solution, and second because the fitness function (3.2.8) can have several stationary points that do not correspond to the lowest value of the cost function (so-called local minima). *Local search methods*, like Levenberg-Marquardt (LM) [26], easily get trapped in one of the local minima rather than finding the global minimum. To explore the whole search space one needs *global search methods*, like Evolution Strategy (ES) or Simulated Annealing (SA). Unfortunately, these methods

converge very slowly once near a minimum. In contrast, gradient-based methods are efficient optimizers [2] for nonlinear least-squares problems once a sufficiently good initial guess for the parameter values is available. Therefore, for large scale problems, such as a gap gene system, it is efficient to use a global search method followed by a local gradient-based technique. In this way, the chance of missing the global minimum is reduced and the determination of the minima is precise and fast.

In this chapter we use the LM method for local optimization. For the initial parameter values we use the parameter estimates obtained by Johannes Jaeger (EMBL/CRG, Barcelona) with SA global search.

Levenberg-Marquardt Method In general, any gradient-based optimization procedure seeks a correction $\delta\theta$ for the parameter vector, such that $S(\theta+\delta\theta) \leq S(\theta)$ holds. The LM method [26] determines the correction as the solution of the equations

$$\left(J^T(\theta)J(\theta)+\lambda I_m\right)\delta\theta = -J^T(\theta)\mathrm{Y}(\theta), \tag{3.2.9}$$

where $\lambda \geq 0$ is a control parameter (see below), I_m is the identity matrix of size m and the Jacobian $J(\theta) = \frac{\partial \mathrm{Y}(\theta)}{\partial \theta}$ is the so-called 'sensitivity' matrix of size $N \times m$. The entry $J_{i,j}$ in $J(\theta)$ shows how sensitive the model response is at the i-th data point for a change in the j-th parameter. The entries of J can be found by solving the system of variational equations

$$\begin{cases} \dfrac{\partial}{\partial t}\dfrac{\partial \mathbf{y}}{\partial \theta_i} = \dfrac{\partial \mathbf{f}}{\partial \theta_i} + \dfrac{\partial \mathbf{f}}{\partial \mathbf{y}}\dfrac{\partial \mathbf{y}}{\partial \theta_i}, & 0 < t \leq T, \\ \dfrac{\partial \mathbf{y}}{\partial \theta_i}(t,\theta) = 0, & t = 0, \end{cases} \tag{3.2.10}$$

where $i = 1, 2, \ldots, m$, coupled to (3.2.7).

The LM method can be seen as the combination of two gradient-based approaches: Gauss-Newton and steepest descent [2]. If $\lambda = 0$ in (3.2.9), it coincides with the Gauss-Newton method. However, when the matrix $J^T(\theta)J(\theta)$ is (almost) singular, to solve (3.2.9), λ has to be positive and for large λ the LM method approaches the steepest descent method. During the optimization λ is adapted such that the algorithm strives to exploit the fast convergence of the Gauss-Newton method whenever this is possible [26, 34].

In order to solve (3.2.9), the singular value decomposition (SVD) [35] of the matrix $J(\theta)$ can be used, i.e.

$$J(\theta) = U(\theta)\ \Sigma(\theta)\ V^T(\theta), \tag{3.2.11}$$

where $U(\theta)$ is an orthogonal matrix of size $N \times m$, such that $U^T(\theta)U(\theta) = I_m$, $V(\theta)$ is an orthogonal matrix of size $m \times m$, such that $V^T(\theta)V(\theta) = V(\theta)V^T(\theta) = I_m$, and $\Sigma(\theta)$ is a diagonal matrix of size $m \times m$ which contains all singular values in non-increasing order. Then the correction $\delta\theta$ can be found as

$$\delta\theta = -V(\theta)\ \left(\Sigma^2(\theta)+\lambda I_m\right)^{-1}\Sigma(\theta)\ U^T(\theta)\ \mathrm{Y}(\theta). \tag{3.2.12}$$

Numerical integration of (3.2.7) and (3.2.10) requires a fast and reliable ODE solver. Searching in the parameter space may lead to some values of θ such that the systems of ODEs become stiff. It is well known that for stiff ODE systems explicit schemes can give rise to numerical instability or, alternatively, extremely small time steps. Therefore, an implicit scheme is the best choice for time integration for stability reasons. In our simulations we use implicit multistep Backward Differentiation Formulas (BDF) [37]. For numerical and implementational aspects of this method we refer the reader to [37] and [6] and references therein.

Statistical Analysis of Parameter Estimates

Once the parameter vector $\hat{\theta}$ minimizing (3.2.8) is found, it is important to know how reliable the obtained estimate is. This is the subject of *a posteriori* identifiability analysis [1, 23, 24]. The ellipsoidal region around $\hat{\theta}$ in which the 'true' parameter vector θ^* lies with a certain probability $1-\alpha$ is defined by

$$(\theta^* - \hat{\theta})^T \left(J^T(\hat{\theta}) J(\hat{\theta}) \right) (\theta^* - \hat{\theta}) \leq \frac{m}{N-m} S(\hat{\theta}) F_\alpha(m, N-m), \tag{3.2.13}$$

where $F_\alpha(m, N-m)$ is the upper α part of Fisher's distribution with m and $N-m$ degrees of freedom. To remind the reader, here m and N are the number of parameters and measurements, respectively. From (3.2.13) one can derive dependent and independent confidence intervals for parameter estimates $\hat{\theta}_i$ $(i = 1, 2, \ldots, m)$. These are, respectively,

$$\left\{ \theta_i : |\theta_i - \hat{\theta}_i| \leq \frac{r_\sigma}{\sqrt{\left(V(\hat{\theta}) \Sigma^2(\hat{\theta}) V^T(\hat{\theta}) \right)_{ii}}} \right\} \tag{3.2.14}$$

and

$$\left\{ \theta_i : |\theta_i - \hat{\theta}_i| \leq r_\sigma \sqrt{\left(V(\hat{\theta}) \Sigma^{-2}(\hat{\theta}) V^T(\hat{\theta}) \right)_{ii}} \right\}. \tag{3.2.15}$$

Here $V(\hat{\theta})$ and $\Sigma(\hat{\theta})$ are obtained from (3.2.11), $r_\sigma^2 = \frac{m}{N-m} S(\hat{\theta}) F_\alpha(m, N-m)$.

Clearly, small confidence intervals for $\hat{\theta}_i$ indicate that it is well-determined. However, in some cases considering only individual confidence intervals can be misleading. For instance, in the presence of strong correlations between parameters, the dependent confidence intervals underestimate the confidence region while the independent confidence intervals overestimate it. For this reason, in addition to confidence intervals, it is essential to compute correlations between parameters. The correlation coefficient between $\hat{\theta}_i$ and $\hat{\theta}_j$ is given by

$$\rho_{ij} = \frac{B_{ij}}{\sqrt{B_{ii} B_{jj}}}. \tag{3.2.16}$$

where $B(\hat{\theta}) = V(\hat{\theta}) \Sigma^{-2}(\hat{\theta}) V^T(\hat{\theta})$. For detailed explanations of these statistical quantities and their derivations we refer the reader to [6] and references therein.

3.3 Results

We estimated all 44 unknown parameters of the gap gene circuit model (3.2.1), such that the state variables fit the given data (Figure 3.3), subject to the constraints (3.2.5)-(3.2.6). We applied statistical analysis for the final parameter sets to assess the quality of the parameter estimates. Both OLS and WLS were used as a cost function in the data fitting procedure and the statistical analysis. We present here both results and give a detailed comparison between them.

3.3.1 OLS results

Selection of OLS gene circuits

The search with the OLS cost function leads to 740 parameter sets. About 80% of them have good-scoring RMS values, i.e., $RMS < 12.0$, which is below the level of experimental errors. However, a closer look at the model outputs for good-scoring sets reveals that most of them have a common patterning defect. Figure 3.5 shows the patterns obtained with one of those parameter sets (with $RMS = 9.21$) for the expression of gap gene Kr at time points T_3 and T_8 (green lines) compared to data (red lines). The model outputs have an artificial Kr hump in the region where no expression is detected for this gene in the data. This hump arises at the beginning of cycle 14A and remains there until the end of cycle 14A. It is noteworthy that the gap gene network topology, i.e., the signs of regulatory weights in (3.2.1), in the parameter sets possessing such a patterning defect is in contradiction with known experimental evidence. In other words, despite the overall reasonable fit to the data, model (3.2.1) predicts wrong regulations between genes. For instance, in the parameter set for which the patterns in Figure 3.5 are shown, *hb* is repressed by Tll and activated by Hkb, while it is known that Tll activates *hb* and Hkb represses it. We have found that the inferred network topology in good-scoring parameter sets producing an artificial Kr hump has some other artifacts as well (not shown here). Therefore, we exclude those parameter sets.

Although many good-scoring parameter sets obtained from the OLS search have the artificial Kr hump, there are still 39 parameter sets left which do not have that patterning defect. Their RMS values vary between 8.71 and 10.11. None of these parameter sets show any significant patterning defects (see Figure 3.15 in the Appendix). As we shall see, their network topology is in agreement with experimental evidence. We consider only these 39 parameter sets in our analysis.

In conclusion, our selection of OLS parameter sets has been based on two criteria. Firstly, only parameter sets with low RMS values are taken into account. Secondly, only those sets which do not have the artificial Kr hump are manually selected. Importantly, both conditions are necessary and one does not imply the other. Many of the obtained low-scoring parameter sets give overall a reasonable fit but do possess the patterning defect for Kr. This underlines the main drawback of using the OLS measure. Extensive amounts of runs and additionally

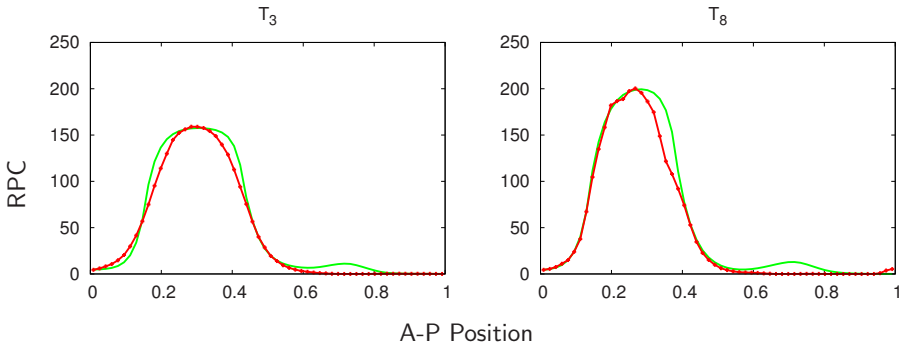


Figure 3.5: Comparison between data (red lines) and patterns obtained with a low-scoring parameter set yielded from the OLS search (green lines) for the expression of gap gene Kr at two different time points. Axes are as in Figure 3.1.

exhaustive manual work of inspection of patterns were needed in order to obtain the parameter sets which correctly describe the gap gene system.

Analysis of OLS gene circuits

Posterior *hb* domain Model outputs for the selected OLS parameter sets reveal the correct set up of the posterior boundary of the posterior *hb* domain by the end of cycle 14A (see Figure 3.15). Figure 3.6a shows the pattern generated with one of those parameter sets compared to the result obtained with the 6-gene gap system from [6]. Clearly, the result for the 4-gene model has a significantly improved fit of the posterior *hb* boundary. As we will see, this is solely due to the inclusion of Hkb in the 4-gene model which is a main repressor of *hb* in that region.

Gap gene domains are established during cycle 13 and the beginning of cycle 14A. Afterwards, there is an anterior shift in the position of these domains. This shift mechanism has been investigated and well understood by using the 6-gene model [12]. It has been noticed that the domain shifts are based only on regulatory interactions between genes and diffusion plays no role in it. The model for the 6-gene network was able to reproduce most of the domain shifts observed in the data. However, for the posterior *hb* domain the shift of its peak and posterior boundary was not present in model outputs [12].

The model outputs with the selected OLS parameter sets for the 4-gene model show the shift in the peak of the posterior *hb* domain. We illustrate it in Figure 3.6b for one of the parameter sets, where the position of the peak on the A–P axis is plotted against time. Despite the slight difference with the corresponding shift in the data, the overall shift in the model output is visible.

Additionally, our results reveal the shift of the posterior boundary of the posterior *hb* domain. Similar to the approach in [12], we performed a graphical analysis

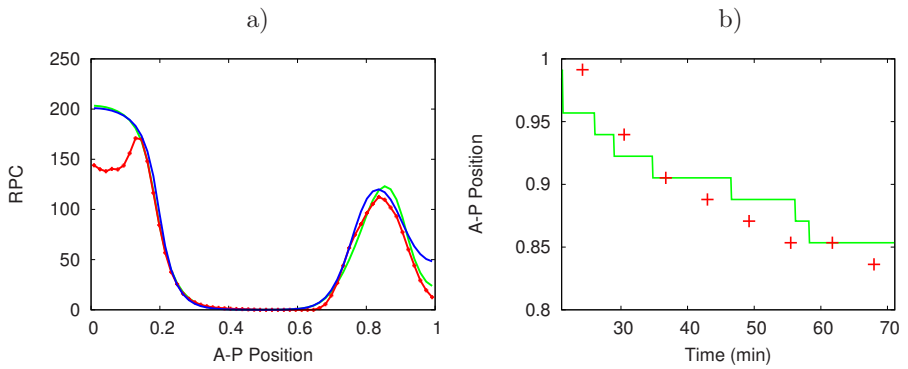


Figure 3.6: a) Comparison between data (red line), the pattern obtained by the parameter set for the 6-gene gap system from [6] (blue line), and the pattern obtained by one of the selected OLS parameter sets for our 4-gene network (green line) for the expression of gap gene *hb* at T_8. Axes are as in Figure 3.1. b) Anterior shift of the peak of the posterior *hb* domain during cycle 14A. Plot shows the A-P position of the peak in the model outputs for one of the selected OLS parameter sets (green line) compared to the corresponding shift observed in the data (+).

of the *hb* regulation over time (cycle 14A) at three different nuclei which lie in the shift zone. Panels **a-c** of Figure 3.7 show a switch from protein synthesis (positive dg_i^{hb}/dt) to decay (negative dg_i^{hb}/dt) of *hb* at the end of cycle 14A. As we can see, diffusion plays no role in it. In fact, diffusion counteracts the boundary shift with an influx of protein into the region where *hb* decays. Note that a lack of smoothness in the protein synthesis term is a consequence of using linear interpolations for time-variable external inputs in the model. Panels **d-f** of Figure 3.7 reveal that the shift is solely driven by the temporal behaviour of the regulatory input for *hb* production (solid black lines). By plotting the individual contributions (coloured areas) we can analyse in detail the regulatory mechanism which underlines the shift. The areas below and above the black line represent the regulatory input from activators (Tll, Cad, Hb, and Gt) and repressors (Kni and Hkb) of *hb*, respectively. Since the regulatory inputs from Kr and Bcd are negligible, they are not plotted here. As we can see, the activating contribution is mainly from Tll and less from Cad and autoactivation of *hb*. Note that insignificant activation by Gt is an artefact of the model. The repressing input from Kni is relatively small because of the low expression of *kni* in that region of the embryo. So, the shift is based on the regulatory input from Hkb, the main repressor of *hb*. This repression increases both in space (posteriorly) and in time.

In conclusions, our model predicts that *hb* in the posterior part of the embryo is mainly activated by Tll. However, this activation is suppressed by increasing repression of *hb* by Hkb yielding eventually the shift of the boundary domain. Contrary to the shifts of other boundaries of gap gene domains, this shift happens

at late stages of cycle 14A.

Network topology A classification of all estimates of the regulatory weights for all 39 parameter sets into 'activating', 'repressing' or 'no interaction' categories is shown in Figure 3.8. The topology is mainly in agreement with the previous findings for the 6-gene model [6]. However, some ambiguities in the network are removed with these results. Namely, the repressive regulations of Hb on Kr and gt, Gt on kni, and Kni on Kr are present in all parameter sets, while previous results for the 6-gene case showed no regulation for these weights in many solutions. Importantly, the activation of hb by Tll is correctly predicted by our model in almost all sets. Note that previously it was found that there exists no regulation for this weight. Repression of gt by Tll is present in almost all parameter sets, while previously many circuits were found with no regulation for this weight. Another remarkable difference is that autoactivation of gt is much weaker than in the 6-gene case. To be more precise, its autoregulation is not required in most of the parameter sets. Finally, we note that the colours in Figure 3.8 do not change if we choose the threshold 0.01 instead of 0.005 for the classification of regulations, except in two regulatory weights. Specifically, the activation of hb by Cad and Tll changes to no regulation category, meaning that these activations in the network topology predicted by the model are weak (almost negligible).

Confidence intervals The network topology shown in Figure 3.8 is based solely on the *values* of estimated parameters. To assess the *quality* of the parameter estimates, we computed dependent and independent confidence intervals for each parameter set using (3.2.14) and (3.2.15), respectively (see Figure 3.17 in the Appendix). We checked if the corresponding confidence intervals for regulatory weights fall entirely into the 'repression', 'no interaction', or 'activation' categories. Results in Figure 3.8 do not change when only dependent confidence intervals are taken into account. However, when including independent confidence intervals, one can no longer make similar qualitative conclusions about some entries in the regulatory weight matrix. For example, Figure 3.9 shows the confidence intervals for regulatory weights W_{Kr}^{gt} (a), E_{hb}^{Bcd} (b), and E_{kni}^{Tll} (c). The independent confidence intervals for W_{Kr}^{gt} lie in the negative part of the plane for almost all parameter estimates and therefore, repression predicted for this weight in Figure 3.8 is confirmed by statistical analysis. The independent confidence intervals for E_{hb}^{Bcd} slightly extend into the negative part of the plane. Therefore, one can make a qualitative conclusion for this weight: the model predicts that Bcd does not repress hb. Note that this is a weaker conclusion than predicting activation for this weight from Figure 3.8. In contrast, we cannot draw any qualitative conclusions about E_{kni}^{Tll}. Thus, statistical analysis does not confirm the repression of kni by Tll inferred from Figure 3.8.

Based on the independent confidence intervals, Table 3.1 summarizes the qual-

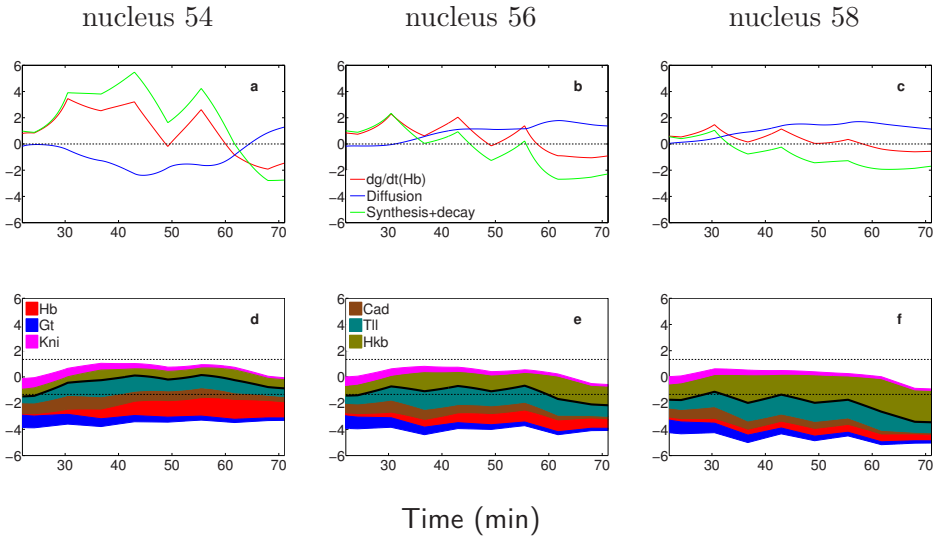


Figure 3.7: Graphical analysis of the shift of the posterior boundary of the posterior *hb* domain in the model outputs for one of the selected OLS parameter sets. Dynamic behaviour is illustrated at thee different nuclei within the shift zone over time (cycle 14A). Plots **a-c** show the rate of change in concentration of *hb* (dg_i^{hb}/dt), as well as individual contributions to it from diffusion and synthesis/decay terms in the right-hand side of (3.2.1). Plots **d-f** show the temporal behaviour of the regulatory input for *hb* production (solid black lines), i.e., $u_i^{hb} = \sum_{b=1}^{N_g} W_{hb}^b g_i^b + \sum_{e=1}^{N_e} E_{hb}^e g_i^e + h_{hb}$. Upper and lower dashed lines indicate 90% and 10% of the maximum rate of protein synthesis, respectively. The sigmoid function (3.2.2) at those values is equal to 0.9 and 0.1, respectively. Coloured areas represent individual contributions to u_i^{hb} from repressors (above black lines) and activators (below black lines) of *hb*. The height of each coloured area is given by $|W_{hb}^b| g_i^b$ or $|E_{hb}^e| g_i^e$.

	hb	*Kr*	*gt*	*kni*	*Bcd*	*Cad*	*Tll*	*Hkb*
hb	0/0/39	2/37/0	0/1/38	39/0/0	0/0/39	0/2/37	1/0/38	37/2/0
Kr	39/0/0	0/1/38	39/0/0	39/0/0	0/0/39	0/0/39	39/0/0	39/0/0
gt	39/0/0	39/0/0	0/35/4	0/0/39	0/0/39	0/0/39	38/1/0	2/2/35
kni	39/0/0	3/36/0	39/0/0	0/1/38	1/0/38	0/0/39	37/0/2	26/9/4

Figure 3.8: Gap gene network topology based on 39 selected OLS parameter sets. Each entry in the table corresponds to regulation of a gap gene given on a row by a gene given in a column. Triplets show the number of parameter sets in which a regulatory weight falls into one of the following categories: repression (values ≤ -0.005)/ no interaction (values between -0.005 and 0.005)/ activation (values ≥ 0.005). Colours: activation (green), no interaction (light-blue), repression (pink).

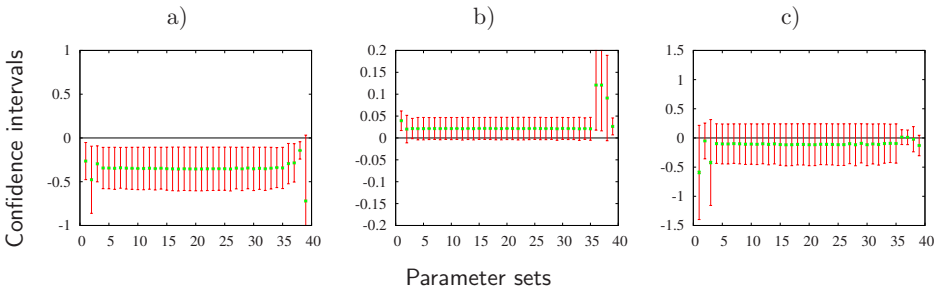


Figure 3.9: Dependent (green lines) and independent (red lines) confidence intervals for regulatory weights W_{Kr}^{gt} (a), E_{hb}^{Bcd} (b), and E_{kni}^{Tll} (c). Confidence intervals are plotted along the vertical axis for the 39 selected OLS parameter sets.

itative conclusions for the regulatory weights in the gap gene model. These conclusions are weaker than those drawn from classifying the parameter values only. Only for 17 regulatory weights out of 32, the confidence intervals confirm the type of the regulation deduced from the network topology in Figure 3.8. For other 4 weights the conclusions in Figure 3.8 are confirmed weakly. Regulations for the remaining 11 weights cannot be qualitatively verified by the confidence interval analysis. However, the conclusions in Table 3.1 show qualitative improvement for a number of regulations in comparison with the corresponding results for the 6-gene gap model [6], where only for 9 regulatory weights the confidence intervals confirmed and for other 5 weights confirmed weakly the type of the regulation deduced from the corresponding network topology.

Note that for all gap genes, promoter strengths R, diffusion coefficients D, and decay rates λ have extremely large independent confidence intervals (not shown here) meaning that all these parameters are not determinable.

tll/hkb mutants The terminal gap genes *tll* and *hkb*, being expressed in the posterior region of the embryo, are responsible for setting up the posterior boundaries of the gap gene domains. In *tll* mutants the expression of *Kr* is normal, the *kni* domain expands posteriorly, the posterior *gt* domain does not retract from the posterior pole, and the posterior *hb* domain is absent (see [10] and references therein). In *hkb* mutants the posterior *hb* domain fails to retract from the posterior pole [45]. We shall investigate here if the gap gene model is capable of reproducing such behaviour in *tll*/*hkb* mutants.

We obtain the model outputs for *tll* mutants by setting $E_a^{Tll} = 0$ for all gap genes and leaving all other parameter estimates unchanged. Figure 3.10 shows the model outputs for *tll* mutants (first row) for the expression of gap genes at time point T_8 compared to wild type data. As we can see, OLS parameter sets mainly fail to produce correct mutant phenotypes. Most of the parameter sets have over-expression of the posterior *hb* domain which contradicts the experimental

	hb	Kr	gt	kni	Bcd	Cad	Tll	Hkb
hb	$+$	$-=$	$\times$	$\times$	$+=$	$+$	$+=$	$\times$
Kr	$-$	$+$	$-$	$-$	$+$	$+$	$\times$	$\times$
gt	$-$	$\times$	$\times$	$\times$	$+$	$+$	$-$	$\times$
kni	$-$	$-=$	$-$	$+$	$+$	$+$	$\times$	$\times$

Table 3.1: Gap gene network topology based on independent confidence intervals of 39 selected OLS parameter sets. Each entry in the table corresponds to regulation of a gap gene indicated on a row by a gene indicated on a column. '$+$' ('$-$') indicates activation (repression) when the confidence intervals for the corresponding regulatory weight fall entirely into the positive (negative) part of the plane for a majority of parameter sets. Similarly, '$+=$' ('$-=$') indicates no repression (no activation) when the confidence intervals for the corresponding regulatory weight fall into the positive (negative) part of the plane and slightly extend to negative (positive) part within 'no regulation' threshold range, i.e., ≥ -0.005 (≤ 0.005). If the confidence intervals significantly extend to both sides of the plane, then no conclusion can be made (denoted by '$\times$').

evidence. In most of the cases the posterior *gt* domain is expanded and only a few model outputs have the correct behaviour when the domain does not retract from the posterior pole. The expression of *kni* has not changed in some sets and an additional domain appears in others, failing to predict the expansion of the posterior boundary. The only consistent result can be stated for *Kr* which has a normal expression in all model outputs.

Similarly, model outputs for *hkb* mutants are obtained by setting $E_a^{Hkb} = 0$ for all gap genes. Figure 3.10 (second row) shows the expression of gap genes at time point T_8 in *hkb* mutants. The posterior *hb* domain almost disappears in some circuits in contradiction to the experimental evidence. Additional expression domains appear for the gap genes *gt* and *kni*, while the expression of *Kr* has not altered.

In conclusion, OLS parameter sets fail to predict the correct behaviour when terminal genes *tll* and *hkb* are knocked out. Model outputs show both ambiguity and inconsistency, the only exception is the gap gene *Kr*.

3.3.2 WLS results

The LM search with the WLS cost function has been performed using as initial points the 39 selected OLS parameter estimates and also the 90 OLS sets with lowest RMS values possessing an artificial *Kr* hump. Additionally, we performed 80 runs starting with parameter estimates obtained from global WLS search (SA). From the obtained results we selected 117 parameter sets with WLS values varying uniformly between 1.08×10^3 and 1.13×10^3. For the comparison with the

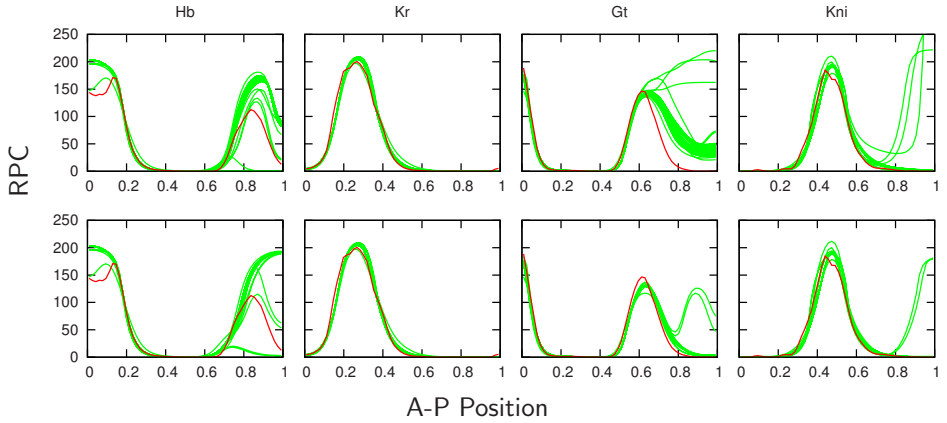


Figure 3.10: Comparison between data (red lines) and model outputs (green lines) with *tll* mutants (first row) and *hkb* mutants (second row) produced by 39 selected OLS parameter sets for the expression of gap genes at T_8. Axes are as in Figure 3.1.

OLS results, we note that these WLS parameter sets have RMS values uniformly varying between 10.41 and 10.67. It suggests that the WLS search leads to less over-fitting compared to OLS search. None of these low-scoring parameter sets show any visible patterning defects (see Figure 3.16 in the Appendix), while the sets with larger WLS values do. As it is difficult to make a distinction between these 117 parameter sets based on WLS values and expression patterns only, we take all of them into consideration. We emphasize that with a significantly less number of WLS runs (209) compared to the OLS case (740) we have obtained three times more WLS parameter estimates than OLS ones. It is also important that the selection of WLS sets is only based on cost function values and the manual inspection of model outputs for patterning defects, as in the OLS case, is not required.

The most important difference between the model outputs generated by the OLS and WLS parameter sets is that the latter do not have a patterning defect for gap gene Kr (hump). This can be expected because the standard deviations in that region of the embryo are small and subsequently the corresponding weights in WLS are relatively large which prevents the rising of the Kr hump. We note that the model outputs generated by WLS parameter sets have one slight problem which does not show up in the OLS case. Model outputs for gap gene Kr at T_1 have a slight cavity next to the anterior boundary. However, this declination does not exceed the experimental error range and therefore is not considered to be significant.

Patterns for WLS parameter sets (Figure 3.16) at cycle 13 and late time points of cycle 14A show a better fit than the corresponding OLS patterns (Figure 3.15).

Especially, the improvement can be seen at the boundaries of gap gene domains at the end of cycle 14A. This can be explained by a relatively small standard deviation of the experimental error at the domain boundaries at late time points (Figure 3.3).

Additionally, WLS model outputs have less variation than those produced by OLS parameter sets. Thereby, WLS model outputs are more consistent with each other while OLS model outputs reveal discrepancies.

Posterior *hb* domain Similar to OLS results, the posterior boundary of the posterior *hb* domain is set correctly (see Figure 3.16) and the anterior shift in the peak of the posterior *hb* domain can be detected in the model outputs (not shown here). The shift of the posterior boundary of the posterior *hb* domain is illustrated in Figure 3.11 by graphical analysis of the *hb* regulation over time (cycle 14A) at three different nuclei which lie in the shift zone. Similar to the OLS case (Figure 3.7), there is a switch from protein synthesis (positive dg_i^{hb}/dt) to decay (negative dg_i^{hb}/dt) of *hb* at the end of cycle 14A and the shift is solely based on the regulatory mechanism rather than being driven by diffusion. Panels **d-f** of Figure 3.11 show that two major contributions to the regulatory input of *hb* are from the activator Tll and the repressor Hkb. Contrary to the OLS case, these inputs are more stronger than the inputs from other regulators. So, in WLS results *hb* in the posterior region of the embryo is predominantly regulated by terminal genes Tll and Hkb.

Network topology Classification of all parameter estimates for regulatory weights for 117 selected WLS parameter sets into 'activating', 'repressing' or 'no interaction' categories is given in Figure 3.12. There are only two differences in this topology in comparison with the OLS results in Figure 3.8. Hkb represses *gt* and activates *kni*, while in the OLS case it is the other way around. For other regulatory weights all conclusions agree. At the same time, numbers in Figure 3.12 indicate that WLS parameters estimates are more consistent than those obtained by OLS search. For instance, although activation is concluded for regulatory weights E_{hb}^{Tll} and E_{kni}^{Bcd} in Figure 3.8, still in each case there is one circuit showing repression. Those ambiguities are completely cleared in Figure 3.12 (the only exception is E_{kni}^{Hkb}).

Confidence intervals To assess the quality of the parameter estimates, we computed dependent and independent confidence intervals for each parameter set using (3.2.14) and (3.2.15), respectively (see Figure 3.18 in the Appendix). We checked if the corresponding confidence intervals for the regulatory weights fall entirely into the 'repression', 'no interaction', or 'activation' categories. Similar to the OLS case, dependent confidence intervals are small and cannot be trusted. Based on the independent confidence intervals, Table 3.2 summarizes the qualitative conclusions for the regulatory weights in the gap gene model. The qualitative

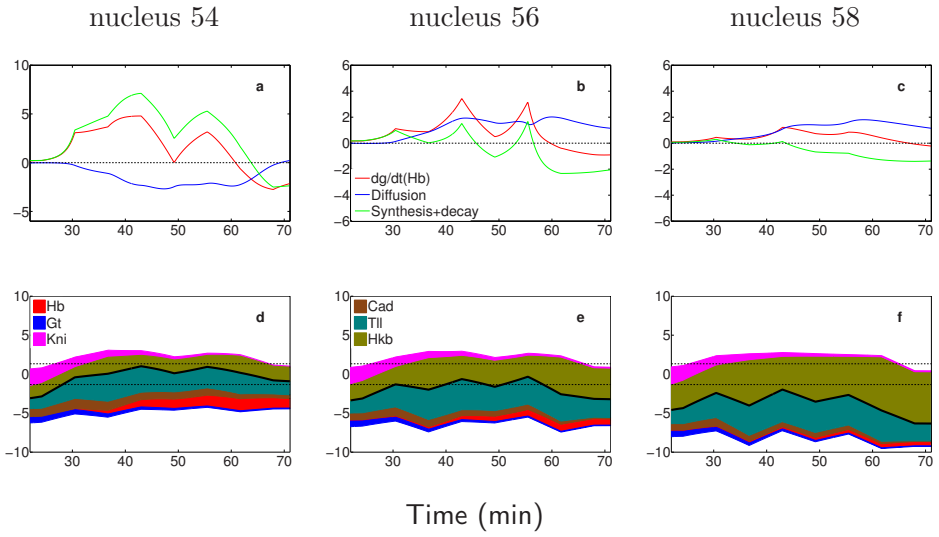


Figure 3.11: Graphical analysis of the shift of the posterior boundary of the posterior *hb* domain in the model outputs for one of the selected WLS parameter sets. Axes, lines, and coloured areas are as in Figure 3.7.

	hb	*Kr*	*gt*	*kni*	*Bcd*	*Cad*	*Tll*	*Hkb*
hb	0/0/117	0/117/0	0/0/117	117/0/0	0/0/117	0/0/117	0/0/117	117/0/0
Kr	117/0/0	0/0/117	117/0/0	117/0/0	0/0/117	0/0/117	117/0/0	117/0/0
gt	117/0/0	117/0/0	0/117/0	0/0/117	0/0/117	0/0/117	117/0/0	117/0/0
kni	117/0/0	0/117/0	117/0/0	0/0/117	0/0/117	0/0/117	117/0/0	2/0/115

Figure 3.12: Gap gene network topology based on 117 selected WLS parameter sets. Numbers and colours are as in Figure 3.8.

conclusions in Table 3.2 show no significant difference from the corresponding OLS results given in Table 3.1. For 17 regulatory weights the confidence intervals confirm and for another 3 weights they confirm weakly the type of the regulation deduced from the network topology in Figure 3.12.

	hb	Kr	gt	kni	Bcd	Cad	Tll	Hkb
hb	$+$	$\times$	$+=$	$\times$	$\times$	$+=$	$+$	$\times$
Kr	$\times$	$+$	$-$	$-$	$+$	$+$	$\times$	$\times$
gt	$-$	$-$	$\times$	$+$	$+$	$+$	$-$	$\times$
kni	$-$	$-=$	$-$	$+$	$\times$	$+$	$\times$	$\times$

Table 3.2: Gap gene network topology based on independent confidence intervals of 117 selected WLS parameter sets. Notations are as in Table 3.1.

In contrast to the OLS case, the confidence interval analysis for WLS solutions suggests that the number of unknown parameters can be reduced in the model. The dependent confidence intervals for all diffusion parameters in the WLS results have a non-empty intersection. This means that for practical reasons they can be fixed to any value in those intersections without giving a difference in the WLS sums. Since the main interest of the gap gene model lies in the inference of the regulatory network topology, the exact value of the diffusion parameters is not important. Correlation analysis shows that the diffusion coefficients are not strongly correlated to other parameters. Therefore, removing them from the parameter space will not change significantly the determinability of the remaining parameters but it will reduce the size of the problem.

tll/hkb mutants The model outputs for *tll* mutants are shown in Figure 3.13 (first row). The expression of posterior *hb* decreases compared to wild type data. Although it is not completely in agreement with experimental evidence (there is no posterior *hb* domain in such embryos), there is still an improvement in comparison with OLS results (Figure 3.10) where over-expression of *hb* is detected. Similar to OLS results, *Kr* has a normal expression which is in agreement with experiments. Expression of *gt* and *kni* at the posterior part of the embryo appears somewhat abnormal as in OLS outputs but they do not produce the behaviour observed in the experiments.

The model outputs for *hkb* mutants are shown in Figure 3.13 (second row). Contrary to the corresponding OLS results (Figure 3.10), they are more consistent with each other. The posterior *hb* domain in all cases fails to retract from the posterior pole of the embryo which is in agreement with the experimental evidence [45]. This confirms again that Hkb is the main repressor of *hb* at the posterior part of the embryo. The expression of gap genes *Kr*, *gt*, and *kni* is not affected in *hkb* mutants. It suggests that Hkb does not regulate these genes exposing an unreliability of corresponding regulations in network topology in Figure 3.12.

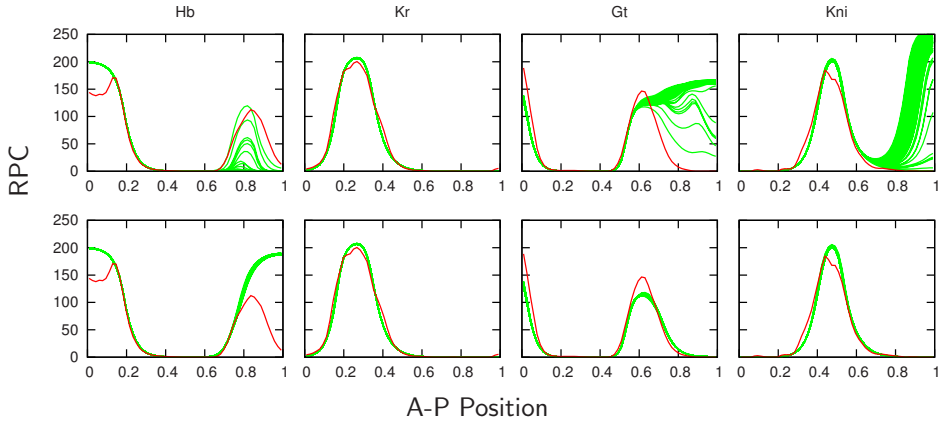


Figure 3.13: Comparison between data (red lines) and model outputs (green lines) with *tll* mutants (first row) and *hkb* mutants (second row) produced by 117 selected WLS parameter sets for the expression of gap genes at T_8. Axes are as in Figure 3.1.

Thereby, we can conclude that in the WLS search the 3 regulatory weights corresponding to the regulation of gap genes *Kr*, *gt*, and *kni* by Hkb can be eliminated from the parameter search by setting up $E_{Kr}^{Hkb} = E_{gt}^{Hkb} = E_{kni}^{Hkb} = 0$. This is also confirmed by the statistical analysis, as their dependent confidence intervals include zero (see Figure 3.18).

WLS results with fixed parameters

We have found that for the WLS search it is possible to reduce the size of the parameter space by fixing all diffusion parameters and the regulatory weights corresponding to the regulation of gap genes *Kr*, *gt*, and *kni* by Hkb. For the diffusion coefficients we computed the averaged values based on the previously found estimates, $D_{hb} = 0.237$, $D_{Kr} = D_{kni} = 0.3$, and $D_{gt} = 0.115$. Note that these averaged values belong to the non-empty intersections of the dependent confidence intervals. So, it leaves us with 37 parameters in the model to be re-estimated. We used LM search with 60 initial parameter sets arbitrarily chosen from previously found 117 WLS parameter sets. Additionally, we performed 20 runs with initial parameter values obtained from global WLS search (SA) with those parameters fixed. From re-estimated parameter sets we select 66 circuits which have low WLS values (about 1.08×10^3). None of them reveals any visible patterning defects (not shown here). The network topology in Figure 3.12 remains unchanged with the new estimates except for the regulations of *Kr*, *gt*, and *kni* by Hkb which are set to zero. Table 3.3 presents the qualitative conclusions for the regulatory weights in the gap gene model based on the independent confidence

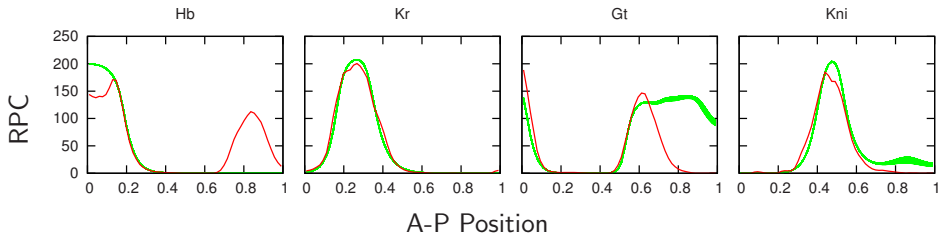


Figure 3.14: Comparison between data (red lines) and model outputs with *tll* mutants (green lines) produced by re-estimated WLS parameter sets for the expression of gap genes at T_8. Axes are as in Figure 3.1.

intervals (Figure 3.19 in the Appendix). These results show an improvement in comparison with Table 3.2. For 20 regulatory weights the confidence intervals confirm and for another 5 weights they confirm weakly the type of the regulation in the network topology and only 4 regulations still remain unclear.

	hb	Kr	gt	kni	Bcd	Cad	Tll	Hkb
hb	$+$	$\times$	$+$	$-$	$\times$	$+=$	$+$	$-=$
Kr	$-=$	$+$	$-$	$-$	$+$	$+$	$\times$	0
gt	$-$	$-$	$\times$	$+$	$+$	$+$	$-$	0
kni	$-$	$-=$	$-$	$+$	$+=$	$+$	$-$	0

Table 3.3: Gap gene network topology based on independent confidence intervals of re-estimated 66 WLS parameter sets. Notations are as in Table 3.1.

tll mutants The model outputs for *tll* mutants are shown in Figure 3.14. As we can see, there is a significant improvement in comparison with the OLS results (Figure 3.10) and preceding WLS results (Figure 3.13). Now, WLS circuits predict correct mutant phenotypes for all gap genes. Namely, the posterior *hb* domain is absent, the expression of *Kr* is normal, there is expansion of the posterior boundary of the *kni* domain, and the posterior *gt* domain does not retract from the posterior pole.

Correlations The qualitative conclusions from Table 3.3 are not completely consistent with the network topology obtained by only considering the value of the parameter estimates. The sizes of the independent confidence intervals (see Figure 3.19) give an indication about the determinability of the corresponding regulatory weights. Note the big difference between the size of the independent confidence intervals for the different regulatory weights indicating a different degree of determinability. The lack of determinability is due to the presence of

correlations between parameter estimates indicated by the large difference between dependent and independent confidence intervals. Individual confidence intervals are not informative for understanding the reason of poor determinability of parameters when their estimates are correlated. Using (3.2.16), we find the correlation matrix for each parameter set. To detect the most significant correlations between parameters present in all correlation matrices, we calculated an averaged matrix—which we call the mean correlation matrix—whose entries are the mean values of the corresponding correlation coefficients in the individual correlation matrices. The obtained mean correlation matrix has a block diagonal structure such that each block corresponds to a given gene and contains the correlation coefficients between parameters for the same gene (not shown here). This is mainly due to the nature of function (3.2.2) used in (3.2.1). The positive and negative inputs in its argument can compensate or complement each other. We identify the most significant parameter correlations which can be interpreted in biological terms with the emphasis on those for which the qualitative conclusions in Table 3.3 are weak or cannot be made at all:

- Activations of *hb* by Bcd and Cad are correlated;
- Activation of *hb* by Bcd is also correlated to its repression by Kni;
- Repression of *hb* by Hkb is correlated to activation by Tll;
- Repression of Kr by Hb is correlated to its activation by Bcd;
- Activation of *kni* by Bcd is strongly correlated to its repression by Hb;

The regulatory weights W_{hb}^{Kr}, W_{gt}^{gt}, and W_{kni}^{Kr} have relatively small independent confidence intervals. Results for these weights in Table 3.3 are based on the threshold 0.005 for classification of regulations. With a larger threshold, such as 0.01, for all 3 weights 'no regulation' type can be concluded confirming the corresponding predictions from Figure 3.12. Finally, we note that E_{Kr}^{Tll} is not correlated to any other weight. Posterior Kr is strongly repressed by Gt and somewhat weaker by Hb and Kni. Apparently, due to these interactions, repression of Kr by Tll is somewhat redundant in the model.

3.4 Conclusions

In this chapter we have investigated the model for spatio-temporal pattern formation of gap gene products (*hb*, *Kr*, *gt*, and *kni*) in early development of *Drosophila*. Previous studies of the gap gene system [6, 10, 12–14] along with these gap genes also included in the model the products of genes *cad* and *tll* as state variables. In our model we have included *cad* and *tll* as time-variable external inputs. This is a more natural way to model that they regulate gap genes, but not vice versa. Contrary to previous studies where protein Bcd was used as external input constant

in time, we have incorporated its temporal behaviour in our model. Finally, new data for *hkb* [46] has allowed us to supplement the gap gene network by including *hkb* as time-variable external input. Note that *hkb*, which is absolutely necessary for correct regulation of posterior *hb* domain, was missing in previous studies of the gap gene system. Thereby, our model describes the spatio-temporal dynamics of 4 gap genes and includes 4 external inputs. It is noteworthy that with our model the complexity of the problem is reduced both with regard to the number of equations and the number of unknown parameters compared to previous models.

The model has a number of unknown parameters among which the most interesting are the regulatory weights, each one representing quantitatively the regulation of one gene by another gene. Following the common way, we have inferred the unknown parameters by fitting model outputs to gap gene data [18, 19]. As cost function to minimize in the parameter estimation procedure we have used both the Ordinary Least Squares (OLS) sum, similar to all previous studies, and the Weighted Least Squares (WLS) sum with weights taken inversely proportional to the corresponding standard deviations of the experimental error distributions. Since the standard deviations are available from [19], the WLS method does not require additional computational work compared to the OLS search.

We have used the gradient-based Levenberg-Marquardt (LM) method in the optimization with the initial parameter values obtained from global search runs using Simulated Annealing (SA). A large amount of runs has been performed to obtain the parameter estimates, 740 and 209 with OLS and WLS search, respectively. From the obtained parameter sets we first selected the low-scoring sets based on the values of OLS and WLS sums only. It gave us 117 WLS and 589 OLS parameter sets. While the network topology based on the values of the estimated regulatory weights in the WLS case (Figure 3.12) shows an agreement with the known genetic evidence, corresponding OLS results reveal a number of contradictions. Interestingly, all those OLS sets which have disagreements with the theory, despite having an overall reasonable fit to the data, do possess one patterning defect (hump) in the expression of *Kr* in the region where this gene is not expressed in the data. By manual inspection of model outputs we have selected 39 OLS parameter sets which do not have that artefact. The network topology based on these sets (Figure 3.8) is in agreement with genetic evidence. Thus, the selection of parameter sets reveals the first drawback of using the OLS rather than the WLS measure. While the selection criterion based on the cost function value is sufficient for the WLS case, an additional check for patterning defects in the OLS model outputs has to be performed. Moreover, with WLS search we have done less estimation runs and still obtained more parameter sets than with OLS search.

The model outputs produced with the selected WLS parameter sets reveal a better fit at the boundaries of the gap gene domains at late stages of cycle 14A than the corresponding OLS patterns. Additionally, WLS patterns are more consistent with each other which is indicated by less variation in the model outputs.

Our results, both OLS and WLS, show a significant improvement in the regulation of the posterior *hb* domain compared to previous results. Namely, the posterior boundary of this domain is set up correctly and the anterior shift in the peak of the domain is present in the model outputs while previous models failed to reproduce such a shift. More importantly, with our network also the shift in time of the posterior boundary of the posterior *hb* domain is detected. We have shown that this shift is solely based on the regulatory mechanism rather than being forced by diffusion (Figures 3.7 and 3.11). Namely, the boundary shift is due to the suppressive repression of *hb* by Hkb. In previous studies, gap gene models failed to show this shift because *hkb* was missing in the network.

Confidence interval analysis for the selected OLS and WLS parameter estimates show no significant difference from each other in terms of their determinability. In both cases qualitative conclusions can be made only for 17 (out of 32) regulatory weights (Tables 3.1 and 3.2). Thus, the network topology based only on the values of parameter estimates is not entirely confirmed by confidence interval analysis. However, there is a significant improvement in comparison with the corresponding results in [6] where qualitative conclusions were deduced only for 9 weights. This improvement is most likely due to the change of genes *cad* and *tll* from state variables to external inputs in our model and decreasing by that the level of uncertainty in the model parameters.

We have used our OLS and WLS parameter sets for qualitative prediction of gap gene expression in *tll* and *hkb* mutants (Figures 3.10 and 3.13). In *tll* mutants both OLS and WLS sets fail to predict correctly the expression of gap genes, except for *Kr* which is not altered. In *hkb* mutants the posterior *hb* domain in WLS outputs does not retract from the posterior pole in agreement with the experiments [45], while OLS results fail to reproduce such behaviour. Additionally, expression of other gap genes in WLS outputs is not changed suggesting that Hkb does not regulate those genes. In OLS results this is observed only for *Kr*.

The confidence intervals for WLS parameter sets show that all diffusion parameters and the regulatory weights corresponding to regulation of *Kr*, *gt*, and *kni* by Hkb can be eliminated from parameter space, i.e., they can be fixed during the search. We have performed additional WLS runs with those parameters fixed and selected 66 low-scoring sets from the obtained results. With the new parameter estimates, firstly, we have achieved an improvement in the qualitative conclusions for some of the regulatory weights (Table 3.3). Secondly, *tll* mutants with those sets give correct qualitative predictions for the expression of all gap genes (Figure 3.14).

To sum up, based on the results of our analysis, we conclude that the WLS sum is a more suitable measure for inferring a gap gene circuit from the experimental data than the OLS sum.

3.5 Appendix: Additional plots

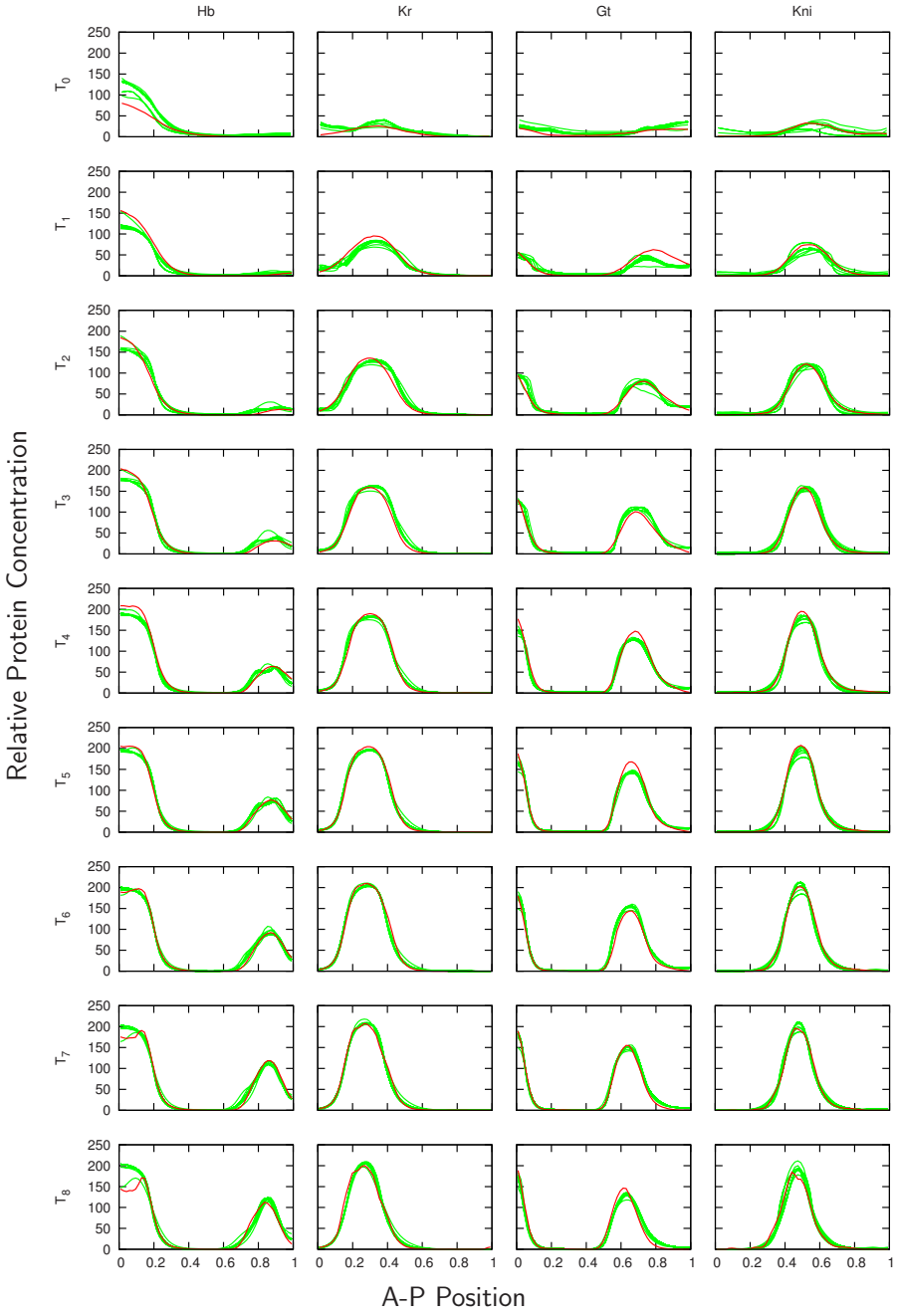


Figure 3.15: Model outputs for the 39 selected OLS parameter sets (green lines) vs data (red lines) for gap genes at all time points. Axes are as in Figure 3.1.

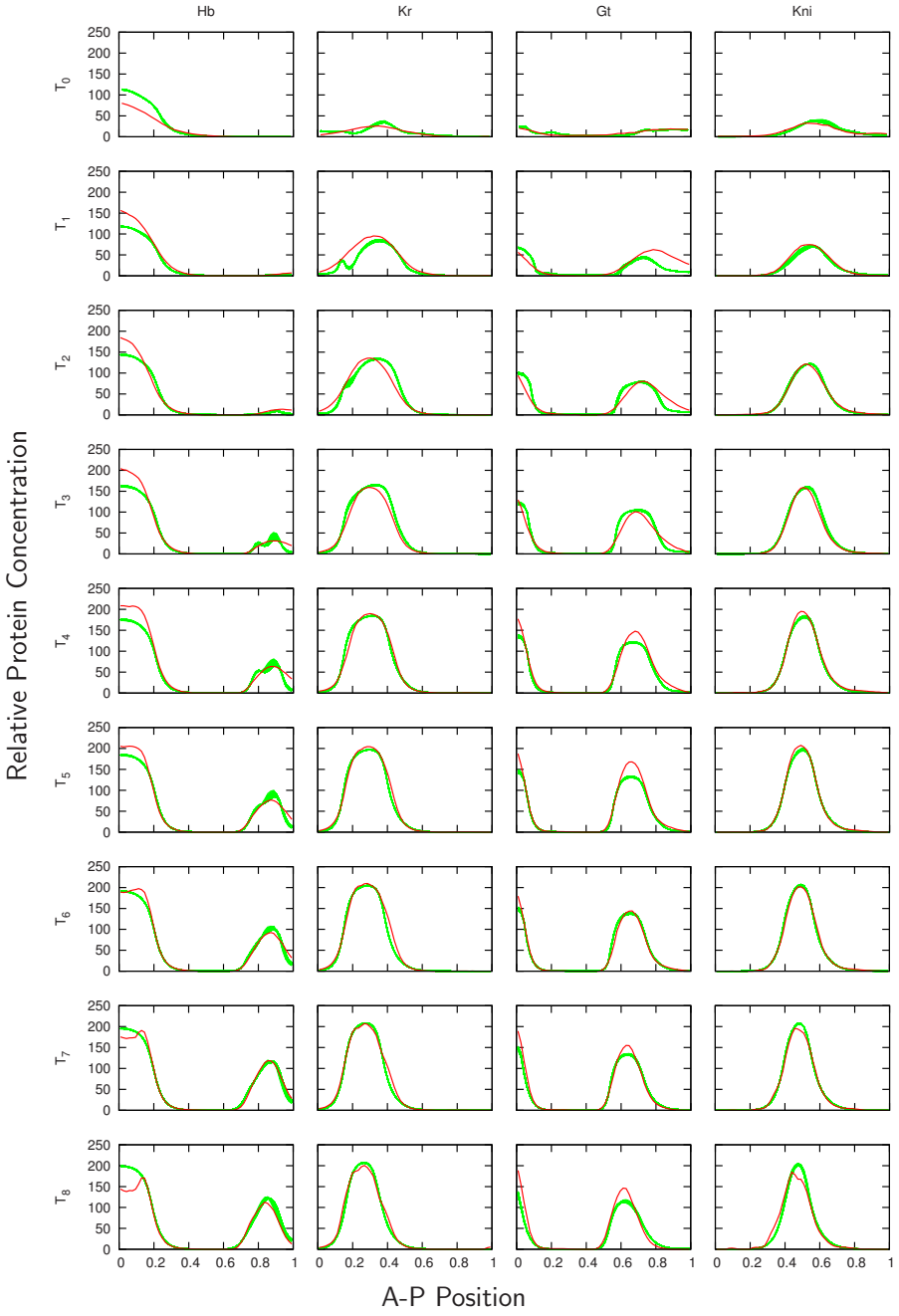


Figure 3.16: Model outputs for the 117 selected WLS parameter sets (green lines) vs data (red lines) for gap genes at all time points. Axes are as in Figure 3.1.

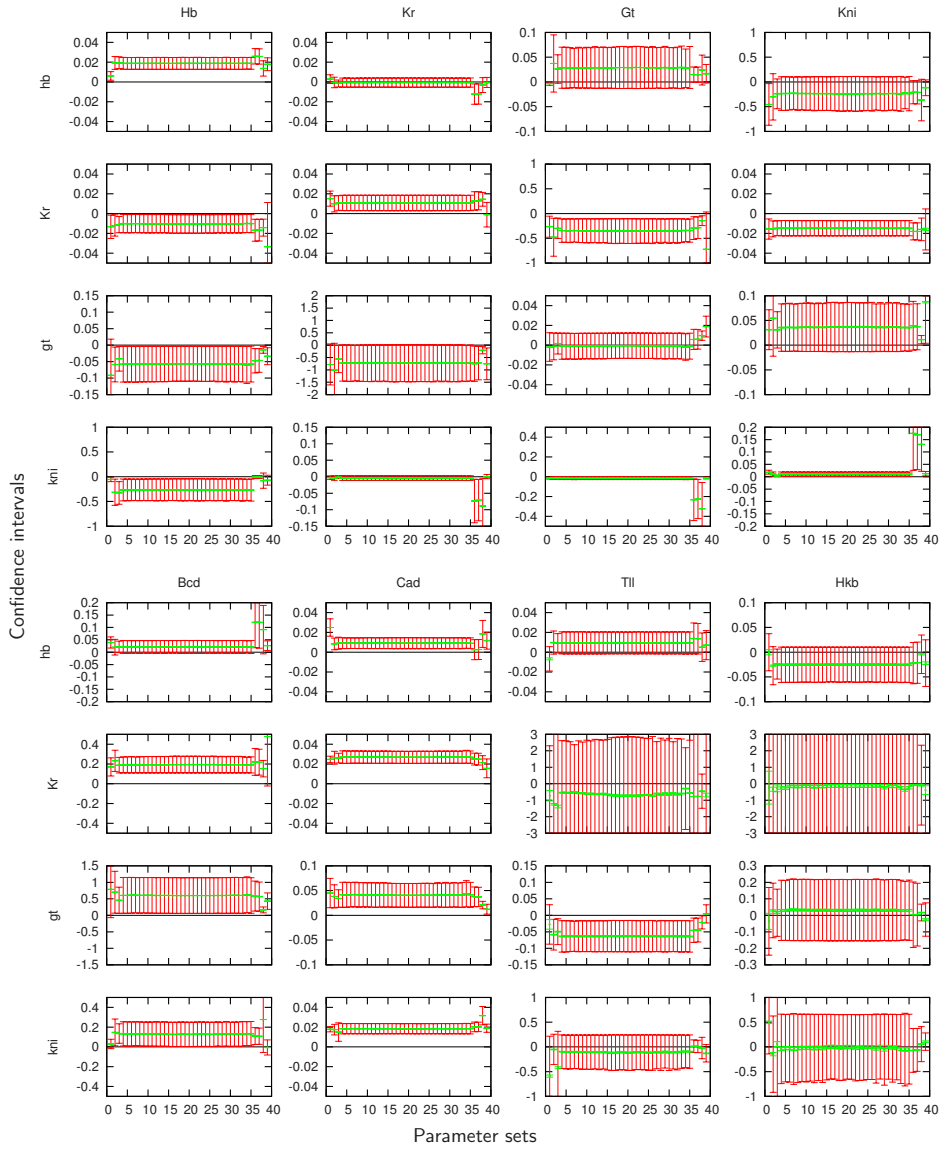


Figure 3.17: Dependent (green lines) and independent (red lines) confidence intervals for all regulatory weights in the gap gene model are plotted along the vertical axis for the 39 selected OLS parameter sets.

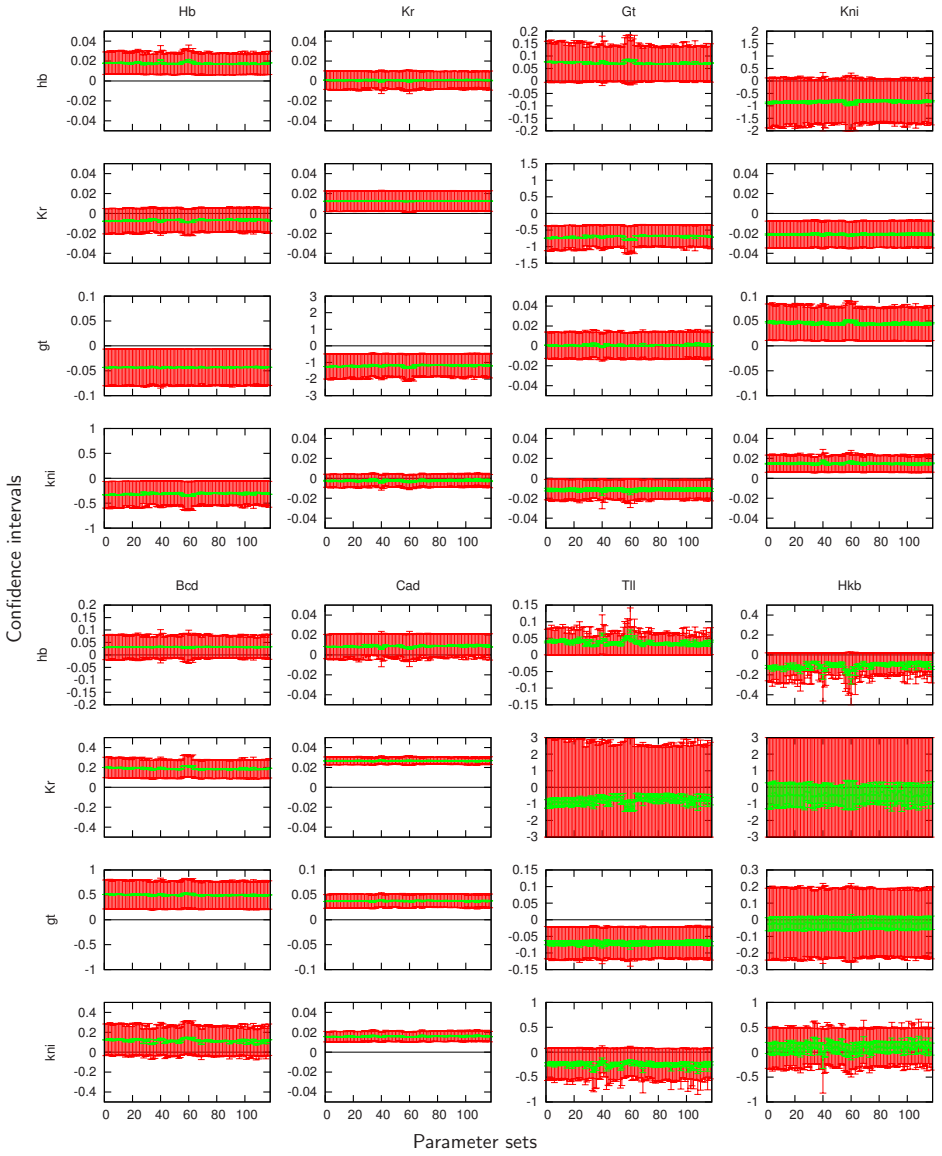


Figure 3.18: Dependent (green lines) and independent (red lines) confidence intervals for all regulatory weights in the gap gene model are plotted along the vertical axis for the 117 selected WLS parameter sets. Note the different scale in y-axis for some of the regulatory weights compared to the corresponding plot in Figure 3.17.

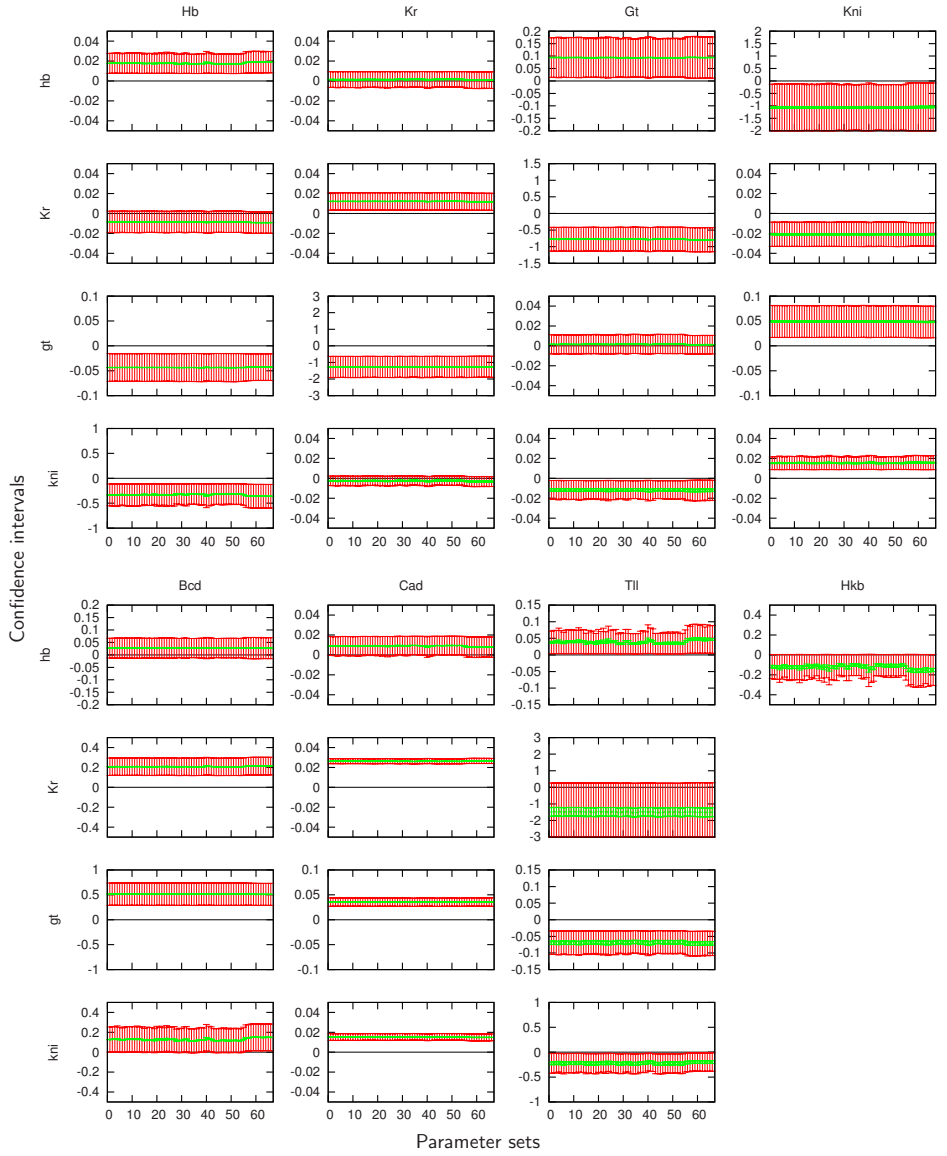


Figure 3.19: Dependent (green lines) and independent (red lines) confidence intervals for all regulatory weights in the gap gene model are plotted along the vertical axis for the 66 selected WLS parameter sets obtained with diffusion parameters and the regulatory weights corresponding to the regulation of gap genes *Kr*, *gt*, and *kni* by Hkb being fixed during the search.

Chapter 4

On the Numerical Solution of Diffusion-Reaction Equations with Singular Source Terms

Abstract

A numerical study is presented of reaction-diffusion problems having singular reaction source terms, singular in the sense that within the spatial domain the source is defined by a Dirac delta function expression on a lower dimensional surface. A consequence is that solutions will be continuous, but not continuously differentiable. This lack of smoothness and the lower dimensional surface form an obstacle for numerical discretization, including amongst others order reduction. In this paper the standard finite volume approach is studied for which reduction from order two to order one occurs. A local grid refinement technique is discussed which overcomes the reduction.

4.1 Introduction

In this paper we discuss the numerical solution of diffusion-reaction problems

$$u_t = L(u) + f \tag{4.1.1}$$

with a singular reaction source term f. Singular means here that within the domain $\Omega \subset \mathbb{R}^d$ of L the source f is defined by a Dirac delta function expression on a lower dimensional surface $\Gamma \subset \Omega$ rather than on the whole of Ω. A consequence is that the solution u is not a solution on Ω in the classical sense because across

Γ the solution u will be continuous, but not continuously differentiable. This lack of smoothness and the lower dimension of Γ form an obstacle for numerical discretization. With any numerical method one has the obvious question how to represent Γ and how to discretize f on a common grid. For finite-difference methods this question is studied in detail in [49] using regularization ideas. Regularization in the sense that the Dirac delta function expression is approximated by a source giving a small but regular support allowing standard finite difference schemes for L. In a close vicinity of Γ the lack of smoothness of u will still be felt with regularization, in the sense that in general the convergence order in the maximum norm is at best equal to one [49].

In this paper we follow the finite volume approach based on the integral form of (4.1.1). We consider this approach more natural than the finite difference one directly based on the differential form, since for the integral form the treatment of the Dirac delta function expression is mathematically clear. However, also with the finite volume approach on the uniform grid the problem of lack of smoothness remains, causing order reduction from two to one for the standard second-order spatial discretization scheme. To reobtain second order convergence we examine the finite volume approach on special locally refined grids.

The paper is organized as follows. In Section 4.2 we study the standard finite volume discretization on a uniform grid for linear and nonlinear test models. We start with linear 1D and 2D test models where the emphasis lies on boundary value problems. These test models are simple but yet significant enough to reveal the essence of a singular source. Further, we turn our attention to initial-boundary value problems having nonlinear source terms. In Section 4.3 we study the finite volume approach on a locally refined grid for 1D and 2D linear test models. The paper is concluded with remarks in Section 4.4.

4.2 The finite volume approach on the uniform grid

4.2.1 The 1D boundary value problem

We begin with the boundary value problem for the 1D equation

$$-u_{xx} = \phi(x)\,, \qquad 0 < x < 1\,, \tag{4.2.1}$$

provided with the homogeneous Dirichlet conditions $u(0) = 0, u(1) = 0$. This simple 1D problem provides a nice test model. In spite of its simplicity it already reveals essential numerical properties for the Dirac delta function source $\phi(x) = \delta(x - \bar{x})$, $\bar{x} \in (0, 1)$. For this ϕ the solution u of (4.2.1) is no longer a classical solution in the sense that it can be explicitly substituted in the differential equation. It can be determined however by the Green's function expression [50]

$$u(x) = \int_0^1 G(x, y)\phi(y)dy\,, \qquad G(x, y) = \begin{cases} x\,(1 - y)\,, & 0 \le x \le y\,, \\ y\,(1 - x)\,, & y < x \le 1\,. \end{cases} \tag{4.2.2}$$

Using the delta function property $\int_0^1 f(x)\delta(x-\bar{x}) = f(\bar{x})$, insertion of $\phi(x) = \delta(x-\bar{x})$ gives

$$u(x) = \begin{cases} x\,(1-\bar{x})\,, & 0 \le x \le \bar{x}\,, \\ \bar{x}\,(1-x)\,, & \bar{x} < x \le 1\,. \end{cases} \tag{4.2.3}$$

Note that u is continuous but not continuously differentiable over $[0,1]$.

In this section we will analyze the standard cell-centered discretization scheme for (4.2.1) obtained through the finite volume approach. For u smooth (sufficiently differentiable) this scheme converges with second order in the maximum norm. However, for the solution defined by the Dirac delta function the scheme becomes locally inconsistent near $\bar{x}$ resulting in a maximum norm order reduction from two to one for the global error. For the sake of insight we will analyze this reduction phenomenon from two points of view, viz. by introducing modified equations as in backward error analysis and by examining the local truncation error as in common (forward) error analysis.

Let $h = 1/N$ where N is the number of uniform grid cells $\Omega_i = [(i-1)h, ih]$ for $i = 1, \ldots, N$ covering $[0,1]$. Let $x_i = (i-1/2)h$ denote the cell center of Ω_i. The finite volume approach for (4.2.1) amounts to first integrating (4.2.1) over Ω_i and dividing by the cell volume,

$$\frac{\int_{\Omega_i} -u_{xx}(x)dx}{\int_{\Omega_i} dx} = \frac{\int_{\Omega_i} \phi(x)dx}{\int_{\Omega_i} dx}\,, \qquad i = 1, \ldots, N\,,$$

followed by applying the divergence (Gauss) theorem [1)],

$$-\frac{u_x(x_{i+1/2})}{\int_{\Omega_i} dx} + \frac{u_x(x_{i-1/2})}{\int_{\Omega_i} dx} = \frac{\int_{\Omega_i} \phi(x)dx}{\int_{\Omega_i} dx}\,, \qquad i = 1, \ldots, N\,, \tag{4.2.4}$$

followed by choosing a difference approximation for u_x and computing the integrals, either exact or by a quadrature rule. After incorporating the boundary conditions this procedure then results in the aimed discretization scheme.

Correct application of the divergence theorem generally requires existence and integrability of u_{xx} which does not hold with a Dirac delta function for ϕ.[2)] To circumvent this problem we will assume, for the sake of analysis only, that we are solving a modified equation defined by a modified source term associated to $\delta(x-\bar{x})$. More specifically, we will associate $\delta(x-\bar{x})$ with a class of source functions $\phi(x)$ leading to twice continuously differentiable solutions u and which are equivalent with $\delta(x-\bar{x})$ in the sense that

$$\int_{\Omega_i} \phi(x)dx = \int_{\Omega_i} \delta(x-\bar{x})dx\,, \qquad i = 1, \ldots, N\,. \tag{4.2.5}$$

[1)] The divergence theorem in 1D is called the fundamental theorem of calculus.

[2)] Solution (4.2.3) forms an exception. For this solution the divergence theorem appears to hold over the cell Ω_i containing $\bar{x}$.

The divergence theorem is then applicable for these twice continuously differentiable solutions and, furthermore, assuming exact integration or a proper quadrature rule, the ϕ-integrals in (4.2.4) are computed as if the source is the Dirac delta function. Hence the resulting difference scheme remains unaltered.

In addition to (4.2.5) we will further assume that any ϕ considered converges uniformly in x to the Dirac delta function with $\mathcal{O}(h)$ in the sense that

$$\int_0^1 G(x,y)\,(\phi(y) - \delta(y-\bar{x}))\,dy = \mathcal{O}(h)\,. \tag{4.2.6}$$

We will show that there exist a function ϕ that satisfies (4.2.5)-(4.2.6). Due to (4.2.2) requirement (4.2.6) immediately leads to first-order convergence of the exact modified solution u to the sought exact solution. This in turn implies first-order convergence of the numerical solution to the sought solution if we have first-order h-convergence of the numerical solution to the assumed modified solution. Below we will illustrate this line of thinking which is reminiscent of backward error analysis as used in numerical linear algebra or numerical differential equations, see e.g. [51].

To set up the difference scheme let us assume that $\bar{x} \in (x_{j-1/2}, x_{j+1/2})$ for a certain $j = j(h)$ at a distance ch of the cell center x_j, i.e., $\bar{x} = x_j + ch$ with $-1/2 < c < 1/2$. Then due to (4.2.5), (4.2.4) becomes

$$-\frac{u_x(x_{i+1/2})}{h} + \frac{u_x(x_{i-1/2})}{h} = \frac{\delta_{ij}}{h}\,, \qquad i = 1,\ldots,N\,, \tag{4.2.7}$$

where δ_{ij} is the Kronecker delta symbol. Next, let $w_i, i = 1,\ldots,N$, denote the numerical solution for $u(x_i)$ resulting from approximating $u_x(x_{i+1/2})$ in (4.2.7) by $(u(x_{i+1}) - u(x_i))/h$, etc. The Dirichlet boundary values are accounted for by extrapolation to auxiliary values $w_0 = 2u(0) - w_1, w_{N+1} = 2u(1) - w_N$ and by insertion of w_0 and w_{N+1} for $i = 1$ and $i = N$, respectively. If we then assemble the w_i in the grid function $w = (w_1,\ldots,w_N)^T$, we get as numerical scheme the $N \times N$ symmetric linear system [3)]

$$-Aw = b\,, \quad A = \frac{1}{h^2}\begin{pmatrix} -3 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -3 \end{pmatrix}, \quad b = \frac{1}{h}\begin{pmatrix} 0 \\ \cdot \\ 1 \\ \cdot \\ 0 \end{pmatrix}, \tag{4.2.8}$$

where b has zero entries except at entry j. The inverse of the difference matrix is bounded uniformly in $h = 1/N$, defining $w = A^{-1}b$ uniquely as the aimed numerical solution.

[3)] The values -3 at the corner entries are due to the fact that we have chosen a cell-centered grid and have Dirichlet boundary values. With a vertex-centered grid (boundary points as cell centers) the common stencil would result with -2 at the corner entries. See Section I.5.3 in [52] for accuracy aspects.

Let $u_h = (u(x_1), \ldots, u(x_N))^T$ denote the restriction of $u(x)$ to the grid. As discussed above, the numerical solution w can be interpreted as an approximation to u_h for a twice differentiable modified solution u defined by an appropriate source function ϕ satisfying (4.2.5). Likewise, once constructed, w may also be directly compared to the actually sought solution lying at a maximum norm distance $\mathcal{O}(h)$ to any appropriate modified solution.

Error analysis through modified solutions Associating a uniquely determined numerical solution with different exact modified solutions for the sake of analysis is the central idea of backward error analysis. This line of thinking may enhance insight in the numerical discretization procedure or, as in our case, even justify the discretization procedure. Here we are in the special situation of being able to find the numerical solution and exact modified solutions in closed form. One can easily check that

$$w_i = \begin{cases} x_i(1-x_j)\,, & i = 1, \ldots, j\,, \\ x_j(1-x_i)\,, & i = j+1, \ldots, N\,, \end{cases} \tag{4.2.9}$$

solves (4.2.8). This numerical solution differs from solution (4.2.3) only in that $\bar{x}$ is replaced by x_j, revealing a small shift in the peak and an error at all grid points. In terms of $\bar{x}$ and ch, with $-1/2 < c < 1/2$, we have

$$w_i = \begin{cases} x_i(1-\bar{x}) + cx_i h\,, & i = 1, \ldots, j\,, \\ \bar{x}(1-x_i) - c(1-x_i)h\,, & i = j+1, \ldots, N\,, \end{cases} \tag{4.2.10}$$

immediately showing $\mathcal{O}(h)$ maximum norm convergence. For $c = 0$, i.e., with the singular point $\bar{x}$ located in the center of cell Ω_j, the scheme returns the sought solution exactly.

Next consider by way of example the continuous source function

$$\phi(x) = \begin{cases} 0\,, & 0 \le x \le x_j - h/2\,, \\ \frac{4}{h^2}(x - x_j + h/2)\,, & x_j - h/2 \le x \le x_j\,, \\ \frac{4}{h^2}(x_j - x + h/2)\,, & x_j \le x \le x_j + h/2\,, \\ 0\,, & x_j + h/2 \le x \le 1\,, \end{cases} \tag{4.2.11}$$

which satisfies (4.2.5)-(4.2.6) and results in the twice continuously differentiable modified solution

$$u(x) = \begin{cases} x(1-x_j)\,, & 0 \le x \le x_j - h/2\,, \\ x(1-x_j) - \frac{2}{3h^2}(x - x_j + h/2)^3\,, & x_j - h/2 \le x \le x_j\,, \\ x_j(1-x) - \frac{2}{3h^2}(x_j - x + h/2)^3\,, & x_j \le x \le x_j + h/2\,, \\ x_j(1-x)\,, & x_j + h/2 \le x \le 1\,. \end{cases} \tag{4.2.12}$$

On the grid this modified solution is closer to the numerical solution (4.2.9) than the sought one as it should be. The u_h and w coincide at all grid points except at x_j where the difference is $h/12$. Observe that (4.2.11) can be interpreted as a regularized form of the singular Dirac delta function as discussed in [49]. Contrary to the approach followed here, in [49] such regularized forms are explicitly used and implemented in the used difference schemes.

Error analysis through the truncation error Following the common approach of (forward) error analysis we will next examine the convergence of (4.2.8) to the sought solution u by analyzing the local truncation error σ and global error e defined by

$$\sigma = -Au_h - b\,, \qquad e = u_h - w\,.$$

There holds $-Ae = \sigma$ so that $\|e\|_\infty \le \|A^{-1}\|_\infty \, \|\sigma\|_\infty$. Hence if $\|\sigma\|_\infty = \mathcal{O}(h^2)$ we immediately have second-order convergence in view of the uniform boundedness of $\|A^{-1}\|_\infty$. However, for the current solution (4.2.3) we find

$$\sigma = \left(0, \ldots, 0, \frac{-c}{h}, \frac{c}{h}, 0, \ldots, 0\right)^T\,, \tag{4.2.13}$$

with nonzero entries for cell j and $j+1$, respectively. Here it is assumed that $\bar{x}$ lies at the right of x_j so that $0 \le c < 1/2$. With $\bar{x}$ at the left the nonzero entries shift to cells $j-1, j$. For $c = 0$, i.e., with the singular point located in the center of cell Ω_j, a zero truncation error results and hence in this special case the scheme returns the exact solution (4.2.3). In all other cases $\|\sigma\|_\infty = \mathcal{O}(h^{-1})$ so that convergence cannot be concluded when the standard argument sketched above is followed.

Through a more subtle local truncation error analysis the correct maximum norm $\mathcal{O}(h)$ convergence can be proven however, gaining two powers of h. A similar situation generally occurs with Dirichlet boundary conditions due to the cell-centered location of x_1 and x_m half a distance h away from the boundary. For a general smooth solution we then would have $\sigma_1 = \mathcal{O}(1), \sigma_N = \mathcal{O}(1)$. In [52], Section I.5.3, it is shown that we then still can expect second-order convergence (with a sufficiently smooth source) due to a favourable local error cancellation and we adopt here the method of proof of [52] to show first-order convergence with σ given by (4.2.13) using the following ansatz: the local truncation error can be decomposed as $\sigma = -Ar + \xi$ such that the grid functions r, ξ are componentwise $\mathcal{O}(h)$. This would immediately prove first-order convergence since the global error then satisfies $e = r - A^{-1}\xi$. The ansatz is verified as follows. Put $\xi = 0$ and $r = h\alpha$. We then have to verify that such a grid function α exists and satisfies $\alpha = \mathcal{O}(1)$ componentwise. The result is

$$\alpha_i = \begin{cases} \frac{1-2i}{2N}\, c\,, & i = 1, \ldots, j\,, \\ \frac{2N-2i+1}{2N}\, c\,, & i = j+1, \ldots, N\,, \end{cases} \tag{4.2.14}$$

which completes the proof. Observe that since $\xi = 0$ we have $e = h\alpha$ connecting this expression with (4.2.9) through $w = u_h - h\alpha$.

Example 4.2.1. As a second illustration of the $\mathcal{O}(h)$ convergence of the cell-centered scheme we consider a slight extension of the 1D test model (4.2.1), viz.,

$$-u_{xx} + u = \delta(x - \bar{x}) \,, \qquad 0 < x < 1 \,, \tag{4.2.15}$$

again with homogeneous Dirichlet boundary conditions. This problem does have as solution

$$u(x) = \begin{cases} \frac{\sinh(1-\bar{x})\sinh(x)}{\sinh(1)} \,, & 0 \le x \le \bar{x} \,, \\ \frac{\sinh(1-x)\sinh(\bar{x})}{\sinh(1)} \,, & \bar{x} \le x \le 1 \,. \end{cases}$$

Figure 4.1 shows u (left plot, solid line) for $\bar{x} = 1/3$ along with the cell-centered solution for $h = 1/20$ (o-marks). The plot at the right nicely reveals the anticipated first-order convergence ($\|u_h - w\|_\infty$ versus h in log-log scale).

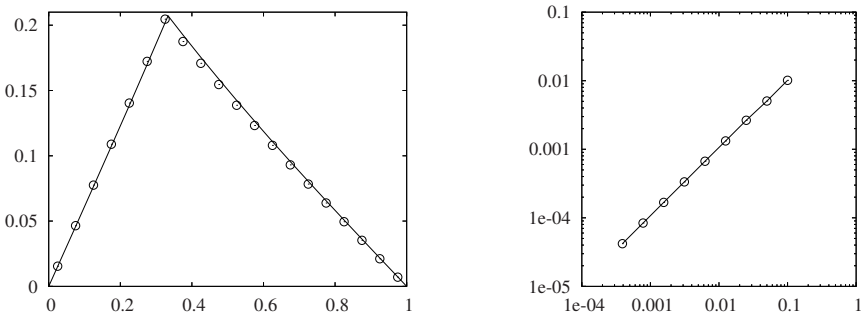


Figure 4.1: Numerical illustration for problem (4.2.15) with $\bar{x} = 1/3$ on the uniform grid.

4.2.2 The 2D boundary value problem

An interesting 2D test model used in [49] is the Poisson equation

$$-\Delta u = \delta(\Gamma, x, y) \,. \tag{4.2.16}$$

Here the source denotes the Dirac delta function along a curve Γ defined by

$$\int_{\mathbb{R}^2} \delta(\Gamma, x, y)\, dxdy = \int_\Gamma d\gamma \,, \tag{4.2.17}$$

with co-ordinate γ on Γ.

As in the 1D test model case, the solution u is continuous but not continuously differentiable across Γ so that the divergence theorem cannot be correctly applied. However, by arguing with assumptions similar to (4.2.5), (4.2.6), the divergence theorem is correctly applied for twice differentiable modified solutions.

Assuming (4.2.16) on a square, and using a uniform $N \times N$ grid the 2D counterpart of (4.2.8) reads

$$-(A \otimes I_N + I_N \otimes A)w = b\,, \tag{4.2.18}$$

where I_N is the identity matrix of size N and $\otimes$ is the direct matrix (Kronecker) product. The entries b_k of the vector $b \in \mathbb{R}^{N\times N}$ are associated to grid cells Ω_{ij} with values emanating from the boundary conditions and the source $\delta(\Gamma, x, y)$. Considering the source contribution, let $\Gamma_{ij} = \Gamma \cap \Omega_{ij}$ and $|\Gamma_{ij}|$ the length of Γ_{ij}. Then, assuming exact integration of the integral along Γ, from (4.2.17) and the finite volume approach follows that either

$$b_k = |\Gamma_{ij}|/h^2 \tag{4.2.19}$$

or $b_k = 0$ (considering only the source term contribution). Because upon intersection $|\Gamma_{ij}|$ is proportional to h, assuming h sufficiently small, b_k is then proportional to $1/h$ or equal to zero, similar as in the 1D case.

For the circle $\Gamma = \{(x, y) : (x - x_c)^2 + (y - y_c)^2 = r^2\}$ problem (4.2.16) has the radial symmetric solution [49]

$$u(x,y) = \begin{cases} u_\Gamma\,, & (x-x_c)^2 + (y-y_c)^2 \le r^2\,, \\ u_\Gamma - r\log\left(\frac{\sqrt{(x-x_c)^2+(y-y_c)^2}}{r}\right), & (x-x_c)^2 + (y-y_c)^2 \ge r^2\,, \end{cases} \tag{4.2.20}$$

where u_Γ is a given constant value on Γ. For $u_\Gamma = 1$, $r = 1/4$ and $(x_c, y_c) = (1/2, 1/2)$ we have applied scheme (4.2.18) on the unit square $0 < x, y < 1$ with Dirichlet boundary values prescribed from (4.2.20). Like in the 1D case this results in order reduction from two to one. Figure 4.2 shows the corresponding u and nicely illustrates the first-order convergence of (4.2.18). The right figure plots $\|u_h - w\|_\infty$ versus h in log-log scale. We have used exact integration along Γ (for the circle this is straightforward).

Remark 4.2.1. No finite solution exists when using the two-dimensional Dirac delta function at a point. To see this, consider $-\Delta u = \delta(x - 1/2, y - 1/2)$ on the unit square with homogeneous Dirichlet boundary values. Introduce the uniform vertex-centered (cell centers at the boundary) $N \times N$ grid with grid size $h = 1/(N + 1)$ and assume N odd. Applying as above the finite volume approach over all internal $h \times h$ grid cells then leads to a linear system $\mathcal{A}w = b$ of type (4.2.18) where A is defined as in (4.2.8), except that at the corner points -3 is replaced by -2 (the standard stencil for homogeneous Dirichlet boundary values). Further, vector b has all its components equal to zero, except at entry

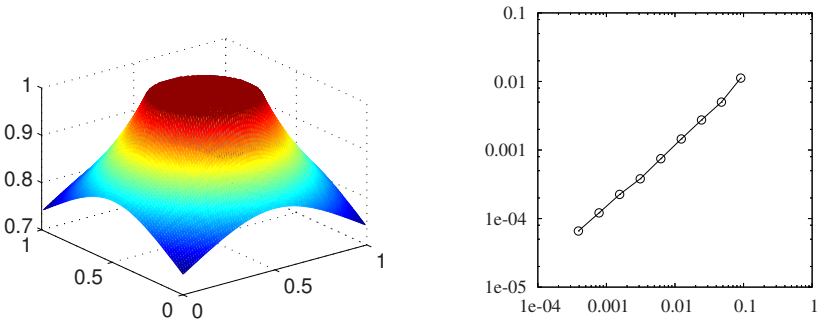


Figure 4.2: Numerical illustration for problem (4.2.16)-(4.2.20) on the uniform grid.

$k = (N^2+1)/2$ associated to the center point $(1/2, 1/2)$. Here $b_k = 1/h^2$ due to

$$\int_{\mathbb{R}^2} \delta(x - 1/2, y - 1/2) = 1\,.$$

Since $\mathcal{A}$ is symmetric, its inverse satisfies $\mathcal{A}^{-1} = XD^{-1}X^T$ with eigenvector matrix X and eigenvalue matrix D. Consequently, $w_k = (r_k \cdot D^{-1} r_k^T)/h^2$ where r_k is the k-th row of X. Using the known expressions for the eigenvectors and eigenvalues [52], Section III.6.2, we then find the expression

$$w_k = \sum_{\text{odd}\, i,j=1}^{N} 1/\mu_{ij}\,, \quad \mu_{ij} = h^{-2}\left(\sin^2\left(\frac{\pi i h}{2}\right) + \sin^2\left(\frac{\pi j h}{2}\right)\right), \quad h = \frac{1}{N+1}$$

for the numerical solution at the center point $(1/2, 1/2)$. Figure 4.3 plots the numerical solution on the 39×39 grid (left plot) and w_k as a function of increasing N-values (right plot; $N = 10 \cdot 2^l - 1$, $l = 0, 1, \ldots, 10$). Even on fine grids w_k is of moderate size, but it is obvious that $w_k \to \infty$ for $N \to \infty$. Since the w_k-sequence is defined by a convergent scheme we conclude that no finite solution exists. It can be shown analytically [53, 54] that $w_k \sim (2\pi)^{-1} \log(N)$ for $N \to \infty$ confirming the growth shown in the figure.

Remark 4.2.2. On any fixed grid no matter how fine, the strength of the Γ-source will vanish upon shrinking Γ due to (4.2.17). The strength is kept by scaling (dividing) the source by $|\Gamma|$, the length of Γ. Such scaling will replace (4.2.19) by $b_k = (|\Gamma_{i,j}|/|\Gamma|)/h^2$ and hence if Γ lies in one or only a few grid cells we have, respectively, $b_k = 1/h^2$ as if we have a point source, or $b_k \approx 1/h^2$. This situation will lead to irregular peak behavior upon grid refinement when initially Γ lies in one cell. First, as long as Γ lies in this same one cell, the peak

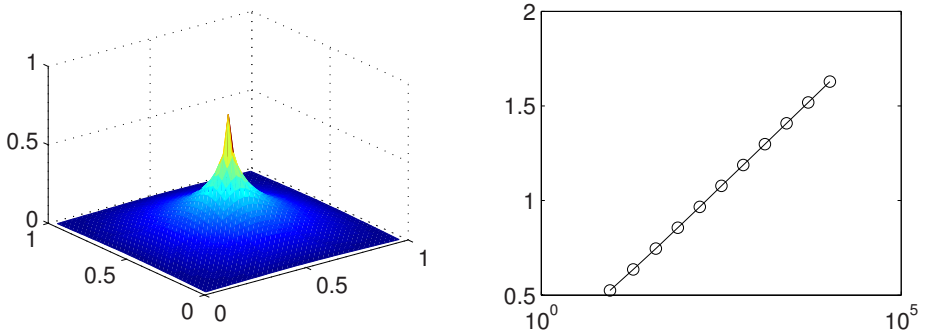


Figure 4.3: Numerical illustration for 2D problem from Remark 4.2.1.

height will increase, cf. Remark 4.2.1. Then, when Γ becomes distributed over a few cells, the height will decrease till it eventually settles down when there are enough intersections. Needless to say that this type of irregular behavior may also occur with unscaled sources, but then with a factor h smaller. Small scaled sources simply require finer grids to achieve the same level of absolute errors.

4.2.3 Linear time dependent problems

The step from boundary value problems to time-dependent problems with singular source terms is not large. Consider in a spatial domain $\Omega \in \mathbb{R}^d$ the general, constant coefficient, second order, linear test model

$$u_t = Lu + \phi(\underline{x}), \qquad t > 0, \quad \underline{x} \in \Omega, \tag{4.2.21}$$

provided with an initial function $u(\underline{x}, 0)$ at time zero, appropriate boundary conditions for $t > 0$, and a singular source term ϕ. The spatial discretization and source treatment through the finite volume approach goes essentially the same as in the boundary value case.

For any volume or grid box $V \in \mathbb{R}^d$ it starts from the integral form

$$\frac{\partial}{\partial t} \frac{\int_V u \, d\underline{x}}{\int_V d\underline{x}} = \frac{\int_V Lu \, d\underline{x}}{\int_V d\underline{x}} + \frac{\int_V \phi(\underline{x}) \, d\underline{x}}{\int_V d\underline{x}}.$$

For the u_t-integral term we will suppose the midpoint quadrature rule approximating the left-hand side by a value $w_V'(t)$ say where $'$ denotes d/dt. As in the boundary value case we will interpret ϕ to be a source function giving rise to a twice continuously differentiable solution $u(\underline{x}, t)$ so that we may apply the divergence theorem for Lu. This ϕ is then supposed to satisfy a counterpart of (4.2.5) for the singular source under consideration, so that we end up with a semi-discrete central difference scheme identical to that for the singular source.

The scheme takes the generic form of a linear, constant coefficient ODE system

$$w'(t) = Aw(t) + b\,, \qquad t > 0\,, \quad w(0) = w_0\,, \tag{4.2.22}$$

the solution of which can be expressed as

$$w(t) = e^{At}w(0) + (e^{At} - I)\,A^{-1}b\,. \tag{4.2.23}$$

If the exponential operator satisfies $\exp(At) \to 0$ for $t \to \infty$, this solution results in the steady state $w = -A^{-1}b$ for $t \to \infty$. The spatial error analysis for finite t is almost identical to the analysis for the stationary case. The local truncation error σ and global error e are now defined by

$$\sigma(t) = u_h'(t) - Au_h(t) - b\,, \qquad e(t) = u_h(t) - w(t)\,,$$

and come together in the global error equation

$$e'(t) = Ae(t) + \sigma(t)\,, \qquad t > 0\,, \quad e(0) = u_h(0) - w(0)\,.$$

Assuming bounds on the exponential matrix, $e(t)$ then can be expressed in bounds on $\sigma(t)$. Hereby one should use a refined error analysis similar as shown above to cater for local order reduction coming from a singular source. Such a refined error analysis can be found in [52], Section I.5.3, for a similar reduction coming from the boundary. These results carry over to reduction caused by singular sources.

The final assumption is that a counterpart of (4.2.6) is satisfied so that the modified solution lies at a maximum norm distance $\mathcal{O}(h)$ of the solution u generated by the singular source. First-order maximum norm convergence of $w(t)$ for the modified solution then results automatically in first-order maximum norm convergence to u. What then remains is to turn the continuous time solution $w(t)$ in a fully discrete solution by numerical time integration (method of lines). There exists a great deal of choices of integrators, depending on issues like stiffness, stability, consistency and efficiency [52]. In the next section we will pick one in a numerical illustration for a time-dependent problem with a nonlinear source.

4.2.4 Nonlinear time dependent problems

We will next consider a nonlinear extension of the linear test models discussed in the previous section. Our aim is to include singular nonlinear reaction terms, singular in the sense that the chemical reactions are confined to a lower dimensional surface Γ in the space domain $\Omega \subset \mathbb{R}^d$, similar as before. The nonlinear test model has the time-dependent form

$$u_t = Lu + \delta(\Gamma, \underline{x})\,R(\underline{x}, t, u)\,, \qquad \underline{x} \in \Omega\,, \quad t > 0\,. \tag{4.2.24}$$

The definition of L is here of secondary importance. For simplicity of presentation we assume that Lu is the linear elliptic form $Lu = \nabla \cdot (D\nabla u)$ with $D \in \mathbb{R}^{d\times d}$

diagonal and dependent on $\underline{x}$ only. The source term is supposed to satisfy the Dirac delta function relation [49]

$$\int_{\mathbb{R}^d} \delta(\Gamma, \underline{x}) R(\underline{x}, t, u) d\underline{x} = \int_{\Gamma} R(\underline{x}(\gamma), t, u) d\gamma \,, \tag{4.2.25}$$

with co-ordinate γ on Γ. The dependent variable $u(\underline{x}, t)$ is supposed to represent s concentrations, hence u is a vector $u = (u_1, \ldots, u_s)$ and accordingly the nonlinear vector function $R(\underline{x}, t, u)$ has also s components. Providing (4.2.24) with appropriate initial and boundary conditions yields the initial-boundary value problem we wish to solve.

The focus of our investigation lies in the singular source term treatment and for this purpose we take $d = 2$ with $\Omega = (-1, 1)^2$ and assume for space discretization the finite volume approach with centered finite differences on a uniform $N \times N$-grid with grid size $h = 2/N$, similar to the linear case presented in Section 4.2.2. This space discretization leads to the following nonlinear counterpart of system (4.2.22),

$$w'(t) = Aw(t) + b\,(t, w(t)) \,, \qquad t > 0 \,, \quad w(0) = w_0 \,. \tag{4.2.26}$$

With a smooth source term defined on the whole of Ω we would have second-order consistency. However, the singular source term will lead to first-order consistency and what remains to show is how this term enters the nonlinear vector function $b(t, w)$. Consider a grid cell Ω_{ij} and recall the derivation in Section 4.2.2. If $\Gamma \cap \Omega_{ij} = \emptyset$, the corresponding contributions to b are zero. If $\Gamma \cap \Omega_{ij} \neq \emptyset$, these contributions are unequal zero and are obtained from the first-order approximation

$$|\Gamma_{ij}|\, R\left(\underline{x}_{ij}, t, u(\underline{x}_{ij}, t)\right) \approx \int_{\Gamma_{ij}} R\left(\underline{x}(\gamma), t, u\right) d\gamma \,.$$

Associating the index pair (i, j) with an index k for b, we thus get

$$b_k = |\Gamma_{ij}|\, R(\underline{x}_k, t, w_k)/h^2 \,. \tag{4.2.27}$$

The computation of b thus goes essentially the same as in the linear case with regard to the singular source term treatment, the only difference being that b_k is nonlinear in the cell-center value w_k. This might complicate the numerical integration in time, but should form otherwise no obstacle for obtaining a fully discrete numerical solution.

Remark 4.2.3. Two invariants for u are positivity (componentwise non-negativity) and mass conservation. We distinguish molecular and spatial mass conservation. The first emanates from the mass action law of chemical kinetics [52] and amounts to the existence of constant, nonnegative weight vectors $v = (v_1, \ldots, v_s)^T$, such that for any solution of the ODE system $u' = R(\underline{x}, t, u)$ the inner product $v^T u(\underline{x}, t)$ is constant in time (the molecular mass defined by v).

Trivially, this holds iff $v^T R(\underline{x}, t, w) = 0$ for any $w \in \mathbb{R}^s$. The spatial conservation depends on the boundary conditions, as usual. Combining these two properties will reveal conservation of the total mass

$$M(t) = \int_\Omega v^T u(\underline{x}, t) d\underline{x}$$

associated to a given v. To see this, we compute

$$M'(t) = \int_\Omega v^T \left(\nabla \cdot (D\nabla u) + \delta(\Gamma, \underline{x})\, R(\underline{x}, t, u)\right) d\underline{x}\,,$$

and due to (4.2.25), molecular mass conservation, and the divergence theorem, we have

$$M'(t) = \int_\Omega v^T \left(\nabla \cdot (D\nabla u)\right) d\underline{x} = \sum_{i=1}^{s} \int_{\delta\Omega} (D\nabla u_i) \cdot \mathbf{n}\, d\underline{x}$$

and get as usual that the fluxes over the boundary determine conservation of the total mass.

Since we use the finite volume approach for spatial discretization this argument applies to any grid box and the standard result will then be that for the semi-discrete solution $w(t)$, $M(t)$ will be approximated over the space grid (to at least first-order in space) by a linear functional $M_h(t) = Q^T w(t)$ which mimics the time evolution of $M(t)$. In particular, $M_h(t)$ will be constant in time if this holds for $M(t)$ and any Runge-Kutta or linear multistep method will mimic this in the time integration because these methods conserve linear invariants, see [51], Section IV.1.5.

Positivity of (4.2.24) under discretization depends on the spatial discretization and the time integration. The central scheme we favor here for approximating (4.2.24) by (4.2.26) is positive [52], Section I.7. To guarantee unconditional positivity in time, that is, for any step size $\tau > 0$, we are bound to the first-order implicit Euler method [52]. We prefer however to sacrifice this guarantee in favor of more time accuracy and will instead use a second-order IMEX version of the second-order BDF method.

Example 4.2.2. We will solve (4.2.24)-(4.2.27) for the reversible reaction $2[u_2] \overset{k}{\rightleftharpoons} [u_1] + [u_2]$, giving

$$R(u) = \begin{cases} -ku_1u_2 + ku_2^2 \\ -ku_2^2 + ku_1u_2 \end{cases}$$

The corresponding ODE system $u' = R(u)$ has the exact solution

$$u_1(t) = \frac{s_0}{2}\,\frac{1 - \alpha + 2\alpha\, e^{-s_0 kt}}{1 - \alpha + \alpha\, e^{-s_0 kt}}\,, \quad u_2(t) = s_0 - u_1(t)\,,$$

where $\alpha = (2u_1(0) - s_0)/s_0$ and $s_0 = u_1(t) + u_2(t)$ which is constant in time (linear invariant). For $t \to \infty$ the ODE solution (u_1, u_2) approaches the steady state $(s_0/2, s_0/2)$ exponentially fast. Here we take $k = 1$ and a finite time interval since we are interested in transient behavior. For the PDE system we choose for the diffusion part the Laplace operator, $u_1(x, y, 0) = 0$ and $u_2(x, y, 0) = s_0 = 1$ for initial functions, and homogeneous Neumann (no flux) for boundary conditions. Due to the no-flux condition, $u_1(x, y, t) + u_2(x, y, t) \equiv s_0$. For Γ we define two circles $(x - x_0)^2 + (y - y_0)^2 = r^2$ with respectively the center points $(.5, -.5), (-.5, .5)$ and radii $.05, .25$. As in Section 4.2.2 the $|\Gamma_{ij}|$ are computed exactly.

For reasonable grid sizes h the parabolic linear part Aw in (4.2.26) will readily be stiff. The reaction constant $k = 1$ does not introduce stiffness for the nonlinear term $b(t, w)$, but due to its singular nature this term has entries proportional to $|\Gamma_{ij}|/h^2 \sim 1/h$ which causes it to be mildly stiff. For time integration we can therefore use the following implicit-explicit (IMEX) version of the BDF2 scheme [52], Section IV.4,

$$w_{n+1} = \frac{4}{3}w_n - \frac{1}{3}w_{n-1} + \frac{2}{3}\tau A w_{n+1} + \frac{2}{3}\tau\left(2b(t_n, w_n) - b(t_{n-1}, w_{n-1})\right) . \quad (4.2.28)$$

This IMEX scheme treats the linear part implicitly and the nonlinear part explicitly and thus avoids nonlinear equation solutions. It does retain the second-order of consistency of BDF2 and if we take τ proportional to h it can deal with the modest stiffness introduced by the singular source. For the first integration step, the similar IMEX form of implicit Euler [52] is used to provide the additional starting value w_1.

For $h = 2/199$ and $\tau = 1/200$ Figure 4.4 shows the computed u_1-field at $t = 0.025$ (left), 0.05 (middle), 0.1 (right). Recall that at time $t = 0$ the u_1-field is zero and note the difference in vertical scaling to see that on the current time interval both sources survive diffusion. The growth along the larger circle is larger due to the greater source strength.

4.2.5 Discussion

In this section we have studied the finite volume discretization on a uniform grid for a number of linear and nonlinear test problems. This spatial discretization, being second order convergent in the maximum norm for smooth problems, does suffer here from order reduction to first order maximum norm convergence. Its advantage is that it is straightforward and that it leads to symmetric, well structured matrices which makes it easy to solve a linear system. However, with first order convergence, only modest accuracy levels can be achieved and the question arises whether second order can be obtained in a feasible way through local grid refinement. We will address this question in the next section.

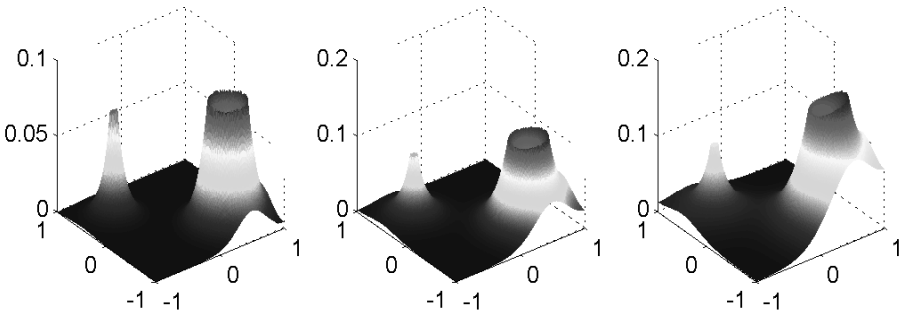


Figure 4.4: Numerical illustration for the problem from Example 4.2.2.

4.3 The finite volume approach on locally refined grids

In this section we will study the finite volume discretization on special locally refined grids for the linear 1D and 2D test models from Section 4.2. Our main aim is to obtain a numerical scheme which converges with second order in the maximum norm.

4.3.1 The 1D boundary value problem

We start again with the boundary value problem for the 1D equation

$$-u_{xx} = \delta(x - \bar{x}), \qquad 0 < x < 1, \qquad \bar{x} \in (0, 1) \tag{4.3.1}$$

provided with the homogeneous Dirichlet conditions $u(0) = 0, u(1) = 0$. The exact solution of this problem is given by (4.2.3). In this section we will consider grid refinement near $\bar{x}$.

Let $h = 1/N$ where N is the number of uniform grid cells $\Omega_i = [(i-1)h, ih]$ for $i = 1, \ldots, N$ covering $[0, 1]$. Let $x_i = (i - 1/2)h$ denote the cell center of Ω_i. The easiest way to refine the grid would be to divide grid cell Ω_j containing $\bar{x}$, now called a coarse grid cell, into two small grid cells in such a way that $\bar{x}$ is the center of one of them. It is not difficult to show that the numerical scheme obtained through finite volume discretization on this locally refined grid returns the sought solution exactly. However, this idea to refine the grid is not extendable to the 2D case. So we introduce a different way to refine the grid.

Let $D_i = [(i-1)h, ih] = \Omega_i$ for $i = 1, \ldots, j-1$ and $D_i = [(i-3)h, (i-2)h] = \Omega_{i-2}$ for $i = j+3, \ldots, N+2$. To define the new grid cells D_j, D_{j+1} and D_{j+2}, we divide the grid cell Ω_j into N small uniform grid cells $\Omega_j^1, \ldots, \Omega_j^N$ with grid

sizes h^2. Assume that $\bar{x} \in \Omega_j^k$. Then we take

$$D_j = \Omega_j^1 \cup \ldots \cup \Omega_j^{k-1}, \quad D_{j+1} = \Omega_j^k, \quad D_{j+2} = \Omega_j^{k+1} \cup \ldots \cup \Omega_j^N.$$

So, grid cells D_i for $i = 1, \ldots, j-1, j+3, \ldots, N+2$ have grid size h, grid cell D_j has grid size $h_l = (k-1)h^2$, grid cell D_{j+1} has grid size h^2 and grid cell D_{j+2} has grid size $h_r = (N-k)h^2$. We will call grid cells D_i for $i = 1, \ldots, j-2$ and $i = j+4, \ldots, N+2$ regular cells and D_i for $i = j-1, \ldots, j+3$ irregular cells. Let y_i denote the cell center of D_i for $i = 1, \ldots, N+2$. Let v_i for $i = 1, \ldots, N+2$ denote the finite volume approximation of the exact solution of problem (4.3.1) on the refined grid.

The finite volume discretization of (4.3.1) for regular cells which are at least one cell away from the boundaries gives us

$$-\frac{1}{h^2}\left(v_{i-1} - 2v_i + v_{i+1}\right) = 0, \qquad i = 2, \ldots, j-2, \qquad i = j+4, \ldots, N+1. \quad (4.3.2)$$

For the boundary grid cells, having homogeneous Dirichlet boundary conditions, we get

$$-\frac{1}{h^2}\left(v_2 - 3v_1\right) = 0, \qquad -\frac{1}{h^2}\left(v_{N+1} - 3v_{N+2}\right) = 0. \quad (4.3.3)$$

Next we derive the finite volume discretization for the irregular cells shown in Figure 4.5.

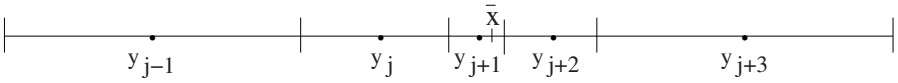


Figure 4.5: The irregular cells $D_{j-1}, \ldots D_{j+3}$ with their cell centers.

Integrating (4.3.1) over D_{j-1}, dividing by the cell volume h and applying the divergence (Gauss) theorem, gives

$$-\frac{u_x(y_{j-1/2})}{h} + \frac{u_x(y_{j-3/2})}{h} = 0.$$

By approximating $u_x(y_{j-3/2})$ by $\left(v_{j-1} - v_{j-2}\right)/h$ and $u_x(y_{j-1/2})$ by $\frac{2}{h+h_l}\left(v_j - v_{j-1}\right)$, we get

$$-\frac{1}{h^2}v_{j-2} + \left(\frac{1}{h^2} + \frac{2}{h(h+h_l)}\right)v_{j-1} - \frac{2}{h(h+h_l)}v_j = 0,$$

or, equivalently,

$$-\frac{1}{h^2}\left[v_{j-2} - \left(1 + \frac{2h}{h+h_l}\right)v_{j-1} + \frac{2h}{h+h_l}v_j\right] = 0. \quad (4.3.4)$$

Despite the fact that the approximation for $u_x(y_{j-1/2})$ is first order as it is non-centered, we will show later that the obtained numerical scheme does converge with second order in the maximum norm.

Working along the remaining irregular cells D_j, D_{j+1}, D_{j+2}, D_{j+3} we find the discretizations

$$-\frac{1}{h^2}\left[\frac{2h^2}{h_l(h+h_l)}v_{j-1}-\left(\frac{2h^2}{h_l(h+h_l)}+\frac{2h^2}{h_l(h^2+h_l)}\right)v_j+\frac{2h^2}{h_l(h^2+h_l)}v_{j+1}\right]=0, \tag{4.3.5}$$

$$-\frac{1}{h^2}\left[\frac{2}{h^2+h_l}v_j-\left(\frac{2}{h^2+h_l}+\frac{2}{h^2+h_r}\right)v_{j+1}+\frac{2}{h^2+h_r}v_{j+2}\right]=\frac{1}{h^2}, \tag{4.3.6}$$

$$-\frac{1}{h^2}\left[\frac{2h^2}{h_r(h^2+h_r)}v_{j+1}-\left(\frac{2h^2}{h_r(h^2+h_r)}+\frac{2h^2}{h_r(h+h_r)}\right)v_{j+2}+\frac{2h^2}{h_r(h+h_r)}v_{j+3}\right]=0, \tag{4.3.7}$$

$$-\frac{1}{h^2}\left[\frac{2h}{h+h_r}v_{j+2}-\left(1+\frac{2h}{h+h_r}\right)v_{j+3}+v_{j+4}\right]=0. \tag{4.3.8}$$

Combining (4.3.2)-(4.3.8), we obtain a $(N+2)\times(N+2)$ tridiagonal linear system

$$-\tilde{A}v=\tilde{b}\,,\quad \tilde{b}=\frac{1}{h^2}\begin{pmatrix}0\\ \cdot\\ 1\\ \cdot\\ 0\end{pmatrix}, \tag{4.3.9}$$

where $\tilde{b}$ has zero entries except at entry $j+1$ and $\tilde{A}$ has the same form as A in (4.2.8) except the entries at rows $j-1,\ldots,j+3$. An elementary calculation then gives the following solution:

$$v_i=\begin{cases} y_i\,(1-y_{j+1})\,, & i=1,\ldots,j\,,\\ y_{j+1}\,(1-y_i)\,, & i=j+1,\ldots,N+2\,. \end{cases} \tag{4.3.10}$$

Since $\bar{x}\in D_{j+1}$, we can write $\bar{x}=y_{j+1}+\gamma h^2$ with $-1/2<\gamma<1/2$, so that (4.3.10) can be written as

$$v_i=\begin{cases} y_i\,(1-\bar{x})\,+\gamma y_i h^2\,, & i=1,\ldots,j\,,\\ \bar{x}\,(1-y_i)\,-\gamma(1-y_i)h^2, & i=j+1,\ldots,N+2, \end{cases} \tag{4.3.11}$$

immediately showing $\mathcal{O}(h^2)$ maximum norm convergence by comparison with (4.2.3), at the expense of only two additional grid cells.

Error analysis based on the truncation error Following the common approach of error analysis as in Section 4.2.1 we will next examine the convergence of (4.3.9) to the sought solution u by analyzing the local truncation error σ and global error e defined by

$$\sigma = -\tilde{A}u_h - \tilde{b}\,, \qquad e = u_h - v\,,$$

where u_h denotes u restricted to the refined grid. Note that matrix $\tilde{A}$ is tridiagonal nonsymmetric. It can be shown that there exist a symmetric tridiagonal matrix B and a diagonal matrix D such that $B = D\tilde{A}D^{-1}$, see [55], p.51. From the uniform boundedness of $\|B^{-1}\|_\infty$, $\|D^{-1}\|_\infty$ and $\|D\|_\infty$ the uniform boundedness of $\|\tilde{A}^{-1}\|_\infty$ can be concluded.

There holds $-\tilde{A}e = \sigma$ so that $\|e\|_\infty \le \|\tilde{A}^{-1}\|_\infty \, \|\sigma\|_\infty$. For the current solution (4.2.3) we find

$$\sigma = \left(0, \ldots, 0, \frac{-2\gamma}{h_r + h^2}, \frac{2\gamma h^2}{h_r(h_r + h^2)}, 0, \ldots, 0\right)^T , \tag{4.3.12}$$

with nonzero entries for cell $j+1$ and $j+2$, respectively. Here it is assumed that $\bar{x}$ lies at the right of y_{j+1} so that $0 \le \gamma < 1/2$. So, in the case where $\gamma \neq 0$ we have $\|\sigma\|_\infty = \mathcal{O}(h^{-2})$ and convergence cannot be concluded.

Through the more subtle local truncation error analysis used in Section 4.2.1 the correct maximum norm $\mathcal{O}(h^2)$ convergence can be proven. We pose the ansatz that the local truncation error can be decomposed as $\sigma = -\tilde{A}r$ with $r = h^2\alpha$. We then have to verify that such a grid function α exists and satisfies $\alpha = \mathcal{O}(1)$ componentwise. The result is

$$\alpha_i = \begin{cases} -\gamma y_i\,, & i = 1, \ldots, j+1\,, \\ \gamma(1 - y_i)\,, & i = j+2, \ldots, N+2\,, \end{cases} \tag{4.3.13}$$

which completes the proof. Observe that we have $e = h^2\alpha$ connecting this expression with (4.3.10) through $v = u_h - h^2\alpha$.

Example 4.3.1. As an illustration of the $\mathcal{O}(h^2)$ convergence of the locally refined approach we again consider problem (4.2.15) with homogeneous Dirichlet boundary conditions.

Figure 4.6 shows u (left plot, solid line) for $\bar{x} = 1/3$ along with the numerical solution on the locally refined grid for $h = 1/20$ (o-marks). The plot at the right nicely reveals the anticipated second-order convergence ($\|u_h - v\|_\infty$ versus h in log-log scale).

4.3.2 The 2D boundary value problem

We consider the boundary value problem (4.2.16)-(4.2.17) on the unit square $0 \le x, y \le 1$ with Dirichlet boundary conditions prescribed from (4.2.20). In this

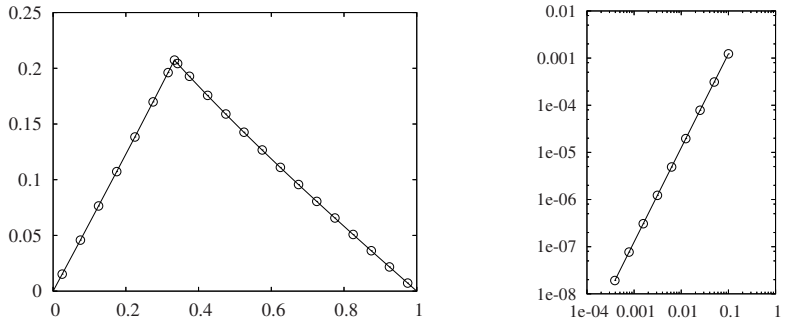


Figure 4.6: Numerical illustration for problem (4.2.15) with $\bar{x} = 1/3$ on the locally refined grid.

section we will examine a number of different approaches for grid refinement near the curve Γ, similar as in Section 4.3.1, so as to obtain a second-order discretization scheme. We start with a uniform $N \times N$ grid and from now on call the grid cells $\Omega_{i,j}$, $i = 1, \ldots, N$, $j = 1, \ldots, N$ coarse grid cells.

First approach

The easiest way to refine the grid is to divide every coarse grid cell $\Omega_{I,J}$ intersected by Γ uniformly into N^2 small grid cells with grid size h^2. Let us denote such cells by $\Omega_{I,J}^{m,k}$, $m = 1, \ldots, N$, $k = 1, \ldots, N$, where I, m are column indices and J, k are row indices. For simplicity we take N to be odd. We will now discuss the finite volume discretization for one particular grid cell $\Omega_{i,j}$.

First assume that $\Omega_{i,j}$ is at least one cell away from the boundary. Also assume that Γ intersects the neighbouring grid cells $\Omega_{i-1,j}$ and $\Omega_{i,j+1}$ and that it does not intersect $\Omega_{i,j}$, $\Omega_{i+1,j}$, $\Omega_{i,j-1}$. So, $\Omega_{i,j}$ has as neighbouring cells $\Omega_{i+1,j}$, $\Omega_{i,j-1}$, $\Omega_{i-1,j}^{N,k}$, $\Omega_{i,j+1}^{k,1}$, $k = 1, \ldots, N$. We need the auxiliary points A_k, B_k, T_k, S_k drawn in Figure 4.7:

$$A_k = (x_i - \frac{h}{2}, a_k) = (x_i - \frac{h}{2}, y_j - \frac{h}{2} + (k - \frac{1}{2})h^2), \qquad k = 1, \ldots, N,$$

$$B_k = (b_k, y_j + \frac{h}{2}) = (x_i - \frac{h}{2} + (k - \frac{1}{2})h^2, y_j + \frac{h}{2}), \qquad k = 1, \ldots, N,$$

$$T_k = (b_k, y_j), \quad S_k = (x_i, a_k), \quad k = 1, \ldots, N.$$

Let $\tilde{w}_k$ denote an approximation of u at the point S_k for $k = 1, \ldots, N$. Further, we define $\tilde{w}_k$ for $k = 1, \ldots, \frac{N+1}{2} - 1$ by linearly interpolating $w_{i,j}$ and $w_{i,j-1}$, which

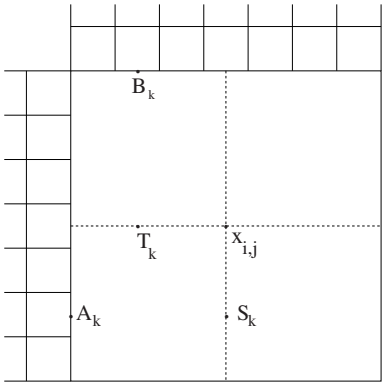


Figure 4.7: Grid cell $\Omega_{i,j}$ with its auxiliary points (first approach).

are the approximations of u at the coarse grid points $x_{i,j}$ and $x_{i,j-1}$, respectively:

$$\tilde{w}_k = \frac{1+(2k-1)h}{2} w_{i,j} + \frac{1-(2k-1)h}{2} w_{i,j-1}, \qquad k = 1, \ldots, \frac{N+1}{2} - 1. \tag{4.3.14}$$

Obviously, $\tilde{w}_{\frac{N+1}{2}} = w_{i,j}$. In order to define $\tilde{w}_k$ for $k = \frac{N+1}{2}+1, \ldots, N$ we linearly interpolate $w_{i,j}$ and $w_{i,j+1}^{\frac{N+1}{2},1}$ which is the approximation of u at the center of the fine grid cell $\Omega_{i,j+1}^{\frac{N+1}{2},1}$. Then we have

$$\tilde{w}_k = \left(2 - \frac{2kh}{1+h}\right) w_{i,j} + \left(\frac{2kh}{1+h} - 1\right) w_{i,j+1}^{\frac{N+1}{2},1}, \qquad k = \frac{N+1}{2}+1, \ldots, N. \tag{4.3.15}$$

Next we will deal with the approximations at the points T_k, $k = 1, \ldots, N$, which we denote by $\tilde{v}_k$. In order to define $\tilde{v}_k$ for $k = 1, \ldots, \frac{N+1}{2} - 1$ we linearly interpolate $w_{i,j}$ and $w_{i-1,j}^{N,\frac{N+1}{2}}$ which is the approximation of u at the center of the fine grid cell $\Omega_{i-1,j}^{N,\frac{N+1}{2}}$. Then we have

$$\tilde{v}_k = \frac{2kh}{1+h} w_{i,j} + \left(1 - \frac{2kh}{1+h}\right) w_{i-1,j}^{N,\frac{N+1}{2}}, \qquad k = 1, \ldots, \frac{N+1}{2} - 1. \tag{4.3.16}$$

Obviously, $\tilde{v}_{\frac{N+1}{2}} = w_{i,j}$. In order to define $\tilde{v}_k$ for $k = \frac{N+1}{2}+1, \ldots, N$ we linearly interpolate $w_{i,j}$ and $w_{i+1,j}$ to obtain

$$\tilde{v}_k = \frac{3-(2k-1)h}{2} w_{i,j} + \frac{(2k-1)h-1}{2} w_{i+1,j}, \qquad k = \frac{N+1}{2}+1, \ldots, N. \tag{4.3.17}$$

We are now ready to set up the finite volume discretization. Integrating (4.2.16) over $\Omega_{i,j}$ and dividing by the cell volume, we get

$$-\frac{1}{h^2}\int_{\Omega_{i,j}} \Delta u(x,y)dxdy = 0\,,$$

because the assumption is that Γ does not intersect $\Omega_{i,j}$. Applying the divergence (Gauss) theorem, gives

$$-\frac{1}{h^2}\left(\sum_{k=1}^{N}\int_{b_k-\frac{h^2}{2}}^{b_k+\frac{h^2}{2}} u_y(x,y_j+\frac{h}{2})dx - \int_{x_i-\frac{h}{2}}^{x_i+\frac{h}{2}} u_y(x,y_j-\frac{h}{2})dx\right.$$

$$\left. + \int_{y_j-\frac{h}{2}}^{y_j+\frac{h}{2}} u_x(x_i+\frac{h}{2},y)dy - \sum_{k=1}^{N}\int_{a_k-\frac{h^2}{2}}^{a_k+\frac{h^2}{2}} u_x(x_i-\frac{h}{2},y)dy\right) = 0, \qquad (4.3.18)$$

revealing the fact that Γ does intersect $\Omega_{i-1,j}$ and $\Omega_{i,j+1}$. By approximating the integrals with the common midpoint rule, the lefthand side is approximated by

$$-\frac{1}{h^2}\left(h^2\sum_{k=1}^{N} u_y(B_k) - hu_y(x_i,y_j-\frac{h}{2}) + hu_x(x_i+\frac{h}{2},y_j) - h^2\sum_{k=1}^{N} u_x(A_k)\right).$$

We use second-order approximations for the fluxes

$$u_y(x_i,y_j-\frac{h}{2}) \simeq \frac{w_{i,j}-w_{i,j-1}}{h}, \qquad u_x(x_i+\frac{h}{2},y_j) \simeq \frac{w_{i+1,j}-w_{i,j}}{h}, \qquad (4.3.19)$$

and for other fluxes first-order approximations, namely

$$\begin{aligned} u_y(B_k) &\simeq \frac{2}{h(1+h)}\left(w_{i,j+1}^{k,1}-\tilde{v}_k\right), \\ u_x(A_k) &\simeq \frac{2}{h(1+h)}\left(\tilde{w}_k - w_{i-1,j}^{N,k}\right), \quad k=1,\ldots,N, \end{aligned} \qquad (4.3.20)$$

Inserting the flux approximations (4.3.19)-(4.3.20), (4.3.18) becomes

$$-\frac{1}{h^2}\left[w_{i,j-1}+w_{i+1,j}-2w_{i,j}+\frac{2h}{1+h}\sum_{k=1}^{N}\left(w_{i,j+1}^{k,1}+w_{i-1,j}^{N,k}\right)\right.$$
$$\left. -\frac{2h}{1+h}\sum_{k=1}^{N}(\tilde{v}_k+\tilde{w}_k)\right] = 0.$$

Finally, using (4.3.14)-(4.3.17), after some calculations we obtain the discretization for the coarse grid cell $\Omega_{i,j}$ which reads

$$-\frac{1}{h^2}\left[\frac{3+h}{4}\left(w_{i,j-1}+w_{i+1,j}\right)-\frac{(3+h)^2}{2(1+h)}w_{i,j}+\frac{2h}{1+h}\sum_{k=1}^{\frac{N+1}{2}-1}\left(w_{i,j+1}^{k,1}+w_{i-1,j}^{N,k}\right)\right.$$

$$\left.+\frac{2h}{1+h}\sum_{k=\frac{N+1}{2}+1}^{N}\left(w_{i,j+1}^{k,1}+w_{i-1,j}^{N,k}\right)-\frac{1-5h}{2(1+h)}\left(w_{i,j+1}^{\frac{N+1}{2},1}+w_{i-1,j}^{N,\frac{N+1}{2}}\right)\right]=0. \tag{4.3.21}$$

In a similar way the discretization for all other cells can be obtained. However, finding the discretization for every cell explicitly in a formula similar to (4.3.21) is a cumbersome task. To avoid that, we suggest Algorithm 1 to generate the discretizations automatically.

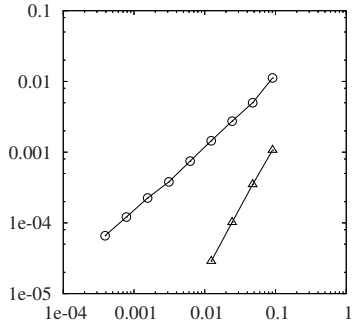
Algorithm 1 Algorithm for the finite volume discretization for a single grid cell.

```
define the neighbouring cells and cell faces
for all cell faces do
  compute the fluxes at the cell face centers:
  if the points used for the flux approximation are equally distanced from the
  cell face center then
    use the second-order approximation, similar to (4.3.19)
  else
    use the first-order approximation, similar to (4.3.20)
  end if
  if a point used in the flux approximation is not the center of the grid cell
  then
    define the approximation at that point by linearly interpolating approxi-
    mations at closest grid cells, similar to (4.3.14)-(4.3.17)
  end if
end for
compute the finite volume discretization for the cell
```

Combining it all, we then obtain an $N_c \times N_c$ linear system, where N_c is the total number of cells. If there are N_Γ coarse grid cells intersected by Γ, then $N_c = N_\Gamma(N^2-1)+N^2$. Note that, if N_Γ is of order N, then N_c is of order N^3, which means that we have gained one order in complexity compared to an overall uniform refinement.

We have applied the current 2D local refinement method to problem (4.2.16), (4.2.17), (4.2.20) on the unit square $0 \le x, y \le 1$ with Dirichlet boundary conditions. The resulting linear systems were resolved to sufficient accuracy by means of

the iterative Bi-CGSTAB method (see [56]) with ILU(0) preconditioning (see [57], p. 294). For comparison we also include the uniform grid approach in the plot.



N	$\|u_h - w\|_\infty$	Order
11	1.07×10^{-3}	
21	3.49×10^{-4}	1.73
41	1.02×10^{-4}	1.84
81	2.87×10^{-5}	1.86

Figure 4.8: Errors in max norm versus h and convergence orders on the refined grid using the first approach.

Figure 4.8 shows the first-order convergence for the uniform grid (o-marks) and second-order convergence for the refined grid approach ($\triangle$-marks). The table gives more quantitative information for the refined grid approach. We note that for $N = 161$ the number of nonzero entries in the resulting matrix is about 4.2×10^7 which makes it impossible to compute the numerical solution on our processor. However, we clearly see that the convergence order is roughly 2. Still the method is quite expensive because N_c is of order N^3. Below we will therefore present ways to refine the grid such that N_c is of order N^2.

Second approach

In our second approach coarse grid cells intersected by Γ are refined as schematically shown in Figure 4.9. In this way Γ intersects only small square cells of size h^2. The idea behind this refinement is that if N_Γ is of order N, then N_c is of order N^2, gaining one order in comparison with the first approach.

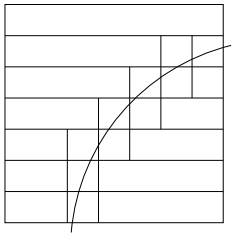


Figure 4.9: Refinement of the coarse grid cell intersected by Γ (second approach).

For the finite volume discretization on the resulting refined grid Algorithm 1 can be used in a similar way. Table 4.1 shows errors in the max norm and convergence orders for the new refined grid approach. As we can see, the refinement fails because the convergence order goes down for increasing N. The maximum error in all cases occurs in those cells where we used first-order approximation for flux computation at points at different sides of Γ. Due to the discontinuity of the first derivatives of u across Γ, such approximation causes serious order reduction. In the next section we will examine how to overcome this.

N	$\|u_h - w\|_\infty$	Order
11	2.19×10^{-3}	
21	7.52×10^{-4}	1.66
41	2.99×10^{-4}	1.38
81	1.40×10^{-4}	1.12

Table 4.1: Errors and convergence orders on the refined grid using the second approach.

Third approach

As we have seen in the previous section, the first-order approximation for the flux computation at points at different sides of Γ results in order reduction. To achieve a second-order flux approximation we introduce a 2-dimensional strip around Γ. In our case, when Γ is the circle, the strip is given by $\{(x, y) : (r-d)^2 \leq (x-x_c)^2 + (y-y_c)^2 \leq (r+d)^2\}$. Now we can refine the grid in a similar way as it was done in the previous section. Figure 4.10 shows schematically the refinement of coarse grid cells intersected by the strip.

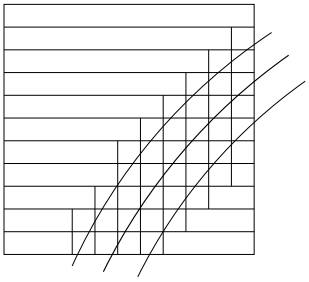


Figure 4.10: Refinement of the coarse grid cell intersected by the strip (third approach).

For the finite volume discretization on such a refined grid again Algorithm 1 introduced in Section 4.3.2 can be used. Note, that if N_Γ is of order N, then

N_c is of order N^2, similar to the second approach. Table 4.2 shows errors in the max norm and convergence orders for a strip with $d = \sqrt{2}h^2$. As we can see, the convergence order approaches 2 when N increases. Necessary for the strip approach to work is that it is wide enough. Apparently, this holds with $d = \sqrt{2}h^2$, while $d = h^2$ still shows order reduction.

N	$\|u_h - w\|_\infty$	Order
11	1.73×10^{-3}	
21	6.22×10^{-4}	1.58
41	2.09×10^{-4}	1.63
81	6.68×10^{-5}	1.68
161	1.92×10^{-5}	1.82

Table 4.2: Errors and convergence orders on the refined grid using the third approach with $d = \sqrt{2}h^2$.

Figure 4.11 summarizes all numerical results. As we have already mentioned, only on the refined grids using the first and third approach second-order convergence is obtained. The third approach is favourable due to the smaller number of nonzero entries in the resulting matrix.

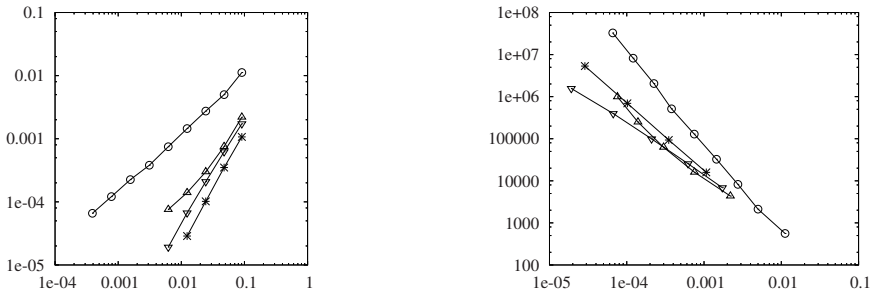


Figure 4.11: Errors in max norm versus h (left) and number of nonzero elements in the matrices versus error (right) on the uniform grid (o-marks), on the refined grid using the first approach ($*$-marks), on the refined grid using the second approach ($\triangle$-marks) and on the refined grid using the third approach with $d = \sqrt{2}h^2$ (∇-marks)

4.3.3 Time dependent problems

The extension from boundary-value problems to time-dependent problems on a locally refined grid is essentially the same as for the uniform grid approach de-

scribed in Section 4.2. Spatial discretization on a locally refined grid also leads to a system of ODEs which can be solved by an appropriate time integrator. For reaction-diffusion time-dependent problems with nonstiff reactions, by using IMEX schemes for time discretization, reaction terms can be treated explicitly and the Jacobian matrices consist of entries from the spatial discretization of the diffusion operator. For time-dependent problems the numerical stability depends on the spectrum of the matrix and the solution of the resulting linear system on the condition number. So, the matrix resulting from spatial discretization is of importance. Here we give condition numbers and spectral radii for the matrices in the uniform and refined grid case using the third approach with $d = \sqrt{2}h^2$. The second and third column in Table 4.3 give the condition numbers C_{uni} and the spectral radii ρ_{uni}, respectively, for the matrices in the uniform grid approach. The fourth and fifth column give the condition numbers C_{ref} and the spectral radii ρ_{ref}, respectively, for the matrices in the refined grid approach before ILU(0) preconditioning. The last column gives the condition numbers C^*_{ref} for the matrices in the refined grid approach after ILU(0) preconditioning. As we can see, the spectral radii are of the same order implying that the same stability constraints hold with respect to time-stepping. But the condition numbers in the local grid refinement case are much larger than in the uniform grid case. By applying the ILU(0) preconditioner, we get condition numbers significantly smaller than in the uniform grid approach.

N	C_{uni}	ρ_{uni}	C_{ref}	ρ_{ref}	C^*_{ref}
11^k	8.63×10^3	1.17×10^5	1.64×10^5	1.11×10^5	1.35×10^3
21^k	1.15×10^5	1.55×10^6	3.40×10^6	1.47×10^6	6.08×10^3
41^k	1.66×10^6	2.26×10^7	8.54×10^7	2.14×10^7	3.69×10^4
81^k	2.54×10^7	3.44×10^8	2.49×10^9	3.27×10^8	2.63×10^5
161^k	3.96×10^8	5.37×10^9	7.65×10^{10}	5.11×10^9	2.00×10^6

Table 4.3: Condition numbers and spectral radii of matrices in uniform ($k = 2$) and refined ($k = 1$) grid cases.

4.4 Concluding remarks

In this paper we have numerically studied 1D and 2D elliptic and parabolic problems with singular source terms. Such sources do contain or are defined by a Dirac delta function expression on a lower dimensional surface. By this type of singularity, solutions are not differentiable across these surfaces which hampers the spatial discretization. We have examined the common second-order finite volume technique which in itself provides a natural spatial discretization for elliptic and parabolic problems having this type of singular source terms. However, the lack of differentiability causes order reduction, that is, on a uniform grid the convergence

rate reduces from two to one in the maximum norm. To overcome this deficiency, we have examined discretization on a number of special locally refined grids, in 1D analytically and in 2D experimentally. We have shown that by an appropriate locally refined grid the maximum norm second-order convergence can be restored, and in such a way that the number of nonzero entries in the resulting difference matrix is of the same order as on the uniform grid. For time-dependent problems with nonstiff reactions integrated with IMEX methods the matrices resulting from spatial discretization of diffusion terms is of importance. We have shown that the local grid refinement approach does not change the spectral radii of these matrices, and thus the time-stepping constraints, but does increase the condition numbers. This problem can be circumvented by preconditioning the matrices. Of course, in 2D it is unavoidable that the sparsity pattern does change and that we loose the simple 5-point structure of the uniform grid. This will unavoidably increase solution costs, whether a direct sparse solver or a preconditioned iterative solver will be used. Therefore, despite the fact that restoring second-order maximum norm convergence is of clear interest in its own, from the practical point of view we advise the locally refined grid technique only when higher accuracy levels are wanted. Figure 4.11 (right plot) illustrates this nicely. In our tests the first-order convergent uniform grid solution is still computed faster up to modest accuracy levels.

Chapter 5

On Parameter Estimation for Delay Models with Discontinuous Right-Hand Sides

Abstract

We study delay models with discontinuous right-hand side. Lack of smoothness in the solutions of such problems may have serious consequences for parameter estimation using gradient-based approaches. Additionally, it may cause ambiguities in the parameter determinability analysis applied on the parameter estimates. In order to overcome these difficulties, we suggest a standard regularization technique to make the model continuous. We prove the convergence of the solution of the regularized model to the solution of the original problem. As a consequence of that, parameter estimates inferred from the regularized model converge to the corresponding estimates of the original problem. We support our findings with numerical illustrations for simple test problems.

5.1 Introduction

Many real-life processes can be modeled by non-linear Ordinary Differential Equations (ODEs) or Partial Differential Equations (PDEs). There is a growing number of applications which use Delay Differential Equations (DDEs). For instance, delay models are widely used in biosciences [58, 59]. Models with DDEs are usually more complicated than corresponding ODE or PDE models, both with regard to theoretical analysis and to numerical simulation. However, for many problems

it has been shown that DDE models provide a better description of reality, as well as a better understanding of the process under consideration. The increasing number of such applications stimulates us to study the common numerical challenges one may encounter while using DDE models. It is well-known that the solutions of DDEs suffer from lack of smoothness [60]. This can be a property of ODE or PDE solutions as well. However, for ODE and PDE cases it is rather an exception, while for DDEs it is more like a rule.

Similar to ODE and PDE models, DDE models may have a number of unknown parameters among which there can be unknown time lags. Sometimes, it is feasible to determine the missing parameters experimentally, but in most cases this is difficult or even impossible. However, one can usually measure other quantities involved in the model. Unknown model parameters can then be found by parameter estimation techniques based on fitting the model solution to the measured data. Lack of smoothness for the solutions of DDEs can have serious consequences for parameter estimation [20]. Gradient-based optimization methods are based on sufficient smoothness of the solution. So, once the solution of a DDE or its derivatives have discontinuities, gradient-based methods may fail to find parameter estimates. Despite the discontinuities parameter inference with gradient-based methods may succeed in some cases, but in general for such models one has to use gradient-free optimization techniques (such as direct search) or methods based on generalized gradients [61].

Once the parameters are found by means of optimization, it is important to assess the quality of the parameter estimates, e.g., by computing confidence intervals. This is especially important if one wants to deduce for the real-life process under consideration quantitative or qualitative conclusions based on parameter estimates. Such conclusions may greatly depend on how reliable the estimates are. One also might be interested in computing the correlations between parameters. The computation of confidence intervals and correlations is based on sufficient smoothness of the solution at the neighborhood of parameter estimates. Therefore, for DDE models, even if one succeeds with parameter estimation, still statistical analysis may be ambiguous in some cases.

In this chapter we will focus our attention on a specific type of non-linear DDE models with application in developmental biology. These models do have a discontinuous right-hand side such that the derivative with respect to parameters is discontinuous. We will discuss how the discontinuities arise in the gradients of the solution. We will demonstrate how these discontinuities may affect the estimation of parameters and the statistical analysis. There is a finite number of such discontinuity points. Unlike in the case of Neutral Delay Differential Equations (NDDEs), they do not propagate in time because it is well-known that the solution of DDEs smooths out [60]. As a natural remedy, we use the regularization technique to make the right-hand side of the DDEs continuous. Since the time lags are constant, this can be done in a straightforward manner. To apply such a technique in practice, one of course has to ask the following questions. How large is the difference between the solutions of the original and the regularized

model? How large is the difference between their gradients? How much differ the parameter estimates inferred from original and regularized models? How does regularization affect the statistical analysis? Finally, it is important to know how all these issues depend on the type of regularization used. We shall answer these questions in this chapter by means of theoretical proofs and numerical illustrations for our test problems. Note that our approach is similar to the regularization method used in [62] for NDDEs having discontinuous solutions. The main significance of our results is that we are able to derive the rate of convergence of the solution of the regularized problem to the solution of the original problem. Namely, we will show that the solution of the regularized problem, which can be also considered as an approximate solution, has second-order convergence to the solution of the original problem almost everywhere in time and first-order convergence only in the vicinity of discontinuity points. Our convergence results do not depend explicitly on the way the model is regularized. Therefore, with a specific choice of regularization we achieve not only the continuity of the right-hand side of the model but also its continuous differentiability gaining extra smoothness in the solution of the regularized model. Finally, our results imply that the parameter values estimated from the regularized model converge to corresponding estimates of the original problem.

The chapter is organized as follows. In Section 5.2 we introduce the notations, describe the problem, outline the main difficulties caused by discontinuities, and define the regularized problem to overcome them. In Section 5.3 we prove the convergence theorem for the solution of the regularized model. Section 5.4 supports our findings with numerical illustrations. We conclude the chapter with a discussion in Section 5.5.

5.2 Notations and problem description

We consider DDEs of the form

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{y}}{\mathrm{d}t}(t,\theta) &= \mathrm{H}(t-\tau)\ \mathbf{f}(t,\mathbf{y}(t-\tau,\theta),\lambda) - \mathrm{A}(\lambda)\ \mathbf{y}(t,\theta), & 0 < t \le T, \\ \mathbf{y}(t,\theta) &= \Psi(t), & t \le 0, \end{cases} \tag{5.2.1}$$

where $\mathbf{y}$ is an n-dimensional state vector, $\mathbf{f}$ is a given nonlinear vector function, continuous and sufficiently differentiable with respect to all arguments, τ is an unknown time lag [1)], λ is an $(m-1)$-dimensional (unknown) parameter vector, $\theta = \{\lambda, \tau\}$ [2)] is a vector of size m, A is a real-valued symmetric matrix, the initial

[1)] For ease of presentation, we assume that there is only one time lag. The presented material is extendable to the case with more than one time lag.

[2)] We distinguish between the parameters τ and λ because they feature in a different way in the analysis. Note that the right-hand side of (2.1) is continuous in λ and discontinuous in τ.

function $\Psi(t)$ is continuous, and H is the Heaviside function

$$\mathrm{H}(t) = \begin{cases} 0, & t < 0, \\ 1, & t \geq 0. \end{cases} \tag{5.2.2}$$

For simplicity of presentation, without loss of generality, we assume that the initial function Ψ does not depend on the parameter vector θ. This type of DDEs finds an application in the modelling of regulatory networks. The non-linear part in the right-hand side of (5.2.1) models protein synthesis, while the linear part includes protein decay and the spatial discretization of protein diffusion [28].

Note that the solution of (5.2.1) does not depend on $\Psi(t)$ $(t < 0)$. In fact, only $\Psi(0)$ is needed for the mathematical definition of the model. However, as we shall see in the next section, if $\Psi(t)$ $(t < 0)$ is available then using that knowledge improves the convergence of the solution of the regularized model to the solution of (5.2.1).

Let us assume that for fitting (5.2.1) there are N measurements available. Each measurement, which we denote by $\tilde{y}_i$, is specified by the time t_i when the c_i-th component of the state vector $\mathbf{y}$ is measured. The corresponding model value obtained from (5.2.1) is denoted by $y_{c_i}(t_i, \theta)$. With the chosen notation, two measurements $y_{c_i}(t_i, \theta)$ and $y_{c_j}(t_j, \theta)$ $(i \neq j)$ can indicate one of the following:

- two different components at two different time points (when $c_i \neq c_j, t_i \neq t_j$);
- the same component at two different time points (when $c_i = c_j, t_i \neq t_j$);
- two different components at the same time point (when $c_i \neq c_j, t_i = t_j$).

We denote the vector of weighted discrepancies between the model outputs and the measured values by $\mathrm{Y}(\theta)$. Then the least squares estimate $\hat{\theta}$ of the parameters is the value of θ that minimizes the sum of squares

$$S(\theta) = \sum_{i=1}^{N} w_i^2 (y_{c_i}(t_i, \theta) - \tilde{y}_i)^2 = \mathrm{Y}^T(\theta)\mathrm{Y}(\theta), \tag{5.2.3}$$

where w_i are positive weights. If the measurement errors in $\tilde{y}_i$ are independent of each other and normally distributed with zero mean and standard deviations σ_i, then the minimization of (5.2.3), with the weights w_i being inversely proportional to the corresponding standard deviations σ_i, yields the Maximum Likelihood Estimate (MLE) [22].

Remark 5.2.1. From the continuity of the solution of (5.2.1) it follows that the objective function (5.2.3) is continuous in the parameters contained in θ.

The gradient-based optimization methods require the objective function $S(\theta)$ to be continuously differentiable with respect to θ. Differentiation of (5.2.3) with

respect to component θ_j, $j = 1, 2, \ldots, m$, gives

$$\frac{\partial S}{\partial \theta_j}(\theta) = 2 \sum_{i=1}^{N} w_i^2 (y_{c_i}(t_i, \theta) - \tilde{y}_i) \frac{\partial y_{c_i}}{\partial \theta_j}(t_i, \theta) = 2\ \mathrm{Y}^T(\theta) \frac{\partial \mathrm{Y}}{\partial \theta_j}(\theta). \tag{5.2.4}$$

Obviously, discontinuities in $\frac{\partial S}{\partial \theta_j}$ can arise only from discontinuities in $\frac{\partial \mathbf{y}}{\partial \theta_j}$. We shall investigate here under which circumstances this may happen. The derivatives $\frac{\partial \mathbf{y}}{\partial \theta_j}$ can be found by solving the so-called variational equations. First, we denote

$$\mathbf{f}_\mathbf{y} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t, \mathbf{y}, \theta), \quad \mathbf{f}_\lambda = \frac{\partial \mathbf{f}}{\partial \lambda}(t, \mathbf{y}, \theta), \quad \mathbf{f}_\tau = \frac{\partial \mathbf{f}}{\partial \tau}(t, \mathbf{y}, \theta). \tag{5.2.5}$$

Differentiating (5.2.1) with respect to λ_i, $i = 1, 2, \ldots, m-1$, we have

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} \dfrac{\partial \mathbf{y}}{\partial \lambda_i}(t, \theta) &= \mathrm{H}(t-\tau) \left[\mathbf{f}_\mathbf{y}\ \dfrac{\partial \mathbf{y}}{\partial \lambda_i}(t-\tau, \theta) + \mathbf{f}_{\lambda_i} \right] \\ & \quad - \dfrac{\partial \mathrm{A}}{\partial \lambda_i}(\lambda)\ \mathbf{y}(t, \theta) - \mathrm{A}(\lambda)\ \dfrac{\partial \mathbf{y}}{\partial \lambda_i}(t, \theta), & 0 < t \le T, \\ \dfrac{\partial \mathbf{y}}{\partial \lambda_i}(t, \theta) &= 0, & t \le 0. \end{cases} \tag{5.2.6}$$

Since the right-hand side of (5.2.1) is continuously differentiable with respect to λ_i, the derivation of (5.2.6) is valid.

Remark 5.2.2. From the continuity of the solution of (5.2.6) it follows that the gradient $\frac{\partial S}{\partial \lambda_i}$ is continuous in the parameters θ_j, $j = 1, 2, \ldots, m$.

Differentiating (5.2.1) with respect to τ, we obtain

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} \dfrac{\partial \mathbf{y}}{\partial \tau}(t, \theta) &= -\delta(t-\tau)\ \mathbf{f} - \mathrm{A}(\lambda)\ \dfrac{\partial \mathbf{y}}{\partial \tau}(t, \theta) \\ & \quad + \mathrm{H}(t-\tau)\ \mathbf{f}_\mathbf{y}\ \left[\dfrac{\partial \mathbf{y}}{\partial \tau}(t-\tau, \theta) - \dfrac{\mathrm{d}\mathbf{y}}{\mathrm{d}t}(t-\tau, \theta) \right], & 0 < t \le T, \\ \dfrac{\partial \mathbf{y}}{\partial \tau}(t, \theta) &= 0, & t \le 0, \end{cases} \tag{5.2.7}$$

where δ is the Dirac delta function. Note that (5.2.7) is defined at $t = \tau$ only in the generalized sense, i.e., the time derivative of the solution does not have to exist at that point. For example, the problem $\frac{\mathrm{d}v}{\mathrm{d}t} = \delta(t - t_0)$, $v(0) = 0$, defined in the generalized sense, has as solution $v(t) = H(t - t_0)$.

The solution of (5.2.7) is discontinuous at $t = \tau$. Although (5.2.7) is a system of NDDEs, the right-hand side of it does not depend on the time derivative of $\frac{\partial \mathbf{y}}{\partial \tau}$ but does depend on the time derivative of $\mathbf{y}$, the solution of (5.2.1), which smooths out in time. Thereby, the discontinuity in $\frac{\partial \mathbf{y}}{\partial \tau}$ at $t = \tau$ does not propagate in time which is a common property in general for solution of NDDEs.

Remark 5.2.3. The discontinuity in $\frac{\partial \mathbf{y}}{\partial \tau}$ at $t = \tau$ enters in $\frac{\partial S}{\partial \tau}$ if this point coincides with one of the data points, i.e., if $\tau = t_i$ for some $i = 1, \ldots, N$.

Example 5.2.1. Consider the simple problem

$$\begin{cases} \dfrac{\mathrm{d}y}{\mathrm{d}t}(t,\tau) = \mathrm{H}(t-\tau)y(t-\tau,\tau), & 0 < t \le 3\tau, \\ y(t,\tau) = 1, & t \le 0. \end{cases}$$

Its solution is

$$y(t,\tau) = \begin{cases} 1, & 0 \le t < \tau, \\ t - \tau + 1, & \tau \le t < 2\tau, \\ \dfrac{(t-2\tau+1)^2}{2} + \tau + \dfrac{1}{2}, & 2\tau \le t \le 3\tau. \end{cases}$$

Therefore,

$$\frac{\partial y}{\partial \tau}(t,\tau) = \begin{cases} 0, & 0 \le t < \tau, \\ -1, & \tau \le t < 2\tau, \\ -2t + 4\tau - 1, & 2\tau \le t \le 3\tau. \end{cases}$$

For this example y is continuous everywhere, while $\frac{\partial y}{\partial \tau}$ is discontinuous at $t = \tau$.

Discontinuities in $\frac{\partial y}{\partial \theta}$ may cause ambiguity for statistical analysis of parameter estimates, such as computing confidence intervals and correlations between parameters. Assume that $\hat{\theta}$ has been found by means of some optimization method. Dependent and independent confidence intervals for each estimate $\hat{\theta}_i$ $(i = 1, 2, \ldots, m)$ are given by

$$\left[\hat{\theta}_i - \triangle_i^D,\ \hat{\theta}_i + \triangle_i^D\right],\ \triangle_i^D = \sqrt{\frac{m}{N-m}\frac{S(\hat{\theta})\ F_\alpha(m, N-m)}{\left(J^T(\hat{\theta})\ J(\hat{\theta})\right)_{ii}}} \tag{5.2.8}$$

and

$$\left[\hat{\theta}_i - \triangle_i^I,\ \hat{\theta}_i + \triangle_i^I\right],\ \triangle_i^I = \sqrt{\frac{m}{N-m}\ S(\hat{\theta})\ F_\alpha(m, N-m)\left(\left(J^T(\hat{\theta})\ J(\hat{\theta})\right)^{-1}\right)_{ii}}, \tag{5.2.9}$$

respectively. Here $J = \frac{\partial \mathrm{Y}}{\partial \theta}$ is the so-called sensitivity matrix and $F_\alpha(m, N-m)$ is the upper α part of Fisher's distribution with m and $N - m$ degrees of freedom. The correlation coefficient between $\hat{\theta}_i$ and $\hat{\theta}_j$ is given by:

$$\rho_{ij} = \frac{B_{ij}}{\sqrt{B_{ii}B_{jj}}}, \tag{5.2.10}$$

where $B(\hat{\theta}) = \left(J^T(\hat{\theta})\ J(\hat{\theta})\right)^{-1}$. See [6] and references therein for explanations of these statistical quantities.

The computation of confidence intervals and correlation coefficients both are based on $J(\hat{\theta})$. If the discontinuities in $\frac{\partial \mathbf{y}}{\partial \theta}$ enter into J and the estimate itself is a discontinuity point, then (5.2.8)-(5.2.10) are not valid and one cannot deduce any conclusions.

So far, we have described two major difficulties if one uses model (5.2.1). Both of them are related to discontinuities in $\frac{\partial \mathbf{y}}{\partial \theta}$. These discontinuities stem from the discontinuous right-hand side of the model. To overcome these problems, we consider a regularized analogy of (5.2.1) given by:

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{z}}{\mathrm{d}t}(t,\theta) &= \mathrm{H}_\epsilon(t-\tau)\ \mathbf{f}(t,\mathbf{z}(t-\tau,\theta),\lambda) - \mathrm{A}(\lambda)\ \mathbf{z}(t,\theta), \quad 0 < t \le T, \\ \mathbf{z}(t,\theta) &= \Psi(t), \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad t \le 0, \end{cases} \tag{5.2.11}$$

where H_ϵ is a regularization of the Heaviside function (5.2.2) on the interval $[-\epsilon, \epsilon]$, with $0 < \epsilon \ll 1$. For instance, the regularization by linear and cubic polynomials are given by:

$$\mathrm{H}_\epsilon(t) = \begin{cases} 0, & t \le -\epsilon, \\ \dfrac{t}{2\epsilon} + \dfrac{1}{2}, & -\epsilon < t < \epsilon, \\ 1, & t \ge \epsilon, \end{cases} \tag{5.2.12}$$

and

$$\mathrm{H}_\epsilon(t) = \begin{cases} 0, & t \le -\epsilon, \\ -\dfrac{t^3}{4\epsilon^3} + \dfrac{3t}{4\epsilon} + \dfrac{1}{2}, & -\epsilon < t < \epsilon, \\ 1, & t \ge \epsilon, \end{cases} \tag{5.2.13}$$

respectively. As we shall see in Section 5.3, under certain assumptions the solution of (5.2.11) converges to the solution of the original problem (5.2.1) as $\epsilon \to 0$. None of those assumptions nor the convergence rate depend on the explicit form of H_ϵ. In fact, any continuous, monotonous and sufficiently differentiable on $(-\epsilon, \epsilon)$ function such that $\mathrm{H}_\epsilon(t) = 0,\ t \le -\epsilon$ and $\mathrm{H}_\epsilon(t) = 1,\ t \ge \epsilon$ is a suitable regularization of the Heaviside function (5.2.2). Any realization of $\mathrm{H}_\epsilon(t)$ has the following properties:

- Using $\mathrm{H}_\epsilon(-\epsilon) = 0,\ \mathrm{H}_\epsilon(\epsilon) = 1$, and Taylor expansion, we have $\mathrm{H}_\epsilon(0) = \frac{1}{2} + O(\epsilon^2)$;
- Monotonicity of $\mathrm{H}_\epsilon(t)$ implies that $0 \le \mathrm{H}_\epsilon(t) \le 1$ for any t.

Note that both (5.2.12) and (5.2.13) are continuous and monotonous. However, (5.2.13) is continuously differentiable, while (5.2.12) is not. Thus, (5.2.13)

adds extra smoothness to the solution of (5.2.11) in comparison with (5.2.12) at no cost. Therefore, in all numerical examples presented in Section 5.4 we use (5.2.13), unless it is mentioned differently.

Remark 5.2.4. Contrary to the solution of the original problem (5.2.1) which does not depend on $\Psi(t),\ t < 0$, the solution of (5.2.11) with the Heaviside function regularized on $[-\epsilon, \epsilon]$ does depend on $\Psi(t),\ -\epsilon \leq t \leq 0$. Therefore, when the initial function $\Psi(t),\ t < 0$ is not known such regularization cannot be practically applied. In such case one has to use a regularization of the Heaviside function (5.2.2) on the time interval $[0, \epsilon]$. For instance, the analogy of (5.2.13) is given by:

$$\mathrm{H}_\epsilon(t) = \begin{cases} 0, & t \leq 0, \\ -\dfrac{2t^3}{\epsilon^3} + \dfrac{3t^2}{\epsilon^2}, & 0 < t < \epsilon, \\ 1, & t \geq \epsilon. \end{cases} \tag{5.2.14}$$

We denote the vector of weighted discrepancies between the model outputs (5.2.11) and the measured values by $\mathrm{Z}(\theta)$. Similar to (5.2.3), we define the function to minimize:

$$R(\theta) = \sum_{i=1}^{N} w_i^2 (z_{c_i}(t_i, \theta) - \tilde{y}_i)^2 = \mathrm{Z}^T(\theta)\mathrm{Z}(\theta). \tag{5.2.15}$$

The value of θ that minimizes (5.2.15) is denoted by $\breve{\theta}$. Similar to (5.2.3), (5.2.15) is continuous due to the continuity of the solution of (5.2.11).

Differentiation of (5.2.15) with respect to θ_j gives

$$\frac{\partial R}{\partial \theta_j}(\theta) = 2 \sum_{i=1}^{N} (z_{c_i}(t_i, \theta) - \tilde{y}_i) \frac{\partial z_{c_i}}{\partial \theta_j}(t_i, \theta) = 2\ \mathrm{Z}^T(\theta) \frac{\partial \mathrm{Z}}{\partial \theta_j}(\theta). \tag{5.2.16}$$

The derivatives $\frac{\partial \mathbf{z}}{\partial \theta_j}$ can be found by solving the corresponding variational equations. The analogy of (5.2.6) for $\frac{\partial \mathbf{z}}{\partial \lambda_i}$ $(i = 1, 2, \ldots, m-1)$ reads

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} \dfrac{\partial \mathbf{z}}{\partial \lambda_i}(t, \theta) &= \mathrm{H}_\epsilon(t - \tau) \left[\mathbf{f}_\mathbf{y}\ \dfrac{\partial \mathbf{z}}{\partial \lambda_i}(t - \tau, \theta) + \mathbf{f}_{\lambda_i} \right] \\ & \quad - \dfrac{\partial \mathrm{A}}{\partial \lambda_i}(\lambda)\ \mathbf{z}(t, \theta) - \mathrm{A}(\lambda)\ \dfrac{\partial \mathbf{z}}{\partial \lambda_i}(t, \theta), & 0 < t \leq T, \\ \dfrac{\partial \mathbf{z}}{\partial \lambda_i}(t, \theta) &= 0, & t \leq 0, \end{cases} \tag{5.2.17}$$

which has a continuous solution $\frac{\partial \mathbf{z}}{\partial \lambda}$. So, similar to the original case, the gradient $\frac{\partial R}{\partial \lambda_i}$ is continuous in the parameters $\theta_j,\ j = 1, 2, \ldots, m$. Differentiating (5.2.11)

with respect to τ, we obtain

$$\left\{\begin{array}{rcll} \dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial \mathbf{z}}{\partial \tau}(t,\theta) & = & -\delta_\epsilon(t-\tau)\ \mathbf{f} - \mathrm{A}(\lambda)\ \dfrac{\partial \mathbf{z}}{\partial \tau}(t,\theta) & \\ & & +\mathrm{H}_\epsilon(t-\tau)\ \mathbf{f}_{\mathbf{y}}\ \left[\dfrac{\partial \mathbf{z}}{\partial \tau}(t-\tau,\theta) - \dfrac{\mathrm{d}\mathbf{z}}{\mathrm{d}t}(t-\tau,\theta)\right], & 0 < t \leq T, \\ \dfrac{\partial \mathbf{z}}{\partial \tau}(t,\theta) & = & 0, & t \leq 0, \end{array}\right. \tag{5.2.18}$$

where δ_ϵ is a regularization of the Dirac delta function. For instance, with regularization (5.2.13) of the Heaviside function the corresponding regularization of the Dirac delta function is given by

$$\delta_\epsilon(t) = \mathrm{H}'_\epsilon(t) = \left\{\begin{array}{ll} -\dfrac{3t^2}{4\epsilon^3} + \dfrac{3}{4\epsilon}, & |t| < \epsilon, \\ 0, & |t| \geq \epsilon. \end{array}\right. \tag{5.2.19}$$

Since the right-hand side of (5.2.11) is continuously differentiable with respect to τ, (5.2.18) is defined for any $0 < t \leq T$ contrary to (5.2.7). The solution of (5.2.18) is continuous. Therefore, for the regularized problem the gradient $\frac{\partial R}{\partial \tau}$ is also continuous contrary to the original case.

Similar to (5.2.8)-(5.2.9), dependent and independent confidence intervals for each estimate $\breve{\theta}_i$ $(i = 1, 2, \ldots, m)$ are given by:

$$\left[\breve{\theta}_i - \triangle_i^D,\ \breve{\theta}_i + \triangle_i^D\right],\ \triangle_i^D = \sqrt{\frac{m}{N-m}\frac{R(\breve{\theta})\ F_\alpha(m, N-m)}{\left(\breve{J}^T(\breve{\theta})\ \breve{J}(\breve{\theta})\right)_{ii}}} \tag{5.2.20}$$

and

$$\left[\breve{\theta}_i - \triangle_i^I,\ \breve{\theta}_i + \triangle_i^I\right],\ \triangle_i^I = \sqrt{\frac{m}{N-m}\ R(\breve{\theta})\ F_\alpha(m, N-m)\left(\left(\breve{J}^T(\breve{\theta})\ J(\breve{\theta})\right)^{-1}\right)_{ii}}, \tag{5.2.21}$$

respectively, where $\breve{J} = \frac{\partial \mathrm{Z}}{\partial \theta}$. The correlation coefficient between $\breve{\theta}_i$ and $\breve{\theta}_j$ is given by:

$$\breve{\rho}_{ij} = \frac{\breve{B}_{ij}}{\sqrt{\breve{B}_{ii}\breve{B}_{jj}}}, \tag{5.2.22}$$

where $\breve{B}(\breve{\theta}) = \left(\breve{J}^T(\breve{\theta})\ \breve{J}(\breve{\theta})\right)^{-1}$. Note that in the regularized case, (5.2.20)-(5.2.22) are unambiguously defined for any value of $\breve{\theta}$.

To conclude, model (5.2.1) has a discontinuous right-hand side. This discontinuity may enter into the gradient of the solution. By regularizing the right-hand

side of (5.2.1), we achieve continuity of the gradient. It allows, firstly, to use gradient-based methods for estimation of unknown parameters and, secondly, to define statistical quantities for parameter estimates in a mathematically correct way.

5.3 Convergence analysis for the regularized problem

In this section we shall prove the convergence of the solution of the regularized problem (5.2.11) to the solution of the original problem (5.2.1). Assume that matrix A in the right-hand side of (5.2.1) is symmetric and row-diagonally dominant with positive diagonal entries [3)]. Then, it is positive definite and thus its eigenvalues are all real and positive.

Remark 5.3.1. For any symmetric and positive definite matrix A the estimate:

$$\|e^{-\mathrm{A}t}\| \leq e^{-\mu t} \tag{5.3.1}$$

holds for any $t > 0$, where μ is the smallest eigenvalue of A and $\|\cdot\|$ is the L_2 norm.

The solution of (5.2.1), $\mathbf{y}(t,\theta)$, and the solution of (5.2.11), $\mathbf{z}(t,\theta)$, are both continuous for any $t \geq 0$ and θ. Therefore,

$$\mathbf{w}(t,\theta) = \mathbf{z}(t,\theta) - \mathbf{y}(t,\theta), \quad 0 \leq t \leq T \tag{5.3.2}$$

is also continuous.

Theorem 5.3.1. Assume that the function $\mathbf{f}$ satisfies

$$\|\mathbf{f}(t,\mathbf{u},\lambda)\| \leq C, \quad \tau - \epsilon \leq t \leq \tau + \epsilon, \tag{5.3.3}$$

$$\|\mathbf{f}(t,\mathbf{u},\lambda) - \mathbf{f}(t,\mathbf{v},\lambda)\| \leq L\,\|\mathbf{u}-\mathbf{v}\|, \quad 0 < t \leq T \tag{5.3.4}$$

for any $\mathbf{u}$ and $\mathbf{v}$. Here $C \equiv const \geq 0$, $L \equiv const > 0$, and $L \leq \mu$, with μ being the smallest eigenvalue of A. Then

$$\max_{|t-\tau|<\epsilon} \|\mathbf{w}(t,\theta)\| = O(\epsilon), \quad \max_{|t-\tau|\geq\epsilon} \|\mathbf{w}(t,\theta)\| \leq O(\epsilon^2)$$

hold for any θ.

[3)] These assumptions are valid for the type of the applications we are interested in. Typically, matrix A includes discretized diffusion and decay terms.

Proof. For $0 < t \leq \tau - \epsilon$ we have

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{w}}{\mathrm{d}t}(t,\theta) = -\mathrm{A}\mathbf{w}(t,\theta), & 0 < t \leq \tau - \epsilon, \\ \mathbf{w}(t,\theta) = 0, & t = 0, \end{cases}$$

which has the solution $\mathbf{w}(t,\theta) = 0$ for $0 \leq t \leq \tau - \epsilon$. Furthermore,

$$\mathbf{y}(t,\theta) = \mathbf{z}(t,\theta) = e^{-\mathrm{A}t}\ \Psi(0), \quad 0 \leq t \leq \tau - \epsilon.$$

For $\tau - \epsilon < t < \tau$ we have

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{w}}{\mathrm{d}t}(t,\theta) = \mathrm{H}_\epsilon(t-\tau)\ \mathbf{f}\left(t, \Psi(t-\tau), \lambda\right) - \mathrm{A}\mathbf{w}(t,\theta), & \tau - \epsilon < t < \tau, \\ \mathbf{w}(t,\theta) = 0, & t = \tau - \epsilon, \end{cases}$$

which has the solution

$$\mathbf{w}(t,\theta) = \int_{\tau-\epsilon}^{t} e^{-\mathrm{A}(t-\xi)}\ \mathrm{H}_\epsilon(\xi-\tau)\ \mathbf{f}\left(\xi, \Psi(\xi-\tau), \lambda\right) d\xi, \quad \tau - \epsilon \leq t < \tau.$$

Using the trapezoidal rule for the integral and $\mathrm{H}_\epsilon(-\epsilon) = 0$, we have

$$\mathbf{w}(t,\theta) = \frac{t-\tau+\epsilon}{2}\ \mathrm{H}_\epsilon(t-\tau)\ \mathbf{f}(t, \Psi(t-\tau), \lambda) + O((t-\tau+\epsilon)^3),\ \tau - \epsilon \leq t < \tau.$$

Therefore,

$$\max_{\tau-\epsilon<t<\tau} \|\mathbf{w}(t,\theta)\| = O(\epsilon).$$

Using $\mathrm{H}_\epsilon(0) = \frac{1}{2} + O(\epsilon^2)$ and the continuity of $\mathbf{w}$, we have $\mathbf{w}(\tau,\theta) = \frac{\epsilon}{4}\ \mathbf{f}\left(\tau, \Psi(0), \lambda\right) + \Delta_1$, where $\Delta_1 = O(\epsilon^3)$.

For $\tau < t < \tau + \epsilon$ we have

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{w}}{\mathrm{d}t}(t,\theta) = (\mathrm{H}_\epsilon(t-\tau) - 1)\ \mathbf{f}\left(t, e^{-\mathrm{A}(t-\tau)}\Psi(0), \lambda\right) - \mathrm{A}\mathbf{w}(t,\theta), & \tau < t < \tau + \epsilon, \\ \mathbf{w}(t,\theta) = \dfrac{\epsilon}{4}\ \mathbf{f}\left(\tau, \Psi(0), \lambda\right) + \Delta_1, & t = \tau, \end{cases}$$

which has the solution

$$\mathbf{w}(t,\theta) = e^{-\mathrm{A}(t-\tau)}\left(\frac{\epsilon}{4}\ \mathbf{f}\left(\tau, \Psi(0), \lambda\right) + \Delta_1\right)$$

$$+ \int_{\tau}^{t} e^{-\mathrm{A}(t-\xi)}\ (\mathrm{H}_\epsilon(\xi-\tau) - 1)\ \mathbf{f}\left(\xi, e^{-\mathrm{A}(\xi-\tau)}\Psi(0), \lambda\right) d\xi, \quad \tau \leq t < \tau + \epsilon.$$

Using the trapezoidal rule for the integral and $\mathrm{H}_\epsilon(0) = \frac{1}{2} + O(\epsilon^2)$, we have

$$\mathbf{w}(t,\theta) = e^{-\mathrm{A}(t-\tau)}\left(\frac{\tau+\epsilon-t}{4}\ \mathbf{f}\left(\tau,\Psi(0),\lambda\right) + \Delta_1\right)$$
$$+\frac{t-\tau}{2}\ \left(\mathrm{H}_\epsilon(t-\tau)-1\right)\ \mathbf{f}\left(t, e^{-\mathrm{A}(t-\tau)}\Psi(0),\lambda\right) + O((t-\tau)^3), \quad \tau \le t < \tau+\epsilon.$$

So,

$$\max_{\tau\le t<\tau+\epsilon} \|\mathbf{w}(t,\theta)\| = O(\epsilon).$$

Using $\mathrm{H}_\epsilon(\epsilon) = 1$ and the continuity of $\mathbf{w}$, we have $\mathbf{w}(\tau+\epsilon,\theta) = \Delta_2$, where $\Delta_2 = e^{-\mathrm{A}\epsilon}\Delta_1 + O(\epsilon^3) = O(\epsilon^3)$.

For $\tau+\epsilon < t \le 2\tau-\epsilon$ we have

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{w}}{\mathrm{d}t}(t,\theta) &=& -\mathrm{A}\mathbf{w}(t,\theta), & \tau+\epsilon < t \le 2\tau-\epsilon, \\ \mathbf{w}(t,\theta) &=& \Delta_2, & t = \tau+\epsilon, \end{cases}$$

which has the solution

$$\mathbf{w}(t,\theta) = e^{-\mathrm{A}(t-\tau-\epsilon)}\ \Delta_2, \quad \tau+\epsilon \le t \le 2\tau-\epsilon.$$

Thus,

$$\max_{\tau+\epsilon\le t\le 2\tau-\epsilon} \|\mathbf{w}(t,\theta)\| = O(\epsilon^3),$$

and $\mathbf{w}(2\tau-\epsilon,\theta) = e^{-\mathrm{A}(\tau-2\epsilon)}\ \Delta_2$.

For $2\tau-\epsilon < t < 2\tau+\epsilon$ we have

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{w}}{\mathrm{d}t}(t,\theta) &=& \mathbf{f}\left(t,\mathbf{z}(t-\tau,\theta),\lambda\right) - \mathbf{f}\left(t,\mathbf{y}(t-\tau,\theta),\lambda\right) & \\ & & -\mathrm{A}\mathbf{w}(t,\theta), & 2\tau-\epsilon < t < 2\tau+\epsilon, \\ \mathbf{w}(t,\theta) &=& e^{-\mathrm{A}(\tau-2\epsilon)}\ \Delta_2, & t = 2\tau-\epsilon, \end{cases}$$

giving the solution

$$\mathbf{w}(t,\theta) = e^{-\mathrm{A}(t-\tau-\epsilon)}\ \Delta_2$$
$$+\int_{2\tau-\epsilon}^{t} e^{-\mathrm{A}(t-\xi)}\left[\mathbf{f}\left(\xi,\mathbf{z}(\xi-\tau,\theta),\lambda\right) - \mathbf{f}\left(\xi,\mathbf{y}(\xi-\tau,\theta),\lambda\right)\right]d\xi.$$

Using (5.3.1) and (5.3.4), we have

$$\|\mathbf{w}(t,\theta)\| \le e^{-\mu(t-\tau-\epsilon)}\ \|\Delta_2\| + L\int_{2\tau-\epsilon}^{t} e^{-\mu(t-\xi)}\|\mathbf{w}(\xi-\tau,\theta)\|d\xi$$
$$\le \|\Delta_2\| + 2L\epsilon \max_{\tau-\epsilon<\xi<\tau+\epsilon}\|\mathbf{w}(\xi,\theta)\| = O(\epsilon^2), \quad 2\tau-\epsilon < t < 2\tau+\epsilon.$$

Thus,

$$\max_{2\tau-\epsilon<t<2\tau+\epsilon} \|\mathbf{w}(t,\theta)\| \le O(\epsilon^2),$$

and $\mathbf{w}(2\tau+\epsilon,\theta) = \delta_1$, where $\|\delta_1\| \le O(\epsilon^2)$.

In the similar way, for $2\tau+\epsilon < t \le 3\tau$ we have

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{w}}{\mathrm{d}t}(t,\theta) &= \mathbf{f}\left(t,\mathbf{z}(t-\tau,\theta),\lambda\right) - \mathbf{f}\left(t,\mathbf{y}(t-\tau,\theta),\lambda\right) \\ & \quad -\mathrm{A}\mathbf{w}(t,\theta), & 2\tau+\epsilon < t \le 3\tau, \\ \mathbf{w}(t,\theta) &= \delta_1, & t = 2\tau+\epsilon, \end{cases}$$

giving the solution

$$\mathbf{w}(t,\theta) = e^{-\mathrm{A}(t-2\tau+\epsilon)}\ \delta_1 +$$

$$\int_{2\tau+\epsilon}^{t} e^{-\mathrm{A}(t-\xi)} \left[\mathbf{f}\left(\xi,\mathbf{z}(\xi-\tau,\theta),\lambda\right) - \mathbf{f}\left(\xi,\mathbf{y}(\xi-\tau,\theta),\lambda\right)\right] d\xi.$$

Using (5.3.1) and (5.3.4), we have

$$\|\mathbf{w}(t,\theta)\| \le e^{-\mu(t-2\tau+\epsilon)}\ \|\delta_1\| + L\int_{2\tau+\epsilon}^{t} e^{-\mu(t-\xi)}\|\mathbf{w}(\xi-\tau,\theta)\|d\xi$$

$$\le \|\delta_1\| + L(\tau-\epsilon)\max_{\tau+\epsilon<\xi\le 2\tau}\|\mathbf{w}(\xi,\theta)\| = O(\epsilon^2), \quad 2\tau+\epsilon < t \le 3\tau.$$

Thus,

$$\max_{2\tau+\epsilon<t\le 3\tau} \|\mathbf{w}(t,\theta)\| \le O(\epsilon^2).$$

The rest we prove by induction. We assume that

$$\max_{(k-1)\tau\le t\le k\tau} \|\mathbf{w}(t,\theta)\| \le O(\epsilon^2)$$

holds for some $k \ge 3$. We have already proven it for $k = 3$. Then for $k\tau < t \le (k+1)\tau$ we get

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{w}}{\mathrm{d}t}(t,\theta) &= \mathbf{f}\left(t,\mathbf{z}(t-\tau,\theta),\lambda\right) - \mathbf{f}\left(t,\mathbf{y}(t-\tau,\theta),\lambda\right) \\ & \quad -\mathrm{A}\mathbf{w}(t,\theta), & k\tau < t \le (k+1)\tau, \\ \mathbf{w}(t,\theta) &= \delta_k, & t = k\tau, \end{cases}$$

where $\|\delta_k\| \le \max\limits_{(k-1)\tau\le\xi\le k\tau} \|\mathbf{w}(\xi,\theta)\| \le O(\epsilon^2)$ and we find the solution

$$\mathbf{w}(t,\theta) = e^{-\mathrm{A}(t-k\tau)}\ \delta_k + \int_{k\tau}^{t} e^{-\mathrm{A}(t-\xi)} \left[\mathbf{f}\left(\xi,\mathbf{z}(\xi-\tau,\theta),\lambda\right) - \mathbf{f}\left(\xi,\mathbf{y}(\xi-\tau,\theta),\lambda\right)\right] d\xi.$$

Using (5.3.1), (5.3.4), and $L \leq \mu$, we have

$$\begin{aligned}
\|\mathbf{w}(t,\theta)\| &\leq e^{-\mu(t-k\tau)}\ \|\delta_k\| + L\int_{k\tau}^{t} e^{-\mu(t-\xi)}\|\mathbf{w}(\xi-\tau,\theta)\|d\xi \\
&\leq \max_{(k-1)\tau\leq\xi\leq k\tau}\|\mathbf{w}(\xi,\theta)\|\left(e^{-\mu(t-k\tau)} + L\int_{k\tau}^{t} e^{-\mu(t-\xi)}d\xi\right) \\
&= \max_{(k-1)\tau\leq\xi\leq k\tau}\|\mathbf{w}(\xi,\theta)\|\left(e^{-\mu(t-k\tau)} + \frac{L}{\mu}\left(1-e^{-\mu(t-k\tau)}\right)\right) \\
&\leq \max_{(k-1)\tau\leq\xi\leq k\tau}\|\mathbf{w}(\xi,\theta)\| \leq O(\epsilon^2), \quad k\tau \leq t \leq (k+1)\tau.
\end{aligned}$$

Thus,

$$\max_{k\tau\leq t\leq(k+1)\tau}\|\mathbf{w}(t,\theta)\| \leq O(\epsilon^2).$$

□

Corollary 5.3.2. From (5.2.3) and (5.2.15), we get

$$R(\theta) - S(\theta) = \sum_{i=1}^{N}\left(z(t_i,\theta) - y(t_i,\theta)\right)\left(z(t_i,\theta) + y(t_i,\theta) - 2\tilde{y}_i\right).$$

Then, results of Theorem 5.3.1 imply that $R(\theta) \to S(\theta)$ as $\epsilon \to 0$ for any θ.

Corollary 5.3.3. $\breve{\theta} \to \hat{\theta}$ as $\epsilon \to 0$.

The obtained result in Theorem 5.3.1 holds only when the Heaviside function (5.2.2) is regularized on the time interval $[-\epsilon, \epsilon]$. If one uses the regularization of the Heaviside function on $[0, \epsilon]$, such as (5.2.14), then the convergence result can be derived in the following theorem[4)].

Theorem 5.3.4. If the Heaviside function (5.2.2) is regularized on the time interval $[0, \epsilon]$ and the assumptions of Theorem 5.3.1 hold, then

$$\max_{0<t\leq T}\|\mathbf{w}(t,\theta)\| \leq O(\epsilon)$$

for any θ.

[4)] The proof of this theorem follows the same lines as the proof of Theorem 5.3.1 and therefore, it is omitted here.

5.4 Numerical illustrations

In this section we shall validate our findings by numerical illustrations for simple problems. In order to solve DDE model (5.2.1), its regularized analogy (5.2.11), and the corresponding variational equations (5.2.17)-(5.2.18), we use the numerical solver RADAR5 [63] based on an implicit Runge-Kutta method. We refer the reader to [64] for an extensive description of this numerical method.

For estimation of parameters in the regularized model we use the gradient-based Levenberg-Marquardt (LM) method [26]. For implementational aspects of LM method we refer the reader to [6]. Here, we only notice that the LM method is a local search approach and therefore it can get trapped in a local minimum. In order to find a global minimum we use the LM search 1000 times with initial parameter values randomly chosen from the whole search space (coarse Monte-Carlo sampling). From the obtained minima we select the one with the lowest cost function value. Since the problems we consider here for numerical illustrations have only a few parameters and a few model equations, such an approach is computationally cheap and easy to implement. Obviously, for the large-scale problems it is far more efficient to use initially a global search method in order to provide the LM search with good starting points.

Example 5.4.1. We consider a simple DDE of the form:

$$\begin{cases} \dfrac{\mathrm{d}y}{\mathrm{d}t}(t) = \mathrm{H}(t-\tau)y(t-\tau) - \lambda y(t), & 0 < t \leq 80, \\ y(t) = 1, & t \leq 0. \end{cases} \tag{5.4.1}$$

The analytical solution of (5.4.1) is given by

$$y(t,\tau,\lambda) = \sum_{j=0}^{k} \frac{(t-j\tau)^j}{j!} e^{-\lambda(t-j\tau)}, \quad k\tau \leq t < (k+1)\tau, \quad k = 0, 1, \ldots. \tag{5.4.2}$$

The plots on the first row of Figure 5.1 show (5.4.2) as a function of time for $\tau = 5$ and $\lambda = 0.5$ (left), $\lambda = 1$ (middle), and $\lambda = 1.5$ (right). As the time increases, the solution (5.4.2) vanishes when $\lambda > 1$, goes to non-zero steady state when $\lambda = 1$, and exponentionally grows in the case $\lambda < 1$.

We compute numerically the solution of the regularized problem, $z(t,\tau,\lambda)$, with $\epsilon = 0.01$. The plots on the second row of Figure 5.1 show $|z(t,\tau,\lambda) - y(t,\tau,\lambda)|$ as a function of time for the same parameter sets with the regularization of the Heaviside function by the cubic polynomial (5.2.13) on the interval $[-\epsilon, \epsilon]$. The plots confirm the results of Theorem 5.3.1. Note that for this problem we have $L = 1$. For $\lambda \geq L$ the absolute difference between the solution of the original problem (5.4.1) and its regularized analogy as predicted is of order $O(\epsilon)$ in the ϵ-neighborhood of $t = \tau$ and at most of order $O(\epsilon^2)$ elsewhere. For $\lambda < L$ the absolute difference grows in time which can be expected due to exponential growth

of solution itself. However, the relative difference $|z(t,\tau,\lambda) - y(t,\tau,\lambda)|/y(t,\tau,\lambda)$ is bounded and has the same order as the absolute difference (not shown here).

The plots on the third row of Figure 5.1 show $|z(t,\tau,\lambda)-y(t,\tau,\lambda)|$ as a function of time when the Heaviside function is regularized by the cubic polynomial (5.2.14) on the interval $[0,\epsilon]$. The plots confirm the results of Theorem 5.3.4. For $\lambda \geq L$ the absolute difference is at most of order $O(\epsilon)$ everywhere.

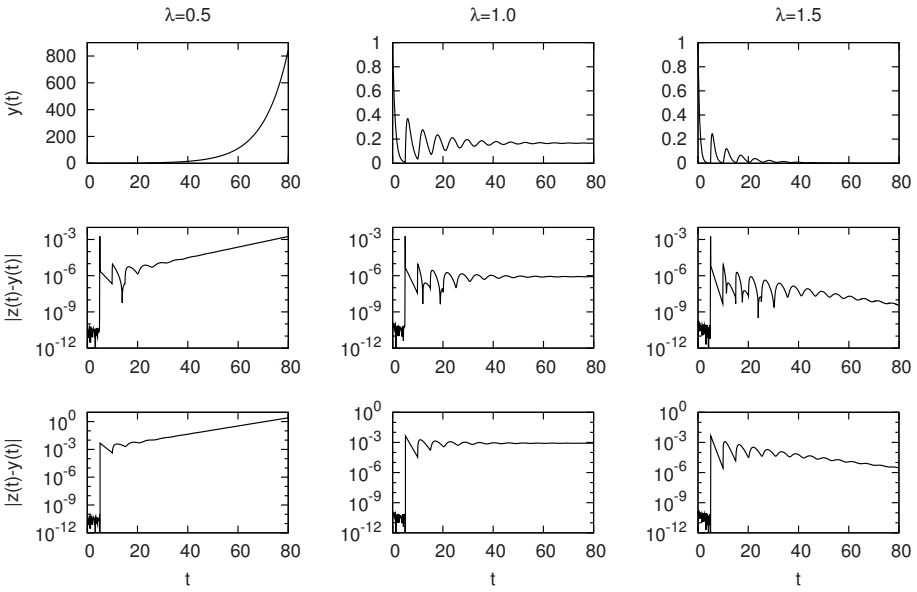


Figure 5.1: Solution (5.4.2) of the original problem (5.4.1) (first row) and the absolute difference between (5.4.2) and the numerical solution of the regularized problems when (5.2.13) (second row) and (5.2.14) (third row) with $\epsilon = 0.01$ are used for approximation of Heaviside function. Columns illustrate different behavior of the solution (5.4.2) and the regularization errors for three different values of λ; $\tau = 5$ is used in all cases.

Differentiating (5.4.2) with respect to λ and τ gives the analytical expression for the gradients of (5.4.1):

$$\frac{\partial y}{\partial \lambda}(t,\tau,\lambda) = -\sum_{j=0}^{k} \frac{(t-j\tau)^{j+1}}{j!} e^{-\lambda(t-j\tau)}, \quad k\tau \leq t < (k+1)\tau, \quad k = 0,1,\ldots \tag{5.4.3}$$

$$\frac{\partial y}{\partial \tau}(t,\tau,\lambda) = \begin{cases} 0, & 0 \le t < \tau, \\ \displaystyle\sum_{j=1}^{k} \frac{(t-j\tau)^{j-1}}{(j-1)!}\left(-j+\lambda(t-j\tau)\right) e^{-\lambda(t-j\tau)}, & k\tau \le t < (k+1)\tau. \end{cases} \tag{5.4.4}$$

We compute the gradients of the regularized problem numerically by solving the corresponding variational equations with $\epsilon = 0.01$. Figure 5.2 shows $|\frac{\partial z}{\partial \lambda} - \frac{\partial y}{\partial \lambda}|$ (first row) and $|\frac{\partial z}{\partial \tau} - \frac{\partial y}{\partial \tau}|$ (second row) as a function of time for $\tau = 5.0$ and $\lambda = 0.5$ (left), $\lambda = 1.0$ (middle), and $\lambda = 1.5$ (right). The plots reveal that for $\lambda < L$ the absolute difference between the gradients grows in time, similar to the difference between the solutions. Additionally, we observe that when $\lambda \ge L$ the difference between the gradients in λ direction is at most of order $O(\epsilon^2)$, while the difference between the gradients in τ direction is of order $O(1)$ in the ϵ-neighborhood of $t = \tau$, of order $O(\epsilon)$ in the ϵ-neighborhood of $t = 2\tau$, and at most of order $O(\epsilon^2)$ elsewhere. Note that the significant difference between the gradients in τ direction around $t = \tau$ is due to the discontinuity in (5.4.4) at that point. However, this discontinuity does not propagate in time. As the time increases, the gradient (5.4.4) smooths out and the order $O(\epsilon^2)$ for the difference between gradients is regained. Similar results both for $|\frac{\partial z}{\partial \lambda} - \frac{\partial y}{\partial \lambda}|$ and $|\frac{\partial z}{\partial \tau} - \frac{\partial y}{\partial \tau}|$ have been obtained with different realizations of ϵ (not shown here). The results suggest that when $\lambda \ge L$ the gradients of the regularized problem pointwisely converge to the corresponding gradients of the original problem as $\epsilon \to 0$, with the only exception for the gradient in τ direction at $t = \tau$. Moreover, they also suggest that the analogy of Theorem 5.3.1 holds for the difference between gradients.

Parameter Estimation We have generated an artificial dataset using solution (5.4.2) of problem (5.4.1) with $\tau^* = 5$ and $\lambda^* = 1$ at 400 time points $t = \{0.2, 0.4, \ldots, 80.0\}$. With this dataset, $\{\hat{\tau}, \hat{\lambda}\} = \{\tau^*, \lambda^*\}$ is the minimizer of the cost function (5.2.3) for problem (5.4.1) and $S(\hat{\tau}, \hat{\lambda}) = 0$. Therefore, the confidence intervals (5.2.8) and (5.2.9) for $\hat{\tau}$ and $\hat{\lambda}$ have zero size. Note that the correlation coefficient (5.2.10) between $\hat{\tau}$ and $\hat{\lambda}$ is not defined because the point $t = \hat{\tau}$ where (5.4.4) is discontinuous coincides with one of the data points and therefore the discontinuity enters into the sensitivity matrix $J(\hat{\tau}, \hat{\lambda})$.

Using this dataset, parameter estimates $\breve{\tau}$ and $\breve{\lambda}$ minimizing the cost function (5.2.15) for the regularized problem are obtained with the LM method. Table 5.1 nicely illustrates the first-order convergence of these parameter estimates as well as the sizes of their independent confidence intervals as $\epsilon \to 0$. The last column shows the correlation coefficient between parameter estimates computed by (5.2.22) for different realizations of ϵ.

Next, we have made a different dataset using solution (5.4.2) with $\tau^* = 5$ and $\lambda^* = 1$ at 40 time points $t = \{0.5, 1.0, \ldots, 20.0\}$ and adding to it a noise

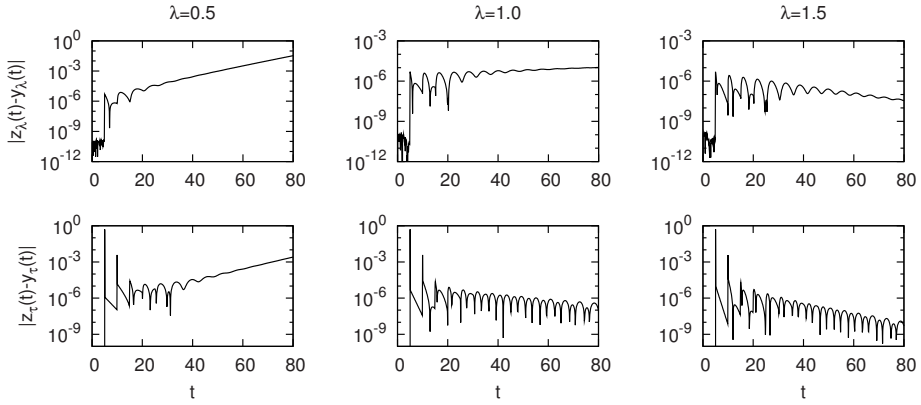


Figure 5.2: Absolute differences between the gradients of the original problem (5.4.1) and the regularized problem (with $\epsilon = 0.01$), $|\frac{\partial z}{\partial \lambda} - \frac{\partial y}{\partial \lambda}|$ (first row) and $|\frac{\partial z}{\partial \tau} - \frac{\partial y}{\partial \tau}|$ (second row), as a function of time with $\tau = 5.0$ and three different values of λ.

ϵ	$\lvert\hat{\tau} - \breve{\tau}\rvert$	$\lvert\hat{\lambda} - \breve{\lambda}\rvert$	$\lvert\triangle^I_{\hat{\tau}} - \triangle^I_{\breve{\tau}}\rvert$	$\lvert\triangle^I_{\hat{\lambda}} - \triangle^I_{\breve{\lambda}}\rvert$	$\breve{\rho}_{\tau\lambda}$
1.0e-01	1.98e-03	2.18e-05	1.21e-03	9.96e-05	-0.16698
1.0e-02	1.78e-04	1.89e-06	1.23e-04	1.02e-05	-0.16682
1.0e-03	1.76e-05	2.31e-07	1.23e-05	1.03e-06	-0.16681
1.0e-04	1.76e-06	2.35e-08	1.24e-06	1.03e-07	-0.16680
1.0e-05	1.65e-07	2.21e-09	1.24e-07	1.03e-08	-0.16679

Table 5.1: Convergence results of parameter estimates and their statistical quantities for the regularized model when the extensive and error-free dataset is used.

generated from a normal distribution with zero mean and standard deviation $\sigma = 0.01$. The left plot of Figure 5.3 shows the obtained data. With this dataset, $\{\hat{\tau}, \hat{\lambda}\} = \{5.001963, 1.001845\}$ is the minimizer of the cost function (5.2.3). It has been obtained by using the LM method for the regularized problem with $\epsilon = 10^{-9}$. Note that the LM search for the original problem (5.4.1) using exact expressions for the solution (5.4.2) and the gradients (5.4.3)-(5.4.4) has resulted in the same minimum. The independent confidence intervals for $\hat{\tau}$ and $\hat{\lambda}$ have sizes $\{\triangle^I_{\hat{\tau}}, \triangle^I_{\hat{\lambda}}\} = \{0.041787, 0.013938\}$, and the correlation coefficient between them is $\hat{\rho}_{\tau\lambda} = 0.171904$.

Using this error-dependent dataset, parameter estimates $\breve{\tau}$ and $\breve{\lambda}$ for the regularized problem are obtained with the LM method. Table 5.2 presents the convergence results of these estimates and the corresponding statistical quantities. Contrary to the previous case, the convergence is not uniform, e.g., results for

smaller values of ϵ show different convergence rate than the results for $\epsilon = 0.1$ and $\epsilon = 0.01$.

ϵ	$\lvert\hat{\tau} - \breve{\tau}\rvert$	$\lvert\hat{\lambda} - \breve{\lambda}\rvert$	$\lvert\triangle^I_{\hat{\tau}} - \triangle^I_{\breve{\tau}}\rvert$	$\lvert\triangle^I_{\hat{\lambda}} - \triangle^I_{\breve{\lambda}}\rvert$	$\lvert\hat{\rho}_{\tau\lambda} - \breve{\rho}_{\tau\lambda}\rvert$
1.0e-01	5.40e-03	4.07e-04	3.10e-03	1.59e-04	2.12e-02
1.0e-02	5.81e-04	2.94e-05	3.08e-03	4.65e-05	1.63e-02
1.0e-03	4.80e-08	5.30e-09	5.10e-09	2.27e-10	5.17e-08
1.0e-04	4.67e-10	5.39e-11	1.57e-11	5.25e-12	3.35e-11

Table 5.2: Convergence results of parameter estimates and their statistical quantities for the regularized model when the data contains measurement errors.

Note that the convergence results we have shown here do not imply that the accuracy of parameter estimates for the regularized model improves as $\epsilon \to 0$. Since the sizes of the independent confidence intervals for $\hat{\tau}$ and $\hat{\lambda}$ in this example are both of order $O(10^{-2})$, the numbers in Table 5.2 indicate that the parameter estimates $\breve{\tau}$ and $\breve{\lambda}$ have the same accuracy for all considered realizations of ϵ. Therefore, if one is interested only in the accuracy of parameter estimates, then for this example the regularization with $\epsilon = 0.1$ is sufficient. In general, however, parameter estimates and their accuracy for the original DDE model are not known and therefore the largest acceptable value of ϵ cannot be priori concluded.

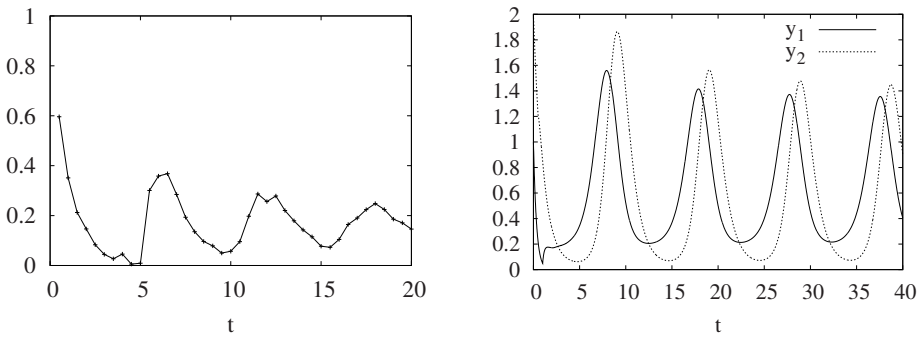


Figure 5.3: Left plot shows the data generated from solution (5.4.2) of problem (5.4.1) with $\tau^* = 5$ and $\lambda^* = 1$ by adding a normal noise to it. Right plot shows the numerical solution of problem (5.4.5) with $\tau_1 = 1.0$ and $\tau_2 = 0.5$.

Example 5.4.2. We consider DDEs of the form:

$$\begin{cases} \dfrac{\mathrm{d}y_1}{\mathrm{d}t}(t) = \mathrm{H}(t-\tau_1)\left(\dfrac{1}{2} + \dfrac{y_1^2(t-\tau_1)}{y_2(t-\tau_1)}\right) - 3y_1(t), & 0 < t \le 40, \\ \dfrac{\mathrm{d}y_2}{\mathrm{d}t}(t) = \mathrm{H}(t-\tau_2)\ y_1^2(t-\tau_2) - y_2(t), & 0 < t \le 40, \\ y_1(t) = 1, \qquad y_2(t) = 2, & t \le 0, \end{cases} \tag{5.4.5}$$

which originates from a Gierer-Meinhard system [30]. On the basis of this example, we shall illustrate that the convergence results hold for models having more than one time lag. The right plot of Figure 5.3 shows the numerical solution of (5.4.5) as a function of time for $\tau_1 = 1$ and $\tau_2 = 0.5$. The regularized analogy of (5.4.5) is solved numerically with three different values of ϵ. Figure 5.4 shows the absolute difference between the solutions of the original and the regularized model, $|z_1(t)-y_1(t)|$ (first row) and $|z_2(t)-y_2(t)|$ (second row). As we can see, the difference is at most of order $O(\epsilon^2)$ everywhere except in the vicinity of time points $t = \tau_1$ and $t = \tau_2$, where the corresponding components of the systems differ from each other by $O(\epsilon)$. These results support the convergence results obtained for the problem in Example 4.1 as well as the theoretical results of Theorem 5.3.1.

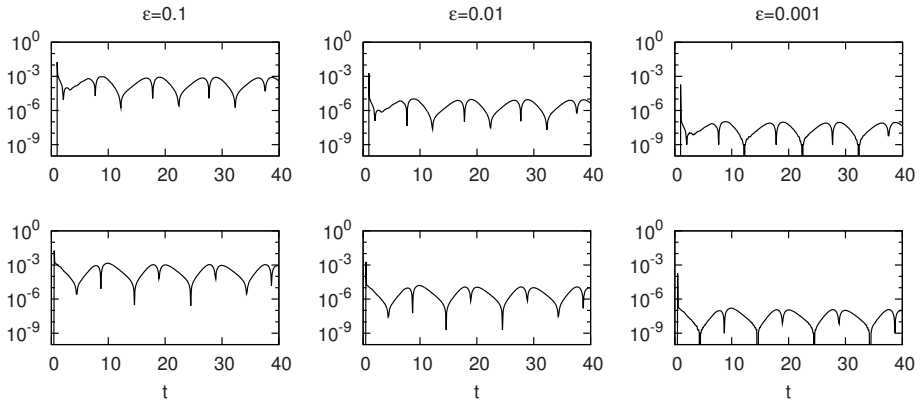


Figure 5.4: Absolute difference between the solution of problem (5.4.5) and the solution of the corresponding regularized problem as a function of time, i.e., $|z_1(t) - y_1(t)|$ (first row) and $|z_2(t) - y_2(t)|$ (second row). Here, $\tau_1 = 1$, $\tau_2 = 0.5$, and the regularized problem is solved with three different values of ϵ.

5.5 Conclusions

In this chapter we have studied DDE models with a right-hand side being discontinuous in time and time-lag parameters. For these type of DDEs the correspond-

ing gradients of the solution are discontinuous as well. This causes difficulties in the parameter estimation procedures or the determinability analysis of parameter estimates when any of those methods is based on gradient information.

We have suggested to regularize the model to make it continuous. The regularization is applied at an ϵ-neighborhood of the discontinuity points. The idea of regularization is to obtain a solution which is sufficiently close to the solution of the original model almost everywhere in time and has extra smoothness in the vicinity of the discontinuity points.

For such a regularization strategy, we have shown that the solution of the regularized model converges to the solution of the original problem as $\epsilon \to 0$ for the DDEs under consideration. Moreover, we have been able to derive the convergence rate. Additionally, our theoretical results imply that the parameter estimates obtained for the regularized model converge to the corresponding estimates of the original problem. We have been able to confirm our findings by numerical illustrations for two simple test problems.

Finally, we note that the type of DDEs considered here finds an application in developmental biology. Namely, our work was inspired by the model of spatio-temporal pattern formation for gene expression at the early developmental time of the fruit-fly *Drosophila* which incorporates the delay in the protein production. The aim of our future work is to use a regularized model for this real-life biological problem, to estimate model parameters using a large-scale dataset, and to apply statistical analysis for the obtained parameter estimates.

Bibliography

[1] L. Ljung (1999), *System Identification – Theory For the User*, Prentice Hall, New Jersey.

[2] J. Nocedal, S.J. Wright (1999), *Numerical Optimization*, Springer, New York.

[3] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi (1983), *Optimization by simulated annealing*, Science 220, pp. 671–680.

[4] M. Rodriguez–Fernandez, P. Mendes, J.R. Banga (2006), *A hybrid approach for efficient and robust parameter estimation in biochemical pathways*, BioSystems 83, pp. 248–265.

[5] R.N. Gutenkunst, J.J. Waterfall, F.P. Casey, K.S. Brown, C.R. Myers, J.P. Sethna (2007), *Universally sloppy parameter sensitivities in systems biology models*, PLoS Computational Biology 3:e189.

[6] M. Ashyraliyev, J. Jaeger, J.G. Blom (2008), *On Parameter Estimation and Determinability for Drosophila Gap Gene Circuits*, BMC Systems Biology 2:83.

[7] A.A. Simcox, J.H. Sang (1983), *When does gastrulation occur in Drosophila embryos?*, Developmental Biology 97, pp. 212–221.

[8] M. Akam (1987), *The molecular basis for metameric pattern in the Drosophila embryo*, Development 101, pp. 1–21.

[9] C. Nüsslein–Volhard, E. Wieschaus (1980), *Mutations affecting segment number and polarity in Drosophila*, Nature 287, pp. 795–801.

[10] J. Jaeger, J. Reinitz (2004), *Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster*, Genetics 167, pp. 1721–1737.

[11] E. Mjolsness, D.H. Sharp, J. Reinitz (1991), *A connectionist model of development*, J. Theor. Biol. 152, pp. 429–453.

[12] J. Jaeger, J. Reinitz (2004), *Dynamic control of positional information in the early Drosophila embryo*, Nature 430, pp. 368–371.

[13] T.J. Perkins, J. Jaeger, J. Reinitz, L. Glass (2006), *Reverse Engineering the Gap Gene Network*, PLoS Computational Biology 2:e51.

[14] Y.F. Nanfack, J.A. Kaandorp, J.G. Blom (2007), *Efficient parameter estimation for spatio–temporal models of pattern formation: Case study of Drosophila melanogaster*, Bioinformatics 23, pp. 3356–3363.

[15] Manu, S. Surkova, A.V. Spirov, V.V. Gursky, H. Janssens, A.–R. Kim, O. Radulescu, C.E. Vanario–Alonso, D.H. Sharp, M. Samsonova, J. Reinitz (2009), *Canalization of Gene Expression in the Drosophila Blastoderm by Gap Gene Cross Regulation*, PLoS Biology 7(3):e49.

[16] V.V. Gursky, J. Reinitz, A.M. Samsonov (2001), *How gap genes make their domains: An analytical study based on data driven approximations*, Chaos 11(1), pp. 132–141.

[17] V.V. Gursky, J. Jaeger, K.N. Kozlov, J. Reinitz, A.M. Samsonov (2004), *Pattern formation and nuclear divisions are uncoupled in Drosophila segmentation: comparison of spatially discrete and continuous models*, Physica D 197, pp. 286–302.

[18] E. Poustelnikova, A. Pisarev, M. Blagov, M. Samsonova, J. Reinitz (2004), *A database for management of gene expression data in situ*, Bioinformatics 20, pp. 2212–2221.

[19] FlyEx Database (http://flyex.ams.sunysb.edu/flyex).

[20] C.T.H. Baker, C.A.H. Paul (1997), *Pitfalls in Parameter Estimation for Delay Differential Equations*, SIAM J. Sci. Comput. 18(1), pp. 305–314.

[21] M. Ashyraliyev, Y.F. Nanfack, J.A. Kaandorp, J.G. Blom (2009), *Parameter estimation for biochemical models*, FEBS J. 276(4), pp. 886–902.

[22] G.A.F. Seber, C.J. Wild (1988), *Nonlinear regression*, John Wiley & Sons, New York.

[23] K. Jaqaman, G. Danuser (2006), *Linking data to models: data regression*, Nature Reviews Molecular Cell Biology 7, pp. 813–819.

[24] R.C. Aster, B. Borchers, C.H. Thurber (2005), *Parameter Estimation and Inverse Problems*, Elsevier, USA.

[25] W.J.H. Stortelder (1998), *Parameter Estimation in Nonlinear Dynamical Systems*, PhD thesis, Univ. of Amsterdam, The Netherlands.

[26] D.W. Marquardt (1963), *An algorithm for least–squares estimation of nonlinear parameters*, SIAM J. Appl. Math. 11, pp. 431–441.

[27] H. de Jong (2002), *Modelling and simulation of genetic regulatory systems: A literature review*, J. Comp. Biol. 9(1), pp. 67–103.

[28] Manu (2007), *Canalization of Gap Gene Expression During Early Development in Drosophila melanogaster*, PhD thesis, Stony Brook Univ., USA.

[29] J. Jaeger, Cambridge & Barcelona, N. Monk, Nottingham (2008), Personal communication.

[30] J.D. Murray (2002), *Mathematical Biology*, Springer, Berlin.

[31] J. Jaeger, J. Reinitz (2006), *On the dynamic nature of positional information*, BioEssays 28, pp. 1102–1111.

[32] N.R. Draper, H. Smith (1981), *Applied regression analysis*, John Wiley & Sons, New York.

[33] P. Mendes, D.B. Kell (1998), *Non–linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation*, Bioinformatics 14, pp. 869–883.

[34] J.C.P. Bus, B. Domselaar, J. Kok (1975), *Nonlinear least squares estimation*, CWI report, NW 17/75.

[35] G.H. Golub, C.F. Loan (1996), *Matrix computations*, Johns Hopkins UP, Baltimore.

[36] P.W. Hemker (1972), *Numerical methods for differential equations in system simulation and in parameter estimation*. In: Analysis and Simulation of biochemical systems (H.C. Hemker & B. Hess eds), North Holland Publ. Comp., pp. 59–80.

[37] C.W. Gear (1971), *Numerical initial value problems in ordinary differential equation*, Prentice Hall, Englewood Cliff.

[38] J. Reinitz, D.H. Sharp (1995), *Mechanism of eve stripe formation*, Mech. Dev. 49, pp. 133–158.

[39] V.E. Foe, B.M. Alberts (1983), *Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in Drosophila embryogenesis*, J. Cell Science 61, pp. 31–70.

[40] E. Myasnikova, A. Samsonova, K. Kozlov, M. Samsonova, J. Reinitz (2001), *Registration of the expression patterns of Drosophila segmentation genes by two independent methods*, Bioinformatics 17, pp. 3–12.

[41] R. Rivera–Pomar, X. Lu, N. Perrimon, H. Taubert, H. Jäckle (1995), *Activation of posterior gap gene expression in the Drosophila blastoderm*, Nature 376, pp. 253–256.

[42] D.E. Clyde, M.S.G. Corado, X. Wu, A. Paré, D. Papatsenko, S. Small (2003), *A self–organizing system of repressor gradients establishes segmental complexity in Drosophila*, Nature 426, pp. 849–853.

[43] J. Jaeger, D.H. Sharp, J. Reinitz (2007), *Known maternal gradients are not sufficient for the establishment of gap domains in Drosophila melanogaster*, Mech. Devel. 124, pp. 108–128.

[44] B.W. Char, K.O. Gedded, G.H. Gonnet, B.L. Leong, M.B. Monagan, S.M. Watt (1991), *Maple V Library Reference manual*, Springer Verlag, New York.

[45] J. Casanova (1990), *Pattern formation under the control of the terminal system in the Drosophila embryo*, Development 110, pp. 621–628.

[46] J. Jaeger, K. Siggens (2009), unpublished results.

[47] B.W. Lindgren, G.W. McElrath, D.A. Berry (1978), *Introduction to Probability and Statistics*, Macmillan Publ., New York.

[48] D.H. Sharp, J. Reinitz (1998), *Prediction of mutant expression patterns using gene circuits*, Biosystems 47, pp. 79–90.

[49] A.–K. Tornberg, B. Engquist (2004), *Numerical approximations of singular source terms in differential equations*, J. Comput. Phys. 200, pp. 462–488.

[50] M.D. Greenberg (1971), *Application of Green's Functions in Science and Engineering*, Prentice Hall, New Jersey.

[51] E. Hairer, C. Lubich, G. Wanner (2002), *Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations*, Springer Series in Computational Mathematics, Vol. 31, Springer, Berlin.

[52] W. Hundsdorfer, J.G. Verwer (2003), *Numerical Solution of Time–Dependent Advection–Diffusion–Reaction Equations*, Springer Series in Computational Mathematics, Vol. 33, Springer, Berlin.

[53] N.M. Temme (1996), *Special Functions: an Introduction to the Classical Functions of Mathematical Physics*, Wiley, New York.

[54] N.M. Temme (2005), Personal communication.

[55] J.M. Ortega, W.C. Rheinboldt (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.

[56] H.A. van der Vorst (1992), *BI–CGSTAB: A fast and smoothly converging variant of BI–CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., Vol. 13(2), pp. 631–644.

[57] Y. Saad (2003), *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia.

[58] G.A. Bocharov, F.A. Rihan (2000), *Numerical modelling in biosciences using delay differential equations*, J. Comput. Appl. Math. 125, pp. 183–199.

[59] P.W. Nelson, J.D. Murray, A.S. Perelson (2000), *A model of HIV pathogenesis that includes an intracellular delay*, Math. Biosciences 163(2), pp. 201–215.

[60] A. Bellen, M. Zennaro (2003), *Numerical methods for delay differential equations*, Clarendon, Oxford.

[61] F.H. Clarke (1983), *Optimization and Nonsmooth Analysis*, Wiley, New York.

[62] C.T.H. Baker, C.A.H. Paul (2006), *Discontinuous solutions of neutral delay differential equations*, Appl. Numer. Math. 56, pp. 284–304.

[63] N. Guglielmi, E. Hairer (2001), *Implementing Radau IIA Methods for Stiff Delay Differential Equations*, Computing 67, pp. 1–12.

[64] N. Guglielmi, E. Hairer (2005), *Users guide for the code RADAR5 – version 2.1*, Technical report, Universita dell Aquila, Italy.

Summary

The research described in this thesis is devoted to different aspects of modelling, simulation, and inferring regulatory networks. We have considered here only deterministic models given by systems of differential equations that are capable of quantitatively reproducing spatio-temporal gene expression patterns.

In the first half of the thesis we have applied a 'connectionist' model given by a system of nonlinear Ordinary Differential Equations for simulating the gap gene network in the early development of the fruit fly *Drosophila melanogaster*. Using this model, the gap gene network has been extensively studied in the literature. However, in all previous studies the main attention has been focused on the estimation of model parameters, among which are most important the regulatory weights representing a regulatory influence of one gene on another, and subsequently on the analysis of the functioning of the gap gene system based on the values of estimated regulatory parameters. The identifiability analysis of inferred parameters has been missing and therefore, the reliability of all previous findings has remained unclear. We have tried to fill this gap by applying *a posteriori* identifiability analysis to assess the quality of the obtained parameters and studying its implications for conclusions deduced from the values of parameter estimates.

Similar to previous studies of the gap gene system, in Chapter 2 we have considered a 6-gene network consisting of gap genes *hunchback* (*hb*), *Krüppel* (*Kr*), *knirps* (*kni*), and *giant* (*gt*), terminal gap gene *tailless* (*tll*), and maternal coordinate gene *caudal* (*cad*), while the other maternal gene *bicoid* (*bcd*) has been implemented as an external input constant in time. The identifiability analysis of inferred regulatory weights has shown that none of them can be determined quantitatively with reasonable accuracy. Although we have been able to draw reliable qualitative conclusions for some of the regulatory weights, it has been found that many other interactions cannot be determined even qualitatively. So, the regulatory topology of the gap gene network deduced by only considering the values of parameter estimates has been confirmed only partially with the parameter determinability analysis. We have illustrated that an overall poor determinability of regulatory parameters is due to the presence of correlations between them. We have shown that these correlations are a property of the model, rather than being originated from the data.

In Chapter 3, we have considered a 4-gene network including gap genes *hb*,

Kr, *kni*, and *gt*. In contrast to the 6-gene network, we have implemented *bcd*, *cad*, and *tll* as time-variable external inputs. Moreover, we have supplemented the gap gene network with the terminal gap gene *huckebein* (*hkb*) also implemented as time-variable external input. The results obtained with this reformulated gap gene network have provided a number of improvements in comparison with the results for the 6-gene network. Firstly, it refers to a correct regulation of the posterior *hb* domain. We have shown that the posterior boundary of this domain is set up correctly and its anterior shift in time is reproduced by model outputs. Secondly, the identifiability analysis has revealed a significant improvement in the qualitative determinability of the regulatory weights.

In Chapter 3, we have also demonstrated that with available gap gene data the Weighted Least Squares sum with appropriately chosen weights is a more suitable measure for data fitting than the Ordinary Least Squares sum which has been used in all previous studies. This has been confirmed by a better fit of the boundaries of the gap gene expression domains, an absence of patterning defects in the model outputs, more reliable qualitative conclusions for a number of regulatory weights, and correct prediction of gap gene expression in *tll* and *hkb* mutants.

The cell-based model for simulating regulatory networks is given by a reaction-diffusion system with singular reaction source terms. Each source term is defined by a Dirac delta function expression on a lower dimensional surface. In Chapter 4, we have numerically studied this type of problems. Due to singularities, their solutions are not differentiable and this lack of smoothness causes order reduction when standard spatial discretization schemes are used on the uniform grid. We have used the finite volume approach based on the integral form such that the numerical treatment of the singular source terms is mathematically clear. We have demonstrated the reduction from order two to order one in the maximum norm when the standard second-order spatial discretization scheme on the uniform grid is applied for a number of 1D and 2D problems with singular source terms. To overcome this reduction, we have examined the discretization on a number of special locally refined grids, in 1D analytically and in 2D experimentally. We have shown that by an appropriate locally refined grid the maximum norm second-order convergence can be regained.

The model of regulatory networks incorporating a delay in the protein production is given by a system of Delay Differential Equations with a right-hand side being discontinuous in time and time-lag parameters. In Chapter 5, we have studied this type of problems. Solutions of such problems have a lack of smoothness in parameters, notably the derivatives with respect to parameters (gradients) can be discontinuous. As a consequence, the correct application of gradient-based optimization methods for parameter estimation as well as the validity of parameter determinability analysis applied on the parameter estimates can be questionable. In order to overcome these difficulties, we have examined a standard regularization technique to make the right-hand side of the model continuous at an ϵ-neighborhood of the discontinuity points. We have proven analytically the

convergence of the solution of the regularized model to the solution of the original problem as $\epsilon \to 0$. Moreover, we have derived the rate of convergence. This result implies that the parameter estimates inferred from the regularized model converge to the corresponding estimates of the original problem as $\epsilon \to 0$. Additionally, we have shown that the convergence results do not depend explicitly on the way the model is regularized. We have supported our findings with numerical illustrations for simple test problems.

Samenvatting

Het onderzoek dat in dit proefschrift wordt beschreven is gewijd aan verschillende aspecten van het modelleren, simuleren en infereren van regulatienetwerken. We hebben hier alleen deterministische modellen beschouwd gegeven door stelsels differentiaalvergelijkingen die genexpressie patronen in ruimte en tijd kwantitatief kunnen reproduceren.

In het eerste deel van het proefschrift hebben we een 'connectionist' model, gegeven door een stelsel niet-lineaire gewone differentiaalvergelijkingen, toegepast om het gap-gen netwerk in de vroege ontwikkeling van de fruitvlieg *Drosophila melanogaster* te simuleren. Met dit model is het gap-gen netwerk uitgebreid bestudeerd in de literatuur. In alle voorgaande studies was de aandacht echter hoofdzakelijk gericht op het schatten van modelparameters, waarvan de regulatiegewichten, die een regulatoire invloed van het ene gen op het andere beschrijven, het belangrijkste zijn, en vervolgens op de analyse van het functioneren van het gap-gen systeem op basis van de waarden van de geschatte regulatieparameters. Omdat de identificeerbaarheidsanalyse van de geschatte parameters ontbrak is de betrouwbaarheid van alle voorgaande resultaten echter onduidelijk. We hebben getracht dit hiaat op te vullen door een a posteriori identificeerbaarheidsanalyse uit te voeren om de kwaliteit van de verkregen parameters te schatten en de gevolgen te onderzoeken voor de conclusies afgeleid uit de waarden van de parameterschattingen.

Evenals in vorige studies van het gap-gen systeem hebben we in Hoofdstuk 2 een 6-gen netwerk beschouwd bestaande uit de gap-genen *hunchback* (*hb*), *Krüppel* (*Kr*), *knirps* (*kni*) en *giant* (*gt*), het eind gap-gen *tailless* (*tll*), en het maternaal coördinaat-gen *caudal* (*cad*), terwijl het andere maternaal gen *bicoid* (*bcd*) is geïmplementeerd als een externe invoer, constant in tijd. De identificeerbaarheidsanalyse van de geïnfereerde regulatiegewichten liet zien dat geen van alle met een redelijke nauwkeurigheid kwantitatief bepaald kan worden. Hoewel we betrouwbare kwalitatieve conclusies konden trekken voor sommige regulatiegewichten, konden veel andere interacties zelfs niet kwalitatief bepaald worden. De regulatietopologie van het gap-gen netwerk, verkregen door alleen de waarden van parameterschattingen te beschouwen, is dus slechts deels bevestigd door de parameter determineerbaarheidsanalyse. We hebben laten zien dat een over het geheel genomen slechte determineerbaarheid van regulatieparameters te wijten is

aan de aanwezigheid van onderlinge correlaties en dat deze correlaties meer een eigenschap zijn van het model dan voortkomen uit de data.

In Hoofdstuk 3 hebben we een 4-gen netwerk beschouwd bestaande uit de gap-genen *hb*, *Kr*, *kni* en *gt*. In tegenstelling tot het 6-gen netwerk hebben we *bcd*, *cad* en *tll* als tijdsvariabele externe invoer geïmplementeerd. Bovendien hebben we het eind gap-gen *huckebein* (*hkb*) aan het gap-gen netwerk toegevoegd, eveneens geïmplementeerd als tijdsvariabele externe invoer. De resultaten verkregen met dit geherformuleerde gap-gen netwerk hebben een aantal verbeteringen opgeleverd vergeleken met de resultaten voor het 6-gen netwerk. Ten eerste geeft het een correcte regulatie in het posterior *hb* domein. We hebben aangetoond dat de posterior grens van dit gebied correct wordt ingesteld en dat zijn anterior verschuiving in tijd wordt gereproduceerd door de model output. Ten tweede heeft de identificeerbaarheidsanalyse een significante verbetering in de kwalitatieve determineerbaarheid van de regulatiegewichten laten zien.

In Hoofdstuk 3 hebben we ook aangetoond dat met de beschikbare gap-gen data de gewogen kleinste-kwadraten som met geschikt gekozen gewichten een passender maat is voor het fitten van data dan de gewone kleinste-kwadraten som die in alle voorgaande studies gebruikt is. Dit is bevestigd door een betere fit van de grenzen van de gap-genexpressie domeinen, door een afwezigheid van patroondefecten in de model output, door betrouwbaardere kwalitatieve conclusies voor een aantal regulatiegewichten en door de correcte voorspelling van gap-genexpressie in *tll* en *hkb* mutanten.

Het op cellen gebaseerde model voor het simuleren van regulatienetwerken wordt gegeven door een reactie-diffusie stelsel met singuliere reactie brontermen. Elke bronterm wordt gedefinieerd door een Dirac-delta functie expressie op een lager-dimensionaal oppervlakte.

In Hoofdstuk 4 hebben we dit type probleem numeriek bestudeerd. Door de singulariteiten zijn de oplossingen niet differentieerbaar en dit gebrek aan gladheid veroorzaakt orde reductie als standaard ruimte-discretisatie schema's worden gebruikt op een uniform grid. We hebben de eindige-volume aanpak gebaseerd op de integraalvorm gebruikt zodat de numerieke behandeling van de singuliere bronterm wiskundig duidelijk is. We hebben voor een aantal 1D en 2D problemen met singuliere brontermen de ordereductie van twee naar één in de maximum norm laten zien als het standaard tweede-orde ruimte-discretisatie schema op een uniform grid wordt toegepast. Om deze reductie te overwinnen hebben we de discretisatie op een aantal speciale lokaal-verfijnde grids onderzocht, in 1D analytisch en in 2D experimenteel. We hebben laten zien dat door een geschikt gekozen lokaal-verfijnd grid de tweede-orde convergentie in de maximum norm kan worden herkregen.

Het model van regulatienetwerken, dat een vertraging in de eiwitproductie omvat, wordt gegeven door een stelsel delay-differentiaalvergelijkingen met een rechterlid dat discontinu is in de tijd en in de time-lag parameters. In Hoofstuk 5 hebben we dit type problemen bestudeerd. Oplossingen van zulke problemen hebben een gebrek aan gladheid in de parameters, met name de afgeleiden naar

de parameters (gradiënten) kunnen discontinu zijn. Het gevolg is dat zowel de correcte toepassing van optimalisatiemethoden voor parameterschatting die op de gradiënt gebaseerd zijn als de geldigheid van de parameter determineerbaarheids-analyse toegepast op de parameterschattingen twijfelachtig kunnen zijn. Teneinde deze problemen te boven te komen hebben we een standaard regularisatietechniek onderzocht om het rechterlid van het model continu te maken in een ϵ-omgeving van de discontinuïteitspunten. We hebben de convergentie van de oplossing van het geregulariseerde model naar de oplossing van het ooorspronkelijke probleem analytisch bewezen voor $\epsilon \to 0$. Bovendien hebben we de convergentiesnelheid afgeleid. Dit resultaat betekent dat de parameterschattingen verkregen met het geregulariseerde model convergeren naar de overeenkomstige schattingen van het oorspronkelijke probleem voor $\epsilon \to 0$. We hebben daarenboven laten zien dat de convergentieresultaten niet expliciet afhangen van de manier waarop het model wordt geregulariseerd. We hebben onze resultaten met numerieke illustraties voor simpele testproblemen ondersteund.

Acknowledgement

This thesis would not have been possible without the support of many people. It is a pleasure for me to convey my sincere gratitude to all of them. There are a few names I would like to mention personally.

With the biggest contribution to this thesis, I would like to gratefully thank Joke Blom for her supervision of my Ph.D. research, for her patience, motivation, advice, inspiration, and persistent help. Her mentorship was paramount from the first days of my research work to the final stages of writing this thesis. Her continuous scientific support enabled me to develop an understanding of the subject. Joke, I am deeply indebted to you for all your help.

I also wish to express my deepest gratitude to my promotor, Jan Verwer, who was always abundantly helpful and offered invaluable assistance, knowledge, encouragement, and continuous support. His involvement with his enthusiasm and immense scientific experience has inspired me and nourished my growth as a young scientist. I owe him special thanks for patiently correcting my English and moral support at the moments when it was needed.

I gratefully acknowledge Johannes Jaeger who has been of great help to my research. He has made available his knowledge in a number of ways. I have benefited greatly from his valuable advices and guidance in studying the gap gene network. Without his crucial contribution that part of the thesis would not have been successful. I would like to thank him for conducting global search optimizations and supplying new data for Huckebein. Yogi, it has been a pleasure to work with you and I hope that our collaboration will continue in the future.

I am sincerely grateful to committee members for their constructive comments on this thesis. I am thankful that in the midst of all their activities, they have found time for joining my Ph.D. committee.

I am grateful to Piet Hemker who introduced me to many aspects of parameter estimation and identifiability analysis. I would like to thank him for his valuable suggestions and stimulating discussions.

I would like to thank the members of the CLS programme for fruitful collaboration. In particular, special thanks go to Jaap Kaandorp for biological insight. I also want to thank Yves Fomekong-Nanfack for sharing his global search results and enlightening me relevant computational issues. I am grateful to Chris Klaassen, Marten Postma, and Nadia Lalam for their valuable advices in scientific

discussions.

I am thankful to John Reinitz for his interesting talks which had everlasting inspiring effect. I am also grateful to Manu for sharing his global optimization code and discussing his results for the delay model. Additionally, I would like to thank people from the Russian FlyEx group. Namely, Maria Samsonova, Konstantin Kozlov, and Ekaterina Myasnikova are acknowledged for assistance regarding the data. I would also like to thank Alexander Samsonov, Konstantin Kozlov, and Vitaly Gursky for exchange of ideas on various aspects of modelling and simulation of the *Drosophila* blastoderm stage.

Collective and individual acknowledgements are owed to my colleagues at CWI. In particular, I would like to thank my former office mates Carolynne Montijn, Valeriu Savcenco, Chao Li, Anna Mozartova, Janis Bajars, and Margreet Nool for creating a pleasant working atmosphere and scientific discussions during coffee breaks. I likewise extend my appreciation to Svetlana Dubinkina, Maciej Dobrzynski, Peter van Heijster, Fabian Brau, Jeroen Wackers, Ivan Zapreev, Yunus Hassen, and Julia Zijlstra for the pleasure of talking to them.

Many thanks to my friends from outside CWI. It is a pleasure to mention: Antonio P. Blazquez and his wife Encarna, Annemarie Cramer, Ilia Korjoukov, Gulyasemin Cakir, Murat Ahat, Saskia Spaninks, Andrew Salt, Eric Pu, Jura, Suze, Jordi V. Rodriguez, Sapar Annayev, Jumamurat Bayjanov, and Marjolijn Voorn. Brian and Susan, thank you for your kind hospitality. Also many thanks to my friends in Turkmenistan: Batyr, Muhammet, Murad, Aylar and their families.

Last but not least, I wish to express my love and gratitude to my beloved family in Turkmenistan, for their patience, understanding and unlimited support through the duration of my research. Foremost, I would like to thank my parents, Allaberen and Govherjan, for their endless love and faith in me throughout my life. I am also grateful to Mehri, Merjen, Maral, Gulruh, Shatlyk, Suray, Annageldy, Akmuhammet, Hatija, and Davud, for being supportive all these years. I am ever-thankful to my grandparents, who are not here anymore but will always be in my heart.

Maksat Ashyraliyev
Amsterdam
September 11, 2009