Model Discrimination Using Markov Chain Monte Carlo Methods

by

Samira Masoumi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Chemical Engineering

Waterloo, Ontario, Canada, 2013

Authors' Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Model discrimination deals with situations where there are several candidate models available to represent a system. The objective is to find the "best" model among rival models with respect to prediction of system behavior. Empirical and mechanistic models are two important categories of models. Mechanistic models are developed based on physical mechanisms. These types of models can be applied for prediction purposes, but they are also developed to gain improved understanding of the underlying physical mechanism or to estimate physico-chemical parameters of interest. When model discrimination is applied to mechanistic models, the main goal is typically to determine the "correct" underlying physical mechanism. This study focuses on mechanistic models and presents a model discrimination procedure which is applicable to mechanistic models for the purpose of studying the underlying physical mechanism.

Obtaining the data needed from the real system is one of the challenges particularly in applications where experiments are expensive or time consuming. Therefore, it is beneficial to get the maximum information possible from the real system using the least possible number of experiments.

In this research a new approach to model discrimination is presented that takes advantage of Monte Carlo (MC) methods. It combines a design of experiments (DOE) method with an adaptation of MC model selection methods to obtain a sequential Bayesian Markov Chain Monte Carlo model discrimination framework which is general and usable for a wide range of model discrimination problems.

The procedure has been applied to chemical engineering case studies and the promising results have been discussed. Four case studies, order of reaction, rate of Fe^{III} formation, copolymerization, and RAFT polymerization, are presented in this study.

The first three benchmark problems allowed us to refine the proposed approach. Moreover, applying the Sequential Bayesian Monte Carlo model discrimination framework in the RAFT problem made a contribution to the polymer community by recommending analysis an approach to selecting the correct mechanism.

Acknowledgements

I would like to thank my supervisor Professor Duever for his continuous guidance and support during this project. I am so thankful to him for providing me the opportunity of studding in the University of Waterloo. He always provide me advices in addition that has always given me great freedom to try and explore my ideas.

I wish to express my sincere gratitude to Professor Park Reilly for his invaluable support and encouragement throughout my Ph.D. studies. In addition, his precious life advices encouraged and help me during difficulties in these years.

Special thanks go to my committee, Professor P. James McLellan, Professor Paul Fieguth, Professor Alexander Penlidis, and Professor Ali Elkamel for their support and helpful suggestions. I would like to give my special thanks to Professor Penlidis for his guidance about RAFT polymerization.

I would also to thank all the staff and faculty members in the Chemical Engineering Department at University of Waterloo.

During my PhD, the moments I shared with my friends and office mates in Waterloo will always be remembered and I am deeply thankful to all of them.

Finally, I take this opportunity to express the profound to my beloved parents and my brother for their love, encouragement and support throughout my life.

Dedication

This thesis is dedicated to both my beloved parents: my mother, Mahin, who has always been my best friend and has always been there for me; and my father, Reza, who has always given his love and support and who has taught me perseverance.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AM	Adaptive Metropolis Hastings method……..	24
AP	Adaptive Proposal Random Walk method……..	23
ARMH	Acceptance-Rejection Metropolis Hastings method ……	19
BFF	Buzzi Ferraris and Forzattie method …….....	29
HR	Hsiang and Reilly method ……..	54
MCMC	Markov Chain Monte Carlo method ...……	7
MH	Metropolis Hastings method ...……	14
RWMH	Random Walk Metropolis Hastings method ...……	16
SBMCMD	Sequential Bayesian Monte Carlo Model Discrimination ……	63

Nomenclature

B_{ij}	----------	Bayes' factor (page 33)
c	----------	Constant in ARMH
c_d	----------	Scaling factor, function of dimension d
$E[.]$	----------	Expectation
$f_k(x)$	----------	Output of k^{th} model
H	----------	Number of samples used to update covariance in AP
i	----------	Index for experimental condition
k	----------	Index for candidate model
K	----------	Number of rival models
K	----------	Matrix of samples in AP
$L(.)$	----------	Marginal likelihood function
$L(M_k\|\mathbf{y})$	----------	Likelihood of the k^{th} model
$l(.)$	----------	Likelihood function
$l(\theta\|M_i, \mathbf{y})$	----------	likelihood of parameter in k^{th} model, given data
$L(M_i\|\mathbf{y})$	----------	Marginal likelihood of M_i
M_i	----------	Index for the i^{th} model
M	----------	Model indicator
$M^{(g)}$	----------	A model indicator
M_k	----------	Model index
m	----------	Model index in RJMCMC method
npar	----------	Number of parameters in the k^{th} model
N	----------	Number of observed data points
N_{sample}	----------	Desired number of samples
$P_{Trans}(x^{t+1}\|x^t)$	----------	Transition kernel
$\mathbf{P}_{\boldsymbol{Trans}}$	----------	Transition matrix

pr_{ij}	----------	denotes the transition probability $P(X_{t+1} = j \mid X_t = i)$
P_{select}	----------	Selection probability criterion
$P(.)$	----------	Probability
$Q(.)$	----------	Target distribution, invariant distribution, stationary distribution
$q(x)$	----------	Proposal distribution
$q(.\mid x^t)$	----------	Proposal distribution
R	----------	A uniform random variable
R_t	----------	Estimated covariance matrix in AP
r	----------	Output dimension
t	----------	Time, iteration, step
$U(0,1)$	----------	Standard uniform distribution
w_k	----------	Weight assigned to each model in model averaging
$w(x)$	----------	importance weight in independent sampler method
x_i	----------	Vector of input variables at the i^{th} experimental trial
x^t	----------	State at time t
x, y	----------	Random states
x_{-i}	----------	present all of *x* except x_i
Y	----------	Candidate point
$\mathbf{y}$	----------	Vector of observed data
$\mathbf{y_n}$	----------	Vector of observed data till step n
y_n	----------	One observed data at step n
y_{ik}	----------	Vector of outputs values, k^{th} model, i^{th} experimental trial
$y(x)$	----------	Output of the averaged model
$\mathbf{y_n} = (y_1, \dots, y_n)$	-------	vector of n observations
z	----------	increment random variable in RWMH

α_{AR} ---------- First acceptance probability in ARMH

α_{MH} ---------- Second acceptance probability in ARMH

$\alpha(x^t, Y)$ ---------- Acceptance probability of moving from sate x^t to Y

ε_{ik} ---------- Error, i^{th} observation, k^{th} model.

θ_k ---------- Vector of parameters in the k^{th} model

θ ---------- Parameters

θ^* ---------- A possible value of parameters

Σ ---------- covariance matrix

π_j ---------- Prior of model i

$\pi(x_i|x_{-i})$ ---------- Conditional distribution

$\pi(.)$ ---------- Prior function

$\pi(\theta|M_i)$ ---------- Prior of parameter in k^{th} model

Chapter 1
Introduction, Motivation and Research Objectives

1.1 Introduction to the Problem of Model Selection

The objective of model discrimination techniques is to select the "best model" among rival candidate models. The model discrimination problem could be interpreted as a regression problem in variable dimension space. The variability in dimension is caused by the different number of parameters in the rival models.

Suppose K models are proposed to describe a process and there are N measurements available. The rival models have the format shown in equation (1.1)

$$y_{ki} = f_k(x_i, \theta_k) + \varepsilon_{ki} \quad \begin{matrix} k = 1,2,\dots,K \\ i = 1,2,\dots,N \end{matrix} \qquad (1.1)$$

where K is the number of rival models, N the number of observations, θ_k the vector of parameters in the k^{th} model, x_i the vector of input variables at the i^{th} experimental trial, ε_{ki} is the error in the i^{th} observation and the k^{th} model, and finally y_{ki} the vector of output values of the k^{th} model at the i^{th} experimental trial.

"Model averaging" is another common term used in the literature (Wasserman, 2000). The goal of model averaging methods is to come up with a predictor for the real system. This predictor is a weighted combination of candidate models when several models have been proposed. To clarify, a model discrimination method selects just one of the K rival models; but, a model averaging method leads to a predictor like equation (1.2), where the w_k values are weights assigned to each rival model.

$$y(x) = \sum_{k=1}^{K} w_k f_k(x) \qquad (1.2)$$

In Bayesian model averaging methods, w_k is the final probability of each model to be the "correct" one.

This research is focused on model discrimination, not model averaging, however the main idea of some of the methods discussed are common between model discrimination and model averaging methods. Thus, the reader may find them under the category of model averaging in the literature.

For the purpose of model selection, the same input condition is applied to all the rival models and the real system, and then their outputs are compared to find which model can predict the real system better. Sufficient input conditions should be tested to provide enough evidence for selecting one of the rival models as the best one (see Figure 1-1).

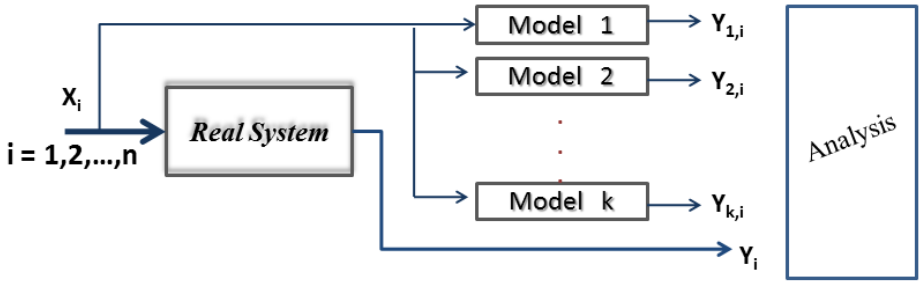

Figure 1-1: Model selection process

Model discrimination methods may be categorized in different ways. In this research, they are classified as sequential or non-sequential approaches, that is, those that include an experimental design step and those that do not. In non-sequential methods, all the available observed data from the real system are input to the analysis algorithm, and then the "best"

model is selected, using an appropriate criterion, from the candidate set of models. No additional experiments are designed or used to augment the initial data set. We refer to this procedure as model selection analysis. On the other hand, in the sequential methods, an experimental design technique is used in conjunction with the model selection method. In this approach, the model discrimination process has an iterative structure in which experiments are designed, carried out, and added to the existing data. Then an analysis is performed to see if the "best" model can be identified. If the "best" model can be selected, the process stops. If not, the next iteration will start. The schematic structure of the sequential model discrimination procedure is shown in Figure 1-2.

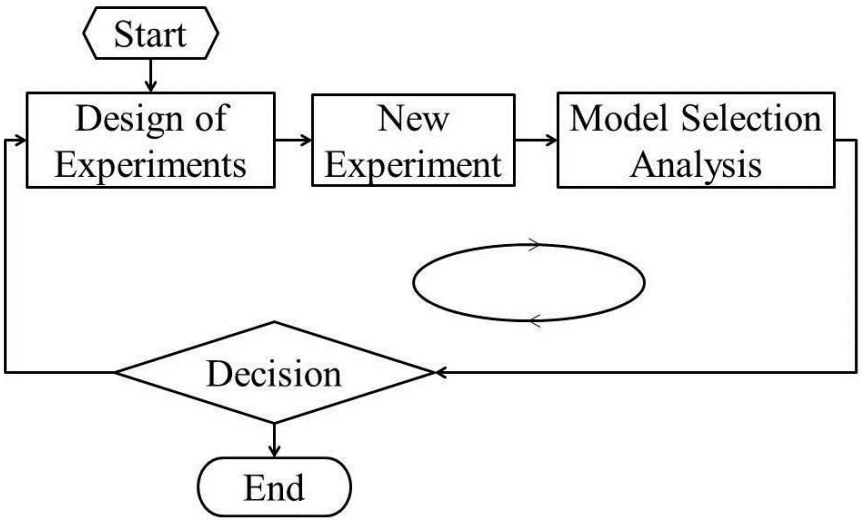

Figure 1-2: Sequential model discrimination scheme

Gathering information from the real system, for example in the case of modeling chemical or industrial systems, could be expensive and time consuming. Consequently, it is highly desirable to get the maximum amount of information from the minimum number of

experiments. This is the goal of sequential model discrimination techniques in which experiments are designed to provide the maximum possible information with respect to discrimination between the models.

Furthermore, modeling is used for two major purposes. In the first one, finding a model that can predict the real system behavior is desired. These models are often used for prediction purposes in process control and optimization applications. Therefore, the model structure is not important; hence, empirical models such as response surface models, neural networks, fuzzy models and other statistical models can be used.

On the other hand, there are mechanistic models, which are developed based on physical-chemical mechanisms. These types of models can also be applied for prediction purposes but are most often developed to gain improved understanding of the underlying physical mechanism or to estimate physico-chemical parameters of interest.

In particular, we are interested in determining chemical reaction mechanisms. Therefore, we deal with mechanistic models, which generally have more complicated structures in comparison to empirical models and in particular are usually nonlinear in the parameters. Markov Chain Monte Carlo (MCMC) methods (Gilks et al., 1996) have been used in the proposed framework to overcome the need for model linearization which is required in most sequential model discrimination methods found in the chemical engineering literature, for example (Buzzi-Ferraris and Forzatti, 1984; 1990; Buzzi-Ferraris et al., 1983; Reilly, 1970). The usage of Markov Chain Monte Carlo (MCMC) techniques in model selection has become popular in the literature (Godsill, 2001) as a solution for model discrimination.

To be clear, in this research the term model selection refers to the mathematical process of selecting a model from a candidate set, given data. The term model discrimination on the other hand indicates a sequential, iterative process, which includes experimental design and model selection.

In this research, a new model discrimination framework will be introduced, the novelty of which is that we combine a well-known model discrimination experimental design procedure with MCMC marginalization likelihood model selection methods, thus yielding a general Bayesian sequential framework of great value to practicing engineers and scientists. Three different implementations of this framework will be explained and applied to case studies.

1.2 Motivation

Burke et al. (1994, 1996, 1997) compared three different procedures for model discrimination in free radical copolymerization reactions. One of the methods used was the method developed by Hsiang and Reilly (HR) (1971), which was not found to work as well as the other two methods investigated. Of the three methods, HR was the only method that does not rely on linearizing the model of the system under investigation. The copolymerization systems were nonlinear. Therefore, the HR method was expected to perform somewhat better than methods that relied on a linearization of the models. One possible explanation for the poor performance of the method might be related to the awkward way in which the posterior parameter distribution had been handled, namely as an array of discrete values. This could pose a serious problem. A much better method of handling the posterior parameter probability distribution would be to use Monte Carlo methods for sampling the distribution. This somewhat puzzling result forms in part the motivation for this project.

1.3 Research Objectives

Based on the research motivation and with information from the literature review, the following research objectives were identified:

- The development of a Bayesian Monte Carlo-based model discrimination framework, which incorporates both an experimental design step and an analysis step.
- Applying the proposed model discrimination method to free radical co-polymerization as a benchmark case study and finding the reason why the Hsiang-Reilly method did not discriminate between the candidate models in a research previously conducted by Burke (1994).
- Applying the Bayesian MCMC model discrimination method to rival Reversible Addition−Fragmentation chain-transfer (RAFT) polymerization models, which is a new application.

This thesis consists of six chapters. In Chapter 2, Markov Chain Monte Carlo (MCMC) techniques are reviewed. Chapter 3 contains a literature review and classification of model discrimination methods. Our proposed framework is presented in Chapter 4. Case studies are presented in Chapter 5 and finally Chapter 6 contains conclusions and recommendations for future work.

In this thesis, all symbols are defined upon their first usage and those that are used more frequency are also listed in the nomenclature section which precedes this chapter.

Chapter 2
Markov Chain Monte Carlo (MCMC) Techniques

2.1 Introduction

The Monte Carlo methods have two major applications (Hammersley et al., 1965). In the first application, used in the simulation of stochastic physical process, random values are usually generated from a well-known distribution to represent a stochastic event in the system; then, using the stochastic and other variables, the system is simulated. The resulting output values from the simulation are random variables due to the randomness in the input variables. Properties of the output variables distributions could be used in studying the behavior of the stochastic system. Some examples of this type of Monte Carlo applications in Chemical Engineering can be found in Tamir (1998).

The second application of Monte Carlo methods is in the estimation of integrals. The integration operation appears in different Bayesian statistical calculations, including calculation of the normalizing constant in the posterior probability, the marginal distributions and expectation estimations with respect to a distribution (Brooks, 1998).

The term Markov chain refers to a sequence of random events such that any state depends only on the last state in the chain. Markov Chain Monte Carlo algorithms (MCMC) are techniques which generate Markov chains converging to the desired distributions when it is impossible to sample directly from the target distribution. Thus, MCMC methods make it possible to estimate the quantities of interest using the sampled values from the target distribution. Consequently, they may be used as a numerical solution for the problem of estimating integrals which occur in Bayesian calculations.

In most of the solutions to the problem of model selection, it is desirable to estimate the marginal likelihood of the model by integrating out the model parameters which is the normalizing constant of the posterior distribution. For the purpose of estimating this integral, different MCMC methods have been proposed in the literature. MCMC model selection methods will be reviewed in the next chapter. They all contain a step for sampling from a target distribution using an MCMC method. Thus a review of MCMC methods is presented in this chapter as a prerequisite for the studying of MCMC Model discrimination methods.

2.2 MCMC Terminology and Definitions

This section introduces the basic concepts needed to explain different MCMC methods. As mentioned earlier, a Markov chain is a sequence of random variables, $\{x^0, x^1, x^2, \ldots\}$, such that at each time step $t \geq 0$, the next state x^{t+1} is sampled from a distribution $P(x^{t+1}|x^t)$. Thus any state depends only on the last state. $P(x^{t+1}|x^t)$ is called a transition kernel, which is assumed to be time independent; thus, the chain is time-homogenous.

Considering a chain of S states, the transition matrix of S states in a $[(S+1) \times (S+1)]$ matrix, is defined by equation (2.1).

$$\mathbf{P_{Trans}} = \begin{bmatrix} pr_{00} & pr_{01} & \ldots & pr_{0S} \\ pr_{10} & pr_{11} & \ldots & pr_{1S} \\ \ldots & \ldots & \ldots & \ldots \\ pr_{S0} & pr_{S1} & \ldots & pr_{SS} \end{bmatrix} \qquad (2.1)$$

where pr_{ij} denotes the transition probability $Pr(X_{t+1} = j \,|X_t = i)$, which is independent of time t. Notice that pr_{ij} values are probabilities, thus, they need to satisfy the equality (2.2).

$$pr_{ij} \geq 0 , \qquad \sum_{j=0}^{S} pr_{ij} = 1 \tag{2.2}$$

Under certain conditions which will be defined later, the MCMC chain will converge to its stationary distribution also called invariant distribution.

The Ergodic theorem (Gilks et al. 1996) guarantees the convergence of chains to the target distribution. To describe the required assumptions, recurrent, nonnull, and irreducible properties will now first be defined.

Starting from any state, if a Markov chain definitely re-enters the same state, the chain is called *recurrent*; otherwise, it is called transient. Also, a recurrent chain is called *nonnull* if its expected time to return to any state is finite, and null if it is infinite. When any set of states can be reached from any other state in finite moves, the related Markov process is called *irreducible*; otherwise, it is reducible. Finally, state i is called periodic if equation (2.3) is true.

$$\gcd\{t: Pr_{ii}^{t} > 0\} = d(i) \qquad d(i) > 1 \tag{2.3}$$

The gcd abbreviation stands for greatest common divisor and pr_{ii}^{t} denotes the probability that, starting from state i, the process re-enters state i after t steps. In contrast, when $d(i) = 1$, the state is called *aperiodic*. When all the states in a chain are aperiodic, the chain is called aperiodic.

A Markov chain is called *ergodic* if it is recurrent *nonnull*, *aperiodic* and *irreducible*. The Ergodic theorem states that an ergodic Markov chain has a limiting or stationary distribution. According to this theorem, equation (2.4) is true for an ergodic Markov Chain.

$$Q_j = \sum_{i=1}^{S} \lim_{n\to\infty} pr_{ij}^{(n)} \tag{2.4}$$

where Q_j shows the stationary probability of state j. In equation (2.4) the summation is over all possible states i. In the proof of the ergodic theorem, aperiodic and recurrent nonnull implies the existence of a limiting distribution, while irreducibility implies the uniqueness of its independence on the initial distribution.

A Markov chain is called reversible if there is a distribution π such that:

$$\pi_i \; P(x_{t+1} = j | x_t = i) = \; \pi_j \, P(x_{t+1} = i | x_t = j) \qquad (2.5)$$

The detailed balance condition determines the stationary distribution. It indicates that, under the reversibility condition, distribution Q is the stationary distribution of the Markov chain. The reversibility condition is a sufficient condition for the convergence of the Markov chain to the stationary distribution but it is not necessary. In other words, it is stronger than the required condition Tierney (1998).

The regularity condition that guarantees convergence to the desired stationary distribution is that the Markov chain must be irreducible, aperiodic, and positive recurrent. Irreducibility ensures that a Markov chain, offer some iterations, can reach any non-empty set with a positive probability from any initial starting points. Aperiodicity stops the Markov chain from oscillating between different sets of states in a regular periodic movement. Lastly, positive recurrency ensures that all subsequent samples will be distributed as a stationary distribution once a previous point is sampled from it. Generally, these conditions are easily met in most MCMC problems. The mathematical details about these conditions can be found in Chib and Greenberg (1995).

Ordering methods are used to pick the suitable Markov chain among those that have the same stationary distribution. In addition they are used to develop new efficient methods. Mira (2001) reviewed the ordering methods on the space of Markov chains.

Two main criteria used to evaluate the performance of the transition matrix used for MCMC simulation are the asymptotic variance (AV) of the resulting estimates and the speed of convergence (SC) to stationary distribution (Mira, 2001). The asymptotic variance represents the accuracy of $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$ as an estimate of $\mu = E_q g(X)$ when X_i are MCMC samples from the $\pi(x)$ distribution and $E(.)$ shows the expectation value over q.

2.3 Sampling Methods

Assume that the objective is sampling values from the distribution $\pi(x)$ and it is desired to generate a chain, $\{x^i\}$, with N_{sample} samples from this distribution. In most applications, $\pi(x)$ has a complex structure and it is not easy to sample from it directly. In the following sections, we review briefly the MCMC sampling methods which we have used during different phases of this project or which are used in the methods presented by others referred to in this research. In all these methods, Markov chain samples are generated that converge to the target distribution.

2.3.1 Rejection Method

Let's assume that $Q(x) < Mg(x)$ where $M > 1$. Thus $g(x)$ is an upper limit of $Q(x)$ and it is a distribution which we know how to sample from. In the rejection method (Karandikar, 2006), a candidate sample Y is drawn from $g(x)$ and this candidate value will be accepted if $R < \frac{Q(Y)}{M\, g(Y)}$, where R is independently generated from $U(0,1)$. If Y is not accepted, a new sample

should be generated from $g(x)$. The efficiency of this algorithm is determined by how close the proposal distribution $Mg(x)$ is to $Q(x)$. For example, in the diagram shown in Figure 2-1, the candidate point x_1, will be accepted with a probability equal to a/b, but this probability is c/d, which is higher, for point x_2. Algorithm 2.1 shows the steps in the rejection sampling method.

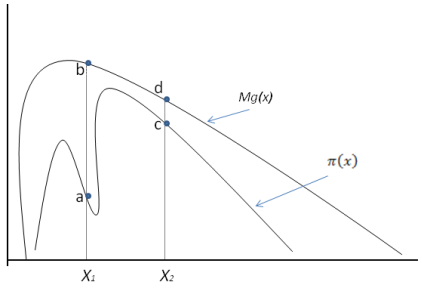

Figure 2-1: Acceptance-Rejection sampling

Algorithm 2.1: Rejection sampling method

Input: Proposal distribution and M such that: $Q(x) < Mg(x)$

Initialize x^0

for i=0… $N_{sample} - 1$

Sample $Y \in g(Y)$

Sample $R \in U(0,1)$

if $R < \frac{Q(Y)}{M\, g(Y)}$ then

$x^{i+1} = Y$

else

Sample a new Y

2.3.2 Gibbs Method

The Gibbs sampler generates samples from a multi-dimensional target distribution by sampling from conditional distributions, so this property is beneficial in cases where it is easier to sample from the conditional distributions than from the joint distribution. Conditional distributions are often known in statistical models or can be obtained by assuming other variables in the joint distribution fixed.

To describe the Gibbs sampling method (Spiegelhalter et al., 1994), suppose that x is partitioned into r blocks: $x = (x_1, \dots, x_r)$. Let $x^t = (x_1^t, \dots, x_r^t)$ denote the current state of x. Then in the next step of the Gibbs iterations, x^{t+1} will be generated block by block from the conditional distribution of that block, updating the x value for block 1 to the current one and keeping the last iteration value for the current block as the final block. This method is presented in algorithm 2.2.

The Gibbs approach does not include a method for drawing samples from the full conditional distributions. In cases where $Q(x_i|x_{-i})$ has a standard distribution, like the normal distribution, some methods can be found in the literature (Robert and Casella, 1999; Tanner, 1991) to generate samples. Another possible solution involves picking a conjugate prior which combines with the likelihood, produces a standard distribution for the posterior, and the last solution involves using other methods like Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) for sampling from the full conditional distribution, $Q(x_i|x_{-i})$.

Algorithm 2.2: Gibbs sampling method

Input:

r : dimension of x

$Q(x_i|x_{-i})$: conditional distribution

Initialize x^0

for t =0… N_{sample}-1

  for i=1 … r

    sample x_i^{t+1} from $Q(x_i|x_{-i})$ where $x_{-i} = x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_r$

where $Q(x_i|x_{-i})$ shows conditional distribution.

2.3.3 Metropolis-Hastings Method

The Metropolis-Hastings (MH) algorithm (see algorithm 2.3), was introduced by Metropolis and Rosenbluth (1953), to numerically estimate integrals in computing properties of interacting individual molecules. Later, Hastings (1970) generalized this technique as an easily implemented procedure for sampling from distributions.

At time t, the MH algorithm generates a sample from the target distribution $Q(x)$ by sampling a candidate point Y for the next sample x^{t+1} from a proposal distribution $q(.|x^t)$, from which it is known how to sample. Then it applies a criterion to keep or reject the candidate point. Note, that the proposal distribution may depend on the current point x^t. The candidate point Y is accepted with probability $\alpha(x^t, Y)$, which is defined in equation (2.6).

$$\alpha(x, Y) = \min\left(1, \frac{Q(Y)q(x|Y)}{Q(x)q(Y|x)}\right) \tag{2.6}$$

The min operator in this equation represents the selection of the minimum between 1 and $\frac{Q(Y)q(x|Y)}{Q(x)q(Y|x)}$. So, if $\frac{Q(Y)q(x|Y)}{Q(x)q(Y|x)}$ is greater than 1, the candidate point is accepted. Otherwise, it will be accepted with α probability defined in equation (2.6).

If the candidate point is accepted, the next state becomes $x_{t+1} = Y$. If the candidate is rejected, the chain does not move so $x^{t+1} = x^t$.

Algorithm 2.3: Metropolis-Hastings Method

Initialize x^0

for t =0… N_{sample}-1

sample $\mathbf{Y} \in q(Y|x^t)$

calculate $\boldsymbol{\alpha}(x^t, Y) = \min\left(1, \frac{Q(Y)q(x^t|Y)}{Q(x^t)q(Y|x^t)}\right)$

$x^{t+1} = Y$

if $\boldsymbol{\alpha}(x^t, Y) < 1$

sample $\mathbf{R} \in U(0,1)$

if $\mathbf{R} > \alpha(x^t, Y)$

$x^{t+1} = x^t$

As mentioned, the sufficient condition to generate samples from a desired distribution by using a transition kernel $P_{Trans}(x, Y)$ is satisfaction of the reversibility condition shown in equation (2.5). Using the acceptance probability $\alpha(x, y)$ shown in equation (2.6), with a proposal

distribution $q(.|x_t)$ in the MH algorithm, guarantees that the reversibility condition is satisfied, as shown in equation (2.7) (Chib and Greenberg, 1995).

$$Q(x)q(x,y)\alpha(x,y) = Q(y)q(y,x)\alpha(y,x) \qquad (2.7)$$

where x and y are two states in the sampling space. Note that the transition probability for state x to y is $P_{Trans}(y|x) = q(y|x)\alpha(y|x)$. However, this form of acceptance probability is not the only one which causes an ergodic chain (Brooks, 1998), although Peskun (1973) shows that this form is optimal relative to the asymptotic variance (defined in section 2.2).

The proposal distribution $q(.|x)$ can be selected from a number of different standard distributions. For instance, $q(.|x)$ might be a multivariate normal distribution with mean x given by the current state x and constant covariance matrix Σ. However, the rate of convergence to the stationary distribution will depend crucially on the relationship between $q(.|.)$ and $\pi(x)$. The closer the proposal distribution is to the target distribution the faster the convergence will be (Mira and Sargent, 2000).

The original Metropolis algorithm (Metropolis and Rosenbluth, 1953) considers only symmetric proposals having the form $q(Y|x) = q(x|Y)$ for all x and Y. In this case, the acceptance probability becomes

$$\alpha(x,Y) = \min\left(1, \frac{\pi(Y)}{\pi(x)}\right). \qquad (2.8)$$

2.3.4 Random Walk Metropolis-Hastings (RWMH) Method

Another special case of the Metropolis algorithm is the random-walk Metropolis (Gilks et al., 1996), for which $q(Y|x) = q(Y-x) = q(x-Y)$, the acceptance probability related to this symmetric proposal distribution, is same as equation (2.8). Therefore, the increment random

variable z is drawn from the walking distribution, q, and then the candidate point Y is determined according to $Y=x+z$. Algorithm 2.4 shows the steps of random walk MH.

Algorithm 2.4: Random Walk Metropolis-Hastings Method

Initialize x^0

for t = 0… N_{sample}-1

  sample $z \in q(z)$

  $Y = x^t + z$

  calculate $\alpha(x^t, Y) = \min\left(1, \frac{Q(Y)}{Q(x^t)}\right)$

$x^{t+1} = Y$

if $\alpha(x^t, Y) < 1$

  sample $R \in U(0,1)$

  if $R > \alpha(x^t, Y)$

    $x^{t+1} = x^t$

2.3.5 Independence Sampler Method

The independence sampler (Liang et al., 2011) is a Metropolis-Hastings algorithm whose proposal $q(Y|x) = q(Y)$ does not depend on x. So, the acceptance probability is:

$$\alpha(x, Y) = \min\left(1, \frac{w(Y)}{w(x)}\right) \quad w(x) = {\pi(x)}\Big/{q(x)} \tag{2.9}$$

where $q(.)$ should be a good approximation of $\pi(x)$. $w(x)$ is also called the importance weight function. Samples with large weights have more chance to be accepted or re-selected. Thus a very high value for $w(x)$ can make the process stall (Tierney, 1994).

2.3.6 Single Component Metropolis Hastings Method

Single-component Metropolis-Hastings (Gilks et al., 1996) or the one block at a time algorithm, divides the multidimensional x to sub-blocks and samples from sub-blocks instead of sampling the entire vector simultaneously. Let x_{-i} present all of *x* except x_i , x_i^t the state of x_i at the end of iteration t and Y_i the candidate state. So $x_{-i}^t = \{x_1^{t+1}, \dots, x_i^{t+1}, x_{i+1}^t, \dots, x_r^t\}$. Then the acceptance probability of Y_i , which is a candidate value for $x_{i,t+1}$, may be shown as the following equation:

$$\alpha\left(x_{-i,t}, x_i, Y_i\right) = \min\left(1, \frac{Q(Y_i|x_{-i})q_i(x_i|Y_i, x_{-i,t})}{Q(x_i|x_{-i})q_i(Y_i|x_i, x_{-i,t})}\right) \tag{2.10}$$

where $Q(Y_i|x_{-i})$ is the full conditional distribution which is the distribution of the i^{th} component of x conditional on all the remaining components, where x has the Q(.) distribution.

$$Q(x_i|x_{-i}) = \frac{Q(x)}{\int Q(x)dx_i} \tag{2.11}$$

After selecting the proposal distribution, picking a suitable set of parameters for that proposal distribution influences the sampling process. For example, in random walk Metropolis, there is a tradeoff between the acceptance rate and mixing rate, which can be optimized by the proposal distribution parameters.

The Gibbs sampler may be categorized as a special case of single-component Metropolis-Hastings and it is particularly useful for generating n-dimensional random vectors with the following proposal distribution which leads to an acceptance probability equal to 1.

$$q_i(Y_i|x_i, x_{-i}) = Q(Y_i|x_{-i}) \tag{2.12}$$

2.3.7 Acceptance Rejection Metropolis-Hastings (ARMH) Method

If a proposal distribution, $q(x)$ and a constant c can be found to satisfy the dominant condition:

$$D = \{x: Q(x) \leq c\, q(x)\} \tag{2.13}$$

Then samples could be generated from the proposal distribution instead of $Q(x)$. After that, these samples should be accepted or rejected using the $Q(x)/(c\, q(x))$ as the acceptance rate. The dominant condition in equation (2.13) is a key requirement of the classical accept-reject sampling method. But in the ARMH method the dominant condition may not be satisfied for some x. Let D^c be the complement of D. The one-block acceptance-rejection Metropolis-Hastings algorithm consists of two steps:

A. Accept-reject step:

Generate a candidate sample $Y \sim q(x)$, and then accept Y with probability given by equation (2.14):

$$\alpha_{AR} = \min\left\{1, \frac{Q(Y)}{c\, q(Y)}\right\} \tag{2.14}$$

Continue sampling until a Y is accepted.

B. Metropolis step:

Given the current value x and the accepted proposed value in step A, Y:

If $x \in D$ $\alpha_{MH}(x, Y) = 1$

If $x \in D^c$ and $Y \in D$ $\alpha_{MH}(x, Y) = (c\, q(x))/Q(x)$

If $x \in D^c$ and $Y \in D^c$ $\alpha_{MH}(x, Y) = \min\{1, (Q(x)\, q(x))/(Q(Y)\, q(Y))\}$

Accept Y with probability α_{MH}, otherwise keep x.

ARMH steps are shown in algorithm 2.4.

Algorithm 2.4: Accept – Reject Metropolis-Hastings Method

Initialize x^0

for t =0… N_{sample}-1

sample $Y \in q(x)$

calculate $\alpha_{AR} = \min\left(1, \frac{Q(Y)}{c \times q(Y)}\right)$

sample $R \in U(0,1)$

if $R > \alpha(x^t, Y)$

sample new Y

$x^{t+1} = Y$

if $x \notin D$

if $Y \in D$ $\alpha_{MH}(x, Y) = \frac{c \times q(x)}{Q(x)}$

if $Y \in D^c$ $\alpha_{MH}(x, Y) = \min\left\{1, \frac{Q(x)\, q(x)}{Q(Y)\, q(Y)}\right\}$

sample $R \in U(0,1)$

if $R > \alpha(x^t, Y)$ $x^{t+1} = x^t$

2.3.8 Random Samples from Discrete Distributions

Assume that there are r different discrete states and the target is to generate n samples from a discrete space. Equation (2.15) shows the discretized space of the states and their probabilities by:

$$\Phi = \{\varphi_1, \varphi_2, \dots, \varphi_r\}$$
$$\text{pr}(\varphi_i) \quad i = 1, \dots, r \tag{2.15}$$

To generate random samples, the cumulative probability of the states should be calculated by

$$C_i = \sum_{j=1}^{i} pr(\varphi_j) \qquad i = 1, \dots, r \tag{2.16}$$

After that, N_{sample} uniform random numbers are generated from $U(0,1)$ and for each of these values, the first discrete state with a cumulative value less than or equal to the u value is selected as one of the selected discrete states.

One application of sampling from discrete probability distributions is the sampling from model indicators, which will be discussed in the next chapter.

2.4 Adaptive MCMC

The main challenge in applications of MH method is selecting and tuning the proposal distribution in such a way that the sampling procedure will be efficient. Adaptive MCMC methods may be categorized as internal and external (Moulines and Priouret, 2010). External methods use extra processes which interact with the main process. These types of MCMC methods have been proposed more recently in comparison to internal adaptive ones. They offer efficient sampling methods from complicated distributions which other methods fail to sample efficiently. For instance, the Small-World Chain (SWC) method (Guan et al., 2006) is a non-adaptive external method which was introduced to overcome the problem of stalling in one mode of a multi-modal distribution by using a mixture of local proposals in MH. The adaptive Parallel Tempered (PT) MCMC method is another example of external methods. The PT-MCMC method has been proposed to overcome the problem of sampling from distributions which have low probability regions between high probability areas. This method generates samples from some auxiliary distributions in addition to the main one Q, which is also called

the cool one. The target distributions of these extra chains are $Q^{1/T}$, where T must be positive and greater than 1. In addition, according to an acceptance probability, samples sweep between different chains. The sweeping steps help to provide samples from areas with low probability, since $Q^{1/T}$ tempered distributions are more flattened. This method needs some extra tuning parameters in comparison to the MCMC method. Thus Adaptive PT MCMC methods (Miasojedow et al., 2012) propose the auto-tuning method for the parameters. External adaptive MCMC methods have not been used in this research; thus, no more details will be discussed. However, this class of methods may be useful for some applications of the proposed framework, which deal with distributions with discrete high probability areas, and they could be used to make the proposed package more general. The other category of adaptive methods is internal adaptive methods. These methods use the history of the chain to tune parameters in the process.

Roberts and Rosenthal (2001) presented the scaling method that tries to adapt to the target distribution. To preserve the ergodicity property Gilks et al. (1998) performed adaptation only at the time of recurrence to an atom. They defined a proper atom as a subspace of the sampling space such that when the chain enters this subspace, it becomes independent of its history.

Adaptive MH methods have been used in the proposed procedure so they will be explained in detail in the following subsections.

2.4.1 Adaptive MH

The main challenge in the application of the MH method is the selection and tuning of the proposal distribution in such a way that the sampling procedure is efficient. In the following section, three adaptive Metropolis Hastings methods will be reviewed.

2.4.1.1 Adaptive Proposal (AP)

Harrio et al. (1999) proposed an Adaptive Proposal (AP) algorithm as an efficient tool for reasonably low dimensional problems in random walk MH (RWMH). The walking distribution is tuned by adapting the covariance matrix calculated from a fixed number of previous states. This procedure tries to locally adapt the walking distribution to the target distribution. In this way, the procedure is kept efficient at all times. They suggested the adaptive proposal distribution shown in equation (2.17).

$$q_t(.\,|X_1, \dots, X_t) \sim N(X_t, c_d^2 R_t) \tag{2.17}$$

where d represents the dimension and R_t is the $d \times d$ covariance matrix determined from the H points $X_{t-H+1}, X_{t-H+2}, \dots, X_t$ and the scaling factor c_d, which depends only on the dimension d. H is the number of previous states used for tuning the proposal distribution. K is an $H \times d$ matrix, where each row represents one sampled point from $X_{t-H+1}, X_{t-H+2}, \dots, X_t$. Then R_t is calculated using equation (2.18).

$$R_t = \frac{1}{H-1}\widetilde{K}^T\widetilde{K}$$
$$\widetilde{K} = K - E[K] \tag{2.18}$$

An easy way to obtain a sample from G: $N(X_t, c_d^2 R_t)$ is presented in equation (2.19).

$$G \sim X_t + \frac{c_d}{\sqrt{H-1}}\widetilde{K}^T N(0, I_H) \tag{2.19}$$

The basic choice for c_d is $2.4/\sqrt{d}$ from Gelman et al. (1996). They presented optimal acceptance rates for Gaussian target distributions in different dimensions when the proposal distribution is also Gaussian. It is also common to use this optimal acceptance rate for other

proposal distributions. Using adaptive approaches, Roberts et al. (1997) proved that the optimal acceptance probability is 1/4 as had been suggested earlier by Gelman et al. (1996).

The proposal distribution can be tuned more easily in single component Metropolis since; we get an individual acceptance rate value for each component. However, it is well known that this method converges slowly to the target distribution if there are correlations between different components.

Due to the adaptive nature of the AP algorithm, it is neither Markovian nor reversible. But its ergodicity and convergence have been proven (Roberts and Rosenthal, 2001).

2.4.1.2 Adaptive Metropolis Hasting

Harrio et al (2001) also presented an adaptive MH method (AM), which is an alternative to AP. In it, the proposal distribution is a Gaussian distribution centered at the current state and the covariance matrix is updated using all of the previous states. By comparison, the AP method uses a finite number of previous states to update the covariance matrix. They showed that the AM method has the ergodicity property.

The proposal distribution in AM method is $N(x_{t-1}, \Sigma_t)$, which is a normal distribution whose mean is the last state and the covariance is updated by

$$\Sigma_t = \begin{cases} \Sigma_0 & t \leq t_0 \\ c_d \mathrm{cov}(X_0, \dots, X_{t-1}) + c_d\, \varepsilon\, I_d & t > t_0 \end{cases} \qquad (2.20)$$

where Σ_0 is the initial positive definite covariance, c_d is a function of dimension d, $\varepsilon > 0$ is a constant value, I_d is a d-dimensional identity matrix, t_0 is an initial period, and $\mathrm{cov}(X_0, \dots, X_{t-1})$ is defined in equation (2.21).

$$\mathrm{cov}(X_0, \ldots, X_{t-1}) = \frac{1}{t}\left(\sum_{i=0}^{t} x_i x_i^T - (t+1)\bar{x}_t \bar{x}_t^T\right) \tag{2.21}$$

where $\bar{x}_k = \frac{1}{k+1}\sum_{i=0}^{k} x_i$ and x_i are column vectors.

2.4.1.3 Delayed Rejection Adaptive Metropolis (DRAM)

Harrio et al. (2006) introduced the DRAM method, which combines the adaptive Metropolis sampler and delaying rejection. The ergodicity of this method is proven. Delaying rejection (DR) is a way to modify MH according to the asymptotic variance ordering by Peskun (1973). The idea is to generate a second or even more samples when the candidate is rejected. The probability for accepting the second candidate is dependent both on the rejected candidate and the last accepted one.

2.5 Practical Implementation Issues

The most important implementation issues regarding Markov Chain generation are the initial state x_0, length of the chain, burn-in length and its convergence to the stationary point. These subjects are discussed briefly in the following sections.

2.5.1 Initial State

Subject to regularity conditions, the chain will gradually 'forget' its initial state and a chain starting from x_0, will eventually converge to a unique stationary (or invariant) distribution that does not depend on time or x_0 (Brooks, 1998). This means that, after a sufficiently long burn-in period, points will be independent samples approximately from $Q(.)$ and the chain converges to the stationary distribution.

2.5.2 Burn-in Size

Assume that the chain has been run for n iterations, and the sample points x^t where $t = m + 1, \dots, n$ are distributed according to the stationary distribution. Then the first m samples should be discarded and the remaining n-m samples can be used to estimate the expectation or the point estimate as follows:

$$\mathbf{E}[f(x)] \approx \frac{1}{n-m} \sum_{t=m+1}^{n} f(x_t) \qquad (2.22)$$

The number of discarded samples is called the "burn-in" period, and the size of the "burn-in" period is a subject of research in MCMC methods. It is suggested (Geyer, 1992) that the size of the "burn-in" period should be between 1% and 2% of n, where n is large enough to obtain accurate estimation of the expectation.

2.5.3 Chain Length

Besides the initial state and the" burn-in" period, another important issue is the total chain length, n. An informal way of finding n is to run several Markov chains from different starting values in parallel; then the estimated expected values are compared. If the deviation between the expected value of chains is high, then the chain length should be increased.

2.6 Summary

In this chapter, first the Markov Chain Monte Carlo (MCMC) concept is defined and a review of MCMC sampling methods is presented with a focus on adaptive MCMC methods. MCMC sampling methods are used in MCMC Bayesian Model selection methods to sample from the

parameter probability distributions of the candidate models. Model selection methods will be reviewed in the next chapter.

Chapter 3
Model Discrimination

3.1 Introduction

As mentioned in Chapter 1 and shown in Figure 1-2, page 3, sequential model discrimination methods include two main steps: model selection and design of experiments. Different model selection methods have been proposed in the literature but there are few sequential model discrimination methods. In this chapter, a literature review on the available methods for model selection, design of experiments, and sequential model discrimination are provided.

Model selection methods determine the "best" model out of a set of candidate models to represent a real system according to available observed data from the target system. In our work we are focusing on choosing the best mechanistic model. Of course there is not any guarantee that the "best" model is the "correct" model. The methods usually assign a probability to each rival model. These probabilities are actually the probability of being the "best" model. Thus when there is not sufficient evidence to support any of the candidate models none of them could be used as the model. In other words, more information from the system is needed to be able to discriminate or a different candidate model is needed. As shown in Figure 1-1, the comparison between candidates is based on analysis of the predicted model outputs and the real system output under the same input conditions.

Model selection methods can be classified into two main categories: Bayesian and Non-Bayesian. Then, Bayesian ones can be categorized as MCMC and non-MCMC. Figure 3-1 shows the categorization of the model selection methods reviewed during this research. Chib's

method is highlighted in this figure as it is used in our proposed procedure. Chib's method will be explained later in this chapter.

3.2 Non-Bayesian Methods

The most common non-Bayesian criterion in model selection is the likelihood ratio, which is sometimes called the odds ratio. The likelihood ratio compares models, two at a time. The original format of the likelihood ratio is just appropriate for comparing the probability of a set of parameters according to a specific model with another set and model, $l(\theta_1|M_1)/l(\theta_2|M_2)$. But in model selection, we are interested in the likelihood ratio which applies to two rival models regardless of their parameters value. One approach for this comparison is usage of the maximum likelihood values of each rival model. For instance, equation (3.1) shows this ratio for model 1 with two parameters and model 2 with three parameters.

$$\lambda_{12} = \frac{\underset{\theta_1, \theta_2}{\text{Max}}\, l_1(\theta_1, \theta_2|\mathbf{y})}{\underset{\theta_1, \theta_{2,} \theta_3}{\text{Max}}\, l_2(\theta_1, \theta_{2,} \theta_3|\mathbf{y})} \qquad (3.1)$$

Another example of a non-Bayesian statistical model selection method, was introduced by Buzzi-Ferraris and Forzatti (1983), who used an F-test to check the adequacy of single output models. The Buzzi-Ferraris and Forzatti's method (BFF) was developed later for multi-response cases by checking model adequacy (Buzzi Ferraris et al., 1984; Buzzi-Ferraris and Forzatti, 1990). They applied a χ^2 test with $Nr - npar_k$ degrees of freedom to the residual sum of squares in the k^{th} model; the variables r and $npar_k$ indicate output dimension and the number of parameters in the k^{th} model, respectively, and N is the number of experimental data points. This method will be explained in more detail in section 3.9.

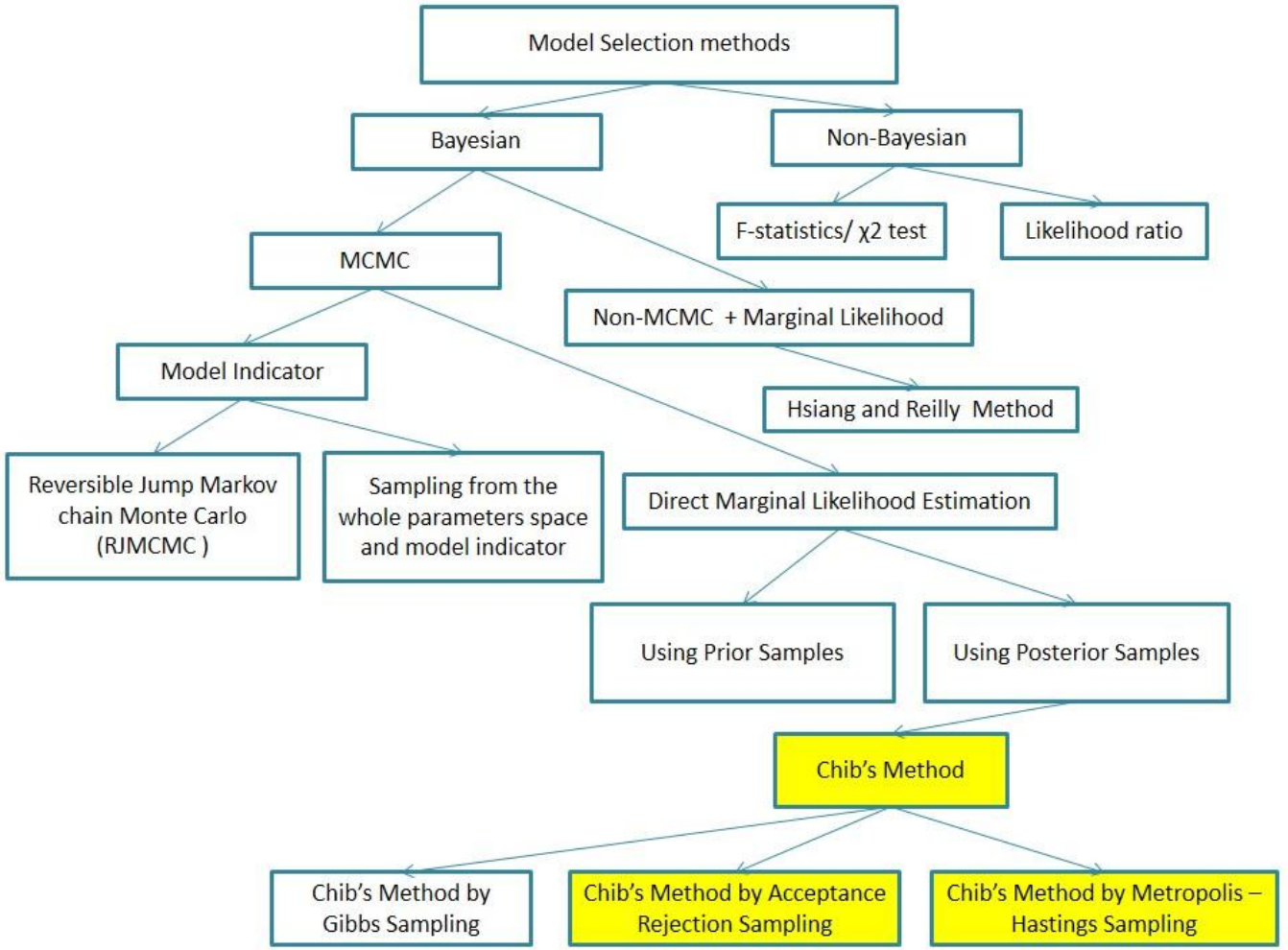

Figure 3-1: Model selection methods classification

3.3 Approaches Using the Likelihood of Rival Models

The two most popular approaches to calculate the likelihood of rival models are marginalization likelihood methods and those based on information criteria. Methods based on information criteria compare rival models according to their maximum likelihood and they penalize the number of parameters. Without a penalty factor, the models with the larger number of parameters would fit better the tuning data points, even though the model with the larger number of parameters does not necessarily predict the system behavior better. Thus the most common format of the information criteria, $\mathcal{L}_{c,k}$ which has a direct relationship with the maximum likelihood and an inverse relation with the number of parameters, is shown in equation (3.2).

$$\ln \mathcal{L}_{c,k} = \ln \mathcal{L}_{k} - f(r, n) \tag{3.2}$$

where $\mathcal{L}_{k}$ is the estimated Maximum Likelihood (ML) of candidate model k and $\mathcal{L}_{c,k}$ is the information criterion. The function $f(r, n)$ is a function of the number of parameters, r, and the number of observed data points, n. For instance, in Akaike's information theory called AIC (Akaike, 1987), $f(r, n) = r$, while in the Bayesian information criteria called BIC (Schwarz, 1978), $f_{BIC} = \frac{1}{2} r \ln n$.

3.4 Bayesian Methods

In Bayesian methods, the posterior probability of a hypothesis is proportional to the product of the likelihood and the prior probability. In model selection, the hypothesis is that the candidate model is the "best" one. Equation (3.3) shows Bayes rule for model selection.

$$P(M_k|\mathbf{y}) = L(M_i|\mathbf{y})\pi(M_k) \tag{3.3}$$

where $P(M_k|\mathbf{y})$ is the posterior probability of the k^{th} candidate model given the observed data $\mathbf{y}$, $L(M_i|\mathbf{y})$ is the likelihood function and $\pi(M_k)$ the prior probability of the model. The likelihood function in a model discrimination problem, $L(M_i|\mathbf{y})$, is independent of the parameter values and it is obtained by integrating out the parameter values.

The marginalization likelihood is another approach to calculating the likelihood of models, which has become more popular especially with advances in computing facilities, since it needs more extensive numerical calculations when dealing with complex nonlinear models in comparison to the information criterion approach. The Marginal likelihood, which is also referred to as the "evidence" of each model, is calculated by integrating out the model parameters as shown in equation (3.4).

$$L(M_k|\mathbf{y}) = \int_{\theta} l(\theta|M_k, \mathbf{y})\pi(\theta|M_k)d\theta \qquad (3.4)$$

where $l(\theta|M_k, \mathbf{y})$ represents the likelihood of the model M_k and its parameters θ . The function $\pi(\theta|M_k)$ is the prior probability of parameters under model M_k and $\mathbf{y}$ is the vector of all data points. Bayesian model selection methods, which use the marginal likelihood are called marginalized methods. Bayesian marginalized model selection methods when applied to nested rival models follow a heuristic principle called Occam's razor. This principle states that "among competing hypotheses favor the simplest one". Jefferys and Berger (1992) illustrated that Bayesian analysis favors the hypothesis with fewer adjustable parameters automatically because of their more narrow priors which lead to enhanced posteriors.

In this research we focus on marginalization methods for calculation of the $L(M_i|\mathbf{y})$; thus the term "marginal likelihood" is sometimes used instead of the more general term "likelihood of the rival model".

3.5 Model Selection Stopping Criteria

In addition to the different approaches for calculating $L(M_i|\mathbf{y})$, model selection methods also differ in their comparison criteria. The first comparison criterion is to use the normalized posterior probability of the models. In this approach, the posterior probabilities of the competing models are checked to see if any of them has a probability higher than the selection probability criteria, P_{select}, which might be set at 95% for example. If the probability of one of the candidate models is higher or equal to P_{select}, it is selected as the "best model".

The second comparison approach is based on comparing models through their Bayes' factor given by equation (3.5).

$$B_{ij} = \frac{P(\mathbf{y_n}|M_i)}{P(\mathbf{y_n}|M_j)} = \frac{L(M_i|\mathbf{y_n})}{L(M_j|\mathbf{y_n})} \tag{3.5}$$

Odds are defined in statistics as $\text{probability}/(1 - \text{probability})$. Using the definition of the Bayes factor when there are only two candidate models, the posterior odds can be defined as

$$\text{Posterior odds} = B_{ij} \times \text{Prior odds} \tag{3.6}$$

If prior probabilities of candidate models are assumed to be equal, then B_{ij} becomes equal to the posterior odds. Jeffreys (1961) suggested the following table for interpreting the evidence for model i compared to model j.

B_{ij}	Evidence for model i
1 to 3.2	Not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
>100	Decisive

In the proposed framework in this study, the normalized posterior probabilities have been used instead of the Bayes' factor to be able to compare more than two rival models.

3.6 MCMC Based Model Selection Methods

Estimation of the marginal likelihood of models is the challenging part of model selection. In what follows we will discuss several specific Bayesian model selection methods that incorporate the use of Markov Chain Monte Carlo (MCMC) methods to solve this difficulty. MCMC model selection methods may be categorized into those which do not estimate marginal likelihoods, and those that do estimate marginal likelihoods directly.

3.6.1 Model Indicator Sampler Methods

The first group of MCMC model selection methods adds a random integer auxiliary variable to the sampling procedure, which acts as a model indicator variable. After the sampling procedure, the ratio of the marginal likelihoods, Bayes' factor in equation (3.5), is estimated from the frequency of the model indicators in the generated samples.

Carlin and Polson (1991) introduced a model indicator sampler method by adding the model indicator variable M, as a random integer number to the Gibbs sampling procedure for model discrimination between non-nested models. Later, Carlin and Chib (1995) developed this method more by introducing pseudo-priors and considering model priors as a tool to improve the convergence conditions.

In Carlin and Chib's method, samples are generated for all the parameters in rival models. Having K rival models and npar_k parameters in each model, each sample in the chain has $\sum_{k=1}^{K} \text{npar}_k + 1$ values. Therefore, samples from θ_j are generated even when $M \neq j$. Because of that, linking densities or pseudo-priors are defined to completely specify the joint model distribution. Pseudo-priors, $\pi(\theta_j|M \neq j)$, do not participate in the calculation of the marginal likelihood, but choosing unsuitable pseudo-priors may cause the procedure to fail. Carlin and Chib used the full conditional distributions shown in equation (3.7) for each set of θ_j and M.

$$P(\theta_j|\theta_{i\neq j}, M, \mathbf{y}) \propto \begin{cases} l(\theta_j, M = j|\mathbf{y})\pi(\theta_j|M = j) & M = j \\ \pi(\theta_j|M \neq j) & M \neq j \end{cases} \tag{3.7}$$

Then the conditional distribution for the model indicator is defined by equation (3.8).

$$P(M = j|\theta, \mathbf{y}) = \frac{l(\theta_j, M = j|\mathbf{y})\{\prod_{i=1}^{K} \pi(\theta_i|M = j)\}\pi_j}{\sum_{k=1}^{K} l(\theta_k|M = k, \mathbf{y})\{\prod_{i=1}^{K} \pi(\theta_i|M = k)\}\pi_k} \tag{3.8}$$

In each Gibbs cycle, θ_j and $\theta_{i\neq j}$ are sampled from the distribution in equation (3.7) and the model indicator is obtained by sampling from a discrete model indicator density. Then the joint posterior distribution and Bayesian factor can be estimated by using equations (3.9) and (3.10):

$$P(M = j|\mathbf{y}) = \frac{\text{number of } M^{(g)} = j}{\text{total number of } M^{(g)}}, \qquad j = 1, \dots, K \tag{3.9}$$

$$B_{ij} = \frac{L(M = j|\mathbf{y})}{LM = i|\mathbf{y})} = \frac{P(M_j|\mathbf{y})/\pi_j}{P(M_i|\mathbf{y})/\pi_i} \tag{3.10}$$

This method can handle discriminating among models that have some common parameters and considering those common parameters just one time in the sampling. Thus, if there are common

parameters, the number of variables in each sample is less than $\sum_{k=1}^{K} npar_k + 1$. The only important thing is that these shared parameters must have exactly the same interpretation in all models. Carlin and Chib (1995) clarified this point in their paper by considering models 1 and 2 shown in equation (3.11). Parameter α shows up in both models but it has two different roles in these two models; the x_i-values are centered far from 0 but the y-values are centered near 0. α in the first model is the mean of y-values, but in the second model it represents the intercept.

$$\begin{array}{lll} M = 1 & y_i = \alpha + \varepsilon_i & \varepsilon_i \sim N(0, \sigma^2) \quad i = 1, \dots, N \\ M = 2 & y_i = \alpha + Bx_i + \varepsilon_i & \varepsilon_i \sim N(0, \tau^2) \quad i = 1, \dots, N \end{array} \tag{3.11}$$

Therefore, α does not have exactly the same role in those two candidate models, although it may be seen so at first. Thus two α values should be considered is the sampling procedure.

Selecting proper linking densities and model priors π_j plays an essential role in the convergence of the Carlin and Polson's method (Carlin and Polson, 1991). In situations where $p(M = j|\mathbf{y})$ is extremely large for one of the models, π_j should be tuned to make the number of samples from each model approximately equal.

Green (1995) introduced another model indicator sampler method. This method is called Reversible Jump Markov Chain Monte Carlo (RJMCMC). It was implemented and improved later (Dellaportas and Forster, 1999; Robert and Casella, 1999). RJMCMC provides a procedure to generate samples that travel across different model parameter subspaces. So, it can handle the model selection problem between rival models with different parameter space dimensions. In each cycle, $npar_k$ parameters are sampled where k is the index of the current subspace and $npar_k$ is the dimension of that subspace (model). Thus the number of variables in all samples is not fixed. RJMCMC creates different types of moves between the subspaces

by random choice between available moves at each transition, in order to jump freely across the combined parameter space.

To illustrate this method, let the current state of the Markov chain, at time t, be (m^t, θ_m^t); where m indicates the model or subspace index. If ρ represents a particular proposed move to model m' and corresponding parameter vector $\theta'_{m'}$, a vector u of dimensionality equal to the difference in dimensionalities between $\theta_m^{(t)}, \theta'_{m'}$ is used to generate $\theta'_{m'}$ by a one-to-one function g as :

$$\theta'_{m'} = g(\theta_m^{(t)}, u) \tag{3.12}$$

Vector u is sampled from a proposal distribution $q_\rho(u)$. On the other hand, if the inverse of the move ρ is used, and $\theta_m^{(t)}$ is the vector of parameters with higher dimension, then $\theta'_{m'}$ is created from $\theta_m^{(t)}$ by applying the inverse transformation $(\theta'_{m'}, u') = g^{-1}(\theta_m^{(t)})$ and discarding u'.

Let the probability of making a move of type ρ, given the current state of the Markov chain $(m^{(t)}, \theta_m^{(t)})$, be $j(\rho, m^{(t)}, \theta_m^{(t)})$, thus $\sum_\rho j(\rho, m^{(t)}, \theta_m^{(t)}) \leq 1$ and with probability $1 - \sum_\rho j(\rho, m^{(t)}, \theta_{m^{(t)}}(t))$, no change to the present state is proposed. For moves that are not available from the starting point, $j\left(\rho, m^{(t)}, \theta_m^{(t)}\right)$ equals to zero.

Jumps considered in this method are classified as increasing and decreasing dimensionality. The probability of accepting a move with each of these jumps is as followings:

- Increasing dimensionality from $\theta_m^{(t)}$ to $\theta_{m'}^{'}$, generated via the appropriate $q_p(u)$, should be accepted as the next realization of the chain, so that $m^{(t+1)} = m'$ with probability $\alpha\left(m^{(t)}, \theta_m^{(t)}, m', \theta_{m'}^{'}\right)$, where

$$\alpha\left(m^{(t)}, \theta_m^{(t)}, m', \theta_{m'}^{'}\right) = \min\left\{1, \frac{p\left(\mathbf{y}\middle|m', \theta_{m'}^{'}\right)\ \pi\left(\theta_{m'}^{'}\middle|m'\right)\ \pi(m')\ j(p, m', \theta_{m'}^{'})}{p\left(\mathbf{y}\middle|m^{(t)}, \theta_m^{(t)}\right)\ \pi\left(\theta_{m^{(t)}}^{(t)}\middle|m^{(t)}\right)\ \pi(m^{(t)})\ j\left(p, m^{(t)}, \theta_m^{(t)}\right)}\right\} \quad (3.13)$$

And rejected otherwise, so that $m^{(t+1)} = m^{(t)}$

- Corresponding accepting probability of downward move from $\theta_m^{(t)}$ to $\theta_{m'}^{'}$ is shown in equation (3.14).

$$\alpha\left(m^{(t)}, \theta_m^{(t)}, m', \theta_{m'}^{'}\right) = \min\left\{1, \frac{P\left(\mathbf{y}\middle|m', \theta_{m'}^{'}\right)\ \pi\left(\theta_{m'}^{'}\middle|m'\right)\ \pi(m')\ j\left(\rho, m', \theta_{m'}^{'}\right)\ q_p\left(\theta_{m^{(t)}\setminus m'}^{(t)}\right)}{P\left(\mathbf{y}\middle|m^{(t)}, \theta_{m^{(t)}}^{(t)}\right)\ \pi\left(\theta_{m^{(t)}}^{(t)}\middle|m^{(t)}\right)\ (m^{(t)})\ j\left(\rho, m^{(t)}, \theta_{m^{(t)}}^{(t)}\right)}\right\} \quad (3.14)$$

Both Carlin and Polson's model (Carlin and Polson, 1991) indicator and reversible jump Markov Chain Monte Carlo methods are based on avoiding estimation of the marginal likelihood for each model by sampling from the model indicator. On the other hand, they both suffer from the disadvantage of depending upon selecting proper prior and pseudo-priors in the model indicator and transition functions in the RJMCMC method.

3.6.2 Marginal Likelihood Methods

The most straightforward approach in calculating the likelihood of models is the estimation of its integral directly. MCMC techniques can be used for the purpose of evaluating the marginal likelihood. The basic idea of MCMC integration (Metropolis et al., 1953) is that $\int_x f(x)\, q(x)dx$ can be approximated as closely as desired by:

$$\int f(x)q(x)dx \approx \lim_{s\to\infty} \frac{1}{s}\sum_{k=1}^{s} f(X_k) \tag{3.15}$$

if the X_i are sampled so that $Pr(X_k = x) = q(x)$. Likewise, in situations where samples are generated from another distribution, $q^*(x_k)$, instead of $q(x)$, the integral may be estimated using equation (3.16):

$$\int_x f(x)q(x)d(x) \approx \frac{\sum_{k=1}^{m} w_k f(X_k)}{\sum_{k=1}^{m} w_k} \tag{3.16}$$

where $w_k = q(x)/q^*(x_k)$.

Comparing equation (3.15) and the marginal likelihood (equation (3.4)), it is concluded that the marginal likelihood can be approximated by the expected value of likelihood of samples from the prior distribution. A problem arises when the posterior distribution and the prior distribution have totally different peaks relative to each other. In this situation, most of the sampled parameters will have small likelihood values and the sampling process will be inefficient.

Because of the problem in estimation of the marginal likelihood by prior samples, using samples from the posterior distribution seems more reasonable. Let us assume $\left\{\theta_k^{(g)}\right\} = \left\{\theta_k^{(1)}, \dots, \theta_k^{(G)}\right\}$ to be G draws from the posterior density of parameters $P(\theta_k|M_k, \mathbf{y})$. Then the marginal likelihood can be calculated by equation (3.17), (Newton and Raftery, 1994). But this

summation may become unstable when the inverse likelihood has infinite variance. In other words, a parameter with small likelihood can have a large effect on the approximation. Therefore, the variance of the estimated marginal likelihood may be large (Chib and Greenberg, 1995).

$$L(M_k|\mathbf{y}) = \left\{\frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{l(\theta_k^{(g)}|M_k,\mathbf{y})}\right)\right\}^{-1} \tag{3.17}$$

Chib (1995) proposed a marginal likelihood estimation method by sampling from the posterior distribution using the Gibbs sampler. This method solves the problem of instability in estimation of the marginal likelihood using posterior samples. Later Chib and Jeliazkov (2001) extended this method by using Metropolis-Hastings sampling. In their latest paper (Chib and Jeliazkov, 2005) the Acceptance Rejection Metropolis Hastings algorithm is applied. Chib and Jeliazkov's method with a Metropolis-Hastings sampling is used in the model selection step of two implementations of SBMCMD in Chapter 4; thus, it is explained in more detail in the next section.

3.7 Chib's Model Selection Methods

Chib (1995) proposed using the logarithm of the likelihood to improve stability. He suggested using equation (3.18) to calculate the Bayes factor for any two models i and j.

$$B_{ij} = \exp\{\ln P(\mathbf{y}|M_i) - \ln P(\mathbf{y}|M_j)\} \tag{3.18}$$

The marginal density function may be written as:

$$P(M_i|\mathbf{y}) = \frac{l(\theta^*|M_i,\mathbf{y})\pi(\theta^*)}{P(\theta^*|M_i,\mathbf{y})} \tag{3.19}$$

Taking the logarithm of both sides of equation (3.19) leads to equation (3.20).

$$\ln P(M_i|\mathbf{y}) = \ln l(\theta^*|M_i, \mathbf{y}) + \ln \pi(\theta^*) - \ln P(\theta^*|M_i, \mathbf{y}) \qquad (3.20)$$

This equation is true for any possible value of the parameters θ^*. Substituting $\ln P(M_i|\mathbf{y})$ from equation (3.20) into equation (3.18), yields the Bayes' factor. The only remaining problem in the calculation of $\ln P(M_i|\mathbf{y})$, is the estimation of the last term, $\ln P(\theta^*|M_i, \mathbf{y})$, which is the posterior density of possible values, θ^*.

What has been described above is the main idea behind Chib's method. Chib and his research group proposed three different versions of this method later. These methods differ in their MCMC methods for sampling from the posterior distribution and the estimation of $P(\theta^*|M_i, \mathbf{y})$. In the remaining of this section, these three versions of Chib's method are discussed.

3.7.1 Chib's Method and Gibbs Sampling

First, Chib used Gibbs sampling to estimate $\ln P(\theta^*|\mathbf{y})$, (Chib, 1995). To explain this method, assume that parameters are divided to B blocks, $\theta = (\theta_1, \dots, \theta_B)$. Thus the posterior probability of any point, θ^* may be given by equation (3.21).

$$P(\theta^*|\mathbf{y}) = P(\theta_1^*|\mathbf{y}) \times P(\theta_2^*|\mathbf{y}, \theta_1^*) \times \dots \times P(\theta_r^*|\mathbf{y}, \theta_1^*, \dots, \theta_{r-1}^*) \times \dots \times P(\theta_B^*|\mathbf{y}, \theta_1^*, \dots, \theta_{B-1}^*) = \prod_{i=1}^{B} P(\theta_i^*|\mathbf{y}, \theta_1^*, \dots, \theta_{i-1}^*) \,. \qquad (3.21)$$

The $P(\theta_1^*|\mathbf{y})$ term, marginal ordinate, is estimated from the samples generated by Gibbs method from $P(\theta|\mathbf{y})$. $P(\theta_r^*|\mathbf{y}, \theta_1^*, \dots, \theta_{r-1}^*)$ is defined by the integral shown in equation (3.22), and can be estimated from samples of the full conditional densities when θ_s is set equal to θ_s^* $(s \leq r-1)$. Let the G draws from the reduced full conditional densities be shown as $\left\{\theta_r^{(j)}, \theta_{r+1}^{(j)}, \dots, \theta_B^{(j)}\right\}$, then an estimated value of $P(\theta_r^*|\mathbf{y}, \theta_1^*, \theta_2^*, \dots, \theta_{r-1}^*)$ can be obtained from equation (3.23).

$$P(\theta_r^*|\mathbf{y}, \theta_1^*, \theta_2^*, \dots, \theta_{r-1}^*)$$

$$= \int P\left(\theta_r^*\middle|\mathbf{y}, \theta_1^*, \theta_2^*, \dots, \theta_{r-1}^*, \theta_l(l > r)\right) dP(\theta_{r+1}, \dots, \theta_B|\mathbf{y}, \theta_1^*, \theta_2^*, \dots, \theta_{r-1}^*) \tag{3.22}$$

$$P\left(\theta_r^*\middle|\mathbf{y}, \theta_s^*(s < r)\right)$$

$$= G^{-1} \sum_{j=1}^{G} p(\theta_r^*|\mathbf{y}, \theta_1^*, \theta_2^*, \dots, \theta_{r-1}^*, \theta_l^j(l > r)) \tag{3.23}$$

So the log of the marginal likelihood is given by the following equation:

$$\ln P(M_i|\mathbf{y}) = \ln f(\mathbf{y}|M_i, \theta^*)$$

$$+ \ln \pi(\theta^*|M_i) - \sum_{r=1}^{B} \ln P(\theta_r^*|\mathbf{y}, \theta_s^*(s < r)) \tag{3.24}$$

Although that selection of θ^* is not critical, it is more efficient to select it from a high density area. Thus it may be selected, after drawing G samples from θ, as the θ with the highest likelihood during sampling.

3.7.2 Chib's Method and Metropolis-Hastings Sampling

Later, Chib and Jeliazkov (2001) used a Metropolis-Hastings sampling procedure to estimate the marginal likelihood on the log scale. If ψ_{i-1}, ψ_{i+1} represent $(\theta_1, \dots, \theta_{i-1}), (\theta_{i+1}, \dots, \theta_B)$ respectively, the transition function of the MH method can be written as:

$$P\left(\theta_i, \theta_i'\middle|\mathbf{y}, \psi_{i-1}, \psi_{i+1}\right)$$

$$= \alpha\left(\theta_i, \theta_i'\middle|\mathbf{y}, \psi_{i-1}, \psi_{i+1}\right) \times q\left(\theta_i, \theta_i'\middle|\mathbf{y}, \psi_{i-1}, \psi_{i+1}\right) \tag{3.25}$$

where α is the acceptance probability and q is the proposal distribution in the MH method. As mentioned in Chapter 2, sub-kernels in the MH algorithm have the reversibility property. Reversibility of sub-kernels in this case is shown by equation (3.26).

$$P(\theta_i, \theta_i^* | \mathbf{y}, \psi_{i-1}, \psi_{i+1}^*)\, \pi(\theta_i | \mathbf{y}, \psi_{i-1}, \psi_{i+1}^*) = \pi(\theta_i^* | \mathbf{y}, \psi_{i-1}, \psi_{i+1}^*) P(\theta_i^*, \theta_i | \mathbf{y}, \psi_{i-1}, \psi_{i+1}) = \quad (3.26)$$

After integrating both sides of equation (3.26) with respect to θ and using an estimation strategy, equation (3.27) is obtained.

$$P(\theta_i^* | \mathbf{y}, \theta_1^*, \dots, \theta_{i-1}^*) = \frac{E_1\{\alpha(\theta_i, \theta_i^* | \mathbf{y}, \psi_{i-1}^*, \psi_{i+1}) q(\theta_i, \theta_i^* | \mathbf{y}, \psi_{i-1}^*, \psi_{i+1})\}}{E_2\{\alpha(\theta_i, \theta_i^* | \mathbf{y}, \psi_{i-1}^*, \psi_{i+1})\}} \quad (3.27)$$

where E_1 is the expectation with respect to $P(\theta_i, \Psi_{i+1} | y, \psi_{i-1}^*)$ and E_2 is the expectation with respect to $P(\psi_{i+1} | \mathbf{y}, \psi_i^*) q(\theta_i^*, \theta_i | \mathbf{y}, \psi_{i-1}^*, \psi_{i+1})$.

To estimate $p(\theta_i^* | y, \theta_1^*, \dots, \theta_{i-1}^*)$, first ψ_{i-1} should be set equal to ψ_{i-1}^*, and then samples from $(\theta_k | \mathbf{y}, \theta_{-k})$ $k = i, \dots, B$ are generated. These samples may be shown by

$\left\{\theta_i^{(g)}, \dots, \theta_B^{(g)}\right\}, g = 1, \dots, M$

Secondly, ψ_i^* is set equal to $(\psi_{i-1}^*, \theta_i^*)$, and then another set of samples is generated from the remaining distribution. Showing these samples by $\left\{\theta_{i+1}^{(j)}, \dots, \theta_B^{(j)}\right\}$ $j = 1, \dots, J$, $P(\theta_i^* | y, \theta_1^*, \dots, \theta_{i-1}^*)$ can be estimated from equation (3.28).

$$P(\theta_i^* | \mathbf{y}, \theta_1^*, \dots, \theta_{i-1}^*) = \frac{M^{-1} \sum_{g=1}^{M} \alpha\left(\theta_i^{(g)}, \theta_i^* | \mathbf{y}, \psi_{i-1}^*, \psi_{i+1,(g)}\right) q\left(\theta_i^{(g)}, \theta_i^* | \mathbf{y}, \psi_{i-1}^*, \psi_{i+1,(g)}\right)}{J^{-1} \sum_{j=1}^{J} \alpha\left(\theta_i^*, \theta_i^{(j)} | \mathbf{y}, \psi_{i-1}^*, \psi_{i+1,(j)}\right)} \quad (3.28)$$

After repeating the above procedure for all blocks, $P(\theta^*|M_i, \mathbf{y})$ is estimated from equation (3.21), and will be substituted in equation (3.20) for estimation of the marginal likelihood.

An specific condition of the above method is when all the parameters can be sampled simultaneously, thus there will be just one block. In this case equations (3.26) and (3.27) become equations (3.29) and (3.30), respectively:

$$P(\theta, \theta^*|M_i, \mathbf{y})\pi(\theta|M_i, \mathbf{y}) = \pi(\theta^*|M_i, \mathbf{y})P(\theta^*, \theta|M_i, \mathbf{y}) \quad (3.29)$$

$$P(\theta^* |M_i, \mathbf{y}) = \frac{E_1\{\alpha(\theta, \theta^*|\mathbf{y})q(\theta^*, \theta|\mathbf{y})\}}{E_2\{\alpha(\theta^*, \theta|\mathbf{y})\}} \quad (3.30)$$

where E_1 is the expectation with respect to $P(\theta|M_i, \mathbf{y})$ and E_2 is with respect to $q(\theta^*, \theta|\mathbf{y})$. Thus showing samples from the posterior distribution and $q(\theta^*, \theta|\mathbf{y})$ with $\{\theta^{(g)}\}$ and $\{\theta^{(j)}\}$, respectively, $P(\theta^*|\mathbf{y})$ is calculated by equation (3.31).

$$P(\theta^* |\mathbf{y}) = \frac{M^{-1}\sum_{g=1}^{M}\alpha(\theta^{(g)}, \theta^*|\mathbf{y})q(\theta^{(g)}, \theta^*|\mathbf{y})}{J^{-1}\sum_{j=1}^{J}\alpha(\theta^*, \theta^{(j)}|\mathbf{y})} \quad (3.31)$$

3.7.3 Chib's Method and Acceptance-Rejection MCMC Sampling

Chib and Jeliazkov (2005) presented the implementation of the Acceptance Rejection Metropolis Hastings method (ARMH) for one-block and multi-block sampling. One-block sampling has been used in our framework, thus it will be explained here. For models which may require multi-block sampling, one should refer to Chib and Jeliazkov (2001).

Like other versions of Chib's method, the problem of estimating the marginal likelihood is reduced to the estimation of the posterior probability of a specific value $P(\theta_i^*|M_i, \mathbf{y})$ given the posterior samples $\{\theta^{(1)}, \ldots, \theta^m\}$.

If $\alpha(\theta,\theta'|y)$ represents the acceptance probability of a move from the current sample θ to θ', and $g(\theta,\theta'|\mathbf{y})$ represents the proposal distribution for the move, then the probability of choosing a value θ' and accepting it, is given by the sub-kernel defined by equation (3.32):

$$Q(\theta,\theta'|\mathbf{y}) = \alpha(\theta,\theta'|\mathbf{y})q(\theta,\theta'|\mathbf{y}). \tag{3.32}$$

Regarding the reversibility property of sub-kernels the following equation is valid:

$$Q(\theta,\theta'|\mathbf{y})P(\theta|M_i,\mathbf{y}) = P(\theta'|M_i,\mathbf{y})Q(\theta',\theta|\mathbf{y})\text{..} \tag{3.33}$$

Integrating both sides and setting $\theta' = \theta^*$ yields:

$$P(\theta^*|M_i,\mathbf{y}) = \frac{\int \alpha(\theta,\theta^*|\mathbf{y})q(\theta,\theta^*|\mathbf{y})P(\theta|M_i,\mathbf{y})d\theta}{\int \alpha(\theta^*,\theta|\mathbf{y})q(\theta^*,\theta|\mathbf{y})d\theta} \tag{3.34}$$

By letting $\theta^* \in D$, $\alpha_{MH}(\theta^*,\theta|\mathbf{y}) = 1$ and $l(\theta^*|M_i,\mathbf{y})\pi(\theta^*|M_i,\mathbf{y}) \leq cq(\theta^*)$, where q is the proposal distribution in ARMH method. Thus, equation (3.34) becomes,

$$P(\theta^*|M_i,\mathbf{y}) = \frac{l(\theta^*|M_i,\mathbf{y})\pi(\theta^*|M_i,\mathbf{y})\int \alpha_{MH}(\theta,\theta^*|\mathbf{y})P(\theta|M_i,\mathbf{y})d\theta}{c\int \alpha_{AR}(\theta|\mathbf{y})q(\theta)d\theta} \tag{3.35}$$

Substituting the above equation into equation (3.19) leads to

$$L(M_i|y) = \frac{c\int \alpha_{AR}(\theta|\mathbf{y})q(\theta)d\theta}{\int \alpha_{MH}(\theta|\mathbf{y})P(\theta|M_i,\mathbf{y})d\theta} \tag{3.36}$$

A numerical estimate of $L(M_i|\mathbf{y})$ could be obtained by equation (3.37).

$$L(M_i|y) = c\,\frac{J^{-1}\sum_{j=1}^{J}\alpha_{AR}(\theta^j|y)}{G^{-1}\sum_{g=1}^{G}\alpha_{MH}(\theta^g,\theta^*|y)} \tag{3.37}$$

.

where θ^j are J samples from $q(\theta)$ and θ^g are G samples from $P(\theta|M_i,\mathbf{y})$.

3.8 Design of Experiments

The term design of experiments (DOE) is more commonly used in regression, where the goal is to design experiments, which lead to more accurate parameter estimates. In model discrimination however, the goal is to find experimental conditions which contain information with which to discriminate between rival models. Burke et al. (1997) classified statistical model discrimination experimental design criteria into maximum divergence-based and maximum entropy-based methods. These two categories seem to be able to cover all criteria which have been proposed in the literature.

3.8.1 Methods Based on Maximum Divergence

Maximum divergence-based methods select experimental conditions where the expected difference between the predicted values of the rival models is maximized. These criteria are based on the fact that the "best" model should be able to predict the model responses more accurately than other rival models. Hence, it is desirable to do the next trial under conditions that lead to the maximum average difference in model outputs. To achieve this, the following criterion is often used:

$$\mathrm{x_{next}} = \max_{\mathrm{x}} \sum_{\mathrm{i=1}}^{\mathrm{K-1}} \sum_{\mathrm{j=i+1}}^{\mathrm{K}} \left(\hat{\mathrm{y}}_{\mathrm{i}}(\mathrm{x}) - \hat{\mathrm{y}}_{\mathrm{j}}(\mathrm{x})\right)^2 \qquad (3.38)$$

where K shows the number of rival models and $\hat{y}_i(x)$ represents the predicted value from the i^{th} model under condition x. A better criterion is obtained if the prediction variance is taken into account. In this way experimental conditions where the prediction variance is large and potentially little information for model discrimination can be gained, can be avoided. So the modified version of this criterion may be written as equation (3.39).

$$x_{next} = \max_{x} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{\left(\hat{y}_i(x) - \hat{y}_j(x)\right)^2}{\mathrm{var}\left(\hat{y}_i(x) - \hat{y}_j(x)\right)} \tag{3.39}$$

Roth (1965) introduced a weighted average of the total separation between the models where weights are the Bayesian posterior probabilities. Let $P(M_k|\mathbf{y_{n-1}})$ denote the posterior Bayesian probability of the k^{th} model after n-1 experiments and $\hat{y}_j(x)$ represent the predicted value of model j, under experimental conditions x. The conditions which maximize equation (3.40) are chosen as the experimental conditions for the n^{th} trial.

$$x_n = \arg\max_{x} \sum_{i=1}^{K} \left[P(M_i|\mathbf{y_{n-1}}) \prod_{\substack{j=1 \\ j\neq i}}^{K} \left|\hat{y}_j(x) - \hat{y}_i(x)\right| \right] \tag{3.40}$$

In its original implementation proposed by Roth the predicted values, $\hat{y}_i(x)$, are calculated using the parameter point estimates obtained based on the n-1 experimental trials. In our work, we refine this method by calculating the predicted values by integrating over all possible parameter values as shown in equation (3.41). This makes the estimates of the predicted values independent of the parameter values.

$$\hat{y}_i(x) = \int_{\theta} y_i(x|\theta)\pi(\theta)d\theta \tag{3.41}$$

Using the *G* samples obtained from the current prior distribution, $\hat{y}_i(x)$ is estimated by equation (3.42)

$$\hat{y}_i(x) = \int_{\theta} y_i(x|\theta)\pi(\theta)d\theta \approx \frac{1}{G}\sum_{g=1}^{G} y_i\left(x\middle|\theta_g\right) \tag{3.42}$$

3.8.2 Methods Based on Maximum Entropy

Entropy-based design criteria provide another alternative to the maximum divergence ones. The entropy definition in experiment selection is related to the concept of entropy in thermodynamics, which is a measure of the randomness or uncertainty in a system. Suppose there are K models, after n-1 experimental trials, the entropy of the system can be defined as equation (3.43):

$$S = -\sum_{i=1}^{K} P(M_i|\mathbf{y_{n-1}}) \ln P(M_i|\mathbf{y_{n-1}}) \tag{3.43}$$

Box and Hill (1967) borrowed the concept of entropy from information theory in experimental design. They assumed that in the beginning, all the models are assumed equally probable, so having K rival models, each with probability equal to $1/K$. In this situation, the entropy will be at a maximum. On the other hand, in the desirable condition, when the probability of one model is 1 and others zero, the entropy value is at a minimum. So in order to obtain the most possible information from the next experiment we can try to maximize the difference between entropy in the n-1st and the nth step. This expected change in entropy is usually shown by R,

R = entropy at input – expected entropy at output

Box and Hill (1967) derived an upper bound on the change in entropy for linearized models under normally distributed errors. They showed that the expected change in entropy resulting from the nth trial made under certain conditions x is

$$R = \sum_{i=1}^{K} P(M_i|\mathbf{y_{n-1}}) \int_{y_n} P(M_i|\mathbf{y_n}) \ln\left[\frac{P(\mathbf{y}_n|M_i)}{q(x, y_n)}\right] dy_n \tag{3.44}$$

where,

$$q(x, y_n) = \sum_{i=1}^{K} P(M_i|\mathbf{y_{n-1}})p(y_n|M_i) \tag{3.45}$$

where $P(y_n|M_i)$ is the probability density function of the n^{th} observation under model i and $P(M_i|\mathbf{y_{n-1}})$ is the probability of model i up to n-1 experiments, so $q(x, y_n)$ is the probability of observing a specific observed value. The predicted posterior probability associated with the k^{th} model is shown in equation (3.46).

$$\widehat{P}(M_i|\mathbf{y_n}) = \frac{P(M_i|\mathbf{y_{n-1}})P(y_n|M_i)}{q(y_n)} \tag{3.46}$$

Then Box and Hill show by Bayesian analysis that if the errors are normally distributed, an upper bound for negative R, D(x) is defined in equation (3.47).

$$D(x) = \frac{1}{2}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} P(M_i|\mathbf{y_{n-1}})\, P(M_j|\mathbf{y_{n-1}}) \left\{ \frac{(\sigma_i^2 - \sigma_j^2)^2}{(\sigma^2 + \sigma_i^2)(\sigma^2 + \sigma_j^2)} + \left[\hat{y}_i(x) - \hat{y}_j(x)\right]^2 \left(\frac{1}{\sigma^2 + \sigma_i^2} + \frac{1}{\sigma^2 + \sigma_j^2}\right)\right\} \tag{3.47}$$

where σ^2 is the known error variance for a single measurement; $\hat{y}_i(x)$ stands for the predicted value of $y_{(x)}$ under model *i* using the first n-1 observations and variance of $\hat{y}_i(x)$ is given by σ_i^2. Therefore, the condition x yielding the largest D(x) will be selected as the next experimental condition.

Reilly (1970) introduced the expected entropy criterion which uses the actual value of the expected entropy as a design criterion instead of an upper bound on the expected entropy change. He presented the integral in equation (3.44) in the following format,

$$R = -\frac{1}{2}[1 + \ln 2\pi\,(\sigma^2 + \sigma_i^2)] - \int_{y_n} P(y_n|M_i)\ln q(x, y_n)dy_n \tag{3.48}$$

The remaining integral in the above equation can be recognized as the expression for the expectation of $\ln q(x, y_n)$ under variation of y_n assuming model i to be correct. This may be written as $\int_{y_n} \ln q(x, y_n)\, dy_n$. It follows that:

$$R(x) = -\frac{1}{2}(1 + \ln 2\pi) - \sum_{i=1}^{K} P(M_i|\mathbf{y_{n-1}}) \left\{ \frac{1}{2} \ln(\sigma^2 + \sigma_i^2) + \int_{y_n} \ln q(x, y_n) \right\} \quad (3.49)$$

All of this is easy to evaluate except for the expectation term. This is the expectation of the function $\ln q(x, y_n)$, where y_n is normally distributed with mean $\hat{y}_n(x)$ and variance $\sigma^2 + \sigma_i^2$. Reilly used Gauss-Hermite quadrature to estimate the expectation term.

Drovandi et al. (2013) in their recent paper used a DOE criterion, which contains the prediction of the posterior probability of models after doing the next experiment. Thus their DOE criterion is like other entropy criteria, which use the predicted values of the posterior probabilities.

3.9 Model Discrimination with T-test (Buzzi-Ferraris and Forzatti method)

Buzzi-Ferraris et al. (1983) described an experiment design criterion based on the maximum variances for linear models or those, which can be assumed approximately linear over the relevant region of the parameter space. Their method can be used for single response methods. After that, Buzzi-Ferraris and Forzatti (1984) modified this method to be applicable to multiple response models. In this method, errors, are assumed to be normally distributed. The next experimental condition is obtained by maximizing a proper indicator of the divergence among the responses relative to the limits of error provided by the data, given by

$$T = S_N^2 / S_D^2 \quad (3.50)$$

where S_N^2 indicates the variations between the deviations (δ_i) and S_D^2 is the mean value of the variances($\hat{\sigma}_{\delta_i}^2$), where δ_i is the deviation between the observation and the expected response from the i^{th} model. By assuming that either we have K rival linear models, or models which can be approximated by linear ones, they can be shown as the following equations:

$$y_i(x) = \hat{\theta}_1 g_{i,1}(x) + \hat{\theta}_2 g_{i,2}(x) + \cdots + \hat{\theta}_{p_i} g_{i,p_i}(x)$$

$$\hat{\sigma}_i^2(x) = g_i^T(x)\left(G_i^T G_i\right)^{-1} g_i(x)\,\hat{\sigma}^2$$

where for nonlinear models: $g_{ih} = \left.\frac{\partial f_i(x,\theta_i)}{\partial \theta_h}\right|_{\theta_i=\hat{\theta}_i^{(n)}}$

$$G_i(x) = \begin{vmatrix} g_{i,1}(x_1) & \cdot & \cdot & \cdot & \cdot & g_{i,p_i}(x_1) \\ g_{i,1}(x_2) & \cdot & \cdot & \cdot & \cdot & g_{i,p_i}(x_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ g_{i,1}(x_N) & \cdot & \cdot & \cdot & \cdot & g_{i,p_i}(x_N) \end{vmatrix} \quad g_i(x)$$

$$= [g_{i,1}(x), \ldots, g_{i,p_i}(x)] \tag{3.51}$$

where $\hat{\theta}_i^{(n)}$ is the point around which the model is linearized around. Using these assumptions, s_N, s_D and T are calculated by using equations (3.52), (3.53) and (3.54) respectively. The criterion requires choosing the vector of experimental conditions that maximizes T.

$$s_N^2 = \frac{\sum_{i=1}^{K}\left(\delta_i(x) - \overline{\delta(x)}\right)^2}{K-1} = \frac{\sum_{i=1}^{K}\left(\hat{y}_i(x) - \overline{\hat{y}(x)}\right)^2}{K-1}$$

$$= \frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\left(\hat{y}_i(x) - \hat{y}_j(x)\right)^2}{K(K-1)} \tag{3.52}$$

$$s_D^2 = \frac{\sum_{i=1}^{K} \nu_i (\hat{\sigma}^2 - \hat{\sigma}_i^2(x))^2}{\sum_{i=1}^{K} \nu_i}$$

$$= \hat{\sigma}^2 + \frac{\hat{\sigma}^2 \sum_{i=1}^{K} g_i^T(x) \left(G_i^T G_i\right)^{-1} g_i(x)}{K} \tag{3.53}$$

$$= \hat{\sigma}^2 + \sum_{i=1}^{K} \frac{\hat{\sigma}_i^2(x)}{K}$$

$$T = \frac{\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \left(\hat{y}_i(x) - \hat{y}_j(x)\right)^2}{(K-1)(K\hat{\sigma}^2 + \sum_{i=1}^{K} \hat{\sigma}_i^2(x))} \tag{3.54}$$

T values below 1 in single output models, indicate that the variance of the divergence among responses can be explained in terms of experimental error variance plus variance of expected response. So if there is no condition with T value above one, the procedure should be stopped. For multiple responses T was defined as equation (3.55):

$$T_{ij}(x) = \left(\hat{y}_i(x) - \hat{y}_j(x)\right)^T \left(S_{ij}(x)\right)^{-1} \left(\hat{y}_i(x) - \hat{y}_j(x)\right) \tag{3.55}$$

where $S_{ij}(x)$ is the covariance matrix of $\left(\delta_i(x) - \delta_j(x)\right)$. In order to discriminate among more than two rival models, it is desired to find the vector of independent variables x that maximize $T_{ij}(x)$ over any pair of models i and j, providing $T_{ij}(x)$ is greater than R (response dimension). The procedure is stopped when no setting of x maximizes $T_{ij}(x)$ above R.

Buzzi-Ferraris and Forzatti (1990) provided another usable design method for multi-response models which can be linearized in the neighborhood of $\hat{\theta}_i^{(n)}$. Let $y'(x_i), y''(x_i)$ stand for the $R \times 1$ vectors of observations recorded in two genuine independent replications of $x = x_i$ with normally distributed error as shown in equation (3.56).

$$y'(x_u) = f(x_u, \theta) + \varepsilon'_u$$

$$y''(x_u) = f(x_u, \theta) + \varepsilon''_u \tag{3.56}$$

$$\varepsilon'_u, \varepsilon''_u \sim N(0, \Sigma)$$

where Σ is the known positive definite constant covariance matrix shown in equation (3.57).

$$\Sigma = \begin{vmatrix} \sigma_{11} & \sigma_{11} & \vdots & \vdots & \vdots & \sigma_{11} \\ \sigma_{11} & \sigma_{11} & \vdots & \vdots & \vdots & \sigma_{11} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{11} & \sigma_{11} & \vdots & \vdots & \vdots & \sigma_{11} \end{vmatrix} = \{\sigma_{hl}\} \tag{3.57}$$

By using linearization the h^{th} output value from the i^{th} rival model may be shown as equation (3.58).

$$f_{ih}(x, \theta_i) = f_{ih}(x, \theta_i^{(n)}) + \sum_{r=1}^{p_i} \left[\theta_{ir} - \hat{\theta}_{ir}^{(n)}\right] \left.\frac{\partial f_{ih}(x, \theta_i)}{\partial \theta_{ir}}\right|_{\theta_i = \hat{\theta}_i^{(n)}} \tag{3.58}$$

$\hat{\theta}_{ir}^{(n)}$ denotes parameter estimates from fitting models using the observed data. In this condition, the estimated response under model i, $\hat{y}_i(x_{n+1})$, for the new design point is normally distributed about $f_i(x_{n+1}, \hat{\theta}_i^{(n)})$ with a covariance matrix equal to $W_i(x_{n+1})$:

$$\hat{y}_j(x_{n+1}) \approx N[f_i(x_{n+1}, \theta_j), W_i(x_{n+1})] \tag{3.59}$$

where Θ_i^{-1}, $G_i(x_{n+1})$ and $W_i(x_{n+1})$ are defined by equations (3.60) and (3.61), respectively.

$$\Theta_i^{-1} = \sum_{h=1}^{k} \sum_{l=1}^{k} \sigma_{hl} G_{ih}^T G_{il} \tag{3.60}$$

$$W_i(x_{n+1}) = G_i(x_{n+1}) \Theta_i^{-1} G_i^T(x_{n+1}) \tag{3.61}$$

Buzzi-Ferraris and Forzatti picked the setting of x_{n+1} which provides the maximum information against the null hypothesis that indicates $E\left\{f_i[x_u, \hat{\theta}_i^{(n)}]\right\} = E\{f_j[x_u, \hat{\theta}_j^{(n)}]\}$, in order

to discriminate most efficiently between two rival models. If we define δ_i , δ_j and δ_{ij} as equation (3.62);

$$\delta_i(x_{n+1}) = [\hat{y}_i^{(n)}(x_{n+1}) - y'(x_{n+1})]$$

$$\delta_j(x_{n+1}) = [\hat{y}_j^{(n)}(x_{n+1}) - y''(x_{n+1})] \qquad (3.62)$$

$$\delta_{ij}(x_{n+1}) = R_{ij}^{-1}(x_{n+1})[\delta_i(x_{n+1}) - \delta_j(x_{n+1})]$$

where R_{ij}^{-1} is a function of Σ, w_i, w_j. Then minimizing the PDF of $\delta_{ij}(x_{n+1})$ over x_{n+1} leads to selecting the next experimental condition that satisfies $E\left\{f_i[x_u, \hat{\theta}_i^{(n)}]\right\} \neq E\{f_j[x_u, \hat{\theta}_j^{(n)}]\}$.

3.10 Hsiang Reilly Method (HR)

In the HR method, the parameters in each of the candidate models are handled separately as a set of discrete arrays in computer storage to facilitate the calculation of the marginal likelihood in this Bayesian approach. Hsiang and Reilly's method is a complete sequential model discrimination framework, including both model selection and the DOE step. The design of experiments step is designed specifically for the discrete space in their framework. The steps in this method will be explained in the following.

Initialization

First, it is required to set the prior probability for each model π_k $k = 1,2,\dots,K$ and $\sum_{k=1}^{K} \pi_k = 1$, where k represents the index of a model and K is the total number of candidate models. Then a discrete set of possible values should be assigned for each parameter. Assuming that the k^{th} model has P_k parameters and discretizing each parameter space to t_{ij} values where $j = 1,\dots,P_k$ $i = 1,\dots,K$, there are $\prod_{j=1}^{p_k} t_{ij}$ possible sets of parameter values in the k^{th} model. After defining the discretized space, a distribution of probabilities

should be considered. For each possible parameter value a probability is set that represents our belief that it is the "true" set of parameter values. To illustrate this method, assume that we have two rival models, each with 2 parameters. In the beginning we may start with non-informative model priors, $P(M_i = 1) = 0.5$, $P(M_i = 2) = 0.5$

Assuming the acceptable range of parameters, $1 \leq P_{11} \leq 2.5, 100 \leq P_{21} \leq 180,\ 10 \leq P_{21} \leq 20,\ 1e5 \leq P_{22} \leq 6e5$, Figure 3-2 and Figure 3-3 show the discretized space and the prior values, respectively.

Parameter 2 :	Parameter 1 : 1	1.5	2	2.5
100	0.05	0.05	0.05	0.05
120	0.05	0.05	0.05	0.05
140	0.05	0.05	0.05	0.05
160	0.05	0.05	0.05	0.05
180	0.05	0.05	0.05	0.05

Figure 3-2: Model 1 priors for the HR example

Parameter 2 :	Parameter 1 : 10	20
1e5	0.08	0.08
2e5	0.08	0.08
3e5	0.08	0.08
4e5	0.08	0.08
5e5	0.08	0.08
6e5	0.08	0.08

Figure 3-3: Model 2 priors for the HR example

Calculation of Model Posterior

When some preliminary measurement from the system are available or after designing and carrying out a new experiment, the likelihood of each cell of the parameter tables will be calculated. The likelihood is a function of the model error variance and the error value, which is the difference between the real system output and the output of the model given the parameters value in the cell and the input condition. Then the prior of that set of parameters multiplied by the likelihood will be put in the table as the posterior value. Figure 3-4 and Figure 3-5 show the posterior probabilities for the illustrating example.

Parameter 2 :	Parameter 1 : 1	1.5	2	2.5
100	0.0001	0.1	0.1	0.05
120	0.0000	0.04	0.05	0.1
140	0.0002	0.3	0.15	0.1
160	0.0003	0.0000	0.0001	0.0000
180	0.0001	0.0002	0.0000	0.0001

Figure 3-4: HR method, model 1 posterior probabilities

Parameter 2 :	Parameter 1 : 10	20
1e5	0.1	0.11
2e5	0.08	0.08
3e5	0.12	0.13
4e5	0.1	0.08
5e5	0.02	0.08
6e5	0.01	0.09

Figure 3-5: HR method, model 2 posterior probabilities

Rescaling is another step in the HR method implemented by Burke (1994). If all the probabilities in a column or a row in the border of the table are insignificant, that part of the table is deleted in the rescaling step. Then, the table will be filled again with the prior values, following by calculation of the likelihood for the new set of parameters. Figure 3-6 shows the rescaled table of parameters in model 1, filled with prior probability values. One column and two rows have been deleted from the table presented in Figure 3-4 to obtain the table in Figure 3-6.

Parameter 2 :	Parameter 1 : 1.5	1.85	2.15	2.5
100	0.05	0.05	0.05	0.05
110	0.05	0.05	0.05	0.05
120	0.05	0.05	0.05	0.05
130	0.05	0.05	0.05	0.05
140	0.05	0.05	0.05	0.05

Figure 3-6: HR method, model 1, rescaled table

Updating Model Probabilities

In this step, the posterior probability of each model is updated using the summation of the prior multiplied by the likelihood over all possible discrete value of parameters, $\sum_{\theta_{j,i}} P(\theta_{j,i}|M_i)\, l(\theta_{j,i}|\mathbf{y_n})$. Equation (3.63) shows the formula for updating the models posterior probability,

$$P(M_i|\mathbf{y}) = \frac{P(M_i)\sum_{\theta_{j,i}} P(\theta_{j,i}|M_i)\, l(\theta_{j,i}|M_i, y_n)}{\sum_{i=1}^{K} P(M_i)\sum_{\theta_{j,i}} P(\theta_{j,i}|M_i) l(\theta_{j,i}|M_i, y_n)} \tag{3.63}$$

where $P(M_i)$ shows the prior of model M_i in the current step and $P(\theta_{j,i}|M_i)$ the prior of $\theta_{j,i}$ in model M_i. Then the tables of parameters are updated by normalized probability values, equation (3.64).

$$P(\theta_i|M_i, \mathbf{y}) = \frac{P(\theta_{j,i}|M_i) l(\theta_{j,i}|M_i, y_n)}{\sum_{\theta_{j,i}} P(\theta_{j,i}|M_i) l(\theta_{j,i}|M_i, y_n)} \tag{3.64}$$

Design of Experiments

The design of experiment criterion used by Hsiang and Reilly is based on the Roth criterion, equation (3.40). The predicted response at the n+1th condition, $\hat{y}_i$, could be easily estimated by weighted summation of predicted values over the table of parameters.

$$\hat{y}_i\,(x) = \sum_{\theta_{j,i}} f_i(\theta_{j,i}, x)\; P(\theta_{j,i}|M_i, \mathbf{y}) \tag{3.65}$$

where $\theta_{j,i}$ shows parameter values and summation is over all cells. Weights are posterior probability of parameters and f_i denotes the ith model output. This procedure is repeated until one model is identified with sufficient certainty.

Chapter 4
Proposed Model Discrimination Method

4.1 Introduction

Assume A and B are two events with probabilities of occurrence equal to $\Pr(A)$ and $\Pr(B) \neq 0$, respectively. In addition, let $\Pr(A|B)$ denote the conditional probability of A given that B has occurred and define $\Pr(B|A)$ analogously. Then Bayes' theorem states that:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \tag{4.1}$$

In other words, for any two random events A and B, the conditional distribution of A given B is proportional to the product of the conditional probability of B given A and the probability of A. The proof of equation (4.1) can be found in Bard (1974).

To illustrate Bayes' theorem to applications in parameter estimation and modeling, suppose that $\mathbf{y_n} = (y_1, \ldots, y_n)$ is a vector of n observations from the system whose probability distribution, $P(\mathbf{y_n}|\theta)$, depends on the values of parameters, θ. Also assume that θ has a probability distribution $P(\theta)$. Then by using equation (4.1), equation (4.2) is obtained.

$$P(\mathbf{y_n}|\theta)\, P(\theta) = P(\mathbf{y_n}, \theta) = P(\theta|\mathbf{y_n})\, P(\mathbf{y_n}) \tag{4.2}$$

Given the observed data $\mathbf{y_n}$, the conditional distribution of θ may be written as equation (4.3).

$$P(\theta|\mathbf{y_n}) = \frac{P(\mathbf{y_n}|\theta)P(\theta)}{P(\mathbf{y_n})} \tag{4.3}$$

By considering $P(\mathbf{y_n}) = \int_\theta P(\mathbf{y_n}|\theta) = c^{-1}$, $P(\theta|\mathbf{y_n})$ may be written as the following equation:

$$P(\theta|\mathbf{y_n}) = c.\, P(\mathbf{y_n}|\theta)\pi(\theta). \tag{4.4}$$

The above equation is referred to as Bayes' theorem, where $\pi(\theta)$, which is equal to $P(\theta)$ in equation (4.3), is called the prior distribution and contains previous knowledge about θ without considering the current data $\mathbf{y_n}$, $P(\mathbf{y_n}|\theta)$ is the probability of the observed data given θ, and the variable c is a normalizing constant necessarily to ensure that the posterior distribution $P(\theta|\mathbf{y_n})$ integrates or sums to one. Bayes' theorem may be written as equation (4.5) (Box and Tiao, 1992).

$$P(\theta|\mathbf{y}) \propto \pi(\theta)l(\theta|\mathbf{y}) \tag{4.5}$$

where $l(\theta|\mathbf{y})$, which equals $P(\mathbf{y_n}|\theta)$, is called the likelihood of θ given the data and $P(\theta|\mathbf{y})$ is the posterior probability. The likelihood function plays a very important role in Bayes' theorem. It is the function through which the data modifies the prior knowledge. Thus, it can be regarded as representing the information about θ coming from the data. Considering the problem of model discrimination and assuming that one of the K candidate models is the true model description of the real system, then equation (4.5) can be cast into a form suitable for model discrimination by replacing the parameter θ, by the model index M_k, where the model index M_k represents the k^{th} rival model. Hence, Bayes' theorem could be re-written as equation (4.6) for any of the k candidate models:

$$P(M_k|\mathbf{y}) = \pi(M_k)L(M_k|\mathbf{y}) \tag{4.6}$$

In the above equation, $P(M_k|\mathbf{y})$ is the normalized posterior probability. The likelihood represented by $L(M_k|\mathbf{y})$ is the likelihood of the model given the data and is usually calculated by integrating out the model parameters as will be discussed later. $\pi(M_k)$ is the prior probability of the model M_k. In the initialization step, prior probabilities which satisfy equation (4.7) should be assigned to candidate models.

$$\sum_{k=1}^{K} \pi(M_k) = 1 \tag{4.7}$$

An important feature of the Bayesian approach is that it allows prior knowledge to be incorporated in the analysis. In addition, the Bayesian approach lends itself for use in sequential statistical procedures and on-line procedures when new information becomes available gradually. Assume each observed point is independent of the previous ones and introduced in a separate step of the sequential procedure. In this case, equation (4.4) may be written as,

$$P_n(\theta|\mathbf{y_n}) = c\prod_{i=1}^{n} P(y_i|\theta)\,\pi(\theta) = c.\underbrace{\prod_{i=1}^{n-1} P(y_i|\theta)\ \pi(\theta)}_{P(\theta|\mathbf{y_{n-1}})}\ P(y_n|\theta) \tag{4.8}$$

$$P_n(\theta|\mathbf{y_n}) = P(\theta|\mathbf{y_{n-1}})\,P(y_n|\theta)$$

where n is the number of observed data points. Equation (4.8) implies that the prior probabilities in each step can be replaced by the posterior probabilities from the previous step. In a sequential procedure, the probability of the parameter values is interpreted as the last updated probability of those values; thus, after the n^{th} observation, it is equal to the product of the (n-1) posterior probabilities, which is the prior probability in the current step, and the probability of the n^{th} observed data given θ.

Non-sequential model selection methods use equation (4.6) just once. By contrast, a sequential model discrimination method uses Bayes' formula repeatedly as new data becomes available until a stopping criterion is satisfied. This process is represented by the pseudo-code in Algorithm 4.1. In this algorithm, $\mathbf{y_n}$ represents the vector of all observed data collected up to and including experiment n and y_n is the observed data at step n.

Algorithm 4.1: A sequential model discrimination algorithm

Inputs:

K : Number of rival models

$N_{max\,experiment}$: Maximum number of experiments

$P_{selection\ criterion}$: Stop criterion probability

$\mathbf{y_0}$: Preliminary information

Algorithm:

Set rival model priors : $P_0(M_k|\mathbf{y_0})$

for n=1 to $N_{max\,experiment}$

  Design the next experiment inputs x_n

  Carry out the experiment to observe the new output data y_n

  for k=1 to K

    Calculate $L(M_k|y_n)$

    Update posterior probability $P_n(M_k|\mathbf{y_n}) = P_{n-1}(M_k|\mathbf{y_{n-1}})\ L(M_k|y_n)$

stop if $\exists\ k \in \{1, \dots, K\}: P_n(M_k|\mathbf{y_n}) > P_{selection\ criterion}$

In this chapter, first our modification to the Hsiang and Reilly's (HR) method (section 3.10) will be presented. After the HR method, the Sequential Bayesian Monte Carlo Model Discrimination (SBMCMD) framework and its three implementations are presented.

4.2 Modification of the Hsiang and Reilly Method

As explained before (section 3.10) in the HR method, the parameter spaces are discretized. Thus the marginal likelihood of each model can be estimated by equation (4.9).

$$L(M_k|\mathbf{y_n}) \sim \sum_{\theta_{j,k}} P(\theta_{j,k}|M_k)\, l(\theta_{j,k}|M_k, \mathbf{y_n})\ \Delta\theta_{j,k} \qquad (4.9)$$

where $\Delta\theta_i$ shows the parameter cell width in each discretized cell. When there are p_k parameters in model M_i, $\Delta\theta_i$ is calculated by, $\Delta\theta_k = \prod_{j=1}^{p_k} \Delta\theta_{k,j}$ where $\Delta\theta_{j,k}$ shows the parameter cell width of the j[th] discretized cell in the k[th] model. Therefore, the posterior probability of each model is directly proportional to:

$$P_{n-1}(M_k|\mathbf{y_n}) \sum_{\theta_{j,k}} P_{n-1}(\theta_{j,k}|M_k, \mathbf{y_n}) l(\theta_{j,k}|M_k, y_n) \Delta\theta_{j,k} \qquad (4.10)$$

Thus, the normalized posterior probability of the models is given by:

$$P(M_i|\mathbf{y}) = \frac{P_{n-1}(M_i|\mathbf{y_n}) \sum_{\theta_{j,k}} P_{n-1}(\theta_{j,k}|M_k, \mathbf{y_n}) l(\theta_{j,k}|M_k, y_n) \Delta\theta_{j,k}}{\sum_{i=1}^{K} P_{n-1}(M_i|\mathbf{y_n}) \sum_{\theta_{j,k}} P_{n-1}(\theta_{j,k}|M_k, \mathbf{y_n}) l(\theta_{j,k}|M_k, y_n) \Delta\theta_{j,} \qquad (4.11)$$

In situations that all candidate models have the same search area, like the example in Hsiang and Reilly (1971), equation (4.11) can be simplified to:

$$P(M_i|\mathbf{y}) = \frac{P_{n-1}(M_i|\mathbf{y_n}) \sum_{\theta_{j,k}} P_{n-1}(\theta_{j,k}|M_k, \mathbf{y_n}) l(\theta_{j,k}|M_k, y_n)}{\sum_{i=1}^{K} P_{n-1}(M_i|\mathbf{y_n}) \sum_{\theta_{j,k}} P_{n-1}(\theta_{j,k}|M_k, \mathbf{y_n}) l(\theta_{j,k}|M_k, y_n)}$$

The above equation was shown on page 57 as equation (3.63). The HR method with equation (4.11) instead of (3.63) is called the modified HR method in this research. The modified method is applicable to candidate models with different parameters search space.

In the copolymerization model discrimination case study, which was previously tested in our research group (Burke, 1994), there were two models with different numbers of parameters. Therefore equation (4.11) should be used for the copolymerization case study but equation (3.64) was used by Burke (1994). This seems to be the reason why the HR method shows poor result in her research. The modified HR method has been applied to the copolymer case study in this research. Results, which will be shown and discussed in Chapter 5 (section 5.3), prove

that the modified equation can address the question of the previously observed unsuccessful results of the HR method for this system.

Although this modification addresses some questions about the HR method, it is not efficient compared to the proposed SBMCMD method for use in more complicated case studies. The reason is that this method uses a discretized parameter space and therefore the accuracy is dependent upon the discretization used, while the Monte Carlo approach, in general, gives better coverage.

Like the HR method, the sequential MC model discrimination framework has the advantage that it is not restricted to linear models. In addition, the use of a discretized parameter space is replaced with sampling from the probability distribution of the parameters by Monte Carlo methods.

4.3 Sequential Bayesian Monte Carlo Model Discrimination (SBMCMD)

As mentioned in the introduction, a sequential Bayesian framework includes both model selection and the design of experiment (DOE) steps. Different implementations of the SBMCMD framework are similar in the design of experiments (DOE) and initialization steps, but they differ in the model selection part.

Three implementations of the framework are presented in this research to show that different Monte Carlo sampling methods can be used in the sampling step. Furthermore, these implementations are different in their marginal likelihood estimation step. The reason is that for different cases, using samples from the prior or posterior distributions may be more convenient or accurate.

During the course of this research, we first started with the posterior and prior based implementation, which is the third implementation described in this chapter. This implementation applies Acceptance-Rejection MH (ARMH) sampling (section 2.3.7). In the first iteration of this implementation, the marginal likelihoods are estimated by Chib & Jeliazkov's method (section 3.7.3), but as of the second iteration, the marginal likelihoods are estimated from the average of likelihoods based on the prior parameters samples as discussed in section 3.6.2. The reason that we started with this implementation was that the Chib & Jeliazkov's method with ARMH sampling is more efficient with respect to the calculation effort compared to the Chib and Jeliazkov's method with RWMH sampling (Chib and Jeliazkov, 2005). After testing more case studies, we came across case studies in which the ARMH sampling algorithm needed too much tuning effort to generate samples with acceptable acceptance ratios. Therefore, we tried the posterior based implementation, which is the second implementation described in this thesis, which uses samples from the posterior distribution of the parameters obtained by the MH method and the Chib and Jeliazkov's model selection method (section 3.7.2) with adaptive MH sampling. Later, we implemented the prior based implementation to be used in cases which require the use of advanced sampling methods to sample efficiently from the parameter distributions. Details of these three implementations will be explained in the remainder of this chapter. After describing the details of these three implementations, they will be compared in section 4.4.

4.3.1 First Implementation of SBMCMD: Prior Based

The advantage of this implementation is that its model selection analysis is not dependent upon the sampling procedure. Thus any MCMC sampling procedure, which is

efficient and covers the whole parameter space can be applied. The disadvantage of this implementation is that the user should decide if there is enough information for updating the model probabilities or if the model updating should be skipped until more experimental data from the real system is obtained.

When preliminary data are available from the system and these data are used to build the initial prior distribution, it is possible to use samples from the parameter prior distributions to estimate the marginal likelihood. In this situation, the criticism that the prior parameter probability distribution may have a completely different shape compared to the shape of the posterior probability of the parameters is not valid any more as discussed in section 3.6.2. This is also true for the second and later iterations, since the prior distribution, being the posterior distribution from the previous iteration, is informed by the data.

Therefore the marginal likelihoods are estimated according to equation (3.15). Considering J samples from the prior distribution $\{\theta_j\}$, which are the saved posterior samples from the last iteration, the marginal likelihood estimation can be rewritten as

$$\hat{L}(M_i|y) \approx \frac{1}{m} \sum_{j=1}^{J} l(\theta_j|M_i, y_i) \tag{4.12}$$

Figure 4-1 shows the SBMCMD implementation with samples from the parameter prior distributions. Steps of this implementation are explained in more detail in the following and summarized in algorithm 4.2, page 69.

Initialization

Before starting model selection on these data, the proper prior distribution of the parameters must be defined. The selected priors must be proper, since improper priors, whose integral over the parameter space is not equal to one, affect the marginal likelihood by a multiplier and can mislead the model selection algorithm. In addition to parameter priors, the prior of the models should be assigned. If there is no prior information available, the models are assumed equally probable, and hence have equal prior probabilities.

Whenever experimental data are already available, they could be used as the preliminary data. Otherwise, a preliminary experiment could be planned using a design criterion. For this purpose, one of the candidate models can be assumed as the "correct" one, and then the experimental conditions can be determined using the D-optimality design criterion. Another useful approach for designing the preliminary experiments is the factorial design method. The last solution is to generate samples from the assumed parameter prior distributions and then move directly to the model discrimination DOE step. After carrying out the designed experiment, a new sample of parameter values should be obtained for the current parameter posterior distributions, which include the designed data. The calculation of the model marginal likelihood and model probabilities will occur after enough model discrimination experiments are carried out to have an informative prior distribution. A rough criterion for deciding to update models probabilities or not, is having at least $\max_k(npar_k)$ experimental points, where $npar_k$ represents the number of parameters in the k^{th} candidate models.

Also, if there are $npar_k$ or more preliminary data points available, parameter estimation methods should be used in this step to find the best parameters. Then estimated parameters

may be used as the starting point of the sampling process, which is the next step of the procedure. This could help to have a more efficient sampling process with reasonable acceptance probability in the Adaptive Proposal Random Walk Metropolis-Hastings (AP) method.

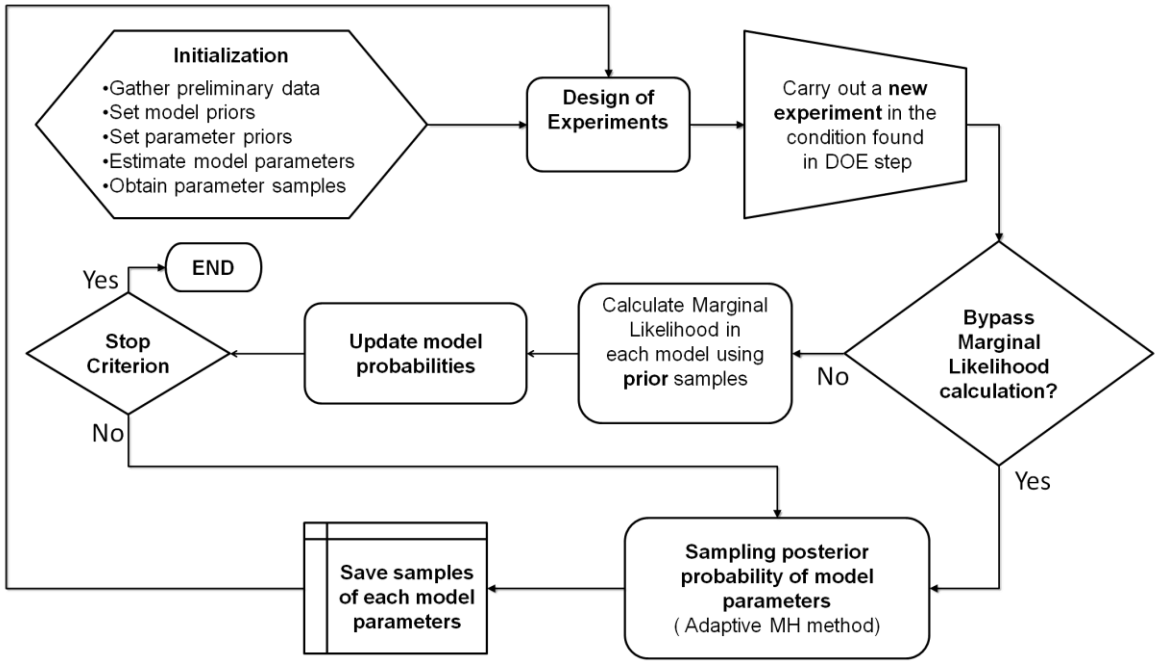

Figure 4-1: Sequential Bayesian MC Model Discrimination - using prior samples

Sampling from the Parameter Distribution

The Random-Walk MH method has been used in this implementation to sample from $P(\theta|M_k, \mathbf{y})$. We used Adaptive Proposal RWMH (AP) method (section 2.4.1.1) to address the problem of tuning the walking distribution. In each sampling step the original prior is used as the prior and the likelihood is calculated as the product of the likelihoods for all observed data including the preliminary ones.

Although using the AP method helps to keep the acceptance rate in the recommended range, 0.15-0.40 (Gelman et al., 1996), cases may occur where AP cannot tune the covariance to force the acceptance range in the desired range. In this situation, our implemented program will stop. Low standard deviation in the likelihood function may decrease the acceptance probability because the likelihood becomes too sharp. By contrast, high values of standard deviation for the likelihood function increase acceptance probability. The size of samples, which are used to update the covariance matrix, H, is another parameter which influences the average of the acceptance rates.

Experimental Design

For the purpose of designing the experiments, the Roth criterion given by equation (3.40) is used. In any design of experiments step, $P(M_i|\mathbf{y_{n-1}})$ which is the prior of the model, is replaced by the posterior probability of the model in the last iteration. The variable $\hat{y}_j(x)$, the predicted value from model i, is evaluated at experimental condition x. Approximating the predicted model outputs is possible by using equation (3.42). Then this new experiment is carried out and the result is added to the available data.

Marginal Likelihood Calculation

As mentioned earlier, the marginal likelihood is estimated by equation (4.12). Note that the likelihood in this equation is calculated only using the last observed experimental point. The reason is that the data points from the preliminary data to the last iteration experiment, have already influenced the prior distribution from which parameter samples are generated. Thus if the likelihood of the parameters was calculated using all the data points, the previous data would affect the marginal likelihood more than once.

Updating Model Probabilities

Model probabilities are updated by equation (4.6). In this equation the prior is the current probability of the model and the likelihood is the estimated marginal likelihood.

Decision Making

In the decision-making step, a criterion is checked to decide whether to stop the procedure or start a new iteration, meaning that an additional experiment is required. Our stopping criterion is to check all the model probabilities to see if one exceeds the stopping probability and to check the number of experiments, to see if the maximum possible number of experiments has been reached. If the stopping criteria are not satisfied, another iteration is required, and then MCMC samples from the current posterior parameter distribution should be generated.

Algorithm 4.2: SBMCMD : First implementation

Inputs:

K : Number of rival models

$N_{max\,experiment}$: Maximum number of experiments

$P_{selection\,criterion}$: Stop criterion probability

y_0 : Preliminary information

$\theta_{0,k}$: Initial guess for parameters in the k^{th} model

c_d: Scaling coefficient

R_1 : Initial covariance matrix

Gather preliminary data points, $\mathbf{y_0}$

Set rival model priors : $P_0(M_k|\mathbf{y_0}) = \pi(M_k)$ so that $\sum_{k=1}^{K} \pi(M_k) = 1$ (equation (4.7))

Set proper prior distributions for the parameters in each model

If (number of preliminary data points > number of parameters in a model)
find the best parameters in the k^{th} model, use $\theta_{0,k}$ as the initial guess. Save the best parameter to the $\theta_{0,k}$.

for n = 1 to $N_{\max\,experiment}$

for k=1,… K

Generate G samples using AP method (section 2.4.1.2),

Initialize θ^0 : if n = 1, $\theta^0 = \theta_{0,k}$ otherwise, $\theta^0 = \theta_{best,n-1,k}$

for j = 1… N_{sample}

Sample $z \in N(0, c_d^2 R_j)$

$\varphi = \theta^{j-1} + z$

Calculate $\alpha(\theta^{j-1}, \varphi) = \min\left(1, \frac{\pi(\varphi)\, l(\varphi|M_k, y_t)}{\pi(\theta^{j-1})\, l(\theta^{j-1}|M_k, y_t)}\right)$ (equation (2.8))

$\theta^j = \varphi$

if $\alpha(\theta^{j-1}, \varphi) < 1$, sample $R \in U(0,1)$

Then if $R > \alpha(\theta^{j-1}, \varphi)$, $\theta^j = \theta^{j-1}$

Update R_n :

If (j\H =0)

$R_n = \frac{1}{H-1}\tilde{K}^T\tilde{K}$ (equation (2.18))

$\tilde{K} = K - E[K]$ K=$[\theta^{j-H+1}, \ldots, \theta^{j+H}]$

else

$R_j = R_{j-1}$

Save them as θ^g $g = 1, \ldots, G$

Save the θ^g with highest $l(\theta^g|M_k, y_t)$ as $\theta_{best,t,k}$

Design the next experiment inputs x_t

$$x_t = \arg\max_{x} \sum_{i=1}^{K} \left[P_{t-1}(M_k|\mathbf{y_{t-1}}) \prod_{\substack{j=1 \\ j\neq i}}^{K} \left|\hat{y}_j(x) - \hat{y}_i(x)\right| \right] \quad \text{(equation (3.40))}$$

where $\hat{y}_i(x) = \frac{1}{G}\sum_{g=1}^{G} y_i(x|\theta_g)$ (equation (3.42))

Carry out the experiment to observe the new output data y_t

for k=1 to K

Calculate $L(M_k|y_t)$ using $\hat{L}(M_i|y) \approx \frac{1}{m}\sum_{g=1}^{G} l(y|M_i, \theta_g)$, (equation (4.12))

Update probability of models $P_t(M_k|\mathbf{y_t}) = P_{t-1}(M_k|\mathbf{y_{t-1}})\, L(M_k|y_t)$, (equation (4.6))

Normalize $P_t(M_k|y_t)$ values

Stop if $\exists\, k \in \{1, \dots, K\}: P_t(M_k|y_t) > P_{\text{selection criterion}}$

4.3.2 Second Implementation of SBMCMD: Posterior Based

The second approach, shown in Figure 4-2, uses the posterior distribution samples in all the iterations to estimate the likelihood. The sampling and the marginal likelihood estimation blocks are different in this implementation from the corresponding blocks in the first implementation. Thus only these two steps will be explained. The advantage of this implementation compared to the first implementation is that it uses the last information available in each step by using the parameter posterior probability samples. However it is limited to using the MH sampling method only. This method works for most of the cases but in situations where multi-modal or discontinuous parameter distributions occur the first implementation should be applied using a suitable MCMC methods that can adequately cover the parameter space.

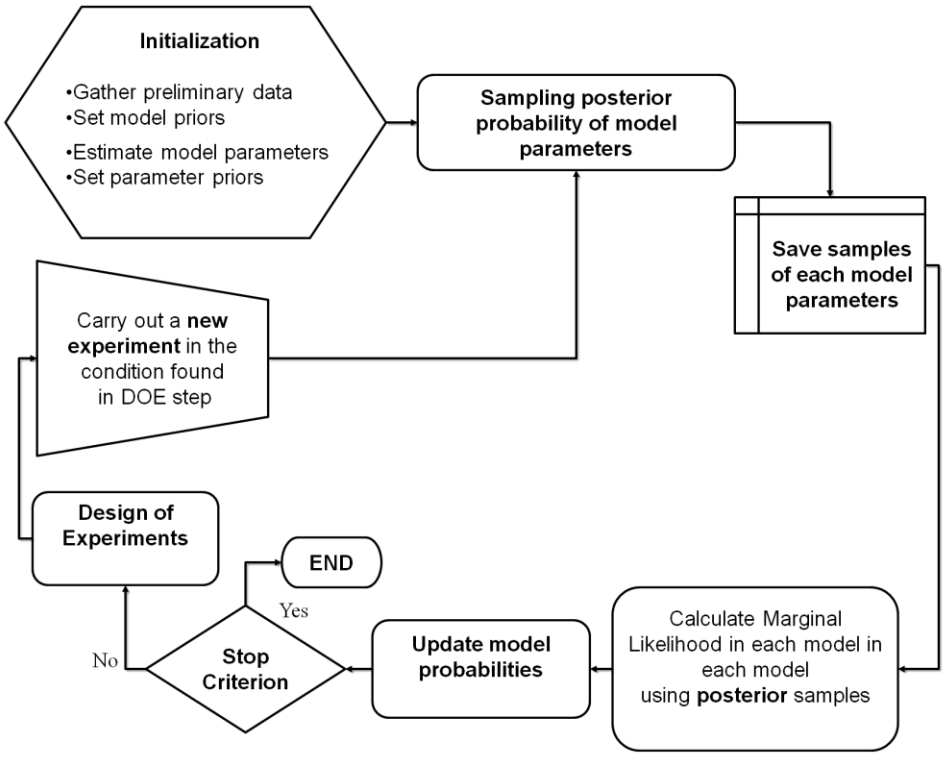

Figure 4-2: Sequential Bayesian MC model discrimination - using posterior samples

MC Sampling

Sampling from the posterior probability has been done by the method in section 2.4.1.1; to address the problem of tuning the walking distribution, an adaptive period of sampling is considered. For this purpose the adaptive proposal (AP) (Haario et al.,1999) method has been used (see section 2.4.1.1).

Chib's method uses the reversibility property of the chain. So it is necessary to preserve this property. Therefore, AP has been applied only during a preliminary adaptive tuning period.

According to the diminishing adaptation property (Rudnick, 2009) of the AP method, the influence of the adaptation tends to zero. We check the convergence of the updated covariance

matrix and continue the adaptation phase until the condition in equation (4.13) is satisfied or the maximum number of iterations is reached:

$$\frac{1}{n}\sum_{i=1}^{n}\Sigma_i - \frac{1}{n+1}\sum_{i=1}^{n+1}\Sigma_i = \frac{\sum_{i=1}^{n}\Sigma_i - n \times \Sigma_{n+1}}{n\,(n+1)} = 0 \qquad (4.13)$$

Estimating the Marginal Likelihood

After generating and saving samples for each candidate model, the marginal likelihood of candidate models is estimated using equations (3.20) and (3.31). Then, the model probabilities are updated by using equation (4.6) . This implementation is summarized in algorithm 4.3.

Algorithm 4.3: SBMCMD : Second implementation

Inputs:

K : Number of rival models

$N_{max\,experiment}$: Maximum number of experiments

$P_{selection\ criterion}$: Stop criterion probability

N_{sample}: Number of samples in MCMC parameters sampling

$\mathbf{y_0}$: Preliminary information

$\theta_{0,k}$: Initial guess for parameters in the k^{th} model

c_d: Scaling coefficient

R_1 : Initial covariance matrix

Gather preliminary data points, $\mathbf{y_0}$

Set rival model priors : $P_0(M_k|y_0) = \pi(M_k)$ so that $\sum_{k=1}^{K}\pi(M_k) = 1$ (equation (4.7))

Set parameters proper prior distributions for the parameters in each model

If (number of preliminary data points > number of parameters in a model)

find the best parameters in the k^{th} model, use $\theta_{0,k}$ as the initial guess. Save the best parameter to the $\theta_{0,k}$.

for n=1 to $N_{max\,experiment}$

for k=1,… K

Generate G samples using AP method (section 2.4.1.2),

Initialize θ^0 : if t=1, $\theta^0 = \theta_{0,k}$ otherwise, $\theta^0 = \theta_{best,n-1,k}$

for j = 1… N_{sample}

Sample $z \in N(0, c_d^2 R_j)$

$\varphi = \theta^{j-1} + z$

Calculate $\alpha(\theta^{j-1}, \varphi) = \min\left(1, \frac{\pi(\varphi)\, l(\varphi|M_k, y_n)}{\pi(\theta^{j-1})\, l(\theta^{j-1}|M_k, y_n)}\right)$ (equation (2.8))

$\theta^j = \varphi$

if $\alpha(\theta^{j-1}, \varphi) < 1$, sample $R \in U(0,1)$

then if $R > \alpha(\theta^{j-1}, \varphi)$, $\theta^j = \theta^{j-1}$

If ((j<burn-in size) & (j\H =0))

$R_t = \frac{1}{H-1} \tilde{K}^T \tilde{K}$ (equation (2.18))

$\tilde{K} = K - E[K]$, K=$[\theta^{j-H+1}, \dots, \theta^{j+H}]$

else

$R_j = R_{j-1}$

save the θ^g with highest $l(\theta^g|M_k, y_t)$ as $\theta_{best,t,k}$

Save them as θ^g $g = 1, \dots, G$

Save final $\theta_{best,t,k}$ as θ^*

Samples from $q(\theta^*, \theta|\mathbf{y})$, save them as $\{\theta^{(j)}\}$ $j = 1, \dots, J$

Calculate $P(\theta^* |\mathbf{y_n}) = \frac{G^{-1} \sum_{g=1}^{G} \alpha(\theta^{(g)}, \theta^*|\mathbf{y}) q(\theta^{(g)}, \theta^*|\mathbf{y})}{J^{-1} \sum_{j=1}^{J} \alpha(\theta^*, \theta^{(j)}|\mathbf{y})}$

Update $P(M_i|\mathbf{y_t})$: $\ln P(M_i|\mathbf{y_t}) = \ln l(\theta^*|M_i, \mathbf{y}) + \ln \pi(\theta^*) - \ln P(\theta^*|M_i, \mathbf{y})$ (equation (3.20))

Normalize $P_t(M_k|y_t)$ values

Stop if $\exists\, k \in \{1, \dots, K\}$: $P_t(M_k|\mathbf{y_t}) > P_{\text{selection criterion}}$

Design the next experiment inputs x_t

$$x_t = \underset{x}{\arg\max} \sum_{i=1}^{K} \left[P_{t-1}(M_k|y_{t-1}) \prod_{\substack{j=1 \\ j\neq i}}^{K} |\hat{y}_j(x) - \hat{y}_i(x)| \right] \qquad \text{(equation (3.40))}$$

where $\hat{y}_i(x) = \frac{1}{G}\sum_{g=1}^{G} y_i(x|\theta_g)$ (equation (3.42))

Carry out the experiment to observe the new output data y_t

4.3.3 Third Implementation of SBMCMD: Posterior and Prior Based Implementation

Figure 4-3 presents the flowchart of SB-MCMD approach with Acceptance –Rejection Metropolis Hastings sampling and Chib's method (Chib and Jeliazkov, 2005). In this implementation the model selection procedure is different for the preliminary iterations from the others. In the preliminary iterations there may be still not enough information available about the system. Thus the marginal likelihood is estimated from the posterior distribution samples. But after the preliminary iterations, the prior distributions are informed by the data, thus the prior samples are substituted in the likelihood to find the marginal likelihood. The advantage of this implementation compared to the first implementation is that updating model probabilities are not skipped during the preliminary step. But it is limited to the MCMC sampling methods, which can be used with Chib and Jeliazkov's method, MH and ARMH. The advantage of this implementation compared to the second implementation is la lower

computation effort after the preliminary steps. In brief, the third implementation tries to benefit from both advantages of the first and the second implementations.

Three blocks from this flowchart are slightly different from the first and the second implementations of the SBMCMD framework but other steps are common between all the implementations of the SBMCMD framework. These three steps are explained in the following and the complete pseudo-code is presented in Algorithm 4.4.

Sampling

After design of the preliminary experiments and data gathering followed by assignment of priors, samples from the posterior distribution of parameters, $P(\theta_i|M_i, \mathbf{y})$, are generated by using the Acceptance Rejection Metrolpolis Hastings method (refer to section 2.3.7). Then these samples are saved as the current posterior samples. The most probable sample in each model is named as θ^*.

One of the challenges in the implementation of this framework is selecting the constant c, in sampling from the posterior probability, equation (2.13). Using the same c value for different iterations is not efficient, since the posterior probability distribution of the parameters, which is a stationary distribution of this sampling process changes in each cycle, thus its upper limit changes. To overcome this problem and to achieve a reasonable acceptance probability in all iterations, we generate a batch of samples of parameters from the posterior distribution before the main sampling process and estimate the c value.

Another important issue is the selection of the prior and the proposal distributions. This implementation is used in the first case study in Chapter 5. In this case, we assumed that the prior and the posterior both are beta distributions. In addition, we updated the prior and the posterior distribution parameters in each cycle to adapt their pick with the best parameter found in the last cycle. This updating occurs before starting a new loop; thus, the MCMC sampling is not affected.

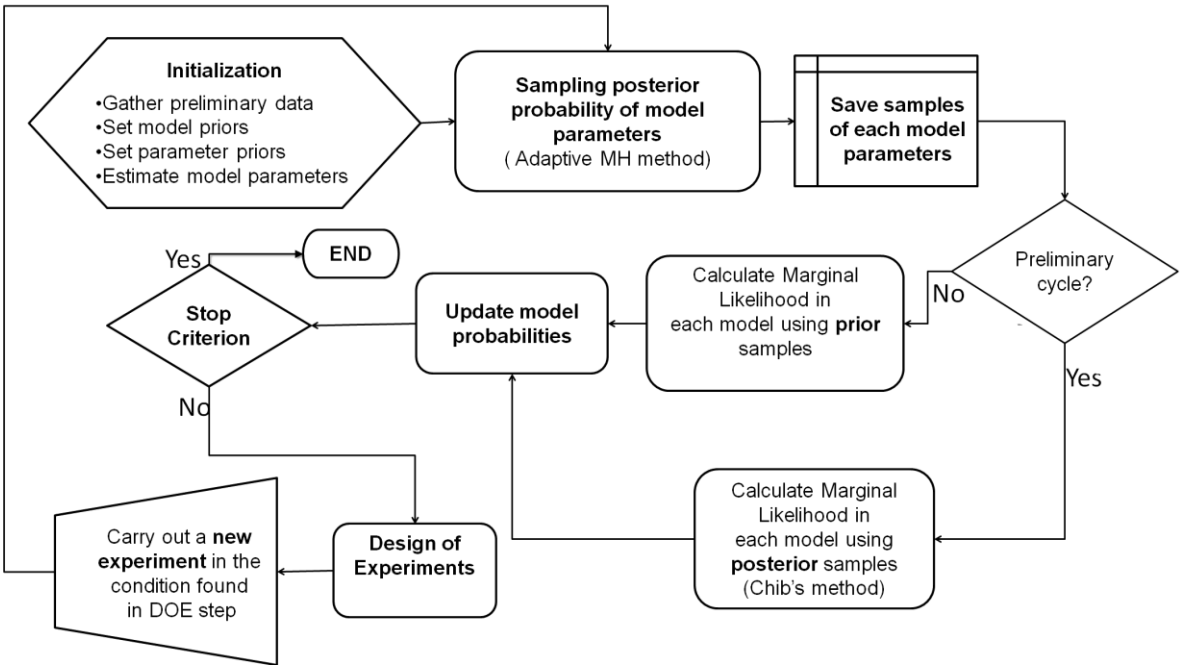

Figure 4-3: Sequential Bayesian MC model discrimination using both prior and posterior samples

Estimation of the Marginal Likelihood Using Posterior Samples

The posterior probability of θ^* is estimated by equation (3.31). Obtaining $p(\theta^*|M_i, \mathbf{y})$, the marginal likelihood of each model is estimated by equation (3.20). After that the posterior probability of each model could be updated using equation (4.6).

Estimation of the Marginal Likelihood Using Prior Samples

From this point, the first iteration of the sequential procedure is over and the second iteration starts; thus, the saved parameter samples are the prior samples.

Algorithm 4.6: SBMCMD : Third implementation

Inputs:

K : Number of rival models

$N_{\max\,experiment}$: Maximum number of experiments

$P_{selection\ criterion}$: Stop criterion probability

$\mathbf{y_0}$: Preliminary information

$\theta_{0,k}$: Initial guess for parameters in the k^{th} model

Gather preliminary data points, $\mathbf{y_0}$

Set rival model priors : $P_0(M_k|\mathbf{y_0}) = \pi(M_k)$ so that $\sum_{k=1}^{K} \pi(M_k) = 1$ (equation (4.7))

Set parameters proper prior distributions for the parameters in each model

if (number of preliminary data points > number of parameters in a model)
find the best parameters in the k^{th} model, use $\theta_{0,k}$ as the initial guess. Save the best parameter to the $\theta_{0,k}$.

for n = 1 to $N_{\max\,experiment}$

for k=1,… K

generate G samples using ARMH method (section 2.3.7),

Initialize θ^0 : if n = 1, $\theta^0 = \theta_{0,k}$ otherwise, $\theta^0 = \theta_{best,t-1,k}$

for j = 1… N_{sample}

Sample $\varphi \in q(\theta)$

Calculate $\alpha_{AR} = \min\left(1, \frac{\pi(\varphi)l(\varphi|M_k,\mathbf{y})}{c \times q(\varphi)}\right)$ (equation (2.14))

Sample $R \in U(0,1)$

if $R > \alpha(\theta^{j-1}, \varphi)$, Sample new φ

$\theta^j = \varphi$

if $\theta^{j-1} \notin D$

if $\varphi \in D$ $\quad \alpha_{MH}(\theta^{j-1}, \varphi|\mathbf{y}) = \frac{c \times q(\theta^{j-1})}{\pi(\theta^{j-1}) l(\theta^{j-1}|M_k, \mathbf{y})}$

if $\varphi \in D^c$ $\quad \alpha_{MH}(\theta^{j-1}, \varphi|\mathbf{y}) = \min\left\{1, \frac{\pi(\theta) l(\theta|M_k, \mathbf{y})\, q(\theta)}{\pi(\varphi) l(\varphi|M_k, \mathbf{y})\, q(\varphi)}\right\}$

Sample $R \in U(0,1)$

if $R > \alpha(\theta^{j-1}, Y)$ $\quad \theta^j = \theta^{j-1}$

Save the θ^g with highest $l(\theta^g|M_k, y_t)$ as $\theta_{best,t,k}$

Save samples as θ^g $g = 1, \ldots, G$

Save final $\theta_{best,t,k}$ as θ^*

Samples from $q(\theta^*, \theta|\mathbf{y})$, save them as $\{\theta^{(j)}\}$ $j = 1, \ldots, J$

Calculate $P(\theta^*|y) = \frac{G^{-1} \sum_{g=1}^{G} \alpha(\theta^{(g)}, \theta^*|y) q(\theta^{(g)}, \theta^*|y)}{J^{-1} \sum_{j=1}^{J} \alpha(\theta^*, \theta^{(j)}|y)}$

Calculate $P(M_i|\mathbf{y})$: $\ln P(M_i|\mathbf{y}) = \ln l(\theta^*|M_i, \mathbf{y}) + \ln \pi(\theta^*) - \ln P(\theta^*|M_i, \mathbf{y})$ (equation (3.20))

Normalize $P_t(M_k|y_t)$ values

Stop if $\exists\, k \in \{1, \ldots, K\}$: $P_t(M_k|y_t) > P_{\text{selection criterion}}$

Design the next experiment inputs x_t

$$x_t = \arg\max_x \sum_{i=1}^{K} \left[P_{t-1}(M_k|y_{t-1}) \prod_{\substack{j=1 \\ j \neq i}}^{K} |\hat{y}_j(x) - \hat{y}_i(x)| \right] \qquad \text{(equation (3.40))}$$

where $\hat{y}_i(x) = \frac{1}{G} \sum_{g=1}^{G} y_i(x|\theta_g)$ (equation (3.42))

Carry out the experiment to observe the new output data y_t

4.4 Comparing the Three Implementations

Table 4.1 summarizes the difference between three implementations of the SBMCMD.

Table 4-1: Comparison between three implementations of SBMCMD

	Implementation 1	Implementation 2	Implementation 3
Sampling from :	Prior parameter distributions	Posterior parameter distributions	Prior and posterior parameter distributions
Implemented sampling method:	Adaptive Random Walk Metropolis Hasting	Metropolis Hasting with preliminary adaptive period	Acceptance Rejection Metropolis Hastings
Marginal likelihood estimation method	Average of sample marginal likelihoods	Chib and Jeliazkov's method	First, Chib and Jeliazkov's method, then Average of sample marginal likelihoods
MCMC sampling method applicable in this framework	Any	Either ARMH or MH with related Chib and Jeliazkov's method	Either ARMH or MH with related Chib and Jeliazkov's method for the preliminary steps. Any for after preliminary steps.

In conclusion when preliminary knowledge are available from the physical system under investigation and it is desired to implement the simplest version of the SBMCMD, the first implementation is recommended. Also in cases where the models have discontinuous parameter spaces, or multi-modal parameter distributions, this implementation is the only choice because it has no limitations with respect to the MCMC sampling method and a suitable sampling method, which can handle the particular parameter space, can be used.

Of the three implementations, the posterior based implementation should produce the most accurate results, since it uses all the updated information in its sampling procedure. The case study presented in section 5.2 shows slightly better results for this implementation compared to the first implementation.

The third implementation is like the first implementation after the preliminary steps. But it is limited to ARMH and MH sampling procedures like the second implementation. Whenever it is possible to tune the proposal distribution in the ARMH sampling, it is computationally more efficient compared to the MH method.

4.5 Computational Implementation

The proposed framework is implemented in C++ programming language, and parallel programming is used to speed up the program by evaluating independent “for loops” in parallel. The framework has been designed based on the object oriented concepts. Some parts of the structure of the implemented program are shown in in UML diagrams in Appendix E.

4.5.1 Parallel Model Discrimination

The sampling procedure and the optimization in the DOE step are two computational time bottle-necks in these procedures. In a multi-thread program, the sampling procedure for different models and calculation of the predicted outputs could be assigned to different threads. For that purpose, OpenMP (Open Multiprocessing), which is an API (application programming interface) library, has been used in the implementation of the SBMCMD framework. This technique speed up the process but is limited to the number of the rival models. When the number of models is much less than the number of available processors, using the parallel MCMC methods will be beneficial.

4.5.2 Parallel MCMC Sampling

Sampling from Markov chains is not easily made parallel although it would be very beneficial to do so. The reason is that the current sample depends on the last one so they should be in the same thread. Making MCMC sampling an efficient parallel algorithm is still a hot topic in the statistical literature. Byrd (2010) introduced a method that is simple and easily implementable although the speed up is not proportional to the number of processors. He showed the application of speculative moves, where different threads generate candidate samples. For example, if there are three threads, three candidate samples are generated by any of them. Let's call them θ', θ'', and θ'''. All these threads calculate their α values. If the first candidate is not accepted, then the second one may be accepted. If the second one will be rejected too, the third one should be considered. This procedure could be implemented with more threads to increase the efficiency of the sampling.

This procedure has been tried during this study but has not been used in the final version of our framework. However, it is suggested for the future work to make the framework parallel and increase its speed.

Chapter 5
Case Studies

The objective of this study is to explore the effectiveness of the Sequential Bayesian Monte Carlo Model Discrimination (SBMCMD) framework. To achieve this objective, the method is applied to several case studies. In this research, computer simulation is used to generate experimental data for all the case studies presented in this chapter by assuming that one of the competing models is the "real system". This allows us to explore the performance of the method in different situations and determine their advantage and disadvantages, which is the focus of this research.

5.1 Order of Reaction

The example presented here has been used previously in Hsiang and Reilly (1971) and Box and Hill (1967) and has been recently used by Masoumi et al. (2012). In this case study, four models are proposed as rival ones for modeling a chemical reaction $B \rightarrow C$. These models are obtained by assuming that the reaction rate is either first, second, third or fourth order. The competing models are shown in the following:

$$\begin{aligned} f_1(t, T, A_1, E_1) &= \exp\left[-A_1 t \exp\left(-\frac{E_1}{T}\right)\right] + \varepsilon_1 \\ f_2(t, T, A_2, E_2) &= \left[1 + A_2 t \exp\left(-\frac{E_2}{T}\right)\right]^{-1} + \varepsilon_2 \\ f_3(t, T, A_3, E_3) &= \left[1 + 2A_3 t \exp\left(-\frac{E_3}{T}\right)\right]^{-\frac{1}{2}} + \varepsilon_3 \\ f_4(t, T, A_4, E_4) &= \left[1 + 3A_4 t \exp\left(-\frac{E_4}{T}\right)\right]^{-1/3} + \varepsilon_4 \end{aligned} \tag{5.1}$$

where $f_i(t, T, A_i, E_i)$ represents the fraction of component B remaining after time t (minutes) at temperature T (Kelvin), E_i is the activation energy divided by the universal gas constant, A_i is the pre-exponential factor, and ε_i is the measurement error which is assumed normally distributed with mean zero and known standard deviation. These models have two input variables, which are time and temperature. The range for time is $0 < t_i \leq 150$ in minutes and the temperature range is $450 \leq T_i \leq 600$ in degrees Kelvin.

Experimental results are simulated by assuming one of the candidate models to be the correct one with the true values of the parameters set to $A_i = 400$ and $E_i = 5000$. The range of parameters is set to $1000 < E_i < 10000$ and $100 \leq A_i \leq 1000$. Note that outside of this range the prior probability is zero. The preliminary experiments are the same as those used by Box and Hill (1976). They obtained the preliminary experimental conditions from a 2^2 factorial design with the levels of t_i set to 25 and 125 minutes and levels of T_i equal to 475 and 575 degrees in Kelvin.

The third implementation of the SBMCMD framework has been applied to this case study. The candidate models have a simpler structure compared to the models in the other case studies reported in this thesis. Therefore, selecting a suitable proposal distribution for the ARMH sampling procedure was not a problem. Thus the third implementation was applied to test the efficiency of this version of the SBMCMD method. In addition different values of the standard deviation were used in our study in order to assess the impact of the error magnitude on our ability to discriminate between the rival models. Here the standard deviation is set to values in the range from 0.01 to 0.1 in steps of 0.01. Each of the rival models was set in turn to be the correct model in order to explore whether or not the model structure affected our ability to

discriminate. The discrimination process was stopped either when the number of designed experiments reached 20 or one of the model posterior probabilities was greater than 95%.

In this case study the normal distribution is used both for the error generation of the simulated real system and for establishing the likelihood function of the candidate models. The standard deviations of the likelihood distributions are set to be higher or equal to the standard deviations used to generate the errors in the simulated data. Normally the true value of the measurement standard deviation is unknown. Therefore, a conservative approach is to set the likelihood standard deviation to a value, which overestimates the real system standard deviation.

The effect of the standard deviation was examined in two separate studies. In the first, the results of which are shown in Table 5-1, the standard error was limited to a maximum of 0.07. The standard deviation of the errors is varied from 0.01 to 0.07 in steps of 0.01 (7 levels). With the likelihood standard deviation set to be higher than or equal to the measurement error standard deviation, a total of 28 cases are tested for each model. For each of the 28 cases 30 independent simulations were run, yielding a total of 840 runs for each model, to get an average result. Each of the rival models was used as the "real" model in turn. So, a total of 112 cases involving 3360 runs were tested. For the range of errors studied the correct model was found in almost all of the cases. There were little to no differences in the success of the algorithm depending upon which model was the "true" model generating the data.

Table 5-2 presents results from a case study in which the upper limit of the error standard deviation is increased to 0.1. Again the error standard deviation of the model is assumed higher or equal to the real system. There are now a total of 220 cases, 55 for each model. With 30 independent runs per case, a total of 1650 runs per model, involving a total of 6600 runs were

carried out. The higher error level had little impact on selecting models 1 and 2 as the correct model. However when higher order kinetics were active, it became more difficult to identify the correct model within 20 experiments. It should be noted that the responses for models 1 and 2 are quite different from models 3 and 4, which helps to explain why they are identified almost always correctly. Comparing Table 5-1 and Table 5-2 shows that the higher error makes the discrimination more difficult, particularly for models 3 and 4.

Table 5-1: Error standard deviation 0.01-0.07 (840 cases)

"Real" model	% of runs where $P(M_i) > 0.95$
Model 1 (i=1)	838/840=99.7%
Model 2 (i=2)	834/840=99.3%
Model 3 (i=3)	821/840=97.8%
Model 4 (i=4)	835/840=99.4%

Table 5-2: Error standard deviation 0.01-0.1 (1650 cases)

"Real" model	% of runs where $P(M_i) > 0.95$
Model 1	1645/1650=99.7%
Model 2	1625/1650=98.5%
Model 3	1427/1650=86.5%
Model 4	1511/1650=91.6%

Table 5-3 focuses on the cases with high level of error. It shows results when the error has three levels: 0.08, 0.09, and 0.1. The second column indicates that the procedure cannot discriminate between models 3 and 4 properly when the error is high and the stopping criterion is set to a posterior probability of 95% or greater. The third column shows that if we decrease the stopping criterion to 80% instead of 95%, the procedure can pick the correct model in significantly more cases. The results are even better when the posterior probability is reduced to 70%. The results show that although for models 3 and 4 the procedure has difficulty picking the correct model properly when the stopping criterion was higher than 95% probability, the

procedure was moving in the right direction and presumably with more experiments would have likely chosen the correct model.

Table 5-3: Error standard deviation 0.08-0.1 (810 cases)

"Real" model	*$P(M_i) > 0.95$*	*$P(M_i) > 0.8$*	*$P(M_i) > 0.7$*
Model 1	0.996296	0.997531	0.997531
Model 2	0.976543	0.991358	0.991358
Model 3	0.748148	0.917284	0.945679
Model 4	0.834568	0.944444	0.965432

Finally, Table 5-4 shows the average number of model discrimination trials that were needed beyond the preliminary experiments to identify the correct model with a probability of 95%. As expected, lower error levels require fewer model discrimination experiments. In addition the procedure generally requires fewer experiments to detect the correct model, when models 1 or 2 are the ones generating the data. This implies that more complex kinetics require more experimental effort to detect the correct model.

Table 5-4: Average number of experiments after the preliminary experiments

	Average number of experiments		
"Real" model	*Error: 0.01-0.1*	*Error: 0.01-0.07*	*Error: 0.08-0.1*
Model 1	2.147879	0.736905	3.611111
Model 2	5.286667	2.295238	8.388889
Model 3	10.0103	5.20119	14.99753
Model 4	8.221212	3.958333	12.64198

The results are also shown in the 3D plots in the following. Figure 5-1 shows the final probability of model 2 on the z-axis, for the cases where this model is the correct model, versus the standard deviation (s.d.) in the real system and the standard deviation in the likelihood function of the models. Cases with probability equal or higher than 80% are shown with red dots and cases which have a probability less than 80% are presented with ×. . This plot shows

that a few of the results are below this surface. Figure 5-2 shows the same results when the correct model is model 3.

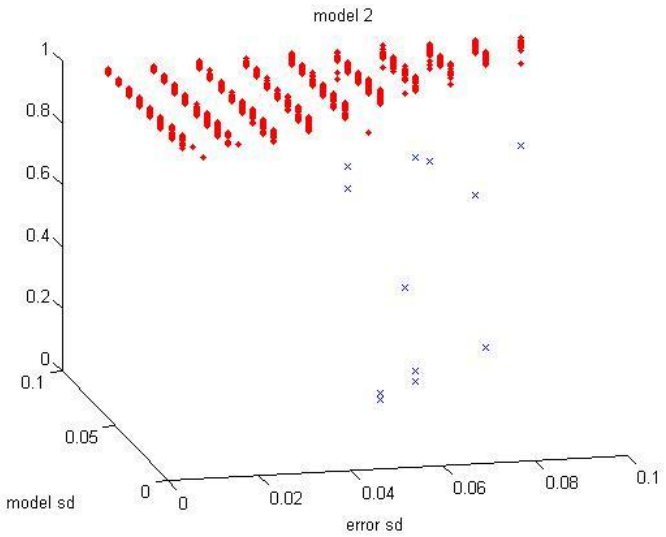

Figure 5-1: Probability of model 2, when it is the correct one (red dots shows probabilities higher than 0.8)

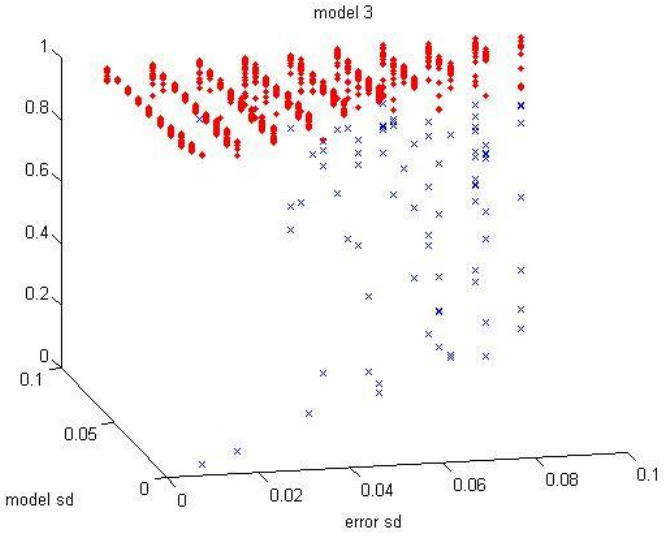

Figure 5-2: Probability of model 3, when it is the correct one (red dots shows probabilities higher than 0.8)

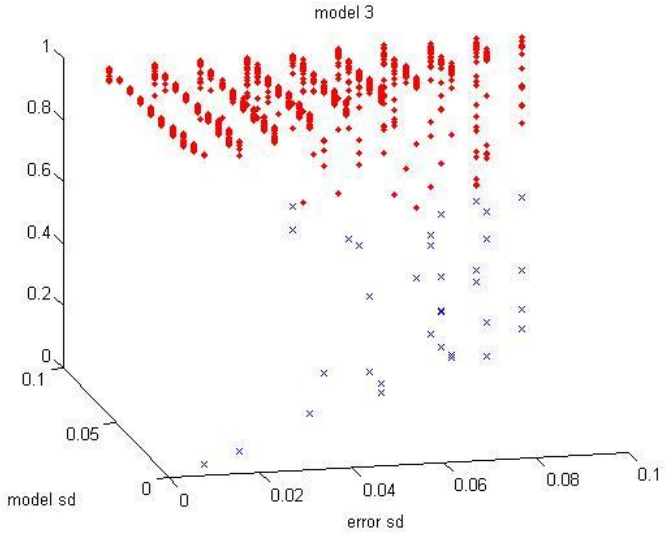

Figure 5-3: Probability of model 3, when it is the correct one (red dots shows probabilities higher than 0.5)

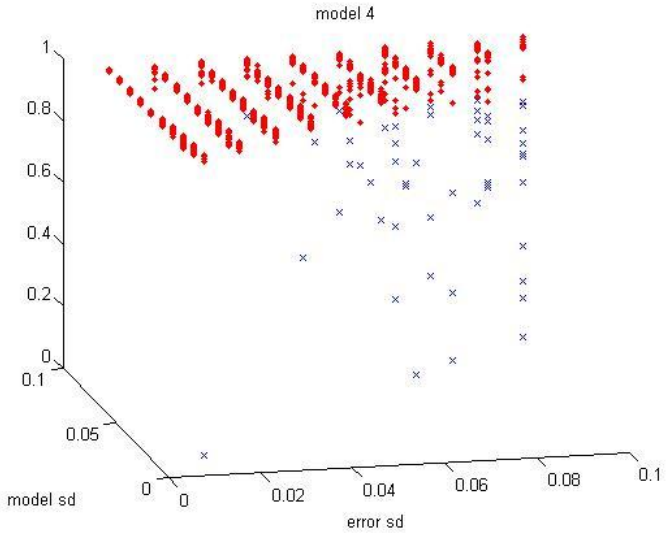

Figure 5-4: Probability of model 4, when it is the correct one (red dots shows probabilities higher than 0.8)

Figure 5-3 presents the same results but cases with 50% or higher probability are shown with red dots. This shows that although some final probabilities are less than 80%, they are still mostly higher than 50%. Figure 5-4 shows the same results when the correct model is model number 4.

Table A-1 (Appendix A) shows the average probability of each model, from 30 replicates, when it is the "correct" one for each combination of the real error and the likelihood function error value. The cells with a probability less than 0.95 are highlighted in this table. The fact that highlighted cells are concentrated in the bottom of the table where the errors are higher, shows that high levels of error make model discrimination more difficult as expected.

5.2 Rate of Fe^{ii} Formation

The reaction studied in this section is the steady-state oxidation of ferrous iron by Ferrobacillus ferrooxidans. The experimental conditions consist of the concentration of the Fe^{II} ion denoted by [A], and the concentration of the Fe^{III} ion denoted by [I] . The parameters in the models are the oxidation rate constant k_{i1} and the inhibition rate constant k_{i2} . There are three functional forms proposed to explain the mechanism. The resulting models are shown in equations (5.2)-(5.4).

$$f_1([A],[I],k_{11},k_{12}) = \frac{k_{11}[A]}{5.5 + [A] + 5.5\,k_{12}[I]} + \varepsilon_1 \qquad (5.2)$$

$$f_2([A],[I],k_{21},k_{22}) = \frac{k_{21}[A]}{(5.5 + [A])(1 + k_{22}[I])} + \varepsilon_2 \qquad (5.3)$$

$$f_3([A],[I],k_{31},k_{32}) = \frac{k_{31}[A]}{5.5 + [A](1 + k_{32}[I])} + \varepsilon_3 \qquad (5.4)$$

where f_i is the rate of formation of the Fe^{III} ion (mol/ L. s) under the i^{th} model and 5.5 is the Michaelis-Menten constant. The measurement error ε_i is assumed to be normally distributed with mean zero and known standard deviation. This case study has been used previously in the assessment of model discrimination techniques by Hsiang and Reilly (1971) and Burke et. al (1997). In this study, data were simulated, by using each of the three models in turn as the "real" model. The "true" parameters used for the data generation are given in Table 5.5. Random experimental errors were added to all simulated data. The experimental error was assumed to be Normal and additive. Errors were generated with standard deviations set to values between 1 and 6 in increments of 1. Different values of the error standard deviation were used in order to determine whether or not the measurement error magnitude affects our ability to discriminate between the models.

Table 5-5: "True" parameter values used for data generation

Model	k_{i1}	k_{i2}
1	1501	22.1
2	1502	22.2
3	1503	22.3

In this case study, the normal distribution is used to generate experimental errors and to establish the likelihood functions of the candidate models. However the standard deviations used in the likelihood functions were set to be higher or equal to the one used to simulate the errors. So for example when the experimental error standard deviation was 3, the standard deviations used in the likelihood function were 3 to 6 in increments of 1. This is done to reflect the fact that the true error variance is normally not known and by using a standard deviation in the likelihood that overestimates the true standard deviation the approach is a conservative one.

In the remainder of this section the results of two tests will be presented where the ranges of the unknown parameter space are different. In both tests, there have been considered $\sum_{i=1}^{6} i = 21$ scenarios with 5 replicates for each model. Thus, 315 cases were tested in total. The procedure stops when one of the model probabilities reaches 95%, or when the number of experiments reaches 20.

5.2.1 Study 1

In this study, the prior probability for each model was set to 1/3 while a beta distribution was used as the prior for the parameters. The ranges of the parameters were set to $100 < k_1 < 9000$ and $1 < k_2 < 1000$. These ranges are chosen to be quite broad to ensure adequate coverage. As shown in Burke et al. (1997), five preliminary data were generated using a two-level factorial experiment with center point in [A] and [I]. These data were used to first establish the posterior probability density of the model parameters before initiating the model discrimination phase of the analysis. In addition, the preliminary data were used to estimate the best parameter in each model. Then the prior distribution of parameters in each model was adapted to fit its mode to the best parameters value found in the parameter estimation step.

The posterior base or second implementation is applied to this case study. We first tried the third implementation with the ARMH sampling method but the sampling process failed frequency and needed to be tuned for each scenario. Therefore we switched to the second implementation with adaptive preliminary sampling period to address the tuning problem.

Table 5-6 shows the results using the SBMCMD. When models 1 and 3 were the model generating the data, the method was able to discriminate and identify the correct model in virtually all cases indicating that experimental error, over the range studied, has little to no

effect on our ability to discriminate. For model 2, the method identifies it as the correct model in 75 out of 105 cases. If we examine the 30 cases where model 2 was not picked with a probability of 95%, there were 20 cases where the maximum number of experiments was reached. In 15 of these cases the model probability for model 2 was 60% indicating that in these cases it is likely that if more experiments than 20 had been carried out eventually mode 2 would be correctly chosen. In the remaining 10 cases model 3 was incorrectly chosen as the right model.

The results show that the method is successful in discriminating between these models and even in most of the cases where a model probability of 95% was not reached, it appears that with more trials the correct model would have eventually been found.

Table 5-6: Cases which the "best" model has selected correctly

"Real" model	***Probability > 0.95***	***Probability > 0.6***
Model 1	105/105=100%	105/105=100%
Model 2	75/105=71.4%	90/105=85.7%
Model 3	103/105=98.1%	105/105=100%
Total	283/315=89.8%	300/315=95.2%

According to the results in Table 5-6 it is more difficult to discriminate between the models when model 2 is the "correct" one. Comparing the denominators of the functions in equation 5.2 - 5.4 shows that model 2 has an extra term, k_2[I][A], in its denominator compared to model 1. In the same way, it has term $5.5\ k_2$[I] in its denominator in addition to the three terms that model 3 has. Thus model 2 can behave like either model 1 or model 2 for some values of [I] and [A]. Therefore the structure of model 2 makes it more difficult to select this model as the "correct" one, whenever it is the one generating the "real" data.

Figure 5-5 shows the joint posterior probability density function for the two model parameters for each of the three candidate models for a case where the true model was selected in four iterations. In each subplot, the x-axis shows k_1 and the y-axis shows k_2 values. The figure shows how with more iterations, in other words more experimental data, the parameter joint distributions become smaller very quickly.

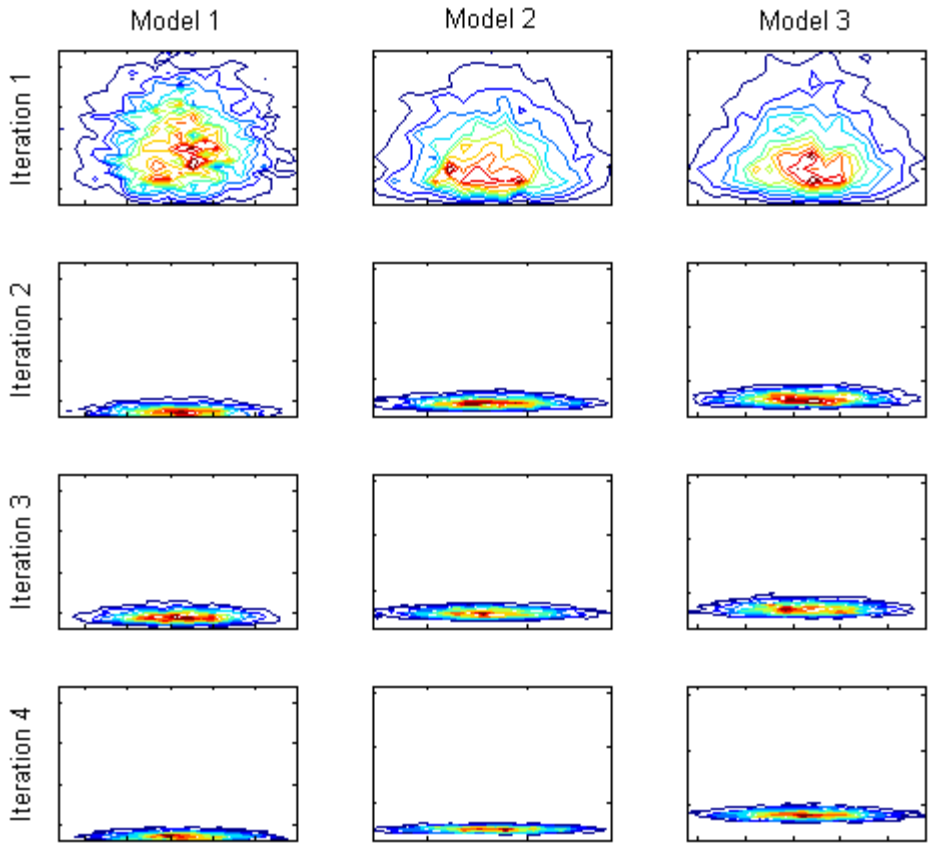

Figure 5-5: The evolution of the joint posterior probability density functions for the two parameters (k1 and k2) in each of models 1 to 3 as a function of the iteration number

Figure 5-6 shows probability of models versus iteration or number of experimental data for a particular case where the second model is the correct one. In this case 14 experiments were

designed and carried before the stopping criterion was satisfied. One can see a very smooth convergence of the model probabilities to their final values.

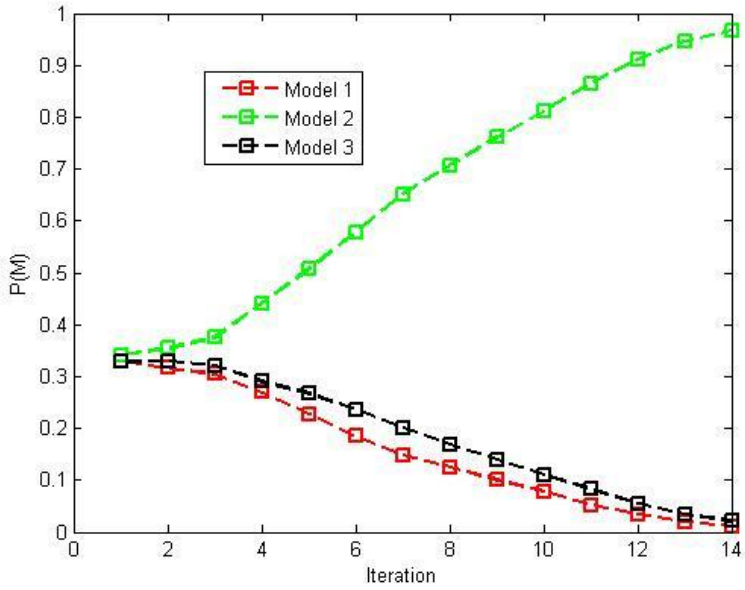

Figure 5-6: The trajectory of model probabilities versus number of iterations

5.2.2 Study 2

In this study the prior based and the posterior based implementations of the SBMCMD framework have been applied to the case study to compare their effectiveness in model discrimination. We first tried to apply the prior-based implementation of SBMCMD without any preliminary experiment but we found that the sampling process is too sensitive to the assumed initial values and the range of the parameters. The parameter distribution plots helped to identify the reason. Figure 5-7 contains 9 sub plots, which present the probability distribution of parameters in three candidate models when model 3 is the true model. These sub-plots are arranged in a 3 by 3 table where each column represents a model. The first row

of subplots shows the posterior distributions using one observed data point. The second row shows the posterior distributions after adding another data point and the third row contains parameters probability distributions from three experimental data points. The range of parameters in these plots are $100 < k_1 < 9000$ and $1 < k_2 < 1000$, which are the same ranges used in the first study.

When there is only one data point available (first row of subplots in Figure 5-7) the distribution of the parameters is a multi-modal one, which is hard to sample from. That explains why our sampling procedure was failing and it was sensitive to the initial point.

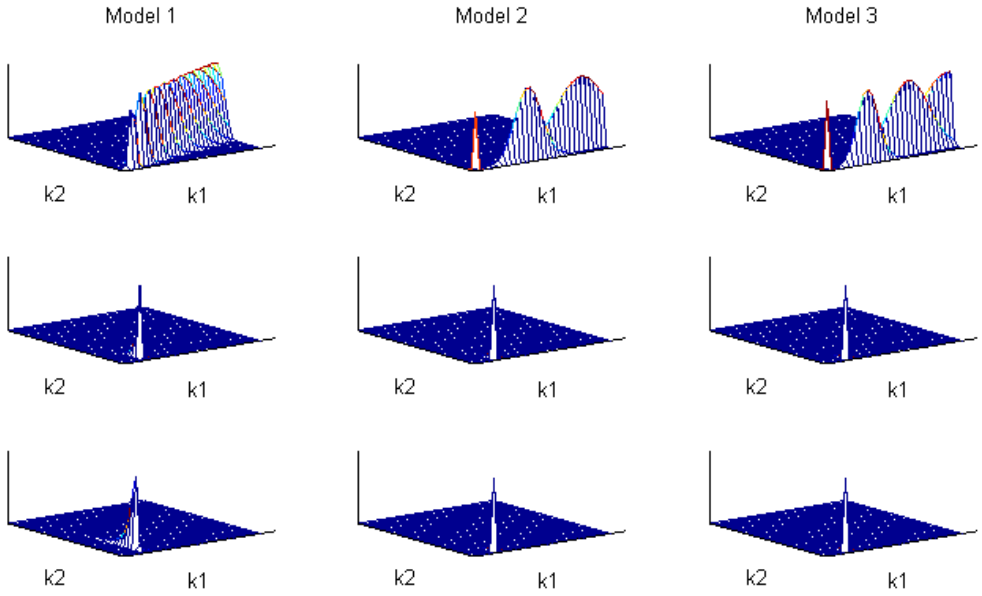

Figure 5-7: Posterior probability of parameters

Based on the preliminary information shown in Figure 5-7, to compare the two implementations of the SBMCMD framework, we limited the search space to $1000 < k_1 <$

2000 and $10 < k_2 < 100$. These narrower ranges imply that we have better prior information about the model parameters. Then we used a uniform distribution as the prior over these ranges. Again the same 315 scenarios from the first study were tried. Table 5.7 and Table 5.8 show the results of applying the prior based and posterior based implementations respectively. In the prior based implementation we first started the procedure with two preliminary data points then updating the models probabilities is skipped until the third model discrimination experiment has been designed yielding a total of five data points, including the two preliminary ones. In order to determine the effect of using fewer preliminary data points, we performed a second analysis in which the model probabilities were updated after the first designed experiment. In other words there was a total of three data points including the two preliminary data. The results are summarized in Table 5-7. The entries in each column show the number of times a particular model was chosen. The entries on the left correspond to the cases where five preliminary data were used, while the entries on the right correspond to the trials with three preliminary data points. For example when model 2 was used as the "true" model (row 2), model 2 was chosen as the correct one 97 times out of 105 when three preliminary data were used and 95 times out of 105 when five preliminary data were used. Overall the results show that fewer preliminary data can be used compared to the studies of Burke et al. (1997) and Hsiang and Reilly (1971). To compare the prior-based sampling approach to the posterior-based sampling approach we repeated the analysis in which five preliminary data were used with the posterior-based sampling approach. Note that for this the five preliminary data were generated using a factorial design.

Table 5-7: SBMCMD - Prior implementation

Selected / Real	Model 1		Model 2		Model 3	
	3 points	*5 points*	*3 points*	*5 points*	*3 points*	*5 points*
Model 1	105	105	0	0	0	0
Model 2	2	3	97	95	6	7
Model 3	3	1	5	6	97	98

Table 5-8 summarizes the results of this study. We see a slight improvement when model two and three are the "true" models. This is expected since the posterior-based approach is supposed to be more accurate and converge faster to the correct model, since it uses the latest available information to estimate the marginal likelihood. The prior based method assumed that the prior and the posterior distribution are in good harmony which is only valid when enough information from the system is available. As shown, compared with the posterior-based approach it performs quite well. In addition it is possible to use different MCMC sampling methods with it. The posterior based method however is limited to using The Random Walk Metropolis Hastings sampling procedure.

Table 5-8: SBMCMD - Posterior implementation

Selected / Real	Model 1	Model 2	Model 3
Model 1	105	0	0
Model 2	0	103	2
Model 3	1	1	103

Note that although here the posterior based one shows slightly better results this might not be true in general. The prior based SBMCMD with 5 and 3 preliminary data points, in table 5-7, shows close results. That implies that for this case study the estimation in equation (4.12) was a good approximation by using two data appoints and the parameters probability distribution did not change dramatically by adding two more data points.

Theoretically, SBMCMD methods can start running without any prior information. But accurate sampling from an unshaped posterior distribution may not be easy. Thus it is always beneficial to start with some preliminary data.

The results for tables 5-6 and 5-8, which are both for the posterior-based sampling method using five preliminary data, differ only in the ranges used for the model parameters. Table 5-8, which is based on narrower parameter ranges, shows somewhat better results for model 2. This demonstrates that the selection of the parameter ranges can be important in some cases.

Table 5-9 shows the average number of total experiments including preliminary ones when applying different implementations. The second method requires slightly fewer experiments and that is expected as it uses more experiments designed using the model discrimination criterion. Results for the second and the third methods used are not significantly different when model 2 and model 3 are used as the "true" model. In all cases when model 1 is the "true" model, the discrimination process easily identifies it to be the correct model.

Table 5-9: Average number of experiments

Method	Model 1	Model 2	Model 3
1. Posterior implementation (5 preliminary points,	5.5	8.7	8.2
2. Prior implementation (5 preliminary points, 3 from model discrimination DOE)	5.0	7.3	6.4
3. Prior implementation (3 preliminary points, 1 from model discrimination DOE)	3.1	6.2	5.6

5.3 Copolymerization

The two main competitive models for copolymerization systems are the terminal and penultimate. The difference between the terminal and penultimate copolymerization models is in their propagation steps. The terminal model assumes that only the last unit in the growing chain influences the rate of monomer addition. On the other hand, the penultimate model assumes that the last two monomer units of the growing radical chain influence the monomer addition reaction. The terminal propagation step is represented by four reactions, which can be shown by:

$$R_{n,i\cdot} + m_j \overset{k_{ij}}{\rightarrow} R_{n+1,j\cdot} \qquad (5.5)$$

where the first subscript on R denotes the radical chain length, i and j are used to denote monomer units (1 or 2), and k_{ij} is the rate constant for addition of monomer j to a radical ending in unit i. m_j shows a j unit monomer. Letting $R_{n,ij}$ represent a radical with terminal

unit j and penultimate unit i, eight propagation reactions in the penultimate model can be summarized in the form shown in equation (5.6).

$$R_{n,ij\cdot} + m_k \xrightarrow{k_{ijk}} R_{n+1,jk\cdot} \tag{5.6}$$

The terminal model is usually expressed in terms of two homo-polymerization rate constants k_{11} and k_{22} and two monomer reactivity ratios which are defined in equation (5.7).

$$r_1 = \frac{k_{11}}{k_{12}}; \quad r_2 = \frac{k_{22}}{k_{21}} \tag{5.7}$$

In the same way, the penultimate model is usually expressed in terms of the two homo-polymerization rate constants k_{111} and k_{222}, along with four monomer reactivity ratios shown in equation (5.8),

$$\begin{aligned} r_{11} &= \frac{k_{111}}{k_{112}}; \quad r_{21} = \frac{k_{211}}{k_{212}} \\ r_{22} &= \frac{k_{222}}{k_{221}}; \quad r_{12} = \frac{k_{122}}{k_{121}} \end{aligned} \tag{5.8}$$

and two radical reactivity ratios,

$$s_1 = \frac{k_{211}}{k_{111}}; \quad s_2 = \frac{k_{122}}{k_{222}} \tag{5.9}$$

The terminal model can be derived from the penultimate one by applying the constraints in equation (5.10). Thus the terminal model is nested within the penultimate model.

$$\begin{aligned} \hat{r}_{11} &= \hat{r}_{21} \\ \hat{r}_{22} &= \hat{r}_{12} \\ \hat{s}_1 &= \hat{s}_2 = 1 \end{aligned} \tag{5.10}$$

5.3.1 Simulation of "Real" System

Three different copolymerization systems are considered, Styrene and Methyl Methacrylate (STY/MMA), Styrene Acrylonitrile (STY/AN), and Styrene and Butyl Acrylate (BBA). In this section first the general model of the "real" system is described; then the simulation conditions for STY/AN, STY/MMA, and STY/BA will be presented.

Table 5-10 shows all the reactions considered in the "general" copolymerization model. "General" in this case means that the model can generate both data for the terminal and for the penultimate model.

Table 5-10: Copolymerization mechanism

Reaction	Reaction step	
Initiator decomposition	$I \xrightarrow{k_d} R^{\cdot}_{in}$	(5.11)
Propagating	$R^{\cdot}_{n,ij} + m_k \xrightarrow{k_{ijk}} R^{\cdot}_{n+1,jk}$	(5.12)
Transfer of radical to monomer	$R^{\cdot}_{n,ij} + m_k \xrightarrow{k_{fM,ijk}} P_n + R^{\cdot}_{1,k}$	(5.13)
Termination	$R^{\cdot}_{n,ij} + R^{\cdot}_{m,kl} \xrightarrow{k_{t,ijkl}} P_{n+m}$ or $P_n + P_m$	(5.14)

where I represents the initiator and $R_{in\cdot}$ is the primary radical from the initiator. $R^{\cdot}_{n,ij}$ is a radical with n monomer units, with the two final monomers being i and j. P_n is a dead polymer with n monomer units and m_k is a monomer of type k (1 or 2). k_d is the initiator decomposition

rate constant, k_{ijk} and $k_{fM,ijk}$ are the propagation and the transfer to monomer rate constants for the radical ending in monomers ij adding monomer k. $k_{t,ijkl}$ is the termination rate constant between the radicals ending in monomers ij and kl.

Table 5-11 presents the mass balance equations in the general modeling of copolymerization. The model obtained from these equations is used as the "real" system simulator.

Table 5-11: Mass balance in copolymerization

Species	Mass balance	
Initiator	$\frac{dN_I}{dt} = -k_d N_I$	(5.15)
Radical concentration	$[R.] = \left(\frac{2fk_d N_I}{k_t V}\right)^{1/2}$	(5.16)
Monomer 1	$\frac{dN_1}{dt} = -\frac{\left[r_{21}N_1^2\left(\frac{r_{11}N_1 + N_2}{r_{21}N_1 + N_2}\right) + N_1N_2\right][R.]}{\frac{r_{21}+N_1}{k_{111}}\left(\frac{r_{11}N_1 + N_2/s_1}{r_{21}N_1 + N_2}\right) + \frac{r_{21}N_2}{k_{222}}\left(\frac{r_{22}N_2 + N_1/s_2}{r_{12}N_2 + N_1}\right)}$	(5.17)
Monomer 2	$\frac{dN_2}{dt} = -\frac{\left[r_{12}N_2^2\left(\frac{r_{22}N_2 + N_1}{r_{12}N_2 + N_1}\right) + N_1N_2\right][R.]}{\frac{r_{21}+N_1}{k_{111}}\left(\frac{r_{11}N_1 + N_2/s_1}{r_{21}N_1 + N_2}\right) + \frac{r_{21}N_2}{k_{222}}\left(\frac{r_{22}N_2 + N_1/s_2}{r_{12}N_2 + N_1}\right)}$	(5.18)
Bound monomer	$\frac{dN_p}{dt} = -\frac{dN_1}{dt} - \frac{dN_2}{dt}$	(5.19)
Conversion	$\frac{dx}{dt} = \frac{dN_p}{dt}\frac{1}{N_1 + N_2 + N_p}$	(5.20)
volume	$\frac{dV}{dt} = \frac{dN_1}{dt}MW_1\left(\frac{1}{\rho_1} - \frac{1}{\rho_p}\right)\frac{1}{1000} + \frac{dN_2}{dt}MW_2\left(\frac{1}{\rho_2} - \frac{1}{\rho_p}\right)\frac{1}{1000}$	(5.21)

where N_I denotes initiator moles, N_1 moles of monomer 1, N_2 moles of monomer 2, and N_p moles of monomer bound in the polymer. MW and ρ are the molecular weight and density and s_1 and s_2 are radical reactivity ratios. [R.] denotes the concentration of radicals, k_d the initiator decomposition rate constant, and k_t the overall termination rate constant.

For simulation of the real system, first the above set of differential equations is solved until the time at which $x(t) = x$, where x is the target conversion, which is considered to be an input to the simulation. Defining the final moles of monomer 1 and monomer 2 by $N_{1,x}$ and $N_{2,x}$, the final mole fractions of the reacting monomers are calculated from:

$$f_{1,x} = \frac{N_{1,x}}{N_{1,x} + N_{2,x}}$$
$$f_{2,x} = 1 - f_{1,x} \tag{5.22}$$

Next the instantaneous triad fractions for the penultimate model, A_{ijk}, are calculated using equations (5.23) to (5.25). A_{222}, A_{122} and A_{221} are obtained by replacing indices 1 and 2 in the following symmetrical equations:

$$A_{111} = \frac{r_{21} r_{11} f_{1,x}^2}{r_{21} r_{11} f_{1,x}^2 + 2 r_{21} f_{1,x} f_{2,x} + f_{2,x}^2} \tag{5.23}$$

$$A_{211} = A_{112} = \frac{r_{21} f_{1,x} f_{2,x}}{r_{21} r_{11} f_{1,x}^2 + 2 r_{21} f_{1,x} f_{2,x} + f_{2,x}^2} \tag{5.24}$$

$$A_{212} = \frac{f_{2,x}^2}{r_{21} r_{11} f_{1,x}^2 + 2 r_{21} f_{1,x} f_{2,x} + f_{2,x}^2} \tag{5.25}$$

Equations (5.23) to (5.25) can be simplified to yield the triad fraction equations for the terminal model by using the relationship shown in equation (5.10). The homo-polymerization constants used for Styrene and Methyl Methacrylate, Acrylonitrile and Butyl Acrylate are shown in

Appendix C, Table C-1, Table C-2, Table C-4 and Table C-7, respectively. The reactivity ratios and other required parameters for STY/MMA, STY/AN and STY/BA are presented in Table C-3, Table C-5 and Table C-6.

The output results of simulating three copolymerization systems have been compared with the real data in the same way that Burke (1994) presented them. To demonstrate that the model can predict the experimental data, conversion versus time and composition versus conversion in STY/MMA are presented in Figure 5-8 and Figure 5-9, respectively. Also the experimental data points, presented by Huang (1988), are shown in these two plots. The symbols in Figures 5-8 and 5-9 represent the experimental data of Huang (1988) at 60 ℃ and [I] =0.01 mol/L. The lines are model predictions using the general model in Table 5-11 with different values for $f_{1,0}$ which are shown in the legend of the plots. The simulated real system outputs produced in this study are exactly the same as those reported by Burke (1994). Therefore, for STY/BA and STY/AN one can refer to Burke (1994).

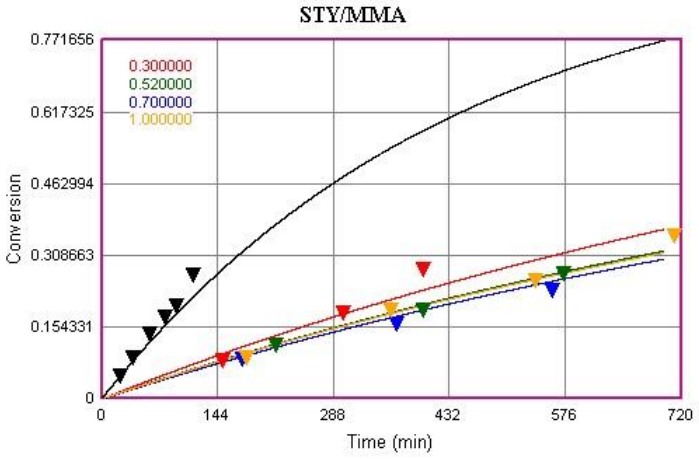

Figure 5-8: Conversion versus time for STY/MMA

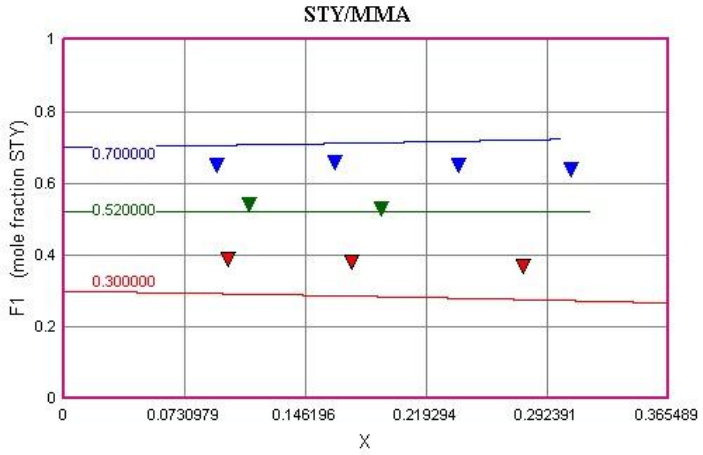

Figure 5-9: Composition versus conversion for STY/MMA.

Comparison between the experimental data and the model predictions shows that the general model predicts the real system behavior well.

5.3.2 Terminal and Penultimate Copolymerization Models

Section 5.3.1 presented the model used to generate the data in this case study. Here we present the competing models which represent the candidate models used in the model discrimination study.

The instantaneous mole fraction of monomer 1 bound in the copolymer, shown by F_1, for the terminal model is calculated from equation (5.26).

$$F_1 = \frac{r_1 f_1^2 + f_1 f_2}{r_1 f_1^2 + 2 f_1 f_2 + r_2 f_2^2} \qquad (5.26)$$

Equation (5.27) presents the corresponding equation for the penultimate model.

$$F_1 = \frac{r_{21}f_1^2\left(\frac{r_{11}f_1 + f_2}{r_{21}f_1 + f_2}\right) + f_1 f_2}{r_{21}f_1^2\left(\frac{r_{11}f_1 + f_2}{r_{21}f_1 + f_2}\right) + 2f_1 f_2 + r_{12}f_2^2\left(\frac{r_{22}f_2 + f_1}{r_{12}f_2 + f_1}\right)} \qquad (5.27)$$

Feed composition, f_1, and molar conversion are related to each other by equation (5.28). This equation can be integrated analytically for the terminal model. The resulting equations are shown in equation (5.29) and (5.30). This equation is called the Meyer-Lowry equation.

$$\ln(1 - x) = \int_{f_{1,0}}^{f_1} \frac{df_1}{F_1 - f_1} \qquad (5.28)$$

$$x = 1 - \left(\frac{f_1}{f_{1,0}}\right)^{\alpha}\left(\frac{1 - f_1}{1\_f_{1,0}}\right)^{\beta}\left(\frac{f_{1,0} - \delta}{f_1 - \delta}\right)^{\gamma} \qquad (5.29)$$

$$\alpha = \frac{r_2}{1 - r_2}; \quad \beta = \frac{r_1}{1\_r_1}$$

$$\gamma = \frac{1 - r_1 r_2}{(1 - r_1)(1 - r_2)}; \quad \delta = \frac{1 - r\_2}{2 - r_1 - r_2} \qquad (5.30)$$

Equation (5.28) should be integrated numerically for the penultimate model.

$$F_1 = \frac{f_{1,0}}{x} - \frac{(1 - x)f_1}{x} \qquad (5.31)$$

Therefore, considering the terminal and penultimate models as competitive ones with the above equations in model discrimination, they have two, $\{r_1, r_2\}$ and four $\{r_{11}, r_{12}, r_{22}, r_{21}\}$, unknown parameters respectively. The parameters s_1 and s_2 shown in equations 5.9 and 5.10 are not used here, since they are required only for the calculation of rate data which has not been considered in this research. For cases in which the triad fraction is the output, equations (5.23) -(5.25) are used.

5.3.3 Model Discrimination between the Terminal and Penultimate Models

The objective of revisiting the work performed by Burke (1994) was to compare the model discrimination method proposed in this research with those used by her for discriminating between the terminal and penultimate models, the two rival models, which have been proposed to describe the propagation mechanism in copolymerization. In her work she reported results using three different model discrimination methods including the Exact Entropy (EE) (Reilly, 1970), the Hsiang and Reilly (HR) (Hsiang and Reilly, 1971) and the Buzzi-Ferraris and Forzatti (BFF) methods (Buzzi-Ferraris and Forzatti, 1984; 1990; Buzzi-Ferraris et al., 1983). These three model discrimination methods were used to discriminate between the two proposed models for three different copolymer systems, namely Styrene and Methyl Methacrylate (STY/MMA), Styrene Acrylonitrile (STY/AN), and Styrene and Butyl Acrylate (STY/BA) using composition data (Burke et al., 1994a), triad fraction (Burke et al., 1994b) and rate data (Burke et al., 1995). A key feature of the problem is that the proposed models are nonlinear in the parameters, as is often the case in chemical engineering applications. Of the three methods, the Exact Entropy and the method of Buzzi-Ferraris and Forzatti rely on a linearization of the nonlinear models, while the HR method does not. Therefore theoretically, the HR method was expected to perform better than the other two. However, Burke's studies showed poorer results for the Hsiang and Reilly method.

They also used three different sets of parameters for the simulation of each real copolymerization system to simulate experimental data with terminal, penultimate and small penultimate models. In addition, three error levels were used in the simulation of the "real" system outputs and three sets of initial guesses of the parameters were used for the parameter

estimation step within the model discrimination method. Thus, 27 case studies were considered by combination of the "real" copolymer system, error levels and initial parameter guesses. These 27 cases are shown in Table C-9, with initial values and simulation parameters shown in Table C-10 to Table C-16. We applied the second implementation of the SBMCMD method in this case study, since the second implementation theoretically should lead to the most accurate results in model discrimination. In addition this method was efficient with respect to the MCMC acceptance rate. Burke et al. (1994a) showed that of the three methods applied to this model discrimination problem, the Buzzi-Ferraris and Forzatti method (BFF) had the best results. For the purpose of comparison, Table 5-12 gives the results of the SBMCMD method and shows the results of the BFF method presented by Burke et al. (1994a) using the copolymer composition data. Columns of this table present the model chosen by the methods and rows show the "true" model used to simulate the data. Each cell shows the number of times a particular model was selected as the "correct" model by a particular model discrimination method. For example, for the STY/MMA system when the terminal model was used to generate the data the BFF method picked the terminal model in 4 cases, the penultimate model in one case and was not able to discriminate between the two with a maximum of 20 experiments in 4 cases.

The BFF method failed to select the "correct" model in cases where the terminal model is the correct one. The reason is that when the models are equally good representation, over-fitting leads to picking the model with the larger number of parameters as the best model (Burke, 1994). On the other hand, SBMCMD easily picked the correct model when it is the terminal one but it had problems determining the correct model in cases where the real system was the

penultimate and especially the small penultimate model. As discussed before, the reason could be that Bayesian methods favor the model with fewer parameters (Jefferys and Berger, 1992). Thus it picks the simpler one when enough data for discrimination is not available. Accordingly, neither of these two methods can properly discriminate between the terminal and penultimate using copolymer composition data because of the small difference in predictions of the competitive models for copolymer composition data. Berger and Kuntz (2003) have observed the same result before experimentally

Table 5-12: Application of the BFF and SBMCMD methods to copolymer composition data

Model Chosen as 'Best' at 95% Confidence (number of simulation runs)						
	Terminal		Penultimate		Neither	
Simulation Model	BFF	SBMC	BFF	SBMC	BFF	SBMCM
Application to STY/MMA						
Terminal	4	8	1	1	4	0
Strong penultimate	1	3	7	6	1	0
Small penultimate	1	5	5	4	3	0
Application to STY/AN						
Terminal	2	9	2	0	5	0
Strong penultimate	0	4	9	5	0	0
Small penultimate	1	7	6	2	2	0
Application to STY/BA						
Terminal	1	9	1	0	7	0
Strong penultimate	2	3	6	6	1	0
Small penultimate	1	7	5	2	3	0

Table 5-13 summarizes results of applying the BFF method presented by Burke et al. (1994b), along with results of applying the SBMCMD method. These two methods were applied to the same 27 cases with triad fraction data used as the model and the "real" system output. In this case, results show that both methods can identify the correct model using triad fraction data. One difference between our application of SBMCMD and the BFF method applied by Burke et al. (1994b) is that we decided to reduce the number of preliminary experimental points to

see if the SBMCMD can discriminate between models with fewer preliminary data. Thus, four preliminary experimental data have been used in the application of the SBMCMD method but Burke started with 8 preliminary experimental data points obtained from a D-optimal design. Note that four is the minimum number of preliminary experiments needed since the penultimate model for triad fractions contains four unknown parameters. The results show that fewer preliminary experiments still lead to successful model discrimination. It is not clear however that this would not have been the case with the BFF method.

Table 5-13: Application of the BFF and SBMCMD methods to triad fraction data

Model Chosen as 'Best' at 95% Confidence (number of simulation runs)						
	Terminal		Penultimate		Neither	
Simulation Model	BFF	SBMCMD	BFF	SBMCMD	BFF	SBMCMD
Application to STY/MMA						
Terminal	9	9	0	0	0	0
Strong penultimate	0	0	9	9	0	0
Small penultimate	0	0	9	9	0	0
Application to STY/AN						
Terminal	9	9	0	0	0	0
Strong penultimate	0	0	9	9	0	0
Small penultimate	0	1	9	8	0	0
Application to STY/BA						
Terminal	9	9	0	0	0	0
Strong penultimate	0	0	9	9	0	0
Small penultimate	0	1	9	8	0	0

Table 5-13 presents the results obtained by applying the HR method to triad data by Burke et al. (1994b) and the results that we obtained using our modified HR method but with the same initial discretization. The difference between these two implementations is that they used equation (3.64) to calculate the posterior probability of parameters but we used equation (4.10),

which is the correct one. In addition we used four preliminary data points instead of eight. Our modified implementation of the HR method showed better results compared to the implementation by Burke (1994) for STY/AN and STY/BA but not for STY/MMA when the terminal model is the "true" model.

Comparing the results of the HR and the SBMCMD method (Table 5-13 and Table 5-14) shows that the SBMCMD method can discriminate significantly better. SBMCMD has the advantage of working with nonlinear models just like the HR method, but in addition the result are as good as the BFF method in discrimination.

Table 5-14: Application of the original and the modified HR method to the triad fraction data

Model Chosen as 'Best' at 95% Confidence (number of simulation runs)						
Simulation Model	Terminal		Penultimate			
	HR – Burke	HR (Modified)	HR – Burke	HR (Modified)	HR – Burke	HR (Modified)
Application to STY / MMA						
Terminal	6	3	3	6	0	0
Strong penultimate	0	0	9	9	0	0
Small penultimate	0	0	9	9	0	0
Application to STY/AN						
Terminal	7	9	2	0	0	0
Strong penultimate	0	0	9	9	0	0
Small penultimate	2	0	7	9	0	0
Application to STY/BA						
Terminal	4	9	4	0	1	0
Strong penultimate	0	0	9	9	0	0
Small penultimate	2	1	7	8	0	0

Table 5-15 compares the average number of experiments used by the different discrimination methods. The averages are between cases with different error and initial values (Table C-9).

The comparison shows that SBMCMD can select the best model with fewer experiments. Details of results from the SBMCMD method are presented in Appendix D.

Table 5-15: Number of experiments in application of different methods to triad fraction data

Simulation Model	HR – Burke	HR (Modified)	BFF	SBMCMD
Application to STY/MMA				
Terminal	9.1	6.2	9	4.4
Strong penultimate	9	4	9	4
Small penultimate	9	4.6	9	4.3
Application to STY/AN				
Terminal	9.3	4	9	4
Strong penultimate	9.1	5.1	9	4
Small penultimate	9.3	6.6	9.1	5.4
Application to STY/BA				
Terminal	11.4	4	10.2	4
Strong penultimate	9	4.1	9	5
Small penultimate	9.2	5.7	9.1	4.7

In the last test we applied the SBMCMD method without any initial preliminary data and with one preliminary cycle in which samples are obtained from the assumed prior to investigate the effect of the preliminary data. Table 5-16 shows the results. SBMCMD successfully discriminated between STY/MMA and STY/AN models but it had problems in model discrimination when the real system was the small penultimate for STY/BA. When the real system is the small penultimate, it is hard to discriminate between models because the parameters are close to the terminal case. These results confirm that at least in some cases it is better to start with some initial experiments rather than just relying on the selected prior.

Figure 5-10 shows joint parameter probability contours for the terminal model during application of the SBMCMD method. The first picture in the upper left hand side presents contours for the terminal parameters (r_1 and r_2) after one experiment and the picture in the

lower right hand side presents the contours in the last iteration. Comparing these plots shows that the probability distribution of the parameters becomes significantly sharper by adding more data points to the observed data set.

Table 5-16: Application of SBMCMD method to the triad fraction data with no preliminary experiments

Model Chosen as 'Best' at 95% Confidence (number of simulation runs)			
Simulation Model	Terminal	Penultimate	Neither
Application to STY/MMA			
Terminal	9	0	0
Strong penultimate	0	9	0
Small penultimate	1	8	0
Application to STY/AN			
Terminal	9	0	0
Strong penultimate	0	9	0
Small penultimate	1	8	0
Application to STY/BA			
Terminal	9	0	0
Strong penultimate	1	8	0
Small penultimate	4	5	0

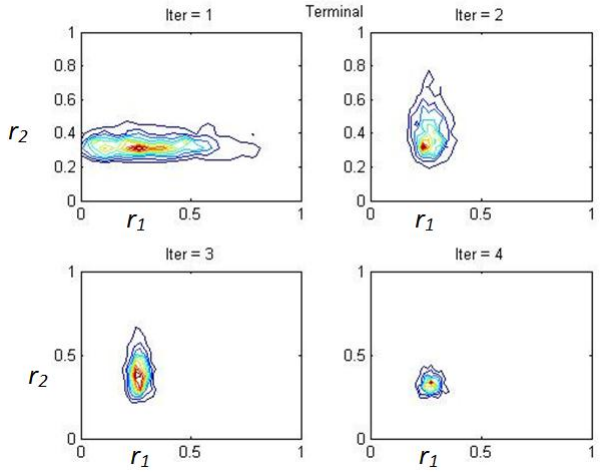

Figure 5-10: The evolution of the marginal joint posterior probability density functions for r1 and r2 in the terminal model (STY/MMA system) as a function of iteration number

Figure 5-11 shows joint parameter contours for the penultimate model. Convergence of the parameter samples to the real parameters shows that Random Walk Metropolis Hastings sampling can generate samples successfully during the procedure.

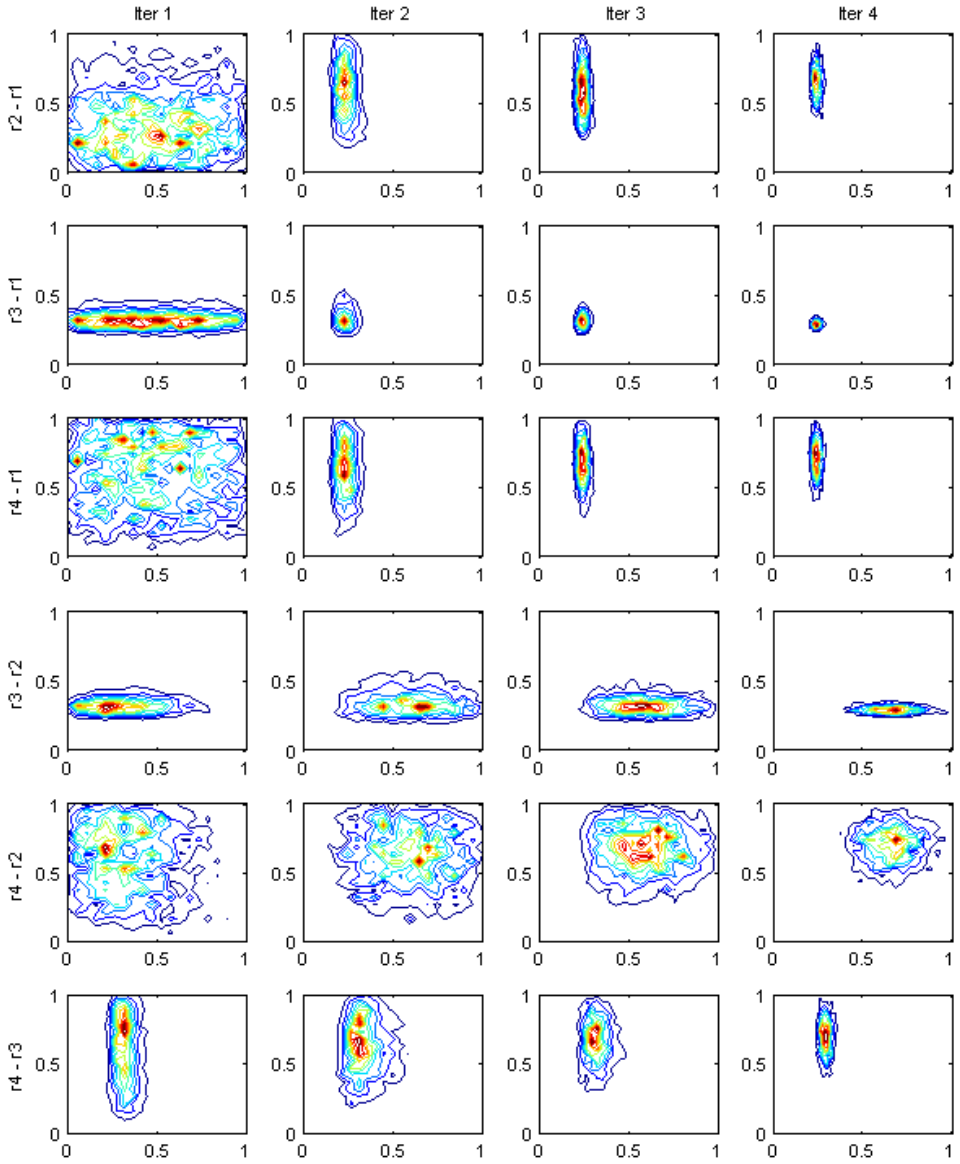

Figure 5-11: The evolution of the marginal joint posterior probability density functions for pairs of penultimate parameters as a function of iteration number STY/MA system

5.4 RAFT Polymerization

Three main groups of controlled polymerization have been introduced: Nitroxide-Mediated Radical Polymerization (NMRP), Atom Transfer Radical Polymerization (ATRP), and Reversible Addition-Fragmentation chain Transfer (RAFT) (Tobita and Yanse, 2007). All of these processes contain a capping and uncapping process, which makes the growing chains active and dormant for a while. In this way, chains grow to a more uniform chain length. As a result, the produced polymers have a narrow molecular weight distribution and their poly-dispersity is close to one. Chiefari et al. (1998) introduced the RAFT process as a controlled/living radical polymerization method (CLRP). A major advantage of the RAFT polymerization process over other living/controlled free-radical polymerization processes is that it can be used with a wide range of monomers and experimental conditions. Since 1998, when RAFT polymerization was introduced, a remarkable amount of research has been done on the RAFT process and its mechanism. A review on RAFT history can be found in Moad et al. (2005).

In addition to initiation, propagation and termination steps, which are included in any free radical polymerization process, controlled/living radical polymerization methods contain another step which avoids termination of growing chains. The addition-fragmentation between RAFT agent and the growing chains avoids termination in RAFT polymerization. Figure 5-12 shows the general structure of a RAFT agent.

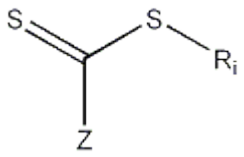

Figure 5-12: General chemical structure of a RAFT agent

The mechanism of the addition – fragmentation step is given in Figure 5-13. It first relies on the chain transfer of active molecule (1), to the RAFT agent (2), followed by fragmentation to active agent and another active molecule, which then reinitiates polymerization. Once the RAFT agent has been consumed, chain equilibrium is established between the active and dormant species (3).

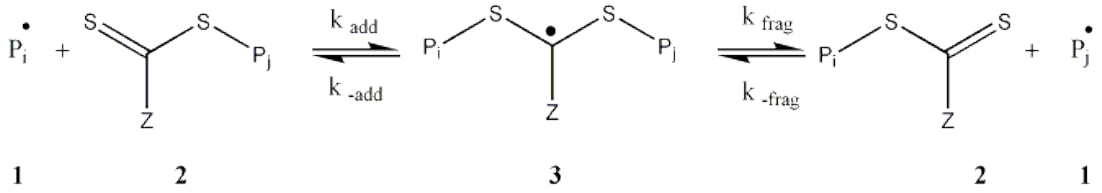

Figure 5-13: Addition – Fragmentation mechanism

Moment equations and different equilibrium mechanisms have been used widely to simulate the RAFT process. As well as the debate about the kinetic mechanism, there was a debate on the validity of these predictions for the RAFT process (Wang et al. 2003; Wang and Zhu 2003; Barner-Kowollik et al. 2001; Barner-Kowollik et al. 2003). Pallares et al. (2006) presented a complete model, which covers most of the mechanisms that had been proposed by that time. Despite this development, the mechanism of the RAFT kinetics is not yet adequately understood. The two main kinetic mechanisms that have been proposed for interpreting the rate retardation are the irreversible termination and the slow fragmentation methods.

The irreversible termination method assumes that the intermediate RAFT radical, produced by fast addition of propagating radicals to the RAFT species, may undergo irreversible termination with other radical species, cross termination, or even self-termination, which both conclude to radical loss and thus retardation (Monteiro and De Brouwer 2001; Zhang and Ray 2001).

On the contrary, the slow fragmentation mechanism assigns rate retardation to slow fragmentation of the intermediate RAFT radical, but ignores participation of this species in irreversible radical-radical termination reactions. The slow fragmentation mechanism causes the intermediate radical to be relatively stable and long-lived (Barner-Kowollik et al., 2001).

The monomer conversion versus time curves measured for a range of input conditions are fitted equally well by all proposed mechanisms. But the experimentally measured concentration of intermediate RAFT radicals is several orders of magnitude below the one predicted by the slow fragmentation model. On the other hand, the products from irreversible termination, three-arm star species, are not found in the product mixture of acrylate polymerizations.

Consequently, a mechanism step called the "missing step" was proposed in the dithiobenzoate agent RAFT process (Buback and Vana 2006; Buback et al. 2007). The missing step contains a reaction between a highly reactive propagation radical and star-shaped product from the termination of a propagating radical and an intermediate RAFT radical.

Also, Konkolewicz et al. (2008) proposed a RAFT kinetic scheme with a very fast cross-termination on the RAFT intermediate with short chains and negligible long chain cross-termination.

In conclusion, there are several proposed models for the RAFT mechanism. So, using model discrimination methods to pick the best measurement and experimental conditions which provide the most amount of information for model discrimination purposes can lead to finding the "best" mechanism that seems to be supported by the process information.

In this study we considered two candidate models, irreversible and slow fragmentation. These are the two main rival mechanistic models in the literature. The reaction mechanism of the irreversible model is shown in Table 5-17.

Table 5-17: Reaction mechanism for the irreversible model

Chemical initiation	$I \xrightarrow{k_d} 2R_{in}^{\cdot}$ $R_{in}^{\cdot} + M \xrightarrow{k_i} RM_1^{\cdot}$	(5.32)
Propagation	$RM_r^{\cdot} + M \xrightarrow{k_p} RM_{r+1}^{\cdot}$	(5.33)
Irreversible chain transfer to RAFT Agent	$RM_r^{\cdot} + AB \xrightarrow{k_{add}} RM_r\dot{A}B \xrightarrow{k_{bd}} RM_rA + B^{\cdot}$	(5.34)
Addition fragmentation	$RM_r^{\cdot} + RM_sA \xrightarrow{k_a} RM_r\dot{A}RM_s \xrightarrow{k_b} RM_rA + RM_s^{\cdot}$	(5.35)
Termination by disproportionation	$RM_r^{\cdot} + RM_s^{\cdot} \xrightarrow{k_{td}} RD_r + RD_s$	(5.36)
Termination by combination	$RM_r^{\cdot} + RM_s^{\cdot} \xrightarrow{k_{tc}} RD_{r+s}R$	(5.37)
Intermediate radical termination	$RM_p\dot{A}RM_q + RM_r^{\cdot} \xrightarrow{k_{tir}} RM_pRM_qRM_r$	(5.38)

Equation (5.32) presents the initiation step; where two radicals are generated from an initiator with rate constant k_d. This initiator primary radical may react with a monomer to generate a chain with one monomer, $RM_1^{\cdot}$. Equation (5.33) shows the propagation step, where monomers are added to the active radicals. Some studies assume that k_i and k_p are equal and they just consider one propagation step regardless of the radical length. Here we assumed that k_p is

constant and independent of the radical length but the propagation reaction for the initial radical is assumed to be different. Equation (5.34) presents the irreversible chain to the RAFT agent, AB. A in equation (5.34) represents the RAFT chain transfer agent. Equation (5.35) shows the addition of the radical to the RAFT agent with rate constant k_a and the fragmentation of the dormant species $RM_r\dot{A}RM_s$ to the RAFT agent and the active radical with rate constant k_b. Equations (5.36) and (5.37) present termination reactions with rate constants k_{td} and k_{tc} for termination by disproportionation and combination, respectively. Equation (5.39) shows the reaction of a two arm-adduct with another radical, thus making a three arm-adduct. This reaction is not considered in the slow fragmentation mechanism which is shown in Table 5-18. Note that those equations which are the same in Table 5-17 and Table 5-18 have the same equation number.

Table 5-18: Reaction mechanism for the slow fragmentation model

Chemical initiation	$I \xrightarrow{k_d} 2R_{in}^{\cdot}$ $R_{in}^{\cdot} + M \xrightarrow{k_i} RM_1^{\cdot}$	(5.32)
Propagation	$RM_r^{\cdot} + M \xrightarrow{k_p} RM_{r+1}^{\cdot}$	(5.33)
Reversible chain transfer to RAFT Agent	$RM_r^{\cdot} + AB \underset{k_{-add}}{\overset{k_{add}}{\rightleftarrows}} RM_r\dot{A}B \underset{k_{-bd}}{\overset{k_{bd}}{\rightleftarrows}} RM_rA + B^{\cdot}$	(5.39)
addition-fragmentation	$RM_r^{\cdot} + RM_sA \underset{k_{-a}}{\overset{k_a}{\rightleftarrows}} RM_r\dot{A}RM_s \underset{k_{-b}}{\overset{k_b}{\rightleftarrows}} RM_rA + B^{\cdot}$	(5.40)
Termination by disproportionation	$RM_r^{\cdot} + RM_s^{\cdot} \xrightarrow{k_{td}} RD_r + RD_s$	(5.41)
Termination by combination	$RM_r^{\cdot} + RM_s^{\cdot} \xrightarrow{k_{tc}} RD_{r+s}R$	(5.42)

In the slow fragmentation model, the equations of addition and fragmentation are considered to be reversible. The models in this study are based on moment equations for these two rival

models. For the parameters in the equations, values mentioned in Table 5-19 have been used which are values used by Wang and Zhu (2003).

Table 5-19: Rate coefficients - Cumyl Dithiobenzoate-Mediated Styrene Homo-polymerization

Parameter	***Irreversible***	***Slow fragmentation***
k_d	1e-5	1e-5
k_i	1e3	1e3
k_p	1e3	1e3
k_{tc}	1e7	1e7
k_{td}	1e7	1e7
k_{tir}	Parameter 3	0
f	0.5	0.5
k_a	Parameter 1	Parameter 1
k_{-a}	0	Parameter 2
k_b	Parameter 2	Parameter 2
k_{-b}	0	Parameter 1
k_{bd}	Parameter 2	Parameter 2
k_{-bd}	0	Parameter 1
k_{add}	Parameter 1	Parameter 1
k_{-add}	0	Parameter 2

As shown in Table 5-19, two parameters are considered unknown in the slow fragmentation model by assuming that $k_a = k_{-b} = k_{-bd} = k_{add}$. On the other hand in the irreversible model three parameters are considered unknown by assuming $ka = k_{add}$ and $k_b = k_{bd}$.

A complete model containing 43 differential equations forms the general model. This general model is presented in pallares et al. (2006). The scheme of this general model is shown in table 5-20.

Table 5-20: Reaction mechanism for the complete model (pallares et al., 2006)

Chemical Initiation	$I \xrightarrow{K_D} 2R_{In}^{\cdot}$ $R_{In}^{\cdot} + M \xrightarrow{K_I} RM_1^{\cdot}$	(5.43)
Thermal Self-Initiation	$M + M \xrightarrow{K_{Dim}} D$ $D + M \xrightarrow{K_{Thi}} D^{\cdot} + M^{\cdot}$	(5.44)
Propagation	$RM_R^{\cdot} + M \xrightarrow{K_P} RM_{R+1}^{\cdot}$	(5.45)
Chain Transfer To Monomer	$RM_R^{\cdot} + M \xrightarrow{K_{Fm}} RD_R + M^{\cdot}$	(5.46)
Chain Transfer To Solvent	$RM_R^{\cdot} + S \xrightarrow{K_{Fs}} RD_R + S^{\cdot}$	(5.47)
Chain Transfer To Chain Transfer Agent (Cta)	$RM_R^{\cdot} + Cta \xrightarrow{K_{Ft}} RD_R + Cta^{\cdot}$	(5.48)
Irreversible Chain Transfer To Raft Agent	$RM_R^{\cdot} + Ab \xrightarrow{K_{Tr}} RM_RA + B^{\cdot}$	(5.49)
Reversible Chain Transfer To Raft Agent	$RM_R^{\cdot} + Ab \underset{K_{-Add}}{\overset{K_{Add}}{\rightleftarrows}} RM_R\dot{A}B \underset{K_{-Bd}}{\overset{K_{Bd}}{\rightleftarrows}} RM_RA + B^{\cdot}$	(5.50)
Re-Initiation First: Re-Initiation Of The Raft Segment	$\left.\begin{array}{l} B^{\cdot} + M \xrightarrow{K_I} BM_1^{\cdot} \\ M^{\cdot} + M \xrightarrow{K_I} MM_1^{\cdot} \\ D^{\cdot} + M \xrightarrow{K_I} DM_1^{\cdot} \\ S^{\cdot} + M \xrightarrow{K_I} SM_1^{\cdot} \\ Cta^{\cdot} + M \xrightarrow{K_I} Ctam_1^{\cdot} \end{array}\right\} RM_1^{\cdot}$	(5.51)
Chain Equilibrium (Addition-Fragmentation)	$RM_R^{\cdot} + RM_SA \underset{K_{-A}}{\overset{K_A}{\rightleftarrows}} RM_R\dot{A}RM_S \underset{K_{-B}}{\overset{K_B}{\rightleftarrows}} RM_RA + B^{\cdot}$	(5.52)
Termination By Disproportionation	$RM_R^{\cdot} + RM_S^{\cdot} \xrightarrow{K_{Td}} RD_R + RD_S$	(5.53)
Termination By Combination	$RM_R^{\cdot} + RM_S^{\cdot} \xrightarrow{K_{Tc}} RD_{R+S}R$	(5.54)
Intermediate Radical Termination	$RM_P\dot{A}RM_Q + RM_R^{\cdot} \xrightarrow{K_{Tir}} RM_PRM_QRM_R$	(5.55)

The definitions of the moments used in the model equations are listed in table 5-21.

Table 5-21: Definition of moments of the polymer distributions (Pallares Et Al. 2006)

Species	***Moment***
Polymer Radicals (Living Polymer)	$Y_M = \sum_{R=1}^{\infty} R^M[RM_R^{\cdot}]$
Dormant Polymer	$Z_M = \sum_{R=1}^{\infty} R^M[RM_RA]$
Dead Polymer From Termination By Disproportionation And Transfer To Small Molecules	$Q_M = \sum_{R=1}^{\infty} R^M[RD_R]$
Dead Polymer From Termination By Combination	$S_M = \sum_{R=1}^{\infty} R^M[RD_RR]$
One Arm Adduct	$E_M = \sum_{R=1}^{\infty} R^M[RM_RA{\cdot}B]$
Two Arm Adduct (Macro Raft Radical)	$F_{Mn} = \sum_{R=1}^{\infty}\sum_{S=1}^{\infty} R^M S^N [RM_RA{\cdot}RM_S]$
Three Arms Dead Polymer	$G_{Abc} = \sum_{P=1}^{\infty}\sum_{Q=1}^{\infty}\sum_{R=1}^{\infty} P^A Q^B R^C [RM_PRM_QRM_R]$

Two Rival Models Can Be Obtained From This General One. The 43 Differential Equations Are Presented In Equations (5.56) To (5.78).

$$\frac{D(V[I])}{Vdt} = -K_D I \tag{5.56}$$

$$\frac{\mathrm{Dv[M]}}{\mathrm{Vdt}} = -\mathrm{K_I[M]}\left(\left[\dot{\mathrm{R}}_{\mathrm{In}}\right] + \left[\dot{\mathrm{B}}\right] + \left[\dot{\mathrm{M}}\right] + \left[\dot{\mathrm{D}}\right] + \left[\dot{\mathrm{S}}\right] + \left[\dot{\mathrm{C}}\mathrm{ta}\right]\right) - \mathrm{K_{Dim}[M]^2} - \mathrm{K_{Thi}[M][D]} - \mathrm{K_P[M][Y_0]} - \mathrm{K_{Fmv}[M][Y_0]} \quad (5.57)$$

$$\frac{\mathrm{D(V[D])}}{\mathrm{Vdt}} = \mathrm{K_{Dim}[M]^2} - \mathrm{K_{Thi}[M][D]} \quad (5.58)$$

$$\frac{\mathrm{D(V[S])}}{\mathrm{Vdt}} = -\mathrm{K_{Fs}[S][Y_0]} \quad (5.59)$$

$$\frac{\mathrm{D(V[Cta])}}{\mathrm{Vdt}} = -\mathrm{K_{Ft}[Cta][Y_0]} \quad (5.60)$$

$$\frac{\mathrm{D(V[R_{In}])}}{\mathrm{Vdt}} = 2\mathrm{fK_D[I]} - \mathrm{K_I[R_{In}][M]} \quad (5.61)$$

$$\frac{\mathrm{D(V[\dot{M}])}}{\mathrm{Vdt}} = \mathrm{K_{Thi}[M][D]} + \mathrm{K_{Fmv}[M][Y_0]} - \mathrm{K_I[M][\dot{M}]} \quad (5.62)$$

$$\frac{\mathrm{D(V[D])}}{\mathrm{Vdt}} = \mathrm{K_{Thi}[D][M]} - \mathrm{K_I[M][\dot{D}]} \quad (5.63)$$

$$\frac{\mathrm{D(V[S])}}{\mathrm{Vdt}} = \mathrm{K_{Fs}[S][Y_0]} - \mathrm{K_I[M][S]} \quad (5.64)$$

$$\frac{\mathrm{D(V[Cta])}}{\mathrm{Vdt}} = \mathrm{K_{Ft}[Cta][Y_0]} - \mathrm{K_I[M][Cta]} \quad (5.65)$$

$$\frac{\mathrm{D(V[Ab])}}{\mathrm{Vdt}} = -\mathrm{K_{Tr}[Ab][Y_0]} - \mathrm{K_{Add}[Ab][Y_0]} + \mathrm{K_{-Add}[E_0]} \quad (5.66)$$

$$\frac{\mathrm{D(V[B])}}{\mathrm{Vdt}} = \mathrm{K_{Tr}[Ab][Y_0]} + \mathrm{K_{Bd}[E_0]} - \mathrm{K_{-Bd}[Z_0][B]} - \mathrm{K_I[M][B]} \quad (5.67)$$

$\mathrm{M} = 0,1,2$

$$\frac{D(V[Y_M])}{Vdt} = 2\,F\,K_D\,[I] + K_I\,[M]([B^{\cdot}] + [M^{\cdot}] + [D^{\cdot}] + [S^{\cdot}] + [CtA^{\cdot}]) + B_M$$
$$- (K_{Fmv}\,[M] + K_{Fs}\,[S] + K_{Ft}\,[Cta] + K_{Tr}\,[Ab])[Y_M]$$
$$- (K_{Td} + K_{Tc})[Y_0][Y_M] - K_{Tir}\,[Y_M]F_{00} + \quad -K_1 Ab * Y(M)$$
$$+ K_2 B_{Dot} Z(M) - K_3\,Z_0\,[Y_M] + K_4\,[Y_0][Z_M] \tag{5.68}$$
$$- K\_Add\,[Ab]\,[Y_M] + K\_Nadd\,[E_M] - K_A\,[Z_0]\,[Y_M]$$
$$+ K_{-A}\,[F_{M0}] + K_B\,[F_{M0}] - K_{-B}\,[Z_0]\,[Y_M]$$

$$\frac{D(V[Z_M])}{Vdt} = K_{Tr}\,[Ab]\,[Y_M] + K_1\,[Ab]\,[Y_M] - K_2\,[B^{\cdot}]\,[Z_M] + K_3\,[Z_0]\,[Y_M]$$
$$- K_4\,[Y_0]\,[Z_M] - K_A\,[Y_0]\,[Z_M] + K_{-A}\,[F_{M0}] + K_B\,[F_{M0}] \tag{5.69}$$
$$- K_{-B}\,[Y_0]\,[Z_M] + K_{Bd}\,[E_M] - K_{-Bd}\,[B^{\cdot}]\,[Z_M]$$

$$\frac{D(V[Q_M])}{Vdt} = (K_{Fmv}\,[M] + K_{Fs}\,[S] + K_{Ft}\,[Cta] + K_{Td}\,[Y_0])[Y_M] \tag{5.70}$$

$$\frac{D(V[E_M])}{Vdt} = K_{Add}\,[Ab]\,[Y_M] - K_{-Add}\,[E_M] - K_{Bd}\,[E_M] + K_{-Bd}\,[B^{\cdot}]\,[Z_M] \tag{5.71}$$

$$\frac{D(VS_1)}{Vdt} = \frac{K_{Tc}}{2} Y_0^2 \tag{5.72}$$

$$\frac{D(VS_2)}{Vdt} = K_{Tc} Y_0 Y_1 \tag{5.73}$$

$$\frac{D(VS_3)}{Vdt} = \frac{K_{Tc}}{2}(2y_0 Y_2 + 2\,Y_1^2) \tag{5.74}$$

$$\frac{D(VF_{Mn})}{Vdt} = K_A Y_M Z_N - K_{-A} F_{Mn} - K_B F_{Mn} + K_{-B} Z_M Y_N \quad \text{Where } F_{Mn}$$
$$= \{F_{10}, F_{01}, F_{00}, F_{20}, F_{02}, F_{11}\} \tag{5.75}$$

$$\frac{D(VG_{Abc})}{Vdt} = K_{Tir} F_{Ab} Y_C \tag{5.76}$$

Where $G_{Abc} = \{G_{000}, G_{001}, G_{010}, G_{100}, G_{002}, G_{020}, G_{200}, G_{011}, G_{101}, G_{110}\}$

After Solving The Set Of Ode Equations, Equations (5.77) To (5.79) Can Be Used To Calculate The Polydispersity Of The Polymer.

$$\overline{M_N} = \frac{Y_1 + Z_1 + Q_1 + S_1 + E_1 + \frac{1}{2}(F_{10} + F_{01}) + \frac{2}{3}(G_{100} + G_{010} + G_{001})}{Y_0 + Z_0 + Q_0 + S_0 + E_0 + \frac{1}{2F_{00}} + \frac{2}{3}\ G_{000}} M_{Rep} \quad (5.77)$$

$$\overline{M_W} = \left[\frac{Y_2 + Z_2 + Q_2 + S_2 + E_2 + \frac{1}{2}(F_{20} + F_{02} + 2F_{11})}{Y_1 + Z_1 + Q_1 + S_1 + E_1 + \frac{1}{2}(F_{10} + F_{01}) + \frac{2}{3}(G_{100} + G_{010} + G_{001})} + \frac{\frac{2}{3}(G_{200} + G_{020} + G_{002} + 2(G_{110} + G_{101} + G_{011})}{Y_1 + Z_1 + Q_1 + S_1 + E_1 + \frac{1}{2}(F_{10} + F_{01}) + \frac{2}{3}(G_{100} + G_{010} + G_{001})}\right] M_{Rep} \quad (5.78)$$

The Polydispersity Index (PDI) Is Calculated By

$$PDI = \frac{\overline{M_W}}{\overline{M_N}} \quad (5.79)$$

In this case study, first the complete model is simulated and compared with the results presented in the literature to make sure that it gives reasonable trends. Figure 5-14 shows one of these outputs for the second model in Pallares et al. (2006), which is the same as the irreversible model in our study. In general our model predictions are exactly the same as the results in Pallares et al. (2006), which are consistent with the experimental data.

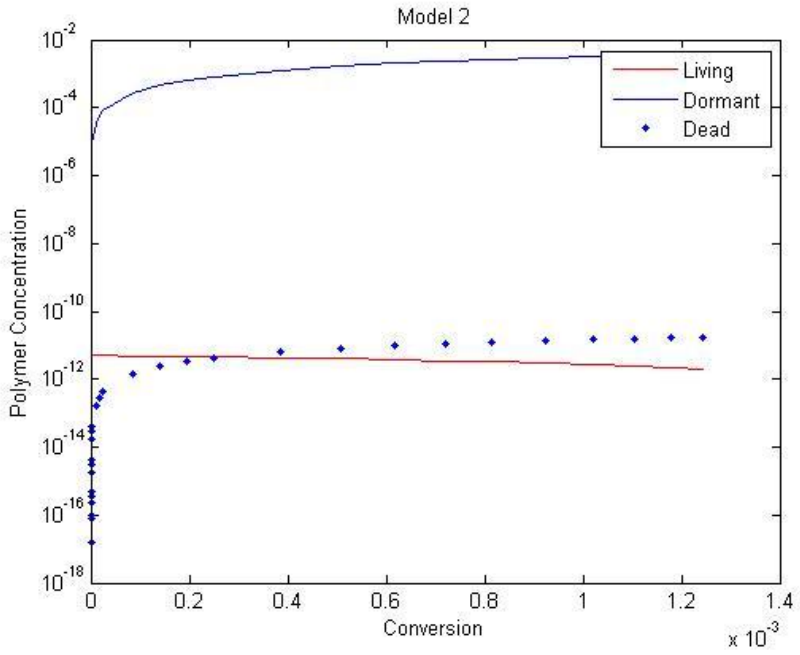

Figure 5-14: Output from the irreversible model

One of the biggest questions in raft polymerization is the order of magnitude of the fragmentation rate constant. Because of that a wide range of values is considered for the parameters. Thus the log of parameters is used in the model discrimination process.

5.4.1 Sensitivity Analysis

In the RAFT models, there are four possible input conditions namely time, $[I_0]$, $[M_0]$, and $[AB_0]$. In this study, time is considered as an input and the three other initial concentrations are assumed fixed at the following values,

$[\mathbf{I_0}] = 0.05$ mol/L
$[M_0] = 5$ mol/L
$[AB_0] = 0.01$ mol/L

We performed a sensitivity analysis to select potential model outputs that could be used for model discrimination purposes. Table 5-22 shows the 7 different models and parameters sets which we used. Cases L1 and L5 are the parameter sets used as real parameters in model discrimination which will be presented later. The other cases are selected to represent values from the lower and the upper ranges of each parameter.

Table 5-22: RAFT, sensitivity analysis conditions

Case	**Model**	**Parameters**
L1	Slow fragmentation	P1=1e6, P2=1e-2, P3=1e4
L2	Slow fragmentation	P1=1e2, P2=1e-2, P3=1e4
L3	Slow fragmentation	P1=1e6, P2=1e4, P3=1e4
L4	Slow fragmentation	P1=1e6, P2=1e-2, P3=1e0
L5	Irreversible	P1=1e6, P2=1e4
L6	Irreversible	P1=1e2, P2=1e4
L7	Irreversible	P1=1e6, P2=1e0

Conversion is the first output tested. Results are shown in Figure 5-15. In this case, L5, L1, and L7 show significant differences but the other cases are not significantly different. According to the small difference between cases in Figure 5-15, the discrimination between models may be possible only if the amount of error in the real conversion measurement is small.

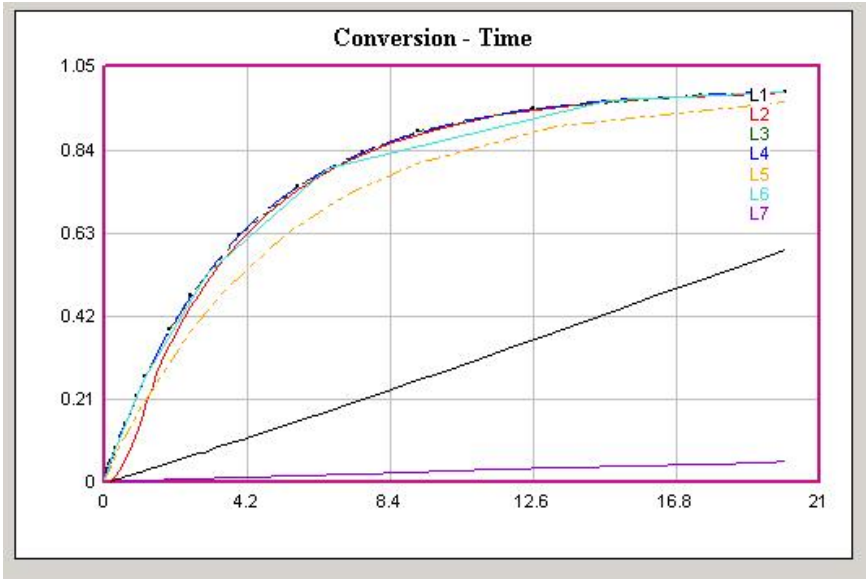

Figure 5-15: RAFT models output, conversion - time (hr)

Adduct concentration is supposed to differ between the two competing models. Figure 5-16, Figure 5-17, Figure 5-18, and Figure 5-19 show adduct concentration versus time. Unfortunately the concentrations of these species are too low to be of practical use for model discrimination.

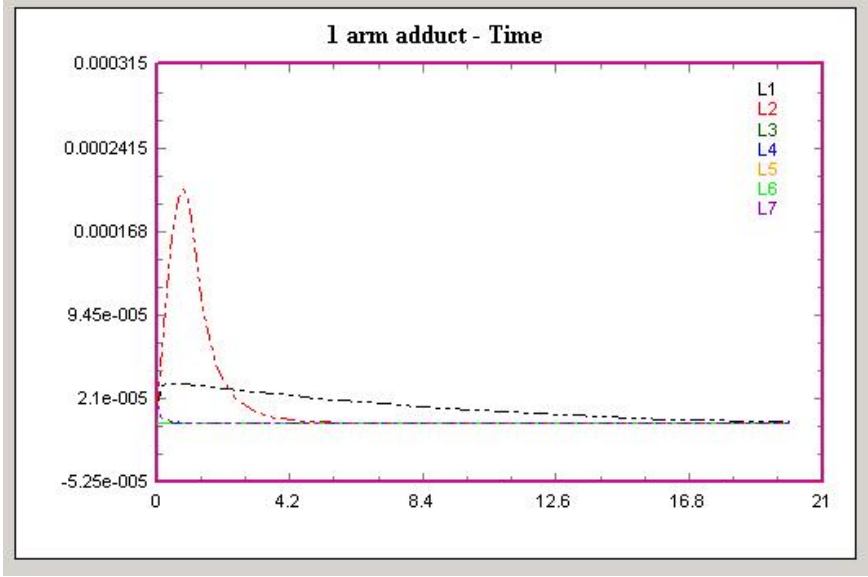

Figure 5-16: RAFT models output, one-arm adduct concentration - time (hr)

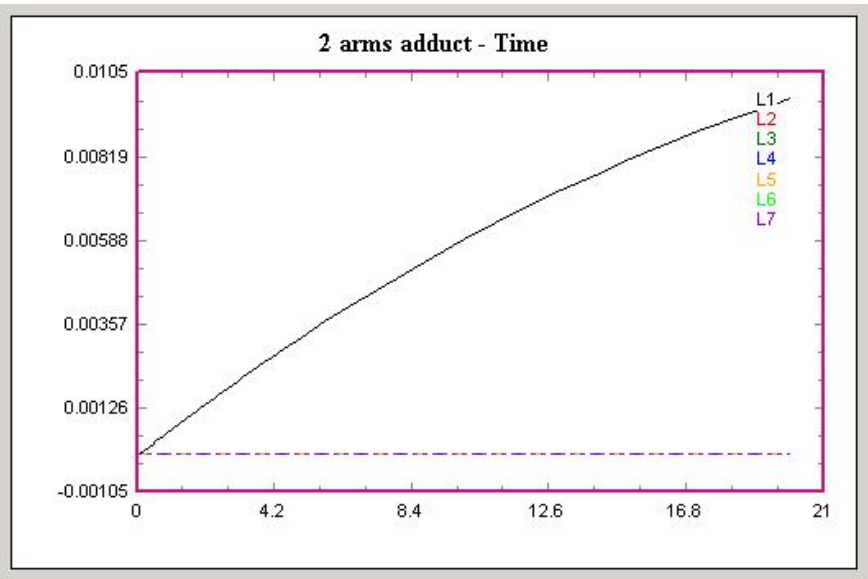

Figure 5-17: RAFT models output, two arm-adduct concentration - time

The log scaled plots are presented in Figures 5-18 and 5-19 to show that the competing models have different outputs but as mentioned the actual concentrations are too low to be of practical use.

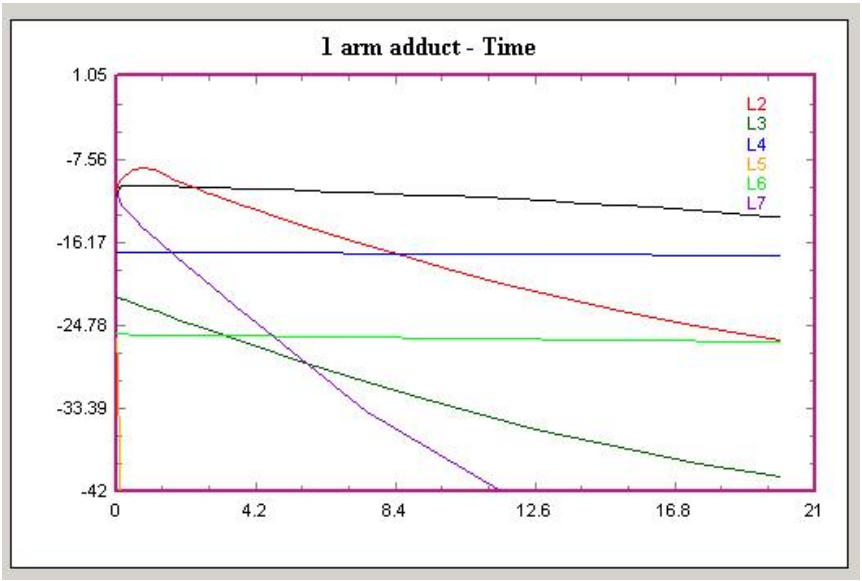

Figure 5-18: RAFT models output, log of one arm adduct concentration - time

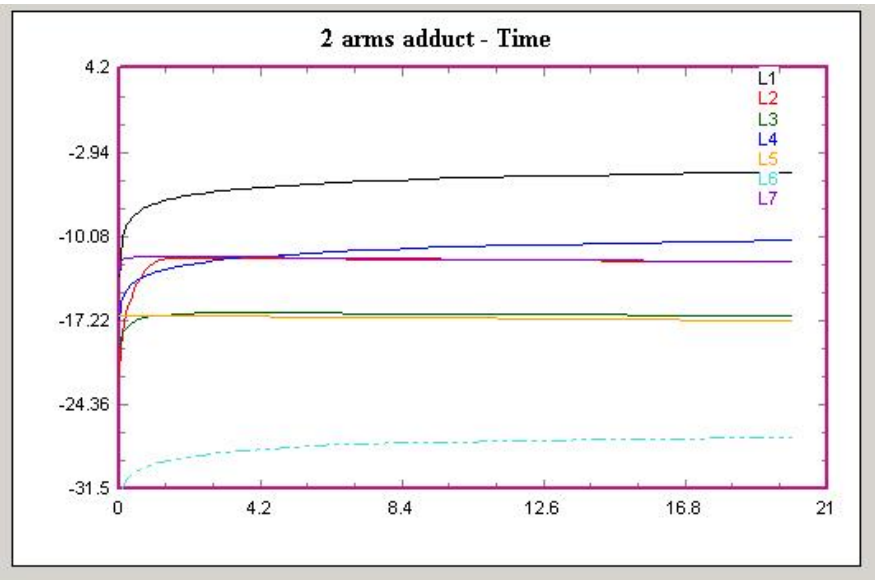

Figure 5-19: RAFT models output, log of two arms adduct concentration - time

The number average molecular weights are shown in Figure 5-20. This plot shows that the L4 and L6 cases are significantly different from the other cases whose outputs are close. Although focusing on the lower M_n values, Figure 5-21 shows that other cases show different behavior too. Thus M_n, could be a suitable choice for model discrimination purposes.

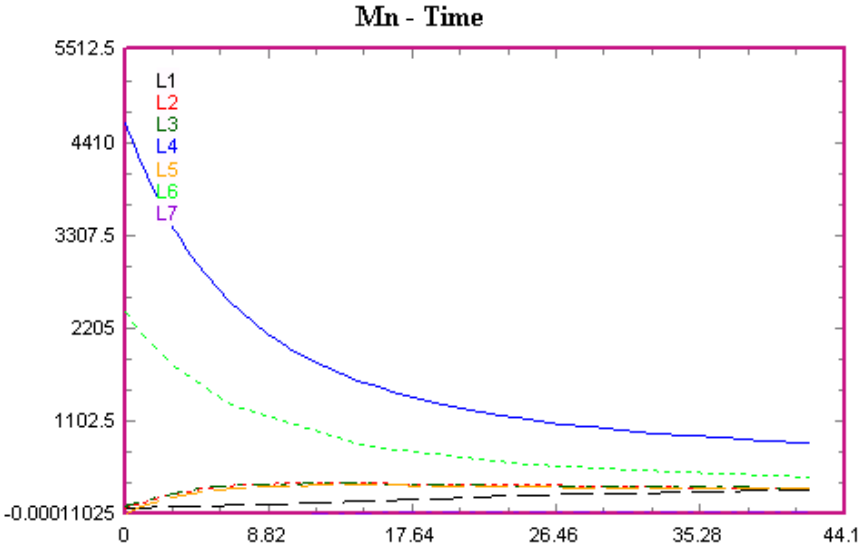

Figure 5-20: RAFT models output, M_n – time

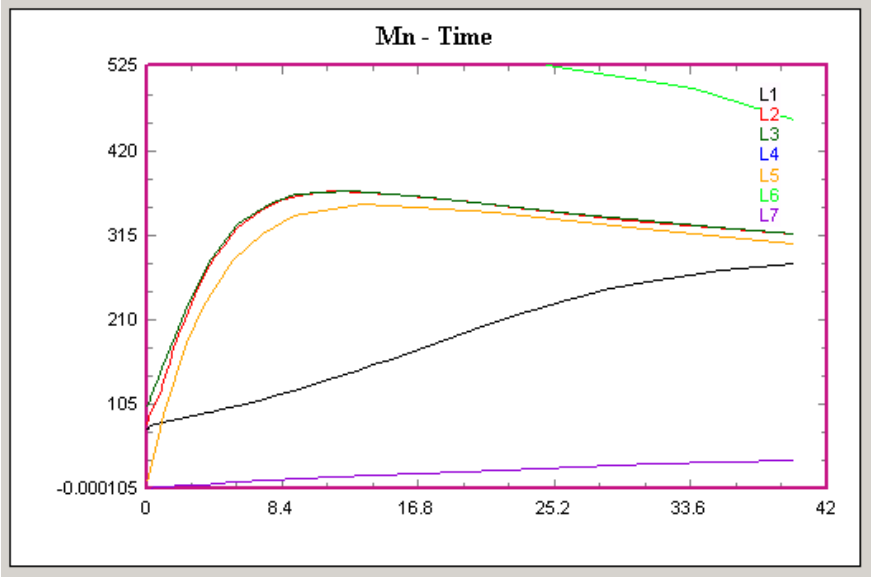

Figure 5-21: RAFT models output, M_n – time

The polydispersity index (PDI) is a function of M_n. Therefore is M_n is useful for model discrimination purposes then PDI may be useful as well. PDI versus time curves are shown in Figure 5-22, and these curves seem significantly different, thereby confirming that PDI could be used as an output for model discrimination purpose.

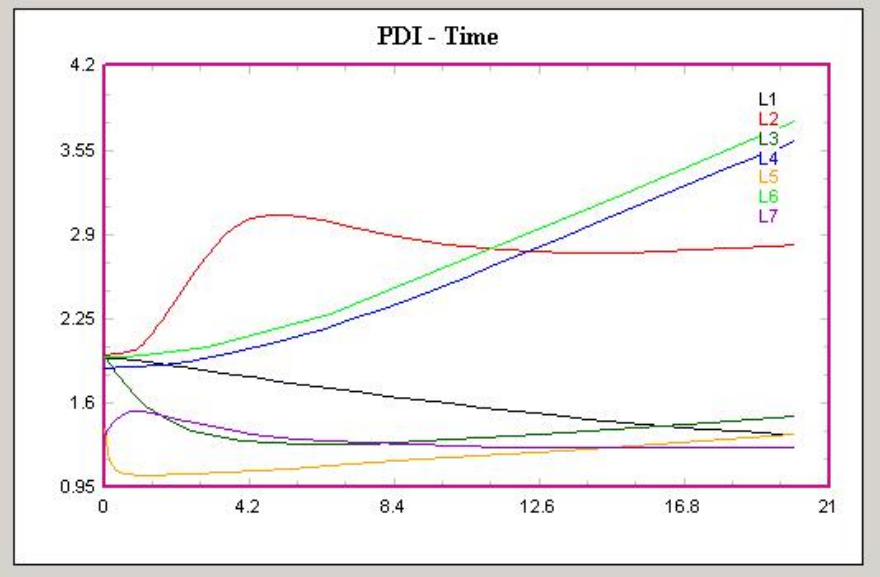

Figure 5-22: RAFT models output, PDI – time

The concentration of the dormant molecule versus time is shown in Figure 5-23. The cases L2, L3, L4 and L6 show close results but the outputs for L1, L5 and L7 are significantly different from the others. Even zoomed plots, for example Figure 5-24, show that L2, L3, L4 and L6 cases output are very close. This leads to the conclusion that the dormant concentration is not be a good output candidate for the model discrimination.

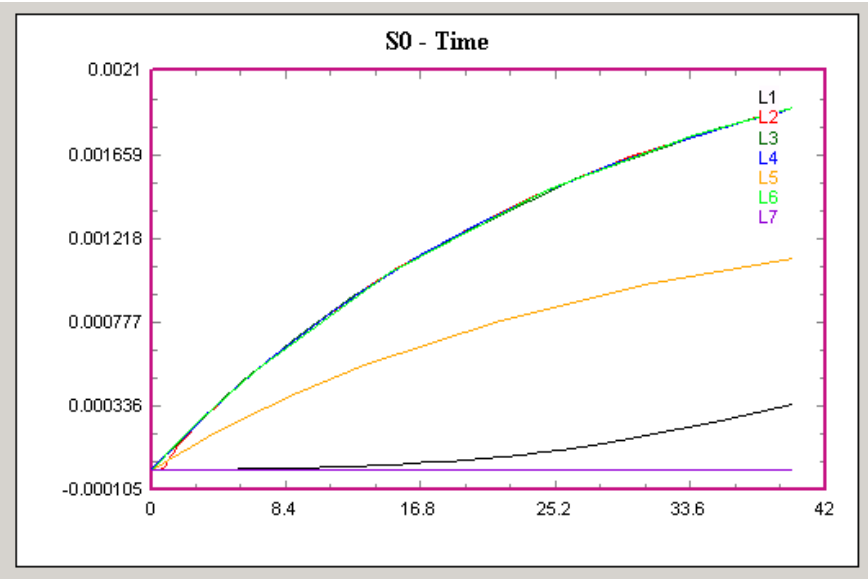

Figure 5-23: RAFT models output, dormant concentration – time

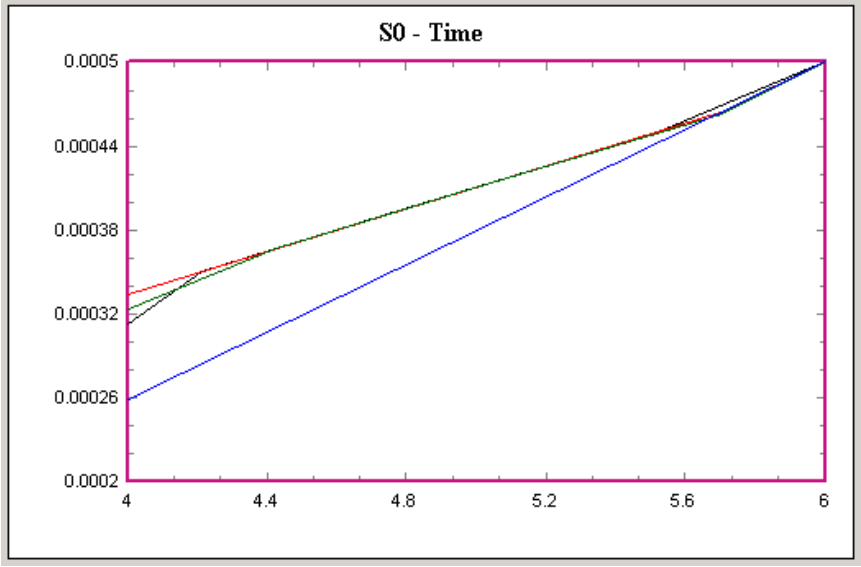

Figure 5-24: RAFT models output, dormant concentration – time

5.4.2 Model Discrimination

According to the sensitivity analysis results, M_n and PDI as well as conversion, x, providing the error is small, could be possible outputs for the purpose of model discrimination between

the slow fragmentation and irreversible models. The first implementation of the SBMCMD framework has been used in this case study,since the competitive models in this case study have more complex structures compared to those in the previous case studies. Therefore it is advantageous to use the adaptive MH sampling to ensure that the entire parameters space is covered with acceptable acceptance ratios.

In this section some different scenarios are considered for model discrimination between the two competing RAFT models using different combinations of the above-mentioned outputs. In the following scenarios, the two competing models will represent the real system in turn. Data from two preliminary experiments are considered at times 2.5 and 5 hours.

The first cases that we tried used conversion with varying amounts of experimental error. We observed that whenever the model discrimination algorithm parameters are tuned so that the acceptance rates are in the recommended range, the procedure can pick the correct model but it is not straightforward to tune these parameters. The model discrimination procedure is sensitive to the level of error considered and the tuning parameters for sampling (H and U), see section 2.4.1.1. In cases where the program stops because of a low acceptance probability, the user can restart the program after tuning the H and U parameters or changing the likelihoods standard deviation. In the restarted program, the last model probabilities should be used as the prior model probabilities and all the observed data should be used as preliminary data in the restarted program. According to the suggestion by Haario et al. (2001) in this study, the first updated covariance matrix is estimated only based on accepted samples. In addition, during the burn-in period, if all the previous H-samples (see equation (2.18)) are the same, the current covariance will be reduced by multiplying it by a constant whose value is less than 1. After,

the burn-in period updating of the covariance is skipped, whenever there are no accepted samples in the last H samples.

The above-mentioned strategies were used to overcome the problem of sampling from the distribution of parameters using the conversion output. While this reduced the number of cases where sampling failed, there were still a significant number of occurrences where samples could not be obtained reliably. While further tuning of the sampling parameters may lead to an acceptable acceptance ratio, MCMC sampling in general was difficult using the conversion output. Therefore we moved on to considering other RAFT output variables.

Table 5-23, Table 5-24 and Table 5-25 show the results of using M_n, as output with and without PDI and conversion. In these tables, "sd. real" and "sd. model", represent the standard deviation value in the normal distribution of errors added to the simulated real data and the likelihood distribution used in the MCMC sampling, respectively. In all scenarios the sd. model value is greater than or equal to the sd. real value. This is done to reflect the fact that the true error variance is normally not known and by using a standard deviation in the likelihood that overestimates the true standard deviation, the approach is a conservative one. In all these scenarios the sampling procedure was tuned easily and the procedure selected the correct model.

Table 5-23: RAFT, Mn output (M1 represents the slow fragmentation and M2 the irreversible model)

Model	sd. Real	sd. model	Num exp.	P_1	P_2
M_1	0	5	1	0.955663	0.044337
M_2	0	5	1	3.81E-27	1
M_1	5	10	2	0.970455	0.029545
M_2	5	10	1	3.85E-05	0.999961
M_1	0	10	2	0.98596	0.01404
M_2	0	10	1	0.019458	0.980542
M_1	10	20	6	0.96412	0.03588
M_2	10	20	4	0.008308	0.991692
M_1	20	20	7	0.953849	0.046151
M_2	20	20	1	0.019814	0.980186

Table 5-24: RAFT, conversion and Mn output (M1 represents the slow fragmentation and M2 the irreversible model)

Model	sd. real	sd. model	Num exp.	P_1	P_2
M_1	[0, 0]	[0.05,10]	1	0.990	0.0099
M_2	[0, 0]	[0.05,10]	1	0	1
M_1	[0.05,10]	[0.05,10]	1	0.999	0.001
M_2	[0.05,10]	[0.05,10]	1	0	1

Table 5-25: RAFT, conversion, Mn and PDI output (M1 represents the slow fragmentation and M2 the irreversible model)

Model	sd. real	sd. model	Num exp.	P_1	P_2
M_1	[0,0,0]	[0.05,10,0.1]	1	0.980	0.020292
M_2	[0,0,0]	[0.05,10,0.1]	1	0	1
M_1	[0.05,10,0.1]	[0.05,10,0.1]	1	0.999	0.000739
M_2	[0.05,10,0.1]	[0.05,10,0.1]	1	0	1

Figure 5-25 to Figure 5-26 show the parameter distributions in the tested scenarios for the correct models. In each figure, the top plot is for the case with no error and the plot in the bottom shows the parameters sample for the test with error added to the outputs. These plots, which are equivalent to confidence interval plots, contain the true parameters value but they are relatively large, implying considerable parameter uncertainty. Thus after model discrimination more experiments are needed for parameter estimation. In Figure 5-25 to Figure 5-27, p_1 and p_2 are normalized values of the parameters. Therefore, they could be substituted to the following equations to find the actual parameter1 and parameter 2 values in Table 5-19.

$$\text{parameter1} = 10^{p_1 * 10}$$
$$\text{parameter2} = 10^{p_2 * 10} \qquad (5.80)$$

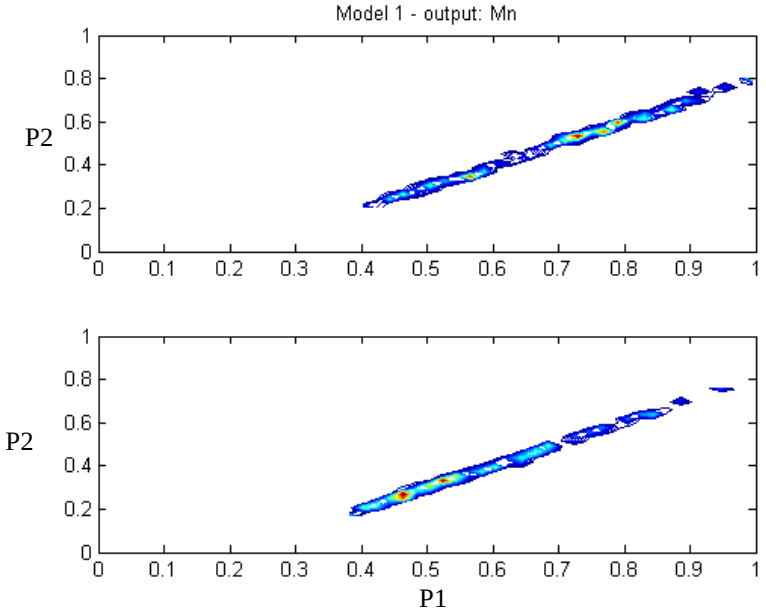

Figure 5-25: Output: M_n, Top: no error, Bottom: with error

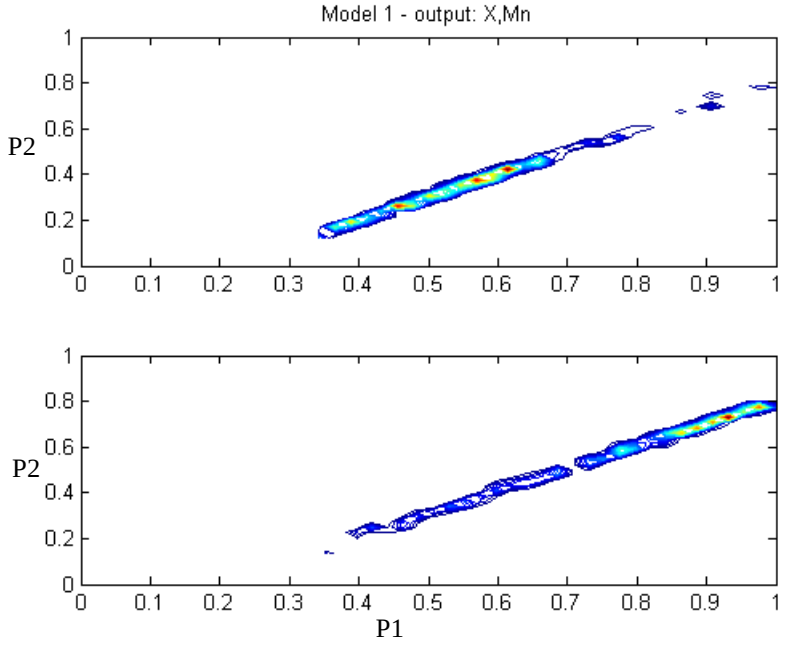

Figure 5-26: Output: x and M_n, Top: no error, Bottom: with error

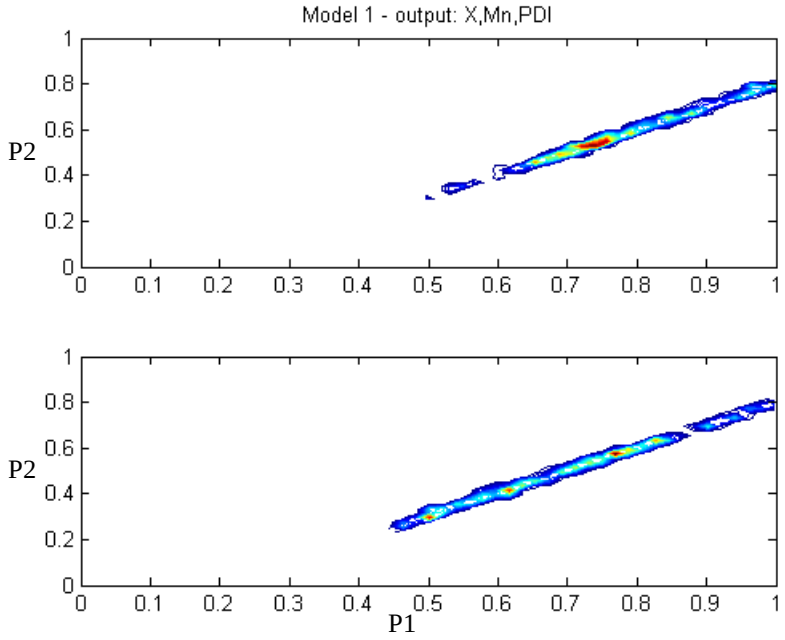

Figure 5-27: Output: x, M_n and PDI, Top: no error, Bottom: with error

Parameters $p_1 - p_3$ in Figure 5-28 to Figure 5-30 must be substituted into equation (5.83) to find parmeter1-parmeter3 in Table 5-19.

$$\begin{aligned} \text{parameter1} &= 10^{p_1 * 10} \\ \text{parameter2} &= 10^{p_2 * 10 - 4} \\ \text{parameter3} &= 10^{p_3 * 10} \end{aligned} \tag{5.81}$$

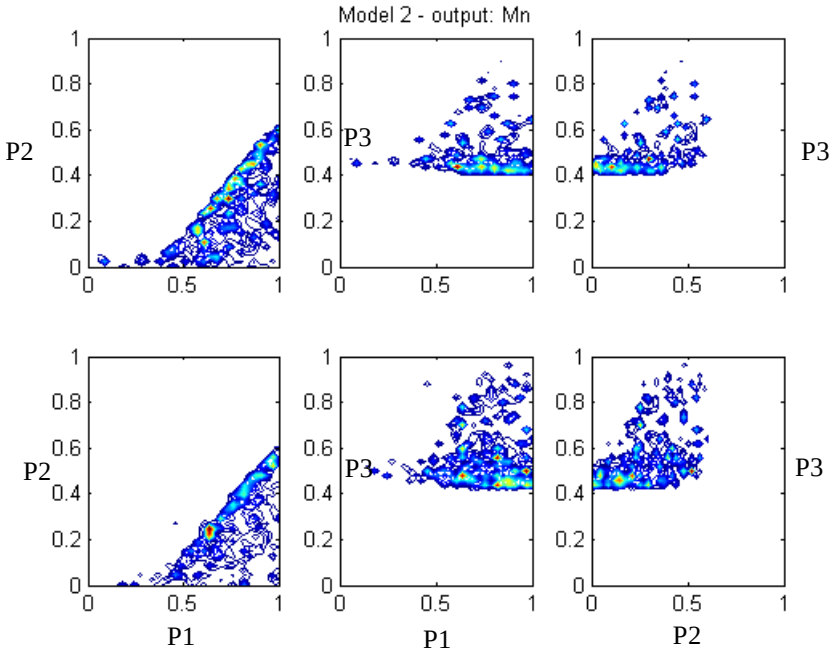

Figure 5-28: Output: M_n, Top: no error, Bottom: with error

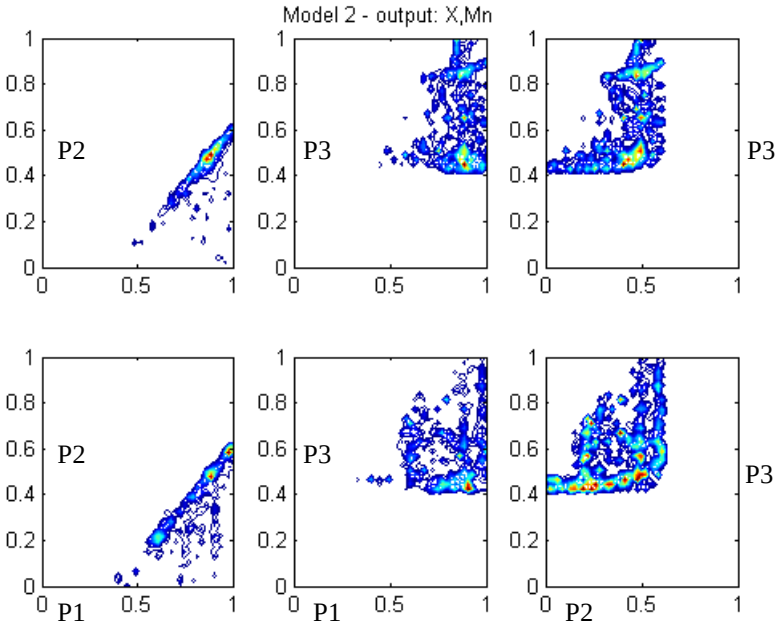

Figure 5-29: Output: x and M_n, Top: no error, Bottom: with error

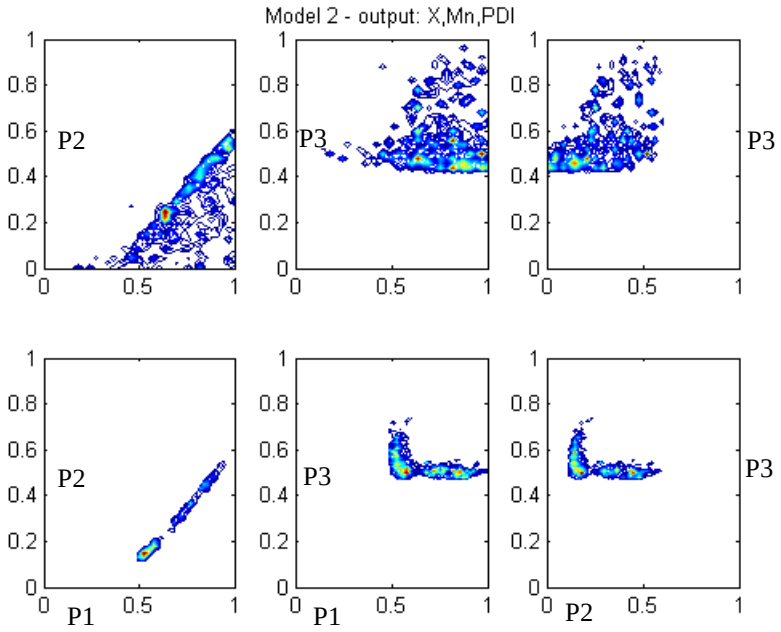

Figure 5-30: Output: x, M_n and PDI, Top: no error, Bottom: with error

According to the results, the SBMCMD procedure can be used for the design of experiments and model discrimination analysis to select the correct model for the RAFT polymerization process. For this purpose, the number average molecular weight is the best output.

Chapter 6
Conclusion and Future Work

6.1 Conclusion

This study discusses a new approach to model discrimination called the Sequential Bayesian Markov Chain Model Discrimination (SBMCMD) procedure. The proposed framework was implemented in three versions. The first version estimates the marginal likelihood using samples from the prior probability distribution of parameters by Adaptive Proposal (AP) Random Walk Metropolis Hastings method. The second and third versions are based on the method introduced by Chib's research group. They sample from the posterior probability distribution of parameters using the Acceptance-Rejection Metropolis Hastings and Random Walk Metropolis Hastings method respectively. The model selection algorithm in the SBMCMD framework is paired with a model discrimination design criterion in a sequential, iterative approach.

The Markov Chain Monte Carlo method samples from the probability distribution of the parameters regardless of whether the model structure is linear or nonlinear. This makes it particularly suitable for model discrimination problems in chemical engineering where mechanistic models are often nonlinear in the parameters. In other words, MCMC model discrimination methods do not require any linearization of the models as many other approaches do. The frameworks worked both for single response and multi response in addition to nested and non-nested models.

This method showed promising results in the case studies presented here. The first two case studies demonstrated the suitability of the SBMCMD as it successfully selected the "correct model". In addition, the effect of error levels was studied in these cases.

Next we applied this method to previously studied copolymerization systems to discriminate between the terminal and the penultimate model which is an ongoing problem in the polymer science community. The SBMCMD successfully discriminated between the two competing models using triad fraction data and showed that the differences between the copolymer composition outputs from the two rival models are too small to be used in model discrimination.

In addition to the SBMCMD, a modified Hsiang and Reilly (1971) method was implemented to address a question arising from a previous study (Burke, 1994).

Finally the SBMCMD framework was applied to the two models proposed for the modeling of RAFT homo-polymerization. The results in this study indicate that the average molecular weight is the best output to use. Also if one of the competing models represents the real mechanism behind the RAFT process, the SBMCMD framework can help to find it.

6.2 Future Work

Future studies to extend this project are in four categories. The first is applying this technique for discriminating between more mechanistic models where there is still ongoing debate about the underlying mechanism or for finding the real mechanism for newly studied systems. In addition, as the case studies results are so promising, the framework can be used in real situations and it seems reliable enough to use with real experiments. So for example two recommendation would be to apply SBMCMD to a real RAFT experimental study and an

experimental study of copolymer systems where there is uncertainty about the underlying model.

The second recommendation relates to developing a criterion to determine in each iteration if model discrimination may be possible with more information or not. The Buzzi-Ferraris method uses a T value for this purpose. An F-test using the variance of differences between the model predicted values and the real output and the variance between model predictions could be used as this criterion. This criterion, which works like a new stopping criterion, should be ignored in the first iteration. It is applicable when the samples of parameter prior become informative. Then the predicted outputs using these prior samples could be used in the newly suggested stopping criterion.

Another recommendation related to the design of experiments step is to study the possibility of the entropy-based DOE methods and comparing the advantage and disadvantage of entropy-based and the maximum divergence-based criteria.

The last recommendation involves making this method available in an easy-to-use and executable program in a parallel computational network. Such a software package could benefit researchers in both universities and industry. In the proposed software package, relatively recent developments in the field of applied statistics involving MCMC methods can be used to solve both the problems of model discrimination and data analysis in a Bayesian approach coupled with DOE. The challenge here will be computation time since, generally, the speed of solving the model within the MCMC cycle should be reasonably fast. Therefore, having to solve complicated models has posed a major obstacle to MCMC implementation of

these problems. This challenge can be addressed by using the recently introduced methods in parallel MCMC (see section 4.3.3) and then improving the already implemented package to fully exploit the advantages of parallel computation. Available statistical packages require a high level of statistical knowledge to be used properly. Providing such a package helps researchers who need to be more focused on their research area rather than on advanced statistical theory and methods. Therefore, these days, most experiments are done without a suitable design even though DOE can save time and cost along with leading to more reliable results.

In the model selection part of this package, more methods than those tried in this research could be used and, in MCMC sampling, new methods could be implemented for the cases where it is not easy to sample from the desired stationary distribution, for example TP MCMC methods (see section 2.3.8)

Appendix A

Results: Order Of Reaction Case Study

Table A-1: Order of reaction case study – average of correct models probability (30 replicates each)

"correct" model :		Model 1	Model 2	Model 3	Model 4
s.d. Real	s.d. lilekihood	Probability of "correct" model			
0.01	0.01	1.00	1.00	0.97	0.97
0.01	0.02	1.00	1.00	0.99	0.99
0.01	0.03	1.00	0.99	0.97	0.98
0.01	0.04	0.99	0.97	0.97	0.97
0.01	0.05	0.98	0.97	0.96	0.96
0.01	0.06	0.97	0.96	0.96	0.96
0.01	0.07	0.97	0.97	0.96	0.96
0.01	0.08	0.97	0.96	0.96	0.96
0.01	0.09	0.96	0.96	0.95	0.95
0.01	0.1	0.96	0.96	0.95	0.95
0.02	0.02	1.00	1.00	0.96	0.99
0.02	0.03	1.00	0.99	0.98	0.98
0.02	0.04	0.99	0.98	0.97	0.97
0.02	0.05	0.99	0.97	0.97	0.97
0.02	0.06	0.98	0.97	0.96	0.96
0.02	0.07	0.97	0.97	0.96	0.96
0.02	0.08	0.97	0.96	0.96	0.96
0.02	0.09	0.96	0.96	0.95	0.96
0.02	0.1	0.96	0.96	0.94	0.95
0.03	0.03	0.97	0.99	0.98	0.99
0.03	0.04	0.99	0.98	0.97	0.98
0.03	0.05	0.99	0.98	0.97	0.96
0.03	0.06	0.98	0.97	0.97	0.97
0.03	0.07	0.97	0.97	0.96	0.96
0.03	0.08	0.97	0.96	0.96	0.96
0.03	0.09	0.97	0.96	0.96	0.96
0.03	0.1	0.96	0.96	0.93	0.95
0.04	0.04	0.99	0.98	0.94	0.97

0.04	0.05	0.99	0.98	0.97	0.97
0.04	0.06	0.98	0.97	0.97	0.97
0.04	0.07	0.97	0.96	0.96	0.96
0.04	0.08	0.97	0.96	0.96	0.96
0.04	0.09	0.97	0.96	0.95	0.96
0.04	0.1	0.97	0.96	0.93	0.95
0.05	0.05	0.95	0.98	0.97	0.98
0.05	0.06	0.98	0.97	0.97	0.96
0.05	0.07	0.98	0.97	0.91	0.96
0.05	0.08	0.97	0.97	0.95	0.95
0.05	0.09	0.97	0.96	0.94	0.93
0.05	0.1	0.97	0.96	0.91	0.95
0.06	0.06	0.98	0.91	0.91	0.95
0.06	0.07	0.98	0.97	0.91	0.95
0.06	0.08	0.97	0.97	0.91	0.94
0.06	0.09	0.97	0.94	0.93	0.92
0.06	0.1	0.97	0.96	0.92	0.93
0.07	0.07	0.98	0.90	0.92	0.93
0.07	0.08	0.98	0.94	0.96	0.94
0.07	0.09	0.97	0.97	0.91	0.87
0.07	0.1	0.97	0.96	0.88	0.93
0.08	0.08	0.98	0.97	0.90	0.89
0.08	0.09	0.98	0.97	0.81	0.95
0.08	0.1	0.97	0.95	0.85	0.89
0.09	0.09	0.94	0.94	0.82	0.89
0.09	0.1	0.97	0.95	0.79	0.90
0.1	0.1	0.96	0.96	0.81	0.78

Appendix B

Results: Oxidation Case Study

Table B-1: Oxidation case study - 315 runs result

#	M	R.E.	M.E.	N	P1	P2	P3	#	M	R.E.	M.E.	N	P1	P2	P3
1	1	1.00	1.00	0.00	1.00	0.00	0.00	159	2	2.00	6.00	19.00	0.00	0.84	0.16
2	1	1.00	1.00	0.00	1.00	0.00	0.00	160	2	2.00	6.00	15.00	0.00	1.00	0.00
3	1	1.00	1.00	0.00	1.00	0.00	0.00	161	2	3.00	3.00	7.00	0.00	0.99	0.01
4	1	1.00	1.00	0.00	1.00	0.00	0.00	162	2	3.00	3.00	6.00	0.00	0.99	0.00
5	1	1.00	1.00	0.00	1.00	0.00	0.00	163	2	3.00	3.00	6.00	0.00	1.00	0.00
6	1	1.00	2.00	0.00	1.00	0.00	0.00	164	2	3.00	3.00	1.00	0.00	0.96	0.04
7	1	1.00	2.00	0.00	1.00	0.00	0.00	165	2	3.00	3.00	2.00	0.00	0.98	0.02
8	1	1.00	2.00	0.00	1.00	0.00	0.00	166	2	3.00	4.00	19.00	0.00	0.95	0.05
9	1	1.00	2.00	0.00	1.00	0.00	0.00	167	2	3.00	4.00	19.00	0.00	0.86	0.14
10	1	1.00	2.00	0.00	1.00	0.00	0.00	168	2	3.00	4.00	14.00	0.00	0.97	0.03
11	1	1.00	3.00	0.00	1.00	0.00	0.00	169	2	3.00	4.00	3.00	0.00	1.00	0.00
12	1	1.00	3.00	0.00	1.00	0.00	0.00	170	2	3.00	4.00	7.00	0.00	0.95	0.05
13	1	1.00	3.00	0.00	1.00	0.00	0.00	171	2	3.00	5.00	19.00	0.00	0.93	0.07
14	1	1.00	3.00	0.00	1.00	0.00	0.00	172	2	3.00	5.00	6.00	0.00	0.97	0.03
15	1	1.00	3.00	0.00	1.00	0.00	0.00	173	2	3.00	5.00	5.00	0.00	1.00	0.00
16	1	1.00	4.00	0.00	1.00	0.00	0.00	174	2	3.00	5.00	19.00	0.00	0.90	0.10
17	1	1.00	4.00	0.00	1.00	0.00	0.00	175	2	3.00	5.00	19.00	0.00	0.91	0.09
18	1	1.00	4.00	0.00	1.00	0.00	0.00	176	2	3.00	6.00	19.00	0.00	0.77	0.23
19	1	1.00	4.00	0.00	1.00	0.00	0.00	177	2	3.00	6.00	6.00	0.00	0.99	0.01
20	1	1.00	4.00	0.00	0.99	0.00	0.00	178	2	3.00	6.00	2.00	0.00	0.95	0.05
21	1	1.00	5.00	0.00	0.96	0.02	0.03	179	2	3.00	6.00	11.00	0.00	1.00	0.00
22	1	1.00	5.00	0.00	0.98	0.00	0.01	180	2	3.00	6.00	15.00	0.00	0.99	0.01
23	1	1.00	5.00	0.00	0.97	0.00	0.02	181	2	4.00	4.00	6.00	0.00	0.96	0.04
24	1	1.00	5.00	1.00	1.00	0.00	0.00	182	2	4.00	4.00	0.00	0.00	0.04	0.95
25	1	1.00	5.00	0.00	0.95	0.02	0.03	183	2	4.00	4.00	2.00	0.00	1.00	0.00
26	1	1.00	6.00	1.00	1.00	0.00	0.00	184	2	4.00	4.00	19.00	0.00	0.71	0.29
27	1	1.00	6.00	1.00	1.00	0.00	0.00	185	2	4.00	4.00	9.00	0.00	0.00	1.00
28	1	1.00	6.00	1.00	1.00	0.00	0.00	186	2	4.00	5.00	14.00	0.00	0.99	0.01
29	1	1.00	6.00	1.00	1.00	0.00	0.00	187	2	4.00	5.00	19.00	0.00	0.65	0.35
30	1	1.00	6.00	1.00	1.00	0.00	0.00	188	2	4.00	5.00	15.00	0.00	0.98	0.02
31	1	2.00	2.00	0.00	1.00	0.00	0.00	189	2	4.00	5.00	19.00	0.00	0.57	0.43
32	1	2.00	2.00	0.00	1.00	0.00	0.00	190	2	4.00	5.00	6.00	0.00	1.00	0.00
33	1	2.00	2.00	0.00	1.00	0.00	0.00	191	2	4.00	6.00	19.00	0.00	0.85	0.15
34	1	2.00	2.00	0.00	1.00	0.00	0.00	192	2	4.00	6.00	10.00	0.00	0.95	0.05
35	1	2.00	2.00	0.00	1.00	0.00	0.00	193	2	4.00	6.00	16.00	0.00	0.97	0.03
36	1	2.00	3.00	0.00	1.00	0.00	0.00	194	2	4.00	6.00	19.00	0.00	0.42	0.58
37	1	2.00	3.00	0.00	1.00	0.00	0.00	195	2	4.00	6.00	5.00	0.00	0.04	0.96
38	1	2.00	3.00	0.00	1.00	0.00	0.00	196	2	5.00	5.00	16.00	0.00	0.95	0.05

39	1	2.00	3.00	0.00	1.00	0.00	0.00	197	2	5.00	5.00	4.00	0.00	1.00	0.00
40	1	2.00	3.00	0.00	1.00	0.00	0.00	198	2	5.00	5.00	3.00	0.00	1.00	0.00
41	1	2.00	4.00	0.00	0.99	0.00	0.00	199	2	5.00	5.00	9.00	0.00	1.00	0.00
42	1	2.00	4.00	0.00	1.00	0.00	0.00	200	2	5.00	5.00	7.00	0.00	0.98	0.02
43	1	2.00	4.00	0.00	0.97	0.00	0.02	201	2	5.00	6.00	3.00	0.00	1.00	0.00
44	1	2.00	4.00	0.00	0.97	0.01	0.01	202	2	5.00	6.00	19.00	0.00	0.34	0.66
45	1	2.00	4.00	0.00	1.00	0.00	0.00	203	2	5.00	6.00	5.00	0.00	1.00	0.00
46	1	2.00	5.00	0.00	0.98	0.00	0.01	204	2	5.00	6.00	4.00	0.00	0.00	1.00
47	1	2.00	5.00	0.00	0.96	0.03	0.01	205	2	5.00	6.00	10.00	0.00	1.00	0.00
48	1	2.00	5.00	0.00	0.99	0.00	0.00	206	2	6.00	6.00	19.00	0.00	0.73	0.27
49	1	2.00	5.00	1.00	1.00	0.00	0.00	207	2	6.00	6.00	15.00	0.00	0.99	0.00
50	1	2.00	5.00	0.00	0.97	0.02	0.02	208	2	6.00	6.00	14.00	0.00	0.96	0.04
51	1	2.00	6.00	1.00	1.00	0.00	0.00	209	2	6.00	6.00	4.00	0.00	0.03	0.97
52	1	2.00	6.00	1.00	1.00	0.00	0.00	210	2	6.00	6.00	3.00	0.00	0.98	0.02
53	1	2.00	6.00	0.00	0.95	0.01	0.03	211	3	1.00	1.00	1.00	0.00	0.00	1.00
54	1	2.00	6.00	1.00	0.96	0.00	0.04	212	3	1.00	1.00	0.00	0.00	0.00	1.00
55	1	2.00	6.00	1.00	1.00	0.00	0.00	213	3	1.00	1.00	1.00	0.00	0.00	1.00
56	1	3.00	3.00	0.00	1.00	0.00	0.00	214	3	1.00	1.00	4.00	0.00	0.00	1.00
57	1	3.00	3.00	0.00	1.00	0.00	0.00	215	3	1.00	1.00	7.00	0.00	0.04	0.96
58	1	3.00	3.00	0.00	1.00	0.00	0.00	216	3	1.00	2.00	2.00	0.00	0.02	0.98
59	1	3.00	3.00	0.00	1.00	0.00	0.00	217	3	1.00	2.00	3.00	0.00	0.00	1.00
60	1	3.00	3.00	0.00	1.00	0.00	0.00	218	3	1.00	2.00	7.00	0.00	0.00	1.00
61	1	3.00	4.00	1.00	1.00	0.00	0.00	219	3	1.00	2.00	10.00	0.00	0.00	1.00
62	1	3.00	4.00	1.00	1.00	0.00	0.00	220	3	1.00	2.00	5.00	0.00	0.00	1.00
63	1	3.00	4.00	1.00	1.00	0.00	0.00	221	3	1.00	3.00	11.00	0.00	0.00	1.00
64	1	3.00	4.00	1.00	1.00	0.00	0.00	222	3	1.00	3.00	3.00	0.00	0.00	1.00
65	1	3.00	4.00	2.00	1.00	0.00	0.00	223	3	1.00	3.00	1.00	0.00	0.00	1.00
66	1	3.00	5.00	0.00	0.97	0.01	0.02	224	3	1.00	3.00	2.00	0.00	0.00	0.99
67	1	3.00	5.00	0.00	0.99	0.00	0.00	225	3	1.00	3.00	1.00	0.00	0.00	1.00
68	1	3.00	5.00	0.00	0.95	0.02	0.03	226	3	1.00	4.00	7.00	0.00	0.04	0.96
69	1	3.00	5.00	0.00	0.97	0.02	0.02	227	3	1.00	4.00	8.00	0.00	0.04	0.96
70	1	3.00	5.00	0.00	0.96	0.02	0.02	228	3	1.00	4.00	11.00	0.00	0.00	1.00
71	1	3.00	6.00	0.00	1.00	0.00	0.00	229	3	1.00	4.00	5.00	0.00	0.00	1.00
72	1	3.00	6.00	0.00	0.97	0.01	0.02	230	3	1.00	4.00	3.00	0.00	0.00	1.00
73	1	3.00	6.00	1.00	1.00	0.00	0.00	231	3	1.00	5.00	12.00	0.00	0.00	1.00
74	1	3.00	6.00	0.00	0.96	0.02	0.03	232	3	1.00	5.00	19.00	0.00	0.07	0.93
75	1	3.00	6.00	1.00	1.00	0.00	0.00	233	3	1.00	5.00	7.00	0.00	0.03	0.97
76	1	4.00	4.00	0.00	0.99	0.00	0.00	234	3	1.00	5.00	6.00	0.00	0.00	1.00
77	1	4.00	4.00	0.00	1.00	0.00	0.00	235	3	1.00	5.00	1.00	0.00	0.00	1.00
78	1	4.00	4.00	0.00	1.00	0.00	0.00	236	3	1.00	6.00	9.00	0.00	0.03	0.97
79	1	4.00	4.00	0.00	1.00	0.00	0.00	237	3	1.00	6.00	10.00	0.00	0.00	1.00
80	1	4.00	4.00	0.00	1.00	0.00	0.00	238	3	1.00	6.00	6.00	0.00	0.03	0.97
81	1	4.00	5.00	0.00	0.98	0.00	0.01	239	3	1.00	6.00	5.00	0.00	0.00	1.00
82	1	4.00	5.00	1.00	1.00	0.00	0.00	240	3	1.00	6.00	8.00	0.00	0.00	1.00
83	1	4.00	5.00	1.00	1.00	0.00	0.00	241	3	2.00	2.00	6.00	0.00	0.00	1.00
84	1	4.00	5.00	1.00	1.00	0.00	0.00	242	3	2.00	2.00	7.00	0.00	0.03	0.97
85	1	4.00	5.00	1.00	1.00	0.00	0.00	243	3	2.00	2.00	3.00	0.00	0.00	1.00
86	1	4.00	6.00	0.00	0.96	0.02	0.02	244	3	2.00	2.00	5.00	0.00	0.00	1.00

87	1	4.00	6.00	1.00	0.99	0.00	0.00	245	3	2.00	2.00	1.00	0.00	0.04	0.96
88	1	4.00	6.00	1.00	1.00	0.00	0.00	246	3	2.00	3.00	11.00	0.00	0.00	1.00
89	1	4.00	6.00	1.00	1.00	0.00	0.00	247	3	2.00	3.00	17.00	0.00	0.04	0.96
90	1	4.00	6.00	0.00	0.98	0.00	0.00	248	3	2.00	3.00	5.00	0.00	0.00	1.00
91	1	5.00	5.00	0.00	0.97	0.02	0.01	249	3	2.00	3.00	7.00	0.00	0.00	1.00
92	1	5.00	5.00	0.00	0.96	0.00	0.03	250	3	2.00	3.00	1.00	0.00	0.03	0.97
93	1	5.00	5.00	0.00	0.99	0.00	0.00	251	3	2.00	4.00	3.00	0.00	0.00	1.00
94	1	5.00	5.00	0.00	0.98	0.00	0.00	252	3	2.00	4.00	8.00	0.00	0.05	0.95
95	1	5.00	5.00	0.00	0.99	0.00	0.00	253	3	2.00	4.00	6.00	0.00	0.05	0.95
96	1	5.00	6.00	0.00	0.98	0.00	0.00	254	3	2.00	4.00	8.00	0.00	0.05	0.95
97	1	5.00	6.00	1.00	1.00	0.00	0.00	255	3	2.00	4.00	12.00	0.00	0.05	0.95
98	1	5.00	6.00	1.00	1.00	0.00	0.00	256	3	2.00	5.00	8.00	0.00	0.02	0.98
99	1	5.00	6.00	1.00	1.00	0.00	0.00	257	3	2.00	5.00	15.00	0.00	0.01	0.99
100	1	5.00	6.00	2.00	1.00	0.00	0.00	258	3	2.00	5.00	5.00	0.00	0.00	1.00
101	1	6.00	6.00	1.00	1.00	0.00	0.00	259	3	2.00	5.00	8.00	0.00	0.01	0.99
102	1	6.00	6.00	1.00	1.00	0.00	0.00	260	3	2.00	5.00	2.00	0.00	0.00	1.00
103	1	6.00	6.00	0.00	0.99	0.00	0.00	261	3	2.00	6.00	5.00	0.00	0.00	1.00
104	1	6.00	6.00	0.00	0.95	0.02	0.03	262	3	2.00	6.00	6.00	0.00	0.04	0.96
105	1	6.00	6.00	0.00	0.97	0.00	0.02	263	3	2.00	6.00	14.00	0.00	0.00	1.00
106	2	1.00	1.00	0.00	0.00	0.03	0.97	264	3	2.00	6.00	6.00	0.00	0.00	1.00
107	2	1.00	1.00	0.00	0.00	0.00	1.00	265	3	2.00	6.00	9.00	0.00	0.01	0.99
108	2	1.00	1.00	0.00	0.00	0.00	1.00	266	3	3.00	3.00	2.00	0.00	0.04	0.96
109	2	1.00	1.00	4.00	0.00	0.95	0.05	267	3	3.00	3.00	8.00	0.00	0.01	0.99
110	2	1.00	1.00	2.00	0.00	1.00	0.00	268	3	3.00	3.00	8.00	0.00	0.03	0.97
111	2	1.00	2.00	0.00	0.00	0.01	0.99	269	3	3.00	3.00	1.00	0.00	0.01	0.98
112	2	1.00	2.00	7.00	0.00	1.00	0.00	270	3	3.00	3.00	3.00	0.00	0.00	1.00
113	2	1.00	2.00	2.00	0.00	0.98	0.02	271	3	3.00	4.00	7.00	0.00	0.01	0.99
114	2	1.00	2.00	11.00	0.00	0.05	0.95	272	3	3.00	4.00	2.00	0.00	0.04	0.96
115	2	1.00	2.00	5.00	0.00	1.00	0.00	273	3	3.00	4.00	6.00	0.00	0.03	0.97
116	2	1.00	3.00	4.00	0.00	0.96	0.04	274	3	3.00	4.00	4.00	0.00	0.00	1.00
117	2	1.00	3.00	8.00	0.00	1.00	0.00	275	3	3.00	4.00	4.00	0.00	0.03	0.97
118	2	1.00	3.00	3.00	0.00	1.00	0.00	276	3	3.00	5.00	3.00	0.00	0.00	1.00
119	2	1.00	3.00	11.00	0.00	1.00	0.00	277	3	3.00	5.00	4.00	0.00	0.00	1.00
120	2	1.00	3.00	13.00	0.00	1.00	0.00	278	3	3.00	5.00	10.00	0.00	0.03	0.97
121	2	1.00	4.00	16.00	0.00	0.96	0.04	279	3	3.00	5.00	8.00	0.00	0.05	0.95
122	2	1.00	4.00	19.00	0.00	0.94	0.06	280	3	3.00	5.00	3.00	0.00	0.02	0.98
123	2	1.00	4.00	19.00	0.00	0.91	0.09	281	3	3.00	6.00	12.00	0.00	0.00	1.00
124	2	1.00	4.00	5.00	0.00	0.96	0.04	282	3	3.00	6.00	15.00	0.00	0.00	1.00
125	2	1.00	4.00	8.00	0.00	0.99	0.00	283	3	3.00	6.00	10.00	0.00	0.04	0.96
126	2	1.00	5.00	19.00	0.00	0.33	0.67	284	3	3.00	6.00	10.00	0.00	0.00	1.00
127	2	1.00	5.00	16.00	0.00	0.99	0.00	285	3	3.00	6.00	16.00	0.00	0.05	0.95
128	2	1.00	5.00	1.00	0.03	0.97	0.00	286	3	4.00	4.00	8.00	0.00	0.02	0.98
129	2	1.00	5.00	9.00	0.00	0.96	0.04	287	3	4.00	4.00	5.00	0.00	0.02	0.98
130	2	1.00	5.00	3.00	0.00	1.00	0.00	288	3	4.00	4.00	2.00	0.00	0.00	1.00
131	2	1.00	6.00	16.00	0.00	0.96	0.04	289	3	4.00	4.00	15.00	0.00	0.04	0.96
132	2	1.00	6.00	19.00	0.00	0.95	0.05	290	3	4.00	4.00	11.00	0.00	0.00	1.00
133	2	1.00	6.00	19.00	0.00	0.90	0.10	291	3	4.00	5.00	9.00	0.00	0.00	1.00
134	2	1.00	6.00	16.00	0.00	1.00	0.00	292	3	4.00	5.00	4.00	0.00	0.00	1.00

135	2	1.00	6.00	10.00	0.00	1.00	0.00	293	3	4.00	5.00	5.00	0.00	0.00	1.00
136	2	2.00	2.00	12.00	0.00	0.99	0.01	294	3	4.00	5.00	6.00	0.00	0.00	0.99
137	2	2.00	2.00	2.00	0.00	1.00	0.00	295	3	4.00	5.00	6.00	0.00	0.05	0.95
138	2	2.00	2.00	2.00	0.00	1.00	0.00	296	3	4.00	6.00	19.00	0.00	0.07	0.93
139	2	2.00	2.00	8.00	0.00	1.00	0.00	297	3	4.00	6.00	6.00	0.00	0.00	1.00
140	2	2.00	2.00	1.00	0.00	0.97	0.03	298	3	4.00	6.00	7.00	0.00	0.02	0.98
141	2	2.00	3.00	7.00	0.00	1.00	0.00	299	3	4.00	6.00	5.00	0.00	0.00	1.00
142	2	2.00	3.00	5.00	0.00	0.99	0.01	300	3	4.00	6.00	8.00	0.00	0.05	0.95
143	2	2.00	3.00	11.00	0.00	1.00	0.00	301	3	5.00	5.00	2.00	0.00	0.02	0.98
144	2	2.00	3.00	17.00	0.00	0.97	0.03	302	3	5.00	5.00	1.00	0.00	0.00	1.00
145	2	2.00	3.00	5.00	0.00	0.98	0.02	303	3	5.00	5.00	4.00	0.00	0.00	1.00
146	2	2.00	4.00	1.00	0.00	1.00	0.00	304	3	5.00	5.00	5.00	0.00	0.04	0.96
147	2	2.00	4.00	2.00	0.00	1.00	0.00	305	3	5.00	5.00	1.00	0.02	0.00	0.97
148	2	2.00	4.00	11.00	0.00	0.96	0.04	306	3	5.00	6.00	7.00	0.00	0.00	0.99
149	2	2.00	4.00	10.00	0.00	1.00	0.00	307	3	5.00	6.00	4.00	0.00	0.00	1.00
150	2	2.00	4.00	8.00	0.00	1.00	0.00	308	3	5.00	6.00	15.00	0.00	0.00	1.00
151	2	2.00	5.00	12.00	0.00	0.97	0.03	309	3	5.00	6.00	7.00	0.00	0.04	0.96
152	2	2.00	5.00	3.00	0.00	1.00	0.00	310	3	5.00	6.00	3.00	0.00	0.00	1.00
153	2	2.00	5.00	6.00	0.00	0.98	0.02	311	3	6.00	6.00	13.00	0.00	0.03	0.97
154	2	2.00	5.00	19.00	0.00	1.00	0.00	312	3	6.00	6.00	1.00	0.00	0.00	1.00
155	2	2.00	5.00	8.00	0.00	0.95	0.05	313	3	6.00	6.00	18.00	0.00	0.00	1.00
156	2	2.00	6.00	2.00	0.00	1.00	0.00	314	3	6.00	6.00	7.00	0.00	0.00	0.99
157	2	2.00	6.00	2.00	0.00	0.99	0.00	315	3	6.00	6.00	6.00	0.00	0.04	0.96
158	2	2.00	6.00	19.00	0.00	0.42	0.58								

Appendix C

Simulation Conditions of Copolymerization Models

Table C-1: Parameter values for the homo-polymerization of Styrene

Constant		
MW	104.1512	g/mol
k_p	$1.09 \times 10^7 \exp(-7051/RT)$	L/(mol. s)
k_t	$1.703 \times 10^9 \exp(-2268/RT)$	L/(mol. s)
k_{fM}	1.096482	L/(mol. s)
ρ_M	$0.924 - 9.18 \times 10^{-4}(T - 273.15)$	g/cm^3
ρ_P	$1.084 - 6.05 \times 10^{-4}(T - 273.15)$	g/cm^3
γ	0.00	-

Table C-2: Parameter values for the homo-polymerization of Methyl Methacrylate

Constant		
MW	100.1162	g/mol
k_p	5.365859	L
k_t	$9.8 \times 10^7 \exp(-701/RT)$	L
k_{fM}	1.557243	L
ρ_M	$0.966471 - 1.16 \times 10^{-3}$(T	g/cm^3
ρ_P	$1.19343 - 2.8 \times 10^{-4}$(T	g/cm^3
γ	0.84	-

Table C-3: Reactivity ratios for the modeling of the STY/MMA [1]

Para	Value
r_{11}	*0.472*
r_{21}	*0.472*
r_{22}	*0.454*
r_{12}	*0.454*
s_1	*0.412*
s_2	*0.170*

[1] Reactivity ratios are taken from O'Driscoll and Huang (1989)

k_{t1}[2]	$\frac{k_{t1}r_{21}N_1^2\frac{(r_{11}N_1+N_2)}{(r_{21}N_1+N_2)}+2(k_{t1}k_{t2})^{0.5}N_1N_2+k_{t2}r_{12}N_2^2\frac{(r_{22}}{(r_{12}}}{r_{21}N_1^2\frac{(r_{11}N_1+N_2)}{(r_{21}N_1+N_2)}+2N_1N_2+r_{12}N_2^2\frac{(r_{22}N_2+N_1}{(r_{12}N_2+N_1}}$
$k_{f,M}$	$F_1k_{f,M_1}+(1-F_1)k_{f,M_2}$
ρ	$F_1\rho_1+(1-F_1)\rho\_2$
ϵ	$\frac{\epsilon_1r_{21}N_1^2\frac{(r_{11}N_1+N_2)}{(r_{21}N_1+N_2)}+2\epsilon_{12}N_1N_2+\epsilon_2r_{12}N_2^2\frac{(r_{22}N_2+N_1)}{(r_{12}N_2+N_1)}}{r_{21}N_1^2\frac{(r_{11}N_1+N_2)}{(r_{21}N_1+N_2)}+2N_1N_2+r_{12}N_2^2\frac{(r_{22}N_2+N_1)}{(r_{12}N_2+N_1)}}$
ϵ_1	-0.170
ϵ_{12}	-0.227
ϵ_2	-0.264
$\frac{dV}{dt}$[4]	$V_0\epsilon\frac{dx}{dt}$

Table C-4: Parameter values for the homo-polymerization of Acrylonitrile

Constant		
MW	53.0634	g/mol
k_p	$1.047\times10^8\exp(-7278.38/RT)$	L
k_t	$2.95\times10^{11}\exp(-5396.88/RT)$	L
k_{fM}	1.090932	L
ρ_M	$0.835549-1.38286\times10^{-3}\exp(T$ $-273.15)$	g/cm^3
ρ_P	1.17	g/cm^3
γ	0.00	-

[2] STY is monomer 1

[3] Homo-polymerization rate constant was decreased by a factor of 1.09 by Burke in order to improve the prediction for the particular data set

[4] instead of using an overall copolymer density, Burke used the volume contraction factor for the simulation of STY/MMA

Table C-5: Parameter values for the modeling of the STY/AN

Parameter	Value	
r_{11}	0.230	
r_{21}	0.634	
r_{22}	0.039	
r_{12}	0.091	
s_1	1.00	
s_2	1.00	
k_d	1.03833	s^{-1}
f	2.47	$-$
k_t	$F_1 k_{t1} + (1 - F_1) k_{t2}$	
k_{fM}	$F_1 k_{fM1} + (1 - F_1) k_{fM2}$	
ρ_P	$F_1 \rho_{P1} + (1 - F_1) \rho_{P2}$	
	$s_1 = s_2 = 1.0$	

Table C-6: Homo-polymerization constants for Styrene in STY/BA models

Constant		
MW	104.1512	g/mol
k_p	0.872	L
k_t	1.703	L
k_{fM}	$1.096482 \times 10^7 \exp(-134268.8/RT)$	L /(mol. s)
ρ_M	$0.924 - 9.18 \times 10^{-4} \exp(T$	g/cm^3
ρ_P	$1.084 - 6.05 \times 10^{-4} \exp(T$	g/cm^3
γ	0.00	-

Table C-7: Homo-polymerization constants for Butyl Acrylate

Constant		
MW	128.17	g/mol
k_p	$1.7 \times 10^8 \exp(^{-7128.46}/_{RT})$	L
k_t	1.156167	L
k_{fM}	3.9374175	L
ρ_M	$0.9191 - 1.012 \times 10^{-3} \exp(T$	g/cm^3
γ	0.00	-

Table C-8: Parameter values for the modeling of the STY/BA

Paramete	Value	
r_{11}	0.551	
r_{21}	0.937	
r_{22}	0.225	
r_{12}	0.130	
s_1	0.405	
s_2	0.505	
k_d	$1.03833 \times 10^{15} \exp(-30706/RT)$	s^{-1}
f	$2.47 \times 10^{-2} \exp(2166/RT)$	$-$
k_t	$\dfrac{k_{t1} r_{21} N_1^2 \frac{(r_{11}N_1 + N_2)}{(r_{21}N_1 + N_2)} + 2(k_{t1}k_{t2})^{0.5} N_1 N_2 + k_{t2} r_{12} N_2^2 \frac{(r_{22}N_2 + N}{(r_{12}N_2 + N}}{r_{21} N_1^2 \frac{(r_{11}N_1 + N_2)}{(r_{21}N_1 + N_2)} + 2N_1 N_2 + r_{12} N_2^2 \frac{(r_{22}N_2 + N_1)}{(r_{12}N_2 + N_1)}}$	
k_{fM}	$F_1 k_{fM1} + (1 - F_1) k_{fM2}$	
ρ_P	$F_1 \rho_{P1} + (1 - F_1) \rho_{P2}$	

Table C-9: Experimental design for simulation runs

case	Model	Initial guess	Error
1	Terminal	Poor	Low
2	Terminal	Poor	Medium
3	Terminal	Poor	High
4	Terminal	Neutral	Low
5	Terminal	Neutral	Medium
6	Terminal	Neutral	High
7	Terminal	Good	Low
8	Terminal	Good	Medium
9	Terminal	Good	High
10	Strong Penultimate	Poor	Low
11	Strong Penultimate	Poor	Medium
12	Strong Penultimate	Poor	High
13	Strong Penultimate	Neutral	Low
14	Strong Penultimate	Neutral	Medium
15	Strong Penultimate	Neutral	High
16	Strong Penultimate	Good	Low
17	Strong Penultimate	Good	Medium
18	Strong Penultimate	Good	High
19	Small Penultimate	Poor	Low
20	Small Penultimate	Poor	Medium
21	Small Penultimate	Poor	High
22	Small Penultimate	Neutral	Low
23	Small Penultimate	Neutral	Medium
24	Small Penultimate	Neutral	High
25	Small Penultimate	Good	Low
26	Small Penultimate	Good	Medium
27	Small Penultimate	Good	High

Table C-10: Error levels

Levels	**Low**	**Medium**	**High**
Feed composition	0.005	0.005	0.005
Polymer composition	0.005	0.010	0.015
Conversion (mole fraction)	0.0067	0.0067	0.0067

Table C-11: Simulation Parameters for STY/AN

Levels	**Terminal**	**Strong Penultimate**	**Small Penultimate**
r_{11}	0.400	0.230	0.331
r_{21}	0.400	0.634	0.533
r_{22}	0.065	0.039	0.052
r_{12}	0.065	0.091	0.078
s_1	0.700	0.700	1.00
s_2	0.700	0.700	1.00

Table C-12: Simulation Parameters for STY/MMA

Levels	**Terminal**	**Strong Penultimate**	**Small Penultimate**
r_{11}	0.47200	0.25050	0.36125
r_{21}	0.47200	0.69354	0.58277
r_{22}	0.45400	0.27240	0.36320
r_{12}	0.45400	0.63551	0.54475
s_1	0.41200	0.41200	0.70000
s_2	0.17000	0.17000	0.60000

Table C-13: Simulation Parameters for STY/BA

Levels	**Terminal**	**Strong Penultimate**	**Small Penultimate**
r_{11}	0.956	0.551	0.648
r_{21}	0.956	0.937	0.841
r_{22}	0.183	0.225	0.201
r_{12}	0.183	0.130	0.154
s_1	0.405	0.405	0.600
s_2	0.405	0.505	0.700

Table C-14: Initial Reactivity Ratio Estimates for STY/AN

	Poor	**Neutral**	**Good**
$\hat{r}_{11}$	0.60	0.35	0.20
$\hat{r}_{21}$	0.20	0.45	0.60
$\hat{r}_{22}$	0.10	0.06	0.05
$\hat{r}_{12}$	0.05	0.08	0.10
$\hat{r}_{1}$	0.10	0.30	0.45
$\hat{r}_{2}$	0.45	0.20	0.10
$\hat{s}_{1}$	1.00	0.85	0.85
$\hat{s}_{2}$	1.00	0.85	0.85

Table C-15: Initial Reactivity Ratio Estimates for STY/MMA

	Poor	**Neutral**	**Good**
$\hat{r}_{11}$	0.75	0.50	0.30
$\hat{r}_{21}$	0.30	0.55	0.75
$\hat{r}_{22}$	0.70	0.40	0.35
$\hat{r}_{12}$	0.35	0.45	0.70
$\hat{r}_{1}$	0.70	0.55	0.50
$\hat{r}_{2}$	0.30	0.45	0.50
$\hat{s}_{1}$	1.00	0.50	0.35
$\hat{s}_{2}$	1.00	0.50	0.35

Table C-16: Initial Reactivity Ratio Estimates for STY/BA

	Poor	**Neutral**	**Good**
$\hat{r}_{11}$	0.90	0.70	0.50
$\hat{r}_{21}$	0.50	0.75	0.90
$\hat{r}_{22}$	0.10	0.15	0.30
$\hat{r}_{12}$	0.30	0.10	0.10
$\hat{r}_{1}$	0.20	0.70	0.90
$\hat{r}_{2}$	0.90	0.50	0.20
$\hat{s}_{1}$	1.00	0.75	0.50
$\hat{s}_{2}$	1.00	0.75	0.50

Appendix D

Results: Copolymer Case Study

Table D-1: SBMCMD method, triad data, STY/BA

Case Number	Number of Experiment	Pr (Terminal)	Pr (Penultimate)
1	0	1.00	0.00
2	0	0.95	0.05
3	0	0.98	0.02
4	0	1.00	0.00
5	0	0.99	0.01
6	0	0.98	0.02
7	0	1.00	0.00
8	0	1.00	0.00
9	0	0.97	0.03
10	0	0.00	1.00
11	1	0.01	0.99
12	7	0.02	0.98
13	0	0.00	1.00
14	0	0.00	1.00
15	0	0.02	0.98
16	0	0.00	1.00
17	0	0.00	1.00
18	1	0.01	0.99
19	0	0.05	0.95
20	1	0.04	0.96
21	1	0.97	0.03
22	0	0.01	0.99
23	0	0.00	1.00
24	1	0.95	0.05
25	0	0.00	1.00
26	2	0.02	0.98
27	1	0.97	0.03

Table D-2: SBMCMD method, triad data, STY/AN

Case Number	Number of Experiment	Pr (Terminal)	Pr (Penultimate)
1	0	0.99	0.01
2	0	1.00	0.00
3	0	0.99	0.01
4	0	0.97	0.03
5	0	0.99	0.01
6	0	0.97	0.03
7	0	0.97	0.03
8	0	0.99	0.01
9	0	0.99	0.01
10	0	0.00	1.00
11	0	0.00	1.00
12	0	0.00	1.00
13	0	0.00	1.00
14	0	0.00	1.00
15	0	0.00	1.00
16	0	0.00	1.00
17	0	0.00	1.00
18	0	0.03	0.97
19	0	0.00	1.00
20	4	0.01	0.99
21	2	0.00	1.00
22	0	0.00	1.00
23	0	0.00	1.00
24	2	0.01	0.99
25	0	0.00	1.00
26	0	0.00	1.00
27	5	0.03	0.97

Table D-3: SBMCMD method, triad data, STY/MMA

Case Number	Number of Experiment	Pr (Terminal)	Pr (Penultimate)
1	0	1.00	0.00
2	1	1.00	0.00
3	1	0.96	0.04
4	0	0.98	0.02
5	0	0.97	0.03
6	1	0.99	0.01
7	0	0.99	0.01
8	0	0.98	0.02
9	1	0.99	0.01
10	0	0.00	1.00
11	0	0.00	1.00
12	0	0.00	1.00
13	0	0.00	1.00
14	0	0.00	1.00
15	0	0.00	1.00
16	0	0.00	1.00
17	0	0.00	1.00
18	0	0.00	1.00
19	0	0.00	1.00
20	0	0.04	0.96
21	1	0.00	1.00
22	0	0.00	1.00
23	0	0.00	1.00
24	1	0.00	1.00
25	0	0.00	1.00
26	0	0.00	1.00
27	1	0.00	1.00

Appendix E

RAFT Sensitivity Analysis Plots (versus conversion)

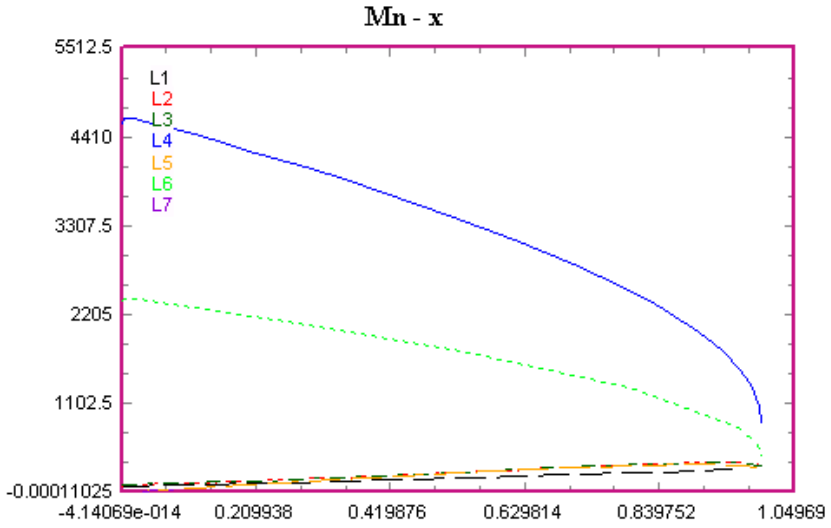

Figure 6E-1: RAFT models output, M_n – conversion

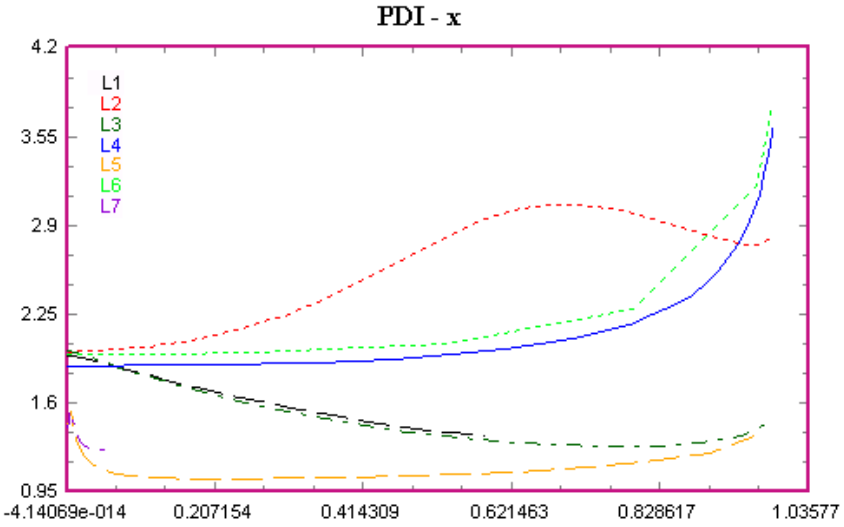

Figure E-2: RAFT models output, PDI – conversion

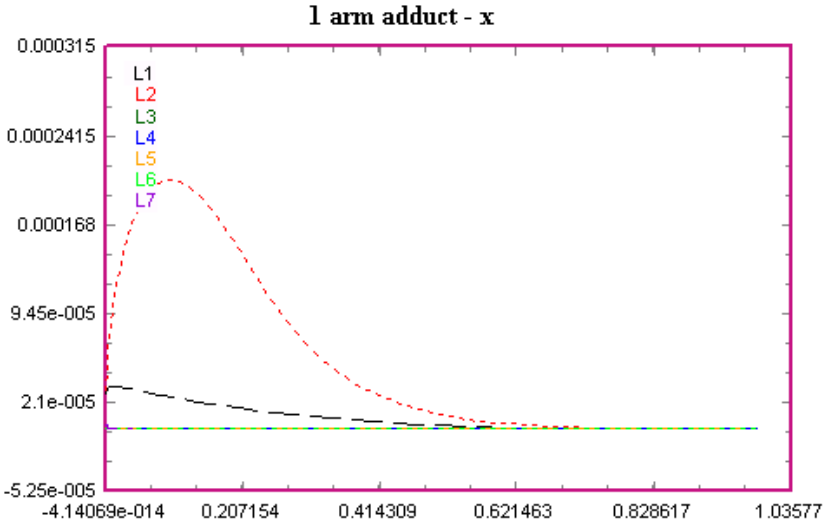

Figure E-3: RAFT models output, one-arm adduct concentration - conversion

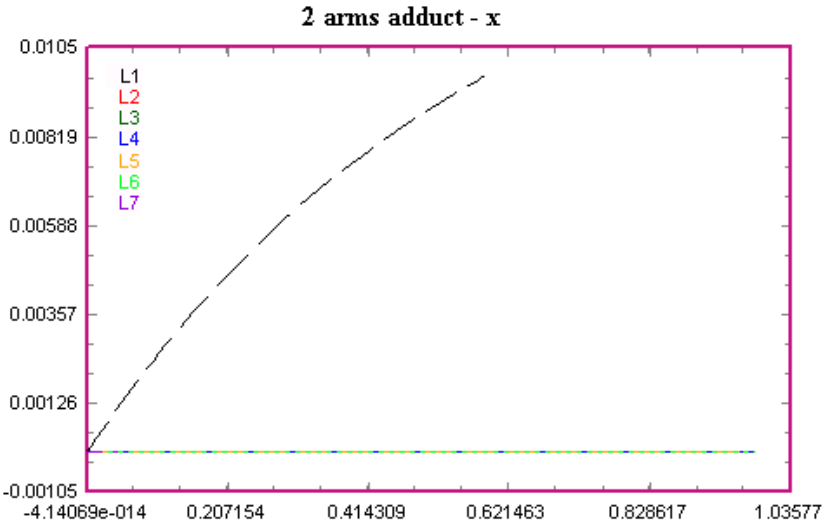

Figure E-4: RAFT models output, two arm-adduct concentration - conversion

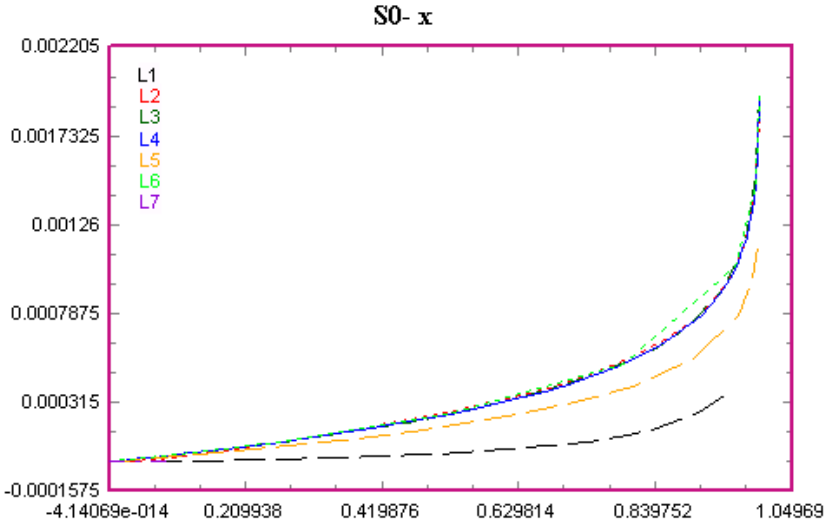

Figure E-5: RAFT models output, dormant concentration – conversion

Appendix F
Structure of the Program

Figure E-1 shows part of the structure of the implemented program.

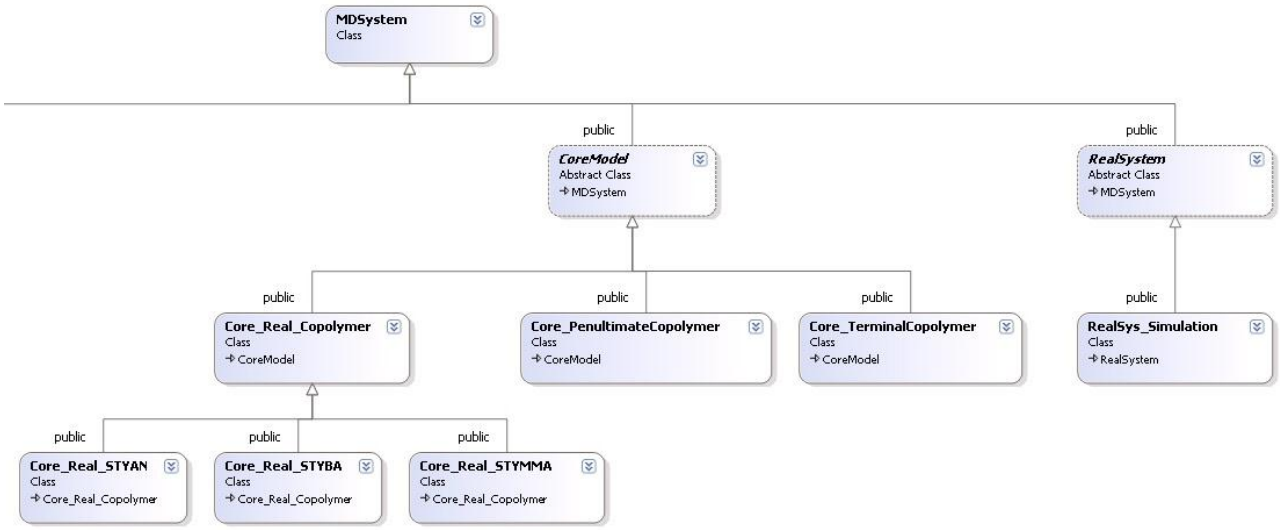

Figure F-1: Object oriented structure - Core object

MDSystem is a basic class in the design, and CoreModel is derived from it. A CoreModel has pure abstract functions that get parameters and input values and then return output of a CoreModel. When a new case study is tested by the framework, the only coding needed is driving a new CoreModel class and overwriting the abstract functions. Figure 4-4 shows the model class derived for the copolymerization case study (section 5.3). In addition to the CoreModel classes for the competitive models, in cases where a computation simulation is

used instead of the real system, a child class from the CoreModel is needed to present the real system. (Core_real_copolymer and derived classes in Figure F-2)

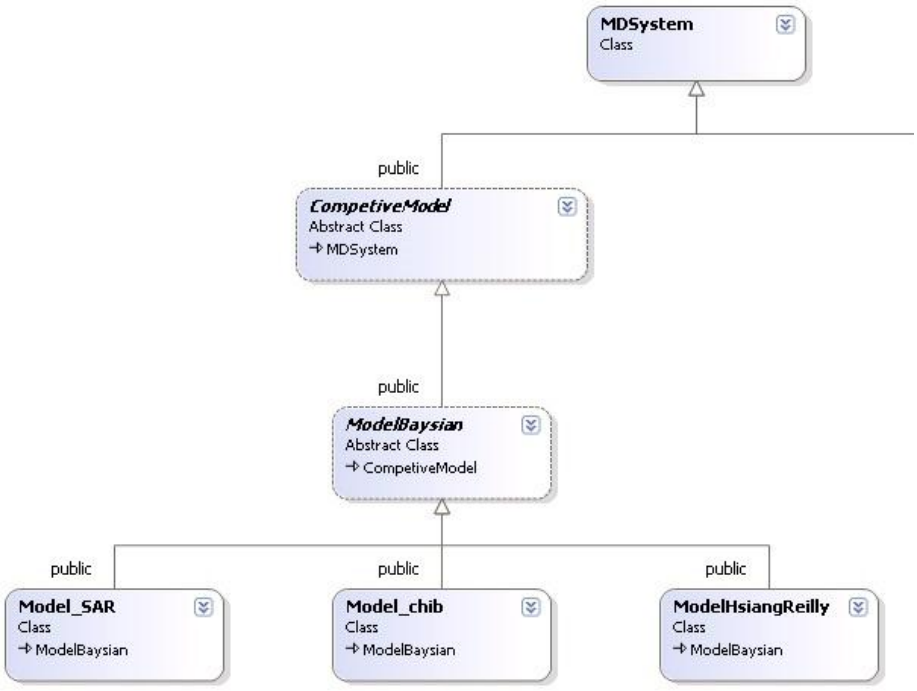

Figure F-2: Object oriented structure - Model object

"Real System" class is another important class in this design; one class derived from it is named RealSystem_Simulation, which contains an object from "CoreModel" representing the simulated real system. Figure 4-5 shows CompetetiveModel class and its derived ones. In this program HR method and three implementations of the SBMCMD approach are coded. Figure 4-6 presents another part of the design for the RandomGenerator. For any of these methods, a class of

Model_Bayesian is derived to handle the specific functions needed in that method. For example, any of these classes (Model SAR, model Chib and Model_Hsiang and Reilly) has a different implementation of the UpdateSamples function, where Updatesamples is an abstract function in Modelbayesian. The functions and features of Model_Chib is shown in Figure F-2.

Figure F-3 presents the CoreClass. One object from this class is necessary to be built for any case study. Case contains an Identifier and List of CompetetiveModels and a realSystem.

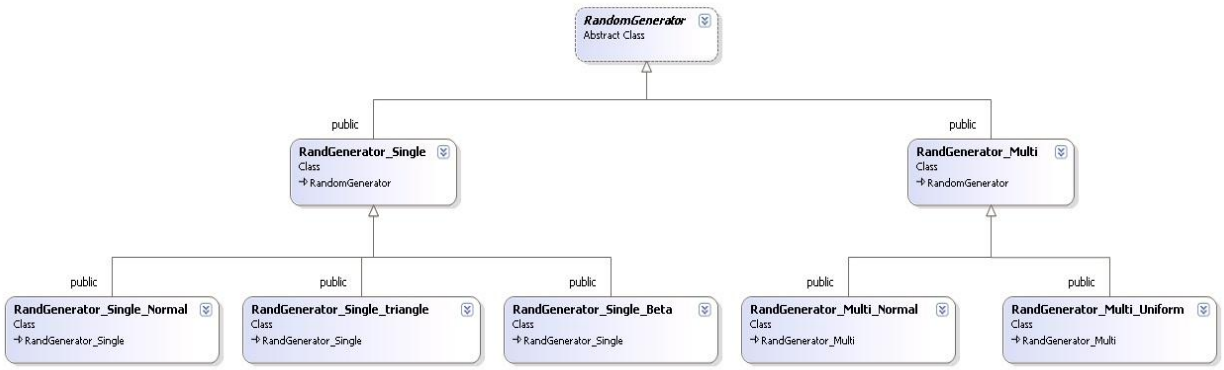

Figure F-3: Object oriented structure - Random Generator object

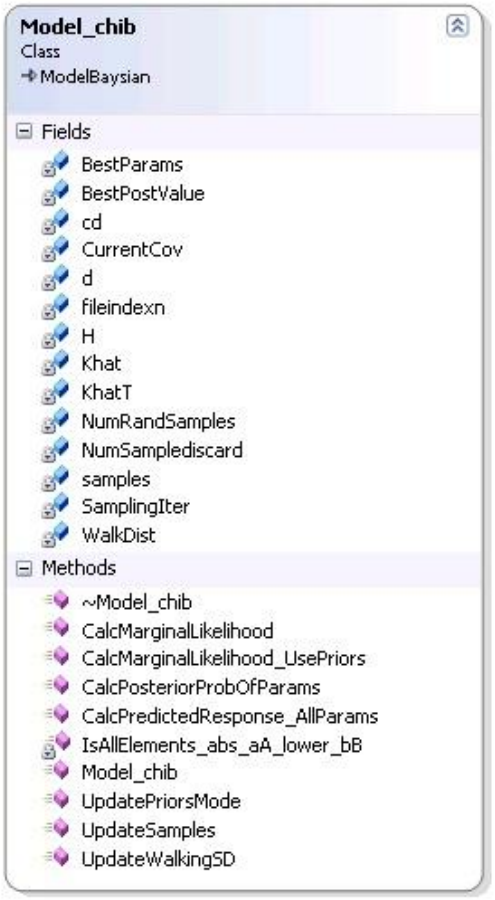

Figure F-4: Object oriented structure – Model_Chib

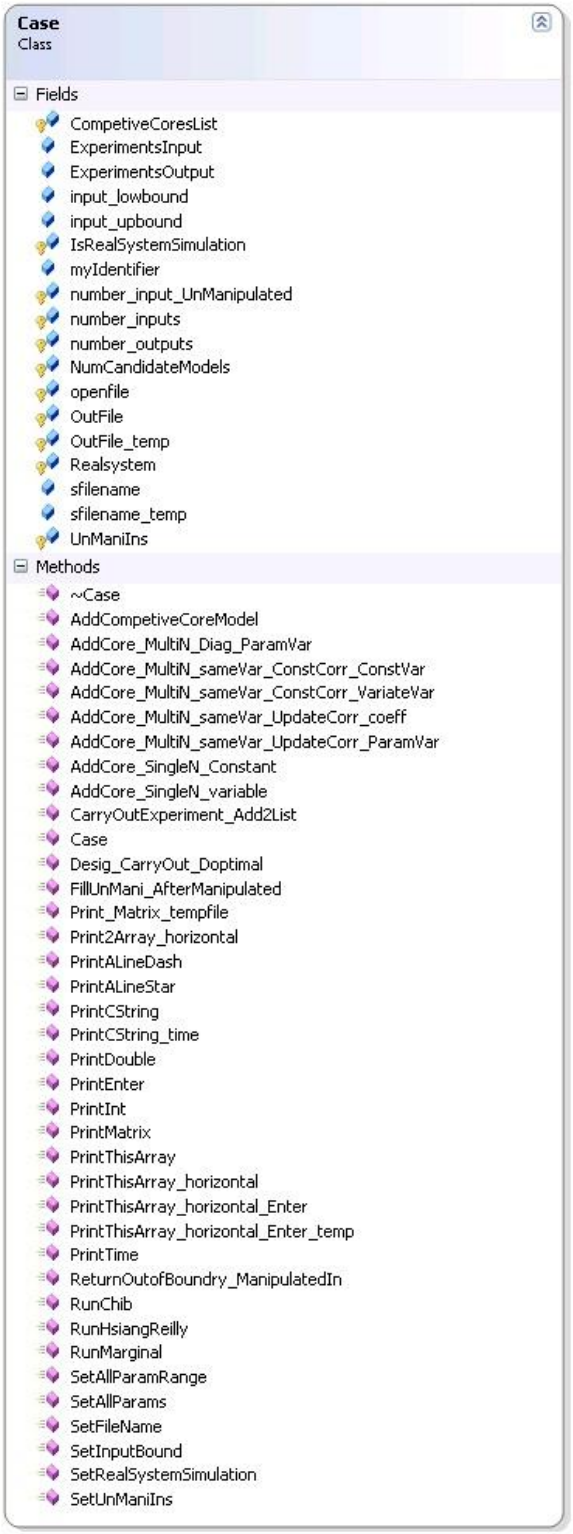

Figure F-5: Object oriented structure - Case class

Bibliography

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52(3), 317-332.

Bard, Y. (1974). Nonlinear parameter estimation (Vol. 477). New York: Academic Press.

Barner-Kowollik, C., Coote, M. L., Davis, T. P., Radom, L., Vana, P. (2003). The reversible addition-fragmentation chain transfer process and the strength and limitations of modeling: Comment on "the magnitude of the fragmentation rate coefficient". Journal of polymer science part A: polymer chemistry, 41(18), 2828-2832.

Barner-Kowollik, C., Quinn, J. F., Morsley, D. R., Davis, T. P. (2001). Modeling the reversible addition–fragmentation chain transfer process in Cumyl Dithiobenzoate-Mediated Styrene homo polymerizations: Assessing rate coefficients for the addition–fragmentation equilibrium. Journal of Polymer Science Part A: Polymer Chemistry, 39(9), 1353-1365.

Berger, M., Kuntz, I. (2003). The distinction between terminal and penultimate copolymerization models. Journal of Polymer Science Part A: General Papers, 2(4), 1687-1698.

Box, G. E., Hill, W. J. (1967). Discrimination among mechanistic models. Technometrics, 9(1), 57-71.

Box, G. E. P.,Tiao, G. C. (1992). Bayesian Inference in Statistical Analysis Wiley. New York.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. Journal of the royal statistical society: series D (the Statistician), 47(1), 69-100.

Buback, M., Janssen, O., Oswald, R., Schmatz, S., Vana, P. (2007, February). A Missing Reaction Step in Dithiobenzoate-Mediated RAFT Polymerization. In Macromolecular symposia (Vol. 248, No. 1, pp. 158-167). WILEY-VCH Verlag.

Buback, M., Vana, P. (2006). Mechanism of Dithiobenzoate-Mediated RAFT polymerization: A missing reaction Step. Macromolecular rapid communications, 27(16), 1299-1305.

Burke, A. L., Duever, T. A., Penlidis, A. (1994). Model discrimination via designed experiments: Discriminating between the terminal and penultimate models on the basis of composition data. Macromolecules, 27(2), 386-399.

Burke, A. L., Duever, T. A., Penlidis, A. (1994). Model discrimination via designed experiments: Discriminating between the terminal and penultimate models based on triad fraction data. Macromolecular theory and simulations, 3(6), 1005-1031.

Burke, A. L., Duever, T. A., Penlidis, A. (1995). Model discrimination via designed experiments: discrimination between the terminal and penultimate models based on rate data. Chemical engineering science, 50(10), 1619-1634.

Burke, A. L., Duever, T. A., Penlidis, A. (1997). Choosing the right model: Case studies on the use of statistical model discrimination experiments. The Canadian journal of chemical engineering, 75(2), 422-436.

Burke, Annette Lynn. (1994). "Model discrimination techniques for the modelling of copolymerization reactions", PhD thesis, Department of chemical engineering, University of Waterloo.

Buzzi Ferraris, G., Forzatti, P., Emig, G., Hofmann, H. (1984). Sequential experimental design for model discrimination in the case of multiple responses. Chemical engineering science, 39(1), 81-85.

Buzzi-Ferraris, G., Forzatti, P. (1990). An improved version of a sequential design criterion for discriminating among rival multi-response models. Chemical engineering science, 45(2), 477-481.

Buzzi-Ferraris, Guido, and Pio Forzatti. 1990. "An improved version of a sequential design criterion for discriminating among rival multiresponse models." Chemical engineering science 45 (6): 477–481.

Buzzi-Ferraris, G., Forzatti, P. (1983). A new sequential experimental design procedure for discriminating among rival models. Chemical engineering science, 38(2), 225-232.

Byrd, J. M. R. (2010). Parallel Markov Chain Monte Carlo, PhD thesis, Department of computer science, University of Warwick.

Carlin, B. P., Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. Journal of the royal statistical society. Series B, 473-484.

Carlin, B. P., Polson, N. G. (1991). Inference for non-conjugate Bayesian models using the Gibbs sampler. Canadian journal of statistics, 19(4), 399-405.

Chib, S., Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. Journal of the American statistical association, 96(453), 270-281.

Chib, S., Jeliazkov, I. (2005). Accept–reject Metropolis–Hastings sampling and marginal likelihood estimation. Statistica Neerlandica, 59(1), 30-44.

Chib, S. (1995). Marginal likelihood from the Gibbs output. Journal of the American statistical association, 90(432), 1313-1321.

Chib, S., Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. The american statistician, 49(4), 327-335.

Chiefari, J., Chong, Y. K., Ercole, F., Krstina, J., Jeffery, J., Le, T. P., Thang, S. H. (1998). Living free-radical polymerization by reversible addition-fragmentation chain transfer: the RAFT process. Macromolecules, 31(16), 5559.

Dellaportas, P., Forster, J. J. (1999). Markov Chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika, 86(3), 615-633.

Drovandi, C. C., McGree, J. M., Pettitt, A. N. (2012). Sequential Monte Carlo for Bayesian sequentially designed experiments for discrete data. Computational statistics and data analysis, 57(1), 320–335.

Fort, G, E Moulines, and P Priouret. (2012). Convergence of adaptive MCMC algorithms : Ergodicity and law of large numbers. Thechnical report. Available at: http://perso.telecom-paristech.fr/~gfort/Publications.html.

Gelman, A., Roberts, G., Gilks, W. (1996). Efficient Metropolis jumping rules. Bayesian statistics, 5, 599-608.

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. Statistical science, 7(4), 473-483.

Gilks, W. R., Richardson, S., Spiegelhalter, D. (1996). Practical Markov Chain Monte Carlo. New York, Chapman-Hall.

Gilks, W. R., Roberts, G. O., Sahu, S. K. (1998). Adaptive Markov Chain Monte Carlo through regeneration. Journal of the American statistical association, 93(443), 1045-1054.

Godsill, S. J. (2001). On the relationship between Markov Chain Monte Carlo methods for model uncertainty. Journal of computational and graphical statistics, 10(2), 230-248.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. Biometrika, 82(4), 711-732.

Guan, Y., Fleißner, R., Joyce, P., Krone, S. M. (2006). Markov Chain Monte Carlo in small worlds. Statistics and computing, 16(2), 193-202.

Haario, H., Laine, M., Mira, A., Saksman, E. (2006). DRAM: efficient adaptive MCMC. Statistics and computing, 16(4), 339-354.

Haario, H., Saksman, E., Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. Computational statistics, 14(3), 375-396.

Haario, H., Saksman, E., Tamminen, J. (2001). An adaptive Metropolis algorithm. Bernoulli, 223-242.

Hammersley, J. M., Handscomb, D. C., Weiss, G. (1965). Monte Carlo methods. Physics today, 18, 55.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1), 97-109.

Hsiang, T., Reilly, P. M. (1971). A practical method for discriminating among mechanistic models. The Canadian journal of chemical engineering, 49(6), 865-871.

Huang, J. (1988). The kinetics of Styrene/Methyl Methacrylate free radical copolymerization. PhD thesis, Department of chemical engineering, University of Waterloo.

Jefferys, W. H., Berger, J. O. (1992). Ockham's razor and Bayesian analysis. American scientist, 64-72.

Jeffreys, H. (1961). Theory of Probability, New York, US.

Jitjareonchai, J. (2006). Implementation of Markov Chain Monte Carlo techniques in parameter estimation for engineering models. PhD thesis, Department of chemical engineering, University of Waterloo.

Karandikar, R. L. (2006). On the Markov Chain Monte Carlo (MCMC) method. Sadhana (Academy proceedings in engineering sciences) (Vol. 31, No. 2, pp. 81-104).

Konkolewicz, D., Hawkett, B. S., Gray-Weale, A., Perrier, S. (2008). RAFT polymerization kinetics: combination of apparently conflicting models. Macromolecules, 41(17), 6400-6412.

Liang, F., Liu, C., Carroll, R. (2011). Advanced Markov Chain Monte Carlo methods: learning from past samples (Vol. 714). Wiley, New York.

Masoumi, S., Duever, T. A., Reilly, P. M. (2012). Sequential Markov Chain Monte Carlo (MCMC) model discrimination. The Canadian journal of chemical engineering. DOI: 10.1002/cjce.21711

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953). Equation of state calculations by fast computing machines. The journal of chemical physics, 21, 1087.

Miasojedow, B., Moulines, E., Vihola, M. (2012). Adaptive parallel tempering algorithm. arXiv preprint arXiv:1205.1076.

Mira, A. (2001). Ordering and improving the performance of Monte Carlo Markov chains. Statistical science, 340-350.

Mira, A., Sargent, D. J. (2000). Strategies for speeding Markov chain Monte Carlo algorithms. Statistical Methods and Applications, to appear. Technical report. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.8113&rep=rep1&type=pdf.

Moad, G., Rizzardo, E., Thang, S. H. (2005). Living radical polymerization by the RAFT process. Australian journal of chemistry, 58(6), 379-410.

Monteiro, M. J., de Brouwer, H. (2001). Intermediate radical termination as the mechanism for retardation in reversible addition-fragmentation chain transfer polymerization. Macromolecules, 34(3), 349-352.

Newton, M. A., Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. Journal of the royal statistical society: Series B, 3-48.

O'Driscoll, K. F., Huang, J. (1989). The rate of copolymerization of Styrene and Methyl Methacrylate. Low conversion kinetics. European polymer journal, 25(7), 629-633.

Pallares, J., Jaramillo-Soto, G., Flores-Cataño, C., Lima, E. V., Lona, L. M., Penlidis, A. (2006). A Comparison of reaction mechanisms for Reversible Addition-Fragmentation Chain Transfer polymerization using modeling tools. Journal of Macromolecular science part A: pure and applied chemistry, 43(9), 1293-1322.

Peskun, P. H. (1973). Optimum Monte- Carlo sampling using Markov chains. Biometrika, 60(3), 607-612.

Reilly, P. M. (1970). Statistical methods in model discrimination. The Canadian journal of chemical engineering, 48(2), 168-173.

Robert, C. P., Casella, G., Robert, C. P. (1999). Monte Carlo statistical methods (Vol. 2). New York: Springer.

Roberts, G. O., Gelman, A., Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. The annals of applied probability, 7(1), 110-120.

Roberts, G. O., Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. Statistical science, 16(4), 351-367.

Roth, P. M. (1965). Design of experiments for discrimination among rival models, PhD thesis, Princeton University.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 6(2), 461-464.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Lunn, D. J. (2003). WinBUGS: Bayesian inference using Gibbs sampling, manual version 1.4. Cambridge: Medical Research Council Biostatistics Unit.

Tamir, A. (1998). Applications of Markov chains in chemical engineering. Elsevier Science.

Tanner, M. A. (1991). Tools for statistical inference: observed data and data augmentation methods. New York: Springer-Verlag.

Tierney, L. (1994). Markov chains for exploring posterior distributions. The annals of statistics, 1701-1728.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. Annals of applied probability, 1-9.

Tobita, H., Yanase, F. (2007). Monte Carlo simulation of controlled/living radical polymerization in emulsified systems. Macromolecular theory and simulations, 16(4), 476-488.

Wang, A. R., Zhu, S. (2003). Modeling the Reversible Addition–Fragmentation Transfer polymerization process. Journal of polymer science part A: Polymer chemistry, 41(11), 1553-1566.

Wang, A. R., Zhu, S., Kwak, Y., Goto, A., Fukuda, T., Monteiro, M. S. (2003). A difference of six orders of magnitude: A reply to "the magnitude of the fragmentation rate coefficient". Journal of polymer science part A: Polymer chemistry, 41(18), 2833-2839.

Wasserman, L. (2000). Bayesian model selection and model averaging. Journal of mathematical psychology, 44(1), 92-107.

Zhang, M., Ray, W. H. (2001). Modeling of "living" free-radical polymerization with RAFT chemistry. Industrial & engineering chemistry research, 40(20), 4336-4352.