ADVANCING SEQUENTIAL MONTE CARLO FOR MODEL CHECKING, PRIOR SMOOTHING AND APPLICATIONS IN ENGINEERING AND SCIENCE

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Lixin Lang, M.S.

* * * * *

The Ohio State University

2008

Dissertation Committee:

Prem K. Goel, Co-Adviser

Bhavik R. Bakshi, Co-Adviser

Mark Berliner

Mario Peruggia

Approved by

_____

Co-Adviser

_____

Co-Adviser

Graduate Program in Statistics

ABSTRACT

The research effort in this dissertation is targeted to investigate theoretical properties of some key statistics used in the sequential Monte Carlo (SMC) sampling, and to extend SMC to model checking, prior smoothing, and constrained state estimation. A novel application of SMC estimation to population pharmacokinetic models is also introduced.

Asymptotic properties of two key statistics in the SMC sampling, importance weights and empirical effective samples size, are discussed in the dissertation. The sum-normalized nature of importance weights makes it extremely difficult, if not impossible, to analytically investigate their properties. By using expectation-normalized importance weights, we are able to show the theoretical estimate of empirical effective sample size under various situations. In addition, the superiority of optimal importance function over prior importance function is verified based on the expectation-normalized weights.

The usage of SMC is also demonstrated for checking incompatibility between the prior and the data, using observation's predictive density value. When the prior is detected to be incompatible with the data, prior smoothing is proposed with a popular numerical method, Moving Horizon Estimation (MHE), to obtain a better estimate of the initial state value. Specifically, the incorporation of MHE smoothing into SMC estimation is among the first efforts to integrate these two powerful tools.

Convergence of constrained SMC (Chen, 2004) is verified and its performance is further illustrated with a more complex model.

SMC estimation is applied to a multi-dimensional population pharmacokinetic (PK) model. It is shown that the SMC sampling is faster than Markov Chain Monte Carlo (MCMC), and it doesn't suffer from the lack of convergence concern for MCMC.

Dedicated to my wife, lijie

ACKNOWLEDGMENTS

I cannot thank my co-advisors, Dr. Prem Goel and Dr. Bhavik Bakshi, more than enough. They were always ready to discuss new ideas and applications, and helped me go through the manuscripts of our papers paragraph by paragraph. My research and the dissertation could not go far without their inspiration, guidance, encouragement and patience.

I am indebted to my family for their unwavering support while I was writing the dissertation. My wife lijie took care of our baby almost all on her own.

Financial support from the National Science Foundation, under CTS-0321911, is gratefully acknowledged.

VITA

1994 ..... B.S. Xi'an Jiaotong University

2003 ..... M.S. The Ohio State University

2000-present ..... Graduate Research Associate,
The Ohio State University.

PUBLICATIONS

Research Publications

Lixin Lang, Wen-Shiang Chen, Bhavik R. Bakshi, Prem K. Goel, and S. Ungarala. "Bayesian estimation of constrained nonlinear dynamic systems via sequential monte carlo sampling". *Automatica*, 43:1615–1622, 2007.

Prem K. Goel, Lixin Lang, and Bhavik R. Bakshi. "Sequential Monte Carlo in Bayesian inference for dynamic models: an overview". In S. K. Upadhyay, U. Singh, and D. K. Dey, editors, *Bayesian Statistics and Its Applications.* Anamaya Publishers, New Delhi, India, 2006.

Lixin Lang and Benjamin Coifman. "Identifying lane mapping errors at freeway detector stations". *Journal of Transportation Research Board*, 1945:89- 99, 2006.

FIELDS OF STUDY

Major Field: Statistics

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION TO SEQUENTIAL MONTE CARLO

Sequential Monte Carlo (SMC), also known as particle filtering, is essentially a recursive importance sampling/resampling method primarily for Bayesian inference of dynamic models. During the simulation of the dynamic process, samples, the so called *particles*, are drawn from importance function and their weights are updated according to model specifications and observations to approximate the model's underlying posterior distribution. Such sampling and updating repeat whenever new data are observed from the model. In the dissertation, the dynamic models are represented in a state-space form.

Particle filtering has many advantages over other approximating methods. One of the most important would be its straightforward application to and accurate estimation of nonlinear non-Gaussian dynamic models. Extended Kalman filter (EKF), for example, is often found to be unreliable and the Moving Horizon Estimate (MHE) method tends to be less accurate and time-consuming (Chen et al., 2004). The particles generated through the SMC simulation scheme can be shown to converge asymptotically to the underlying posterior distribution, under certain conditions, as the number of particles goes to infinity (Künsch, 2005). Furthermore, several central limit theorems exist for the convergence of point estimates for SMC (Chopin, 2004,

Künsch, 2005). Besides its versatility, SMC is more suitable for online estimation than Markov chain Monte Carlo (MCMC), which needs to run from scratch whenever a new observation becomes available.

Another powerful estimation method for state-space models, the *Moving Horizon Estimation*, is a complex numerical optimization method discussed in Chapters 3 and 4. The combination of MHE and SMC for prior smoothing in Chapter 3 is the first ever effort to employ both together to exploit their advantages.

The rest of this chapter is organized as follows. In Section 1.1, a general introduction to Bayesian inference is given under state-space models. The SMC methodology, including some specific concerns, are discussed in Section 1.2. In the last section, some extensions to and applications of SMC estimation are briefly summarized, and the detailed discussions are presented in their respective chapters.

1.1 Bayesian Inference under state-space models

The state-space model is defined as:

$$x_k = f(x_{k-1}, \omega_k), \tag{1.1a}$$

$$y_k = h(x_k, \nu_k), \tag{1.1b}$$

where (1.1a) is called state equation and (1.1b) measurement equation. Random variable x_k, the unknown system state at time point k, evolves through the state equation; while observation y_k is a stochastic function of state x_k as described in the measurement equation. Random variable ω_k is called the state noise, and ν_k is called the measurement noise. Usually it is assumed that $\{\omega_j\}$ and $\{\nu_k\}$ are mutually independent for all $j, k \in \mathbb{N}$. Equivalently, the above state-space model can be considered as a discrete-time random process $\{(x_k, y_k); k \in \mathbb{N}\}$. The sequence of unknown states,

$\{x_k\}$, constitute a Markov chain with transition density function defined by the state equation; observations, $\{y_k\}$, are conditionally independent given the states $\{x_k\}$.

We use the notation $x_{i:j} \equiv \{x_i, x_{i+1}, \ldots, x_j\}$, and similarly, $y_{i:j}$, for $i \leq j$ and $i, j \in \mathbb{N}$. A superscript counterpart, $\{x_k^{(i)}, i = 1, \ldots, N\}$ represents a set of (independent) samples of x_k. With an abuse of notation, let $p(\cdot)$ represent the density function of the random variable(s) explicitly specified as its argument. If the density function itself is not of interest, notation $[\cdot]$ is used to denote the distribution following the convention of Wakefield et al. (1994). For example, the random variable x_k conditional on the observations $y_1, \ldots, y_k$ has a density function $p(x_k|y_{1:k})$, or equivalently, is simply referred to as $[x_k|y_{1:k}]$. Specially, denote the prior distribution on the initial state x_1 as π.

Under the framework of the state-space model, either the marginal posterior, $[x_\ell|y_{1:k}]$, or the joint, $[x_{1:\ell}|y_{1:k}]$, could be of interest from the Bayesian point of view and is called the target distribution. Depending on the relationship between ℓ and k above, the posterior is often further distinguished as filter for $\ell = k$, prediction for $\ell > k$, and smoothing for $\ell < k$ respectively. Various solutions to the posterior distribution have been proposed for different specific forms of state-space models.

The simplest form of a state-space model is perhaps the linear Gaussian model as illustrated below,

$$x_k = Fx_{k-1} + \omega_k, \tag{1.2a}$$

$$y_k = Hx_k + \nu_k, \tag{1.2b}$$

where F and H are scalars or matrices compatible with the state space and the measurement space, process noise $\omega_k \sim \mathcal{N}(0, Q)$, and measurement noise $\nu_k \sim \mathcal{N}(0, R)$. The given prior, π on x_1, is also Gaussian and distributed as $\mathcal{N}(\mu_1, P_1)$. Under

such settings, the posterior distributions of any type—joint, marginal, filtering, prediction, or smoothing, are all Gaussian, thus it suffices to find their mean and the covariance values. For simplicity, only marginal filtering posterior at a single time point is discussed below. Let $\mu_{\ell|k}$ and $P_{\ell|k}$ denote the mean and the covariance of state x_ℓ conditional on the observations $y_{1:k}$ for any $k, \ell \in \mathbb{N}$. Note that $\mu_{1|0} = \mu_1$ and $P_{1|0} = P_1$ are implicitly indicated in the above settings. Then the Kalman filter recursively gives the optimal filtering estimates of states x_k, *i.e.*, $\mu_{k|k}$ and $P_{k|k}$, for the observations $y_{1:k}$, where $k \geq 1$. Details are given in Appendix A.

Other than the linear Gaussian model in (1.2), which has an analytically tractable form of posterior distribution, approximating the target distribution as accurately as possible is the best that can be done due to the generality of state-space models. Various approximating methods are proposed in the literature.

The Extended Kalman Filter is a popular method in analytic approximation. Under the assumption of additive Gaussian noises and Gaussian prior, it approximates the nonlinear functions in (1.1) by the first-order Taylor series expansion and then applies the Kalman filter to the linearized system; see, for example, Anderson and Moore (1979). However, EKF introduces errors to the mean and covariance of the states when strong nonlinearity exists, possibly leading to a divergence of state estimates from the true values.

Another algorithm, Unscented Kalman filter (UKF) (Julier and Uhlmann, 1995, van der Merwe et al., 2000), uses a small set of deterministically chosen points to estimate the mean and covariance value of the system state without linearizing model equations. UKF is claimed to substantially outperform the EKF at the same order of computational complexity.

Numerical approximations to the posterior density are discussed in many papers (e.g., see Geweke (1988), Naylor and Smith (1982), Tierney and Kadane (1986), Kitagawa (1987)). These approaches can be very accurate, but implementation either requires sophisticated numerical skills or is hardly feasible for high dimensions.

Simulation methods are becoming more popular with the increase of computing power and advances in statistical theory. Markov chain Monte Carlo is a powerful tool to generate dependent samples that converge to the desired distribution if sampling function is connivently available. As a matter of fact, its applications in state-space models are often restricted to some specific models due to availability of sampling function. Carlin et al. (1992) suggest state space augmentation, and similarly, Carter and Kohn (1996) discuss a particular conditional Gaussian state-space model to facilitate the use of Gibbs sampling. For the general state-space models, Hürzeler and Kunsch (1998) present rejection sampling and Geweke and Tanizaki (2001) recommend the Metropolis-Hastings algorithm within Gibbs sampling. However, these two methods often suffer from small acceptance rate. In summary, MCMC methods are suitable for joint posterior estimation, but unfortunately, require a full iteration each time when a new observation is available.

1.2 Sequential Monte Carlo

Sequential Monte Carlo is essentially a recursive importance sampling and resampling scheme for dynamic models. SMC is also known as particle filter, perhaps stemming from the fact that Monte Carlo is heavily used in the engineering research. The word particle was first seen in Kitagawa (1996). Historically, similar algorithms were reported under different names in independent efforts, see bootstrap filter or

sequential importance resampling in Rubin (1987), Gordon et al. (1993), sequential imputation in Kong et al. (1994), Monte Carlo filter in Kitagawa (1996) and condensation algorithm in Isard and Blake (1998). A unified framework covering many of these methods is proposed in Doucet et al. (2000), and is described in the following section. Compared to the MCMC, sequential Monte Carlo is inferior at joint state estimation due to the strong correlation between the generated samples at adjacent time points. However, its fast marginal state estimation in general state-space models has made it a more favorite choice than MCMC and other methods. We first introduce importance sampling and then its recursive implementation in dynamic models with specially structured importance function.

Importance sampling is a useful technique to simulate complex distributions and make Monte Carlo estimates. Suppose that random variable X is distributed with density function $p(x)$ and one needs to evaluate the integral $\mathbb{E}[g(x)] = \int g(x)p(x)dx$. It is well known that a Monte Carlo approximation is the sample mean of $\{g(x^{(i)}), i = 1, \ldots, N\}$, where $\{x^{(i)}, i = 1, \ldots, N\}$ are samples from $p(x)$. However, when it is difficult to draw samples from $p(x)$, one can choose another density function, $q(x)$, the so called importance function, draw samples $x^{(i)}, i = 1, \ldots, N$ from $q(\cdot)$, and use the weighted mean,

$$\mathbb{E}[g(x)] \approx \frac{1}{N} \sum_{i=1}^{N} \tilde{w}^{(i)} g(x^{(i)}), \tag{1.3}$$

where

$$\tilde{w}^{(i)} = \frac{p(x^{(i)})}{q(x^{(i)})}, \tag{1.4}$$

as an approximation. The unnormalized importance weights $\tilde{w}$ can be normalized and Equation (1.3) is further approximated by:

$$\mathbb{E}\left[g(x)\right] \approx \sum_{i=1}^{N} w^{(i)} g(x^{(i)}), \tag{1.5}$$

where

$$w^{(i)} = \frac{\tilde{w}^{(i)}}{\sum_{j=1}^{N} \tilde{w}^{(j)}}. \tag{1.6}$$

From now on, $w^{(i)}$ is referred to as sum-normalized weight when it is necessary to distinguish $w^{(i)}$ from a similar one, expectation-normalized weight introduced in Chapter 2. When there is no confusion under the context, we simply call $w^{(i)}$ normalized weight. An advantage of the normalized formulation in Equation (1.6) is that one needs to know $p(\cdot)$ and $q(\cdot)$ only up to proportionality constants. This advantage is important in the dynamic models, where the proportionality constants vary from time to time and have to be approximated through Monte Carlo simulations.

Particle filtering is a recursive way to do importance sampling and resampling. For general state-space models in Equation (1.1), however, some special form of importance function has to be used to implement recursive importance sampling. We need to understand how state information is propagated and updated in the first step. One recursive updating formula for the joint posterior is as follows:

$$p(x_{1:k}|y_{1:k}) \propto p(x_{1:k-1}|y_{1:k-1}) p(x_k|x_{k-1}) p(y_k|x_k). \tag{1.7}$$

As indicated by the above expression, the posterior at time k is known up to some constant given the posterior at time $k-1$, $p(x_{1:k-1}|y_{1:k-1})$, transition density $p(x_k|x_{k-1})$ and likelihood $p(y_k|x_k)$. The latter two can be easily calculated for the given model in Equation (1.1). Based on Equation (1.7), Doucet et al. (2000) suggest to use the

following updating scheme for the importance function:

$$u_k(x_{1:k}; y_{1:k}) = u_{k-1}(x_{1:k-1}; y_{1:k-1}) q(x_k; x_{1:k-1}, y_{1:k}), \tag{1.8}$$

where $q(x_k; x_{1:k-1}, y_{1:k})$ is a density function of x_k with some parameters decided by part or all of $x_{1:k-1}$, as well as $y_{1:k}$. Note that the above sampling scheme implicitly states that new samples of x_k are drawn from $q(\cdot)$ based on the state samples at previous time points and possibly the observations $y_{1:k}$. Using Equation (1.4), the unnormalized importance weights become

$$\tilde{w}_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(y_k|x_k^{(i)}) p(x_k^{(i)}|x_{k-1}^{(i)})}{q(x_k^{(i)}; x_{1:k-1}^{(i)}, y_{1:k})}. \tag{1.9}$$

Thus, the weights can be recursively updated through Equations (1.9) and (1.6). Clearly, the density function $q(\cdot)$ must be easy to sample from and analytically tractable. In practice,

$$q(x_k; x_{1:k-1}, y_{1:k}) = p(x_k|x_{k-1}), \tag{1.10}$$

known as the prior importance function, is conveniently available in a closed form for most of general state-space models. It propagates particles from time $k-1$ to time k through the state equation in (1.1a). Under the prior importance function, Equation (1.9) is further simplified to

$$\tilde{w}_k^{(i)} \propto w_{k-1}^{(i)} p(y_k|x_k^{(i)}). \tag{1.11}$$

Figure 1.1 shows how particles evolve over time under the prior importance function.

1.2.1 Optimal Importance Function

As stated above, the prior importance function can be applied to any state-space models where the state equation defines a transition density function. However,

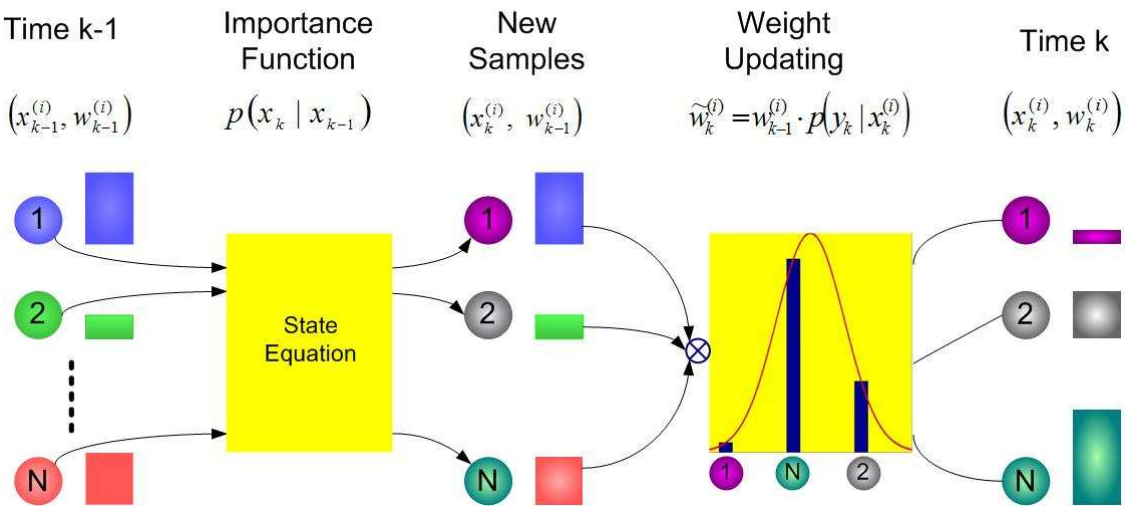


Figure 1.1: Particle evolves through a two-phases process, samples prediction (in the left half) and weight updating (in the right half)

it depends only on previous samples of $x_{k-1}^{(i)}, i = 1, \ldots, N$ and inherently previous observations. Therefore less representative particles could be generated if the current observation y_k is far from what would be expected given the previous samples of x_{k-1}.

A better sampling method is to draw samples also conditional on the current observation y_k, that is, let

$$q(x_k; x_{1:k-1}, y_{1:k}) = p(x_k | x_{k-1}, y_k). \tag{1.12}$$

This is called the optimal importance function in that it minimizes the variance of the importance weights conditioned on $x_{1:k-1}$ and $y_{1:k}$ (Doucet et al., 2000). The unnormalized weights are then

$$\tilde{w}_k^{(i)} = w_{k-1}^{(i)} p(y_k | x_{k-1}^{(i)}) \tag{1.13}$$

Unfortunately, either the optimal importance function in (1.12) is not always available for direct sampling or the density function $p(y_k|x_{k-1})$ does not always have a closed-form expression as required in (1.13). However, the state-space model with a linear measurement equation and additive Gaussian noises in both equations does satisfy the above requirements. Define such model by

$$x_k = f(x_{k-1}) + \omega_k, \tag{1.14a}$$

$$y_k = Hx_k + \nu_k, \tag{1.14b}$$

where $\omega_k \sim \mathcal{N}(0, Q), \nu_k \sim \mathcal{N}(0, R)$ for each k, and H is a compatible matrix with state space as in the linear Gaussian model (1.2). In fact, the matrices $H, QandR$ in these models could vary with time k, without adding any complexity to the posterior distributions. Then it is known that $[x_k|x_{k-1}, y_k]$ is a normal distribution with mean $\mu_{opt,k}$ and variance $\Sigma_{opt,k}$, where

$$\Sigma_{opt,k} = \left[R^{-1} + H'Q^{-1}H\right]^{-1}, \tag{1.15a}$$

$$\mu_{opt,k} = \Sigma_{opt,k}\left(R^{-1}f(x_{k-1}) + H'Q^{-1}y_k\right). \tag{1.15b}$$

The corresponding importance weight is $p(y_k|x_{k-1})$, which is also a Gaussian density function such that

$$p(y_k|x_{k-1}) \propto \exp\{-\frac{1}{2}(y_k - Hf(x_{k-1}))'\left[R + HQH'\right]^{-1}(y_k - Hf(x_{k-1}))\} \tag{1.16}$$

Notice that Equation (14) of Doucet et al. (2000) should be the same as Equation (1.15). However, the former has a typo, where Σ_ν and Σ_w should be exchanged (please be advised that our notation is different from that in Doucet et al. (2000)). A more

computationally efficient expression than (1.15) also exists:

$$\Sigma_{opt,k} = R - RH' \left[HRH' + Q\right]^{-1} HR \tag{1.17a}$$

$$\mu_{opt,k} = f(x_{k-1}) + RH' \left[HRH' + Q\right]^{-1} \left(y_k - Hf(x_{k-1})\right) \tag{1.17b}$$

Though seemingly completely different, the two sets of solutions above are the same since

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}, \tag{1.18}$$

for compatible matrices $A, B, C,$ and D. Since the matrix inverse operation in (1.17) involves only measurement space whose dimension is usually less that of state space, (1.17) is preferred for the sake of computation even though it looks more complex.

1.2.2 Particle degeneracy and resampling

Repeated updating of importance weights will inevitably lead to particle degeneracy (Kong et al., 1994, Doucet et al., 2000), when nearly all the normalized weights are zero. Thus a great deal of computational effort is wasted on updating particles that make very little contribution toward the estimation. Resampling can be used to remove small-weighted particles while keeping the underlying distribution unchanged. However, as unnecessarily frequent resampling would likely cause particle depletion as discussed in the next section, Liu (1996) proposes the Effective Sample Size (ESS) statistic, $\tilde{N}_k$, to control resampling frequency.

$$\tilde{N}_k = \frac{1}{N \sum_{i=1}^{N} (w_k^{(i)})^2} \tag{1.19}$$

It can be easily verified that $\tilde{N}_k \leq 1$. Resampling step is performed whenever $\tilde{N}_k$ gets below a specified threshold value N_{thre}, say, 1/3.

Various resampling schemes have been introduced into SMC simulation process. Simple random resampling, or multinomial resampling, draws samples with the probability equal to their weights. Other methods are developed to reduce Monte Carlo variation or computing time. Stratified resampling (Kitagawa, 1996) aims to avoid drawing more than one sample from a group of weights whose sum is less than $1/N$. Systematic resampling (Kitagawa, 1996) is a slight modification to the stratified resampling in that it is more time-efficient. Residual resampling (Liu and Chen, 1998) guarantees that each sample $x_k^{(i)}$ have at least $\lfloor N w_k^{(i)} \rfloor$ replicates of themselves, where $\lfloor m \rfloor$ is the integer part of m.

It is worth noting that even though the resampling procedure may increase the variance of the estimate at the current time, it may provide a better estimate at some of the future time points (Chopin, 2004).

1.2.3 Particle depletion and moving strategies

Particle depletion or impoverishment happens when a large portions of particles are identical, making the empirical distribution less representative. In the worst case, the empirical distribution becomes a single delta function. Particle depletion usually results from repeatedly resampling the empirical distribution. Particles in dynamic models are able to evolve into distinct ones when propagated through the system equation. Consequently, it is more often for static models that one runs into this problem, when resampling is applied.

Particle moving is designed to reduce this effect, in a way that brings variations to the identical particles without changing the underlying distribution. Berzuini and

Gilks (2001) discuss this topic with resample-move. In brief, each particle after resampling is subject to be moved to a new position in the state space by an appropriate transition kernel. Thus, duplicated particles have a chance to become diversified and the number of identical particles is smaller than without moving. Chopin (2002) suggests using a normal distribution as the transition kernel in a static model with large data set. The author claims that an independent normal kernel with the empirical sample mean and variance is a reasonable choice since posterior tends to be Gaussian asymptotically. Another strategy that moves each particle around itself, as described by Goel et al. (2006), has similar performance. In Chapter 5 where SMC is applied to a static model, particle moving is performed based on a Gibbs sampling scheme.

1.2.4 SMC algorithm

The general SMC algorithm is listed in the following.

The Sequential Monte Carlo algorithm

1. Initialization

Set $w_0^{(i)} = 1/N$, for $i = 1, 2, \ldots, N$

For time $k = 1, 2, \ldots$

2. Importance sampling

For $i = 1, 2, \ldots, N$

Draw samples $\tilde{x}_k^{(i)} \sim q(\cdot | x_{1:k-1}^{(i)}, y_{1:k})$

Compute weight $\tilde{w}_k^{(i)}$

End

3. Resampling

For each $i = 1, 2, \ldots, N$, normalize weight $w_k^{(i)}$

Evaluate effective sample size, $\tilde{N}_k$

If $\tilde{N}_k < N_{thre}$

For $i = 1, 2, \ldots, N$

Draw $x_k^{(i)}$ from $\tilde{x}_k^{(1:N)}$ with probability $w_k^{(1:N)}$ respectively

End

Reset $w_k^{(i)} = 1/N$, for $i = 1, 2, \ldots, N$

Else

Set $x_k^{(i)} = \tilde{x}_k^{(i)}$, for $i = 1, 2, \ldots, N$

End

End

1.2.5 Convergence Properties

Convergence properties of SMC estimation have been studied by many authors, see *e.g.*, Del Moral and Miclo (2000), Crisan (2001), Künsch (2005), and in many references cited therein. A survey of convergence results is provided by Crisan and Doucet (2002). Two theorems from Künsch (2005) are presented below and are used in Chapter 2 for the development of asymptotic properties of some statistics.

Following the convention of Künsch (2005), denote the state transition density function as a_k such that

$$\Pr(x_k \in dx|x_{1:k-1}) = \Pr(x_k \in dx|x_{k-1}) = a_k(x_{k-1}, dx)$$

Also define b_k such that

$$\Pr(y_k \in dy|x_k) = b_k(x_k, dy).$$

For example, in the widely used additive noises state-space model, $\omega_k \sim p_\omega$ and $\nu_k \sim p_\nu$, the expressions of a_k and b_k are given by

$$\begin{aligned} a_k(x_{k-1}, x_k) &= p_\omega(x_k - f_k(x_{k-1})), \\ b_k(x_k, y_k) &= p_\nu(y_k - h_k(x_k)). \end{aligned}$$

Given the measurements $y_{1:k}$, the posterior density of x_k is approximated by the empirical distribution based on N particles and weights at time k, $\{x_k^{(i)}, w_k^{(i)}, i = 1, \ldots, N\}$,

$$\hat{p}^N(x_k|y_{1:k}) = \sum_{i=1}^{N} \delta(x_k - x_k^{(i)}) w_k^{(i)}.$$

Künsch (2005) provides the following theorems.

Theorem 1. If x from $a_k(x, \cdot)$ is continuous, and if for all k, all x and all y,

$$0 < b_k(x, y) \leq C(k, y) < \infty,$$

then for all k and all $y_{1:k}$,

$$\left\| \hat{p}^N(x_k|y_{1:k}) - p(x_k|y_{1:k}) \right\|_1 \to 0$$

in probability as $N \to \infty$.

This result states that at each time point, the empirical distribution of the particles converges to the underlying true posterior density when the number of particles goes to infinity. The conditions imposed above are quite weak. For example, it is straightforward to verify that these conditions are satisfied for the widely used additive Gaussian noise model.

The convergence of SMC sampling shows that the empirical distribution converges to the true distribution. However, the availability of a particle-based approximation to the posterior distribution, though critical, does not answer the question of the approximation error for the particle-based point estimates. The following central limit theorem (Künsch, 2005) shows that under very weak conditions, the SMC approximation of the estimate based on the empirical distribution $\hat{p}^N(x_k|y_{1:k})$ converges to the true estimate at the rate of $1/\sqrt{N}$.

Theorem 2. Under the conditions in Theorem 1, for each finite k, and all $y_{1:k}$ and functions $g(\cdot)$ that are square integrable with respect to the true posterior distribution,

$$\sqrt{N}\left(\sum_{i=1}^{N} g(x_k^{(i)})w^{(i)} - E(g(x_k))\right)$$

is asymptotically normal.

This result is a highly simplified version of the one in Künsch (2005). It is important to note that, as time evolves, the asymptotic variance of the Monte Carlo estimate stay bounded. See Künsch (2005) for details.

1.3 Insight, Improvement, Extension, and Application

The empirical effective sample size in (1.19) is the primary statistic to control resampling frequency. However, its properties are rarely discussed. By introducing

the new expectation-normalized weight, some insights on the empirical effective sample size are investigated theoretically in Chapter 2. In addition, using this newly defined weight, we show the advantage of optimal importance function (1.12) over the prior importance function (1.10). Thus, whenever feasible, it makes sense to use the optimal importance function.

As a simulation method, performance of SMC is subject to prior settings. An incompatible prior, e.g. poor initial guess, could adversely affect SMC performance. We recommend the use of predictive density value to detect the possible existence of an incompatible prior. If the prior is diagnosed to be incompatible with the observed data, we suggest a numerical smoothing method, based on the moving horizon estimation, to find a prior compatible with initial observations. Detailed discussions are in Chapter 3.

Regular SMC methodology does not apply to the estimation of constrained states. Chen (2004) extends SMC to constrained estimation by an extra acceptance/rejection step. In Chapter 4, its asymptotic convergence property is verified and its performance is tested with a more complex model.

SMC has seen many applications in different areas as surveyed in Doucet et al. (2001). A novel application of SMC is introduced for the population pharmacokinetic model estimation in Chapter 5, and its performance is compared with the traditional Gibbs Sampler.

CHAPTER 2

SOME INSIGHTS ON IMPORTANCE WEIGHTS AND EFFECTIVE SAMPLE SIZE

In this chapter some asymptotic properties of the empirical effective sample size, $\tilde{N}_k$, in (1.19), are developed as the particle size N goes to infinity. Two different resampling control strategies, either resampling at every time point or resampling when necessary, are considered here.

The sum-normalized nature of w_k in Equation (1.6) makes it extremely difficult, if not impossible, to analytically track the properties of $\tilde{N}_k$. We introduce a newly defined importance weight, $\bar{w}_k$, in Equation (2.1). To distinguish $\bar{w}_k$ from w_k in Equation (1.6), we call $\bar{w}_k$ the expectation-normalized importance weight. It is shown in the following that $\tilde{N}_k$ is a consistent estimate of $\bar{N}_k$ and its variations based on $\bar{w}_k$. We call $\bar{N}_k$ and its variations the expectation-normalized effective sample size. With the use of expectation-normalized importance weight, it is easy to study the distribution of the effective sample size for the linear Gaussian model, when each new observation is treated as a random variable before being observed.

In addition, the superiority of optimal importance function over the prior importance function is established by showing that $\bar{N}_k$ is larger under the former importance sampling scheme. The new effective sample size does not depend on the generated

samples, thus making it a good criterion to compare the performance of different importance functions. As a matter of fact, when optimal importance function is used, the variance of weight $w_k^{(i)}$ is zero under the distribution $[x_k|x_{k-1}^{(i)}, y_k]$ since the importance weight $w_k^{(i)}$ depends only on $x_{k-1}^{(i)}$ and y_k. However, such an argument does not seem to be convincing. In this chapter, we use the new effective sample size to formally prove that the optimal importance function leads to a larger $\bar{N}_k$.

The rest of this chapter is organized as follows. The expectation-normalized importance weight is defined at first. Then the asymptotic properties of effective sample size are introduced under different resampling-control schemes. In the end, the theoretical behavior of effective sample size is introduced for the linear Gaussian model, if we treat a new observation as a random variable before it is obtained.

2.1 Expectation-normalized Importance Weights

Without further repeating, it is assumed from now on that the requirements in Theorem 1 are satisfied. The new expectation-normalized importance weight $\bar{w}_k$ is given by:

$$\begin{aligned} \bar{w}_k &= \frac{p(x_k|y_{1:k})}{p(x_k|y_{1:k-1})}, \\ &= \frac{p(y_k|x_k)}{p(y_k|y_{1:k-1})}. \end{aligned} \tag{2.1}$$

As seen in its definition (2.1), $\bar{w}_k$ is the importance weight of samples drawn from the importance function $p(x_k|y_{1:k-1})$ for the desired posterior density function $p(x_k|y_{1:k})$. The second step simplification in (2.1) shows that $\bar{w}_k$ is exactly the likelihood value $p(y_k|x_k)$ normalized by its expectation with regard to the importance density function, since $p(y_k|y_{1:k-1}) = \int p(y_k|x_k)p(x_k|y_{1:k-1})dx_k$. And this justifies the name of "expectation normalized".

In effect, Nw_k is a consistent estimate of $\bar{w}_k$ if we use the prior importance function in the SMC sampling and resample particles at every time point. It can be shown that under the above assumptions $\tilde{w}_k^{(i)} = \frac{p(y_k|x_k^{(i)})}{N}$ and $w_k^{(i)} = \frac{p(y_k|x_k^{(i)})}{\sum_{i=1}^N p(y_k|x_k^{(i)})/N}$ using Equation (1.11) and Equation (1.6), where $x_k^{(i)}$ is drawn independently from $p(x_k|y_{1:k-1})$ for $i = 1, \ldots, N$. As $N \to \infty$, $\frac{\sum_{i=1}^N p(y_k|x_k^{(i)})}{N} \xrightarrow{\text{p}} p(y_k|y_{1:k-1})$, where $\xrightarrow{\text{p}}$ denotes convergence in probability. As a result, $Nw_k^{(i)} \xrightarrow{\text{p}} \bar{w}_k^{(i)}$ as $N \to \infty$.

2.2 Asymptotic Properties of ESS Based on Expectation-normalized Weights

Based on the expectation-normalized importance weight, define $\bar{N}_k$ as

$$\begin{aligned} \bar{N}_k &= \frac{1}{1 + \text{Var}^{x_k|y_{1:k-1}}\{\bar{w}_k\}}, \\ &= \frac{1}{\mathbb{E}^{x_k|y_{1:k-1}}\{\bar{w}_k^2\}}, \\ &= \frac{p^2(y_k|y_{1:k-1})}{\mathbb{E}^{x_k|y_{1:k-1}}\{p^2(y_k|x_k)\}}. \end{aligned} \tag{2.2}$$

Lemma 2.2.1. *For the SMC estimation under general state-space models, if resampling is done at every time point and prior importance function is used, then*

$$\tilde{N}_k \xrightarrow{p} \bar{N}_k, \qquad as\, N \to \infty.$$

Proof. With the above discussion, it can be shown that $N \sum_{i=1}^N \left[w_k^{(i)}\right]^2 \xrightarrow{\text{p}} \mathbb{E}^{x_k|y_{1:k-1}}\{\bar{w}_k^2\}$. Then the result follows. □

Lemma 2.2.2. *Assume that no resampling is done before the current time point k and prior importance function is used for a general state-space model. Define $\bar{N}_{1:k}$ as*

$$\bar{N}_{1:k} = \frac{p^2(y_{1:k})}{\mathbb{E}^{x_{1:k}}\{p^2(y_{1:k}|x_{1:k})\}}. \tag{2.3}$$

Then

$$\tilde{N}_k \xrightarrow{p} \bar{N}_{1:k}, \qquad as\, N \to \infty.$$

Proof. When no resampling is performed and prior importance function is used, it is equivalent to drawing joint samples $x_{1:k}^{(i)}, i = 1, \ldots, N$ from $[x_1, \ldots, x_k]$. Then it can be shown that the above result holds. □

Now we extend the Lemma 2.2.1 or 2.2.2 to the case of controlled resampling.

Proposition 2.2.3. *Assume the most recent resampling is done at time point $j-1$ with $j \geq 1$ before the current time k. Define $\bar{N}_{j:k}$ as*

$$\bar{N}_{j:k} = \frac{p^2(y_{j:k}|y_{1:j-1})}{\mathbb{E}^{x_{j:k}|y_{1:j-1}}\{p^2(y_{j:k}|x_{j:k})\}}. \tag{2.4}$$

Then

$$\tilde{N}_k \xrightarrow{p} \bar{N}_{j:k}, \qquad as\, N \to \infty.$$

Proof. When resampling is performed at time point $j-1$, the resampled $x_{j-1}^{(i)}, i = 1, \ldots, N$ is regarded as a realization of the underlying posterior distribution. The further sampling from the prior importance function is equivalent to drawing from the true posterior distribution $p(x_{j:k}|y_{1:j-1})$. Then the above results holds. □

The following propositions establish that the effective sample size based on the expectation-normalized weight for the optimal importance function is larger than that for the prior importance function. By using expectation-normalized weights, the effective sample size does not depend on the generated samples, whereas the sum-normalized effect sample size in Equation (1.19) changes value with different samples. Therefore, it is now feasible to theoretically compare the performance of importance functions using the new formulation. A larger effective sample size generally indicates a better performance in generating samples. It is noted that the effective sample size serves as a performance comparison criterion in this chapter, but not for the purpose of

resampling frequency control as described in the SMC sampling algorithm in Chapter 1.

Proposition 2.2.4. *For general state-space models, if resampling is done at every time point and optimal importance function is used, then*

$$\tilde{N}_k \xrightarrow{p} \bar{N}_k^o, \qquad as\, N \to \infty.$$

where $\bar{N}_k^o$ *is defined as*

$$\bar{N}_k^o = \frac{p^2(y_k|y_{1:k-1})}{\mathbb{E}^{x_{k-1}|y_{1:k-1}}\{p^2(y_k|x_{k-1})\}}. \tag{2.5}$$

Proof. With optimal importance function and resampling every time, the unnormalized weight $\tilde{w}_k^{(i)} = \frac{p(y_k|x_{k-1}^{(i)})}{N}$, which is derived from Equation (1.13) and where $x_{k-1}^{(i)}, i = 1, \ldots, N$ are drawn from $p(x_{k-1}|y_{1:k})$. It can be shown that $\frac{\sum_{i=1}^N w_k^{(i)}}{N} \xrightarrow{\text{p}} p(y_k|y_{1:k-1})$ as $N \to \infty$. By using the sum-normalized equation (1.6), the result can be proved. □

Proposition 2.2.5. *Assume that resampling is done at every time point in SMC sampling. Let* $\bar{N}_k$ *be the expectation-normalized effective sample size resulted from using the prior importance function, and* $\bar{N}_k^o$ *be expectation-normalized effective sample size obtained with optimal importance function. Then*

$$\bar{N}_k^o > \bar{N}_k,$$

for every k.

Proof.

$$
\begin{aligned}
&\mathbb{E}^{x_k|y_{1:k-1}}\left\{p^2(y_k|x_k)\right\} \\
=&\left[\mathbb{E}^{x_k|y_{1:k-1}}\left\{p(y_k|x_k)\right\}\right]^2 + \mathrm{Var}^{x_k|y_{1:k-1}}\{p(y_k|x_k)\}, \\
\geq& p^2(y_k|y_{1:k-1}) + \mathrm{Var}^{x_{k-1}|y_{1:k-1}}\{\mathbb{E}^{x_k|x_{k-1},y_{1:k-1}}\{p(y_k|x_k)\}\}, \\
=& p^2(y_k|y_{1:k-1}) + \mathrm{Var}^{x_{k-1}|y_{1:k-1}}\{\mathbb{E}^{x_k|x_{k-1}}\{p(y_k|x_k, x_{k-1})\}\}, \\
=&\left[\mathbb{E}^{x_{k-1}|y_{1:k-1}}\left\{p(y_k|x_{k-1})\right\}\right]^2 + \mathrm{Var}^{x_{k-1}|y_{1:k-1}}\{p(y_k|x_{k-1})\}, \\
=&\mathbb{E}^{x_{k-1}|y_{1:k-1}}\left\{p^2(y_k|x_{k-1})\right\}.
\end{aligned}
$$

Therefore, $\bar{N}_k^o > \bar{N}_k$ by using the equations in Lemma 2.2.1 and Proposition 2.2.4. Note that the above proof uses the following result

$$
\begin{aligned}
\mathrm{Var}(U) &= \mathrm{Var}(\mathbb{E}(U|V)) + \mathbb{E}(\mathrm{Var}(U|V)), \\
&\geq \mathrm{Var}(\mathbb{E}(U|V)).
\end{aligned}
$$

for any random variables U and V. □

2.3 Expectation-normalized effective sample size in Linear Gaussian Models

In this section, $\bar{N}_k$ is assumed to be a random variable before an observation becomes available at time point k. That is, y_k is treated as a random variable and its predictive distribution is given as $p(y_k|y_{1:k-1})$ with all the previous observations $y_{1:k-1}$ known.

Lemma 2.3.1. *Assume that the prior importance function is used with resampling at every time point. Then the distribution* $-2\log(\bar{N}_k)$ *is equivalent to a weighted sum of* d_y *independent Chi-square random variables, each of which has one degree of freedom. Particularly,* $-2\log(\bar{N}_k)$ *is equivalent to a chi-square distribution with one degree of freedom when* $d_y = 1$.

Proof. Using the result in Equation (C.3) in Appendix C, we have the solution for $\bar{N}_k$ as

$$\bar{N}_k = c_k \cdot e^{-\frac{1}{2}(y_k - \xi'_{k|k-1}\Delta^{-1}_{k|k-1}(y_k - \xi_{k|k-1})}, \tag{2.6}$$

where

$$\Delta_{k|k-1} = \Psi_{k|k-1}\left[1 + \frac{1}{2}\left(HP_{k|k-1}H\right)^{-1}R\right], \tag{2.7}$$

c_k is a proportional constant for time point k, $\xi_{k|k-1} = H\mu_{k|k-1}$, and $\Psi_{k|k-1} = HP_{k|k-1}H' + R$. □

Therefore,

$$-2\log(\bar{N}_k) = -2\log(c_k) + (y_k - H\mu_{k|k-1})'\Delta^{-1}_{k|k-1}(y_k - H\mu_{k|k-1}). \tag{2.8}$$

By Kalman filter theory presented in Appendix A, $[y_k|y_{1:k-1}]$ is $\mathcal{N}(\xi_{k|k-1}, \Psi_{k|k-1})$. Therefore $(y_k - H\mu_{k|k-1})'\Psi_{k|k-1})^{-1}(y_k - H\mu_{k|k-1})$ is a chi-square distribution of d_y degree of freedom. However, the normalized matrix for $(y_k - H\mu_{k|k-1})$ is not its covariance matrix in (2.8). As a result, $-2\log(\bar{N}_k)$ is equivalent to a linear combination of chi-square random variables of 1 degree of freedom. The linear coefficients are eigenvalues of $\Psi_{k|k-1}$.

This property is useful for model checking. A feasible approach is to compute the value of $\bar{N}_k$ for the observed y_k and compare to a threshold value which is decided according to the distribution of $\bar{N}_k$. Detailed discussions are in Chapter 3.

CHAPTER 3

POOR PRIOR SMOOTHING WITH PREDICTIVE DENSITY

Model checking is an important component of statistical inference in practice. It investigates the possibility of the posited model that is assumed to generate the observed data. In practical Bayesian analysis, checking the compatibility of the prior distribution with the data is also a part of model checking. These two components of the model can be tested as a whole with some statistics as illustrated in many papers; see, *e.g.*, Guttman (1967), Box (1980), Bayarri and Berger (2000). One can also assume that one of these two components is correct and test the validity of the other one. For example, Evans and Moshonov (2006) discuss how to detect an incompatible prior, or assess "prior-data conflict", while assuming that sampling model is appropriate. The above methodology is proposed mainly for static models, where closed form of the posterior predictive density of a sufficient statistics is available. When extended to dynamic models, these (sufficient) statistics either do not exist or are hard to compute. Furthermore, appropriate remedies for dealing with incompatible prior are not fully discussed, once it is detected.

In this chapter, we introduce both prior checking and improvement for state-space models. Prior checking may seem unnecessary in that its effect usually diminishes

as the number of observations increase. This happens for static models, when the dimension of parameter space remains fixed. However, for dynamic models, the dimension of state space increases with time and the number of observations at each time step is limited.

Sequential Monte Carlo sampling has been successfully applied to general state-space models for Bayesian inference. Being a simulation method, its performance relies to some extent on the generated samples, or particles. With a poor initial guess, a large fraction of particles will usually be less representative of the underlying distribution of the states, which could cause the SMC estimates to diverge from the true values.

In this Chapter, Moving Horizon Estimate is creatively combined with the sequential Monte Carlo method to obtain a stable estimate of initial states. Inside each window, MHE gives a smoothed estimate of the states at the beginning of the window, since some "future" data in the same window is used in their estimation. We find that combining a Moving Horizon Smoother with SMC is very effective for recovering from a poor prior, and develop an integrated approach that combines these two powerful tools.

A statistic based on the predictive density value, $p(y_k|y_{1:k-1})$, is proposed to evaluate the compatibility of the provided prior distribution with the data model (likelihood). When the statistic goes beyond a threshold value, the smoothing step is triggered to keep SMC estimation on track. Its theoretical properties are discussed in the context of linear Gaussian model and generalized to nonlinear state-space models

The rest of this chapter is organized as follows. In the next section, we briefly describe model validation and smoothing methods. Following the review are discussions of the predictive density and MHE smoothing. Finally, two models are used to demonstrate the performance of this strategy in a reliable manner. In addition, one model illustrates situations in which this method could possible fail to improve the prior. This seems to occur when the measurements do not provide enough information about the underlying states.

3.1 Background

3.1.1 Model Checking

Model checking can be regarded as a preliminary analysis to tell whether, or to what extent, a posited model is compatible with observed data. It is important in that misleading inferences could be obtained under an inappropriate model. Usually, model checking is performed by finding a statistic which gives a measure of surprise associated with the observed data under the posited model. Bayarri and Berger (1997) provide a fairly comprehensive review of the model checking literature. Broadly speaking, frequentist methods aim to provide a threshold for some statistic, while Bayesian methods base the evaluation on the posterior distribution, as discussed in Guttman (1967), Rubin (1984), Bayarri and Berger (2000). Only a brief introduction is given below.

In a simplified setup for model checking, assume random variable $\mathbf{Y}$ is modeled to follow a probability distribution $p(\mathbf{Y})$. A statistic $T(\mathbf{Y})$ is designed to give a small value when the observed event leads to a surprise under the posited model. Then the observed T is compared with a selected threshold value that decides whether or not

the observed data y is compatible with the assumed model. Several statistics have been proposed in the literature, for example, the density value itself. In fact, for a discrete random variable Y, Weaver (1948) defines a surprise index for an observed event $\{Y = y\}$ as $\mathbb{E}\{p(Y)\}/p(y)$, where $p(y) = \text{Prob}\{Y = y\}$. Based on this ratio, Weaver (1948) points out that a rare event is not necessarily a surprising event, and that a surprising event is decided relative to other events. Good (1956) generalized this idea by replacing $p(y)$ with $logp(y)$ in this ratio, and linking the surprise index to entropy. In fact, one can use $g(p(y))$, where $g()$ is a monotone convex function. However, note that since the numerator (expectation value of $g()$) is a constant, it does not matter which $g()$ is chosen, if a threshold value is used to assess the extant of surprise.

In this chapter, we assume that the state-space model (1.1) is valid and the prior on the initial state is a to-be-verified component of the fully specified model. The compatibility between the specified prior and the observations is tested through the frequentist sequential predictive density values, as discussed in section 3.1.5.

3.1.2 Smoothing Methods

Of the many smoothing schemes, the Monte Carlo smoother (Kitagawa, 1994, 1996) is a natural extension to the existing SMC framework in that it resamples some previously obtained samples with appropriate weights. However, the samples are actually implicitly assumed to well represent some filtering distribution. Unfortunately, this is not the case when the prior is poorly specified. Resampling the ill-positioned samples does not improve their reliability.

Chen et al. (2004) suggested an approach based on empirical Bayes sequential importance sampling/resampling (EBSIR). It ignores the given prior and, instead, samples the initial state from a uniform distribution. After the first observation y_1 is available, a point estimate of the initial state is obtained to replace the given prior mean. Then the regular SMC sampling begins with the new prior. This method does compensate for a poor prior, but is limited to models with an invertible observation function. In addition, it can use only the observation y_1 for smoothing, and cannot go further.

Rauch et al. (1965) presented the well known RTS smoother, which gives an analytic solution to the smoothed posterior for linear Gaussian state-space models. Essentially, it performs forward Kalman Filtering or extended Kalman Filtering as time evolves, saves intermittent data and then performs backward smoothing from the latest time point to a desired earlier time point, *i.e.* time 1, in the case of dealing with a poor initial prior. Like EKF, it uses Taylor series expansion to linearize equations (1.1a) and (1.1b) for nonlinear Gaussian models. Such approximation could make the smoothing unreliable as shown via a case study in Section 3.2.2.

Moving Horizon Estimation employs the widely used maximum a posterior (MAP) criterion to find the joint estimate of states. For the sake of computational feasibility, the estimation horizon is set to be of a small fixed length, that is,

$$(\hat{x}_{k-M+1}, \ldots, \hat{x}_k) = \underset{x_{k-M+1:k}}{\arg\max}\, p(x_{k-M+1:k}|y_{1:k}), \tag{3.1}$$

where M is the window length. Under widely used assumptions that state noise and measurement noise are additive Gaussian, this becomes a least squares optimization problem, and can be solved efficiently with optimization algorithms (Robertson et al., 1996).

A poor prior can cause the performance of the SMC method to be significantly worse than MHE and EKF as illustrated by Chen et al. (2004). Theoretically, the prior, good or poor, does not impose any problem in the quality of the posterior calculation. In fact, the forgetting property states that two hidden Markov chains, with the only difference between them being the initial distribution, approach each other with the passage of time. That is, the initial prior can have larger effects only on the estimation of earlier states. However, since SMC is a simulation method under a finite set of particles, it could fail to recover from the poorly located particles drawn from the poor initial prior. As time evolves, the error accumulates and the posterior continues to get worse.

It seems that a poor initial prior could be well compensated by a few early observations because $p(x_1|y_{1:k})$ contains more information than $p(x_1)$ for every value of k when the model is correct. In this chapter, we explore the use of smoothing strategies to search for a more compatible prior for reliable future inferences when an appropriate number of system observations are available.

Since the smoothing operation has to wait for some data to become available, and performs extra extensive computations to find a compatible prior and then re-estimate system states, it should be done only when necessary. That said, a statistic needs to be developed to detect whether or not the given prior is compatible with the observed data. Furthermore, a strategy should also be provided to decide how many data are used for smoothing. Detailed discussions on these issues are presented in the sequel.

3.1.3 Smoothing under a Linear Gaussian Model

In practice, too few observations do not add much information on the system states, while a large number of observations usually waste computing resources unnecessarily, when the forgetting property holds. Some explanations are given for the linear Gaussian model, under which MHE is equivalent to Kalman Filter and analytical solutions are recursively available.

Analytical recursive computation of smoothed posterior, $p(x_\ell|y_{1:k}), k > \ell$, is available for linear Gaussian models; see appendix A for details. For simplicity, a 1-dimensional linear Gaussian model is used, *i.e.*, F and H are both scalers in Equation (1.2). It has less smoothing effects with increasing smoothing length.

Define $\eta_k = \mu_{1|k} - \mu_{1|k-1}$ for any $k > 1$, that is, η_k denotes the additional adjustment with one more observation being used to do smoothing at time k. If it is expected that such adjustment is smaller stochastically, smoothing operation might not help much. It immediately follows from the RTS smoothor in Appendix A that

$$\eta_k = \left(\prod_{i=\ell}^{k-1} C_i\right) K_k(y_k - H\mu_{k|k-1}), \tag{3.2}$$

where C_k is defined in Equation (A.3) and K_k is defined in Equation (A.1). Since $E\{\eta_k\} = 0$ with respect to $y_k|y_{1:k-1}$, its variance is considered next. Let $\tau_k = \mathbb{E}\{\eta_k^2|y_{1:k-1}\}$. Then

$$\tau_k = \left\{\left(\prod_{i=1}^{k-1} C_i\right) K_k\right\}^2 \left(HP_{k|k-1}H^{'} + R\right) \tag{3.3}$$

Therefore,

$$\begin{aligned} \zeta_k &= \frac{\tau_k}{\tau_{k-1}}, \\ &= \frac{R}{P_{k|k-1}H^2 + R} \cdot \frac{R}{P_{k-1|k-2}H^2 + R}, \\ &< 1. \end{aligned} \tag{3.4}$$

Thus τ_k is decreasing and therefore η_k gets more and more concentrated around 0 as k increases. Theoretically, it is possible to calculate τ_k and then decide if it is worth using more data to smooth.

Similarly for general state-space models, one can use smoothing progressively and then decide if it is enough for the smoothing. The next subsection describes a statistic that does not only detect an incompatible prior, but also tells how many observations are needed for smoothing.

3.1.4 MHE Smoothing for Initial State

In this section, we use a robust and accurate smoothing method based on MHE. This smoother, derived in Tenny (2002), enhances the accuracy of MHE by estimating the arrival cost (prior) for a window of data based on all the observations in the window. Current research at the forefront of estimation of nonlinear dynamic systems includes methods based on optimization like MHE, and based on simulation like SMC. We propose a novel combination of MHE smoothing with SMC online filtering. MHE smoothing can guard SMC from slow convergence due to a poor initial guess while still maintaining SMC's fast and accurate estimation performance, particularly after the smoothing operation. The poor initial guess is detected through the use of predictive density values at each time. Such detection is performed for only a limited duration of time from the beginning, since SMC is likely to be on a good track when predictive density values are large for several consecutive time points.

The MHE formulation (Tenny, 2002) will now be tailored to the specific prior smoothing problem. Assuming that measurement equation is additive Gaussian and MHE smoothing is performed to a segment of observations $y_{1:M}$, where M is the

window length, the objective function for MHE smoothing is:

$$\min \Upsilon_e(\rho) + \frac{1}{2}\sum_{k=1}^{M} \mathcal{L}_e(\omega_j, \nu_k),$$

where

$$\begin{aligned} x_1 &= \mu_1 + \rho, \\ \Upsilon_e(\rho) &= \frac{1}{2}\rho P^{-1}\rho, \\ \mathcal{L}_e &= \omega_k^T Q^{-1}\omega_k + \nu_k^T R^{-1}\nu_k, \end{aligned} \tag{3.5}$$

and for each k:

$$\begin{aligned} x_k &= f(x_{k-1}, \omega_{k-1}), \\ y_k &= h(x_k) + \nu_k, \\ Ax_k &\leq a, B\omega_k \leq b, C\nu_k \leq c. \end{aligned}$$

In the above equation, the noises are assumed to be independent Gaussian random variables with Q and R being the variance of the state noise and measurement noise, respectively. The prior $\pi(\cdot)$ on x_1 is also Gaussian, $\mathcal{N}(\mu_1, P_1)$. After MHE smoothing is done over a segment of M observations $y_{1:M}$, the prior mean μ_1 is replaced by an adjusted value, $\mu_{1|M} = \mu_1 + \rho$, which starts the regular SMC estimation. The new prior distribution is denoted $\pi_{1|M}$, which is $\mathcal{N}(\mu_{1|M}, P_1)$.

Since MHE is a numerical optimization method and not a simple resampling process like Monte Carlo smoother, the MHE smoothed prior is robust to a misspecified prior because observations have been used to learn about the prior hyper-parameters. MHE smoother is also likely to be better than a noninformative prior because it incorporates some prior information. Therefore, MHE estimation may provide valuable information for the early period of the observations to help its recovery from a poor prior and using this smoothing estimate presents a good start for simulation methods

like SMC (Lang et al., 2006). Once the prior is of good quality, the SMC filter can be used by itself to avoid the time delay due to smoothing. The currently implemented MHE assumes multivariate Gaussian or other fixed shape of distributions to represent the prior knowledge or arrival cost at the start of a window (Rao and Rawlings, 2000). This assumption does not hurt the performance because the benefit of a prior with a continuous distribution outweighs the disadvantages of having a poor prior. Of course, it is important to be able to detect when smoothing should be used, which the proposed model checking via predictive density is able to accomplish.

3.1.5 Predictive Density

Prior compatibility detection statistic, predictive density γ_k, is defined as

$$\gamma_k(y_k, \pi) = p(y_k|y_{1:k-1}). \tag{3.6}$$

Note that the prior π is explicitly included in the above equation (3.6) to indicate that the posterior predictive density depends on a specific prior, assuming x_1 has been integrated out using the prior. Intuitively, γ_k is expected to be large if the specified prior is compatible with the observed data where the specified model is correct and the observation is not an outlier. Weaver (1948) and Good (1956) used similar statistic for static models. A direct use of their proposed statistic would be the joint density value $p(y_{1:k})$ at time point k. The proposed posterior predictive density is an extension to the dynamic models and works naturally under the framework of SMC. In fact, $\prod_{\ell=1}^{k} \gamma_\ell$ is exactly the joint density $p(y_{1:k})$. In addition, by treating y_k as a random variable prior to its observing at time k, the distribution of Γ_k, the random variable γ_k, provides valuable information for the realized value of $p(y_k|y_{1:k-1})$ after

y_k is observed in the sense that one can assess if the observed value of y_k is surprising under this density.

When combined with MHE smoothing, predictive density value reflects the effect of a new prior provided by MHE smoother. New threshold value may be needed. Intuitively, replacing the original prior distribution with the smoothed tends to increase the predictive density value. This is verified for the linear Gaussian models in Appendix B, *i.e.,*

$$\gamma_1(y_1, \pi_{1|1}) \geq \gamma_1(y_1, \pi),$$

where $\pi_{1|1}$ is the prior based on the smoothing with only y_1. Next, we further discuss its computation under the linear Gaussian model and apply it to nonlinear model based on SMC.

Linear Gaussian Model

Using Kalman filter theory presented in Appendix A, we obtain $Y_k|y_{1:k-1} \sim \mathcal{N}(\xi_{k|k-1}, \Psi_{k|k-1})$, where $\xi_{k|k-1} = H\mu_{k|k-1}$, and $\Psi_{k|k-1} = H'P_{k|k-1}H + R$. So the analytical form of γ_k for any y_k is

$$\gamma_k = \frac{1}{(2\pi)^{d_y/2}|\Psi_k|^{1/2}} \exp\left\{-\frac{1}{2}(y_k - \xi_{k|k-1})'\Psi_{k|k-1}^{-1}(y_k - \xi_{k|k-1})\right\}, \tag{3.7}$$

where d_y is the dimension of the observation y_k.

Assume γ_k to be a random variable before y_k is observed. As we know $(Y_k - \xi_{k|k-1})'\Psi_{k|k-1}^{-1}(Y_k - \xi_{k|k-1}) \sim \chi^2_{d_y}$, we can write the CDF or PDF of Γ_k in a closed form. A threshold value can be selected in its lower tail at a predefined significance level. As a matter of fact, this is equivalent to testing whether the normalized residual $(y_k - \xi_{k|k-1})'\Psi_{k|k-1}^{-1}(y_k - \xi_{k|k-1})$ is greater than a threshold value $\chi^2(1-\alpha, d_y)$, where $\chi^2(p, df)$ is the inverse CDF at the probability value p for a Chi-square random

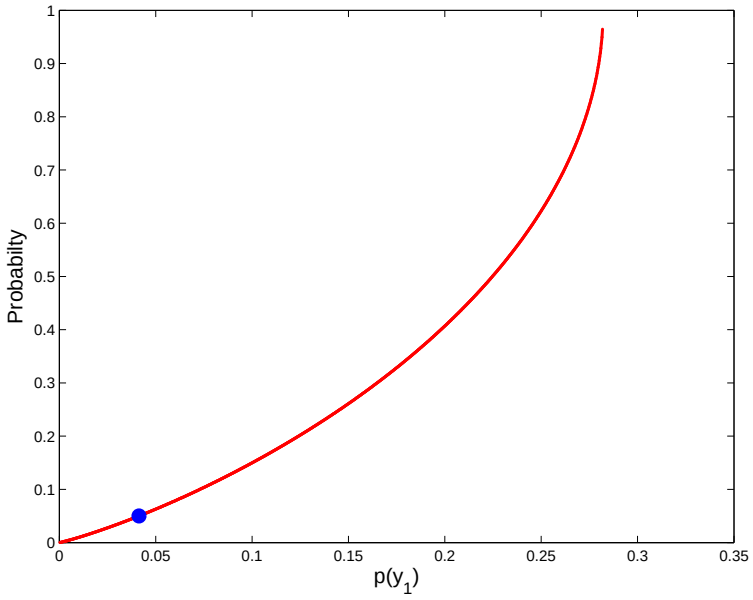


Figure 3.1: Theoretical CDF of Γ_1 under linear Gaussian model

variable with *df* degrees of freedom. That said, if the actual γ_k for the observed y_k is smaller than the selected threshold value, then the given prior is declared to be incompatible with the observed data at time point k. Smoothing via MHE or RTS, is then performed to find $\mu_{1|\ell}$ such that new predictive density values exceed their respective threshold values for all observation $y_{1:\ell}$ after $\mu_{1|\ell}$ replaces the original prior mean. The theoretical CDF of Γ_1, *i.e.* $p(y_1)$, is drawn in Fig. 3.1. The dot on the CDF curve is the threshold value at the 5% level.

An important property of the predictive density γ_k is that its threshold value does not depend on the prior mean for any k in the linear Gaussian model. This also applies to γ_1 under the state-space model with Gaussian prior and a linear measurement equation. The case study of the CSTR model in Section 3.2.1 demonstrates this

feature though the CSTR model has only a truncated Gaussian prior distribution. See Section 3.2.1 for more details.

Nonlinear Non-Gaussian Models

Under general nonlinear models, $\xi_{k|k-1}$ and $\Psi_{k|k-1}$ can be readily computed by the EKF when additive Gaussian noises are assumed in Equation (1.1). A more reliable estimate is also available with the framework of particle filtering through Rao-Blackwellization:

$$\begin{aligned}\gamma_k &= p(y_k|y_{1:k-1}) \\ &= \int p(y_k, x_k|y_{1:k-1})\, dx_k \\ &= \mathbb{E}^{x_k|y_{1:k-1}} \{p(y_k|x_k, y_{1:k-1})\} \\ &= \mathbb{E}^{x_k|y_{1:k-1}} \{p(y_k|x_k)\}\end{aligned}$$

Then γ_k can be approximated as:

$$\hat{\gamma}_k \approx \frac{1}{N} \sum_{i=1}^{N} p(y_k|x_{k|k-1}^{(i)}) \tag{3.8}$$

With the above approximation, we can find empirical distribution of Γ_k by propagating the resampled filtered particles $\{x_{k-1|k-1}^{(i)}, i = 1, \ldots, N\}$ through the state equation to generate samples $\{x_{k|k-1}^{(i)}, i = 1, \ldots, N\}$. Predictions $\{y_k^{(i)}, i = 1, \ldots, N\}$ are sampled from the measurement equation. Value $\gamma_k(y_k^{(i)}), i = 1, \ldots, N$ is obtained. Finally, a threshold value is obtained at some fixed percentile of the resulting empirical CDF of γ_k.

3.1.6 Smoothing Length

Smoothing length is also decided by the same statistic, predictive density. It is done by progressively increasing the number of observations and setting the smoothing

Let $\ell = 0$
FOR times $k = 1, \ldots,$ S
— Make the regular SMC
— Calculate predictive density value γ_k
— Find the threshold value for time k
— IF γ_k is less than its threshold value
– $\ell = \ell + 1$
– IF $\ell > S$
– Exit Smoothing
– END IF
– Use MHE smoothing to find $\mu_{1|\ell}$
– Set the Gaussian prior mean to be $\mu_{1|\ell}$
– Reset $k = 1$ and restart the SMC estimation
— END IF
END FOR

Table 3.1: Algorithm for poor prior smoothing.

length to be the minimum number of observations such that the predictive density values exceed their respective threshold value for all the observation used.

Let S denote the time point until which the poor prior detection is performed, then the algorithm is summarized in Table 3.1. Selection of S needs some considerations. Larger S value provides the opportunity that more observations can be used to perform smoothing. However, prior checking and smoothing are time-consuming operations and may delay SMC from fast on-line estimation. In our case studies in Section 3.2.1 and 3.2.2, S is set to be 15.

3.1.7 Effective Sample Size

Effective sample size $\tilde{N}_k$ is generally used to indicate whether the generated particles can give a good estimate in general. When ESS is large, particles are supposed to fit the model well, thus avoiding unnecessary resamplings as discussed in Section

1.2.2. It is reasonable to expect that ESS may be very small under a bad model specification (including prior).

As stated in Chapter 2, under the linear Gaussian model with one dimensional observation ($d_y = 1$), distribution of $-2\log(\bar{N}_k)$ is equivalent to Chi-square with one degree of freedom before y_k is observed. However, for $d_y \geq 2$, it is distributed as a linear combination of d_y independent Chi-square random variables. Therefore, ESS will be more difficult to use than $-2\log(p(y_k|y_{1:k-1}))$. Furthermore, for general state-space models, predictive density value has an intuitive explanation—the density value of predictive observation conditional on all previous observations. The effective sample size does not have this convincing explanation.

3.2 Successful Case studies

In this section, prior checking and smoothing are demonstrated via simulations of two dynamic systems: a CSTR model (Chen et al., 2004) and a McKeithan reaction network (Chaves and Sontag, 2002). The former case study shows that MHE smoothing has similar performance as RTS smoothing while the latter one demonstrates that MHE smoothing is better than RTS smoothing. Two prior checking statistics, predictive density value and effective sample size are also tested in the McKeithan example. The study shows that predictive density value performs well for general state-space models.

To account for randomness in simulations, a total of L realizations are run for each model with the methods tested on the generated data. Performance of each method is measured by mean-squared error averaged across realizations, MSE_k^R, which is defined

as

$$\mathrm{MSE}_k^R = \frac{1}{L}\sum_{r=1}^{L}(x_{k,r} - \hat{x}_{k,r})^T(x_{k,r} - \hat{x}_{k,r}), \tag{3.9}$$

for all time k. Examining MSE_k^R over time is likely to indicate the long term behavior of the tested method and to provide insight into the distribution of errors over time. In the above equation, $x_{k,r}$ is the true state value at time point k in the r-th realization and $\hat{x}_{k,r}$ is its point estimate. For the SMC and its variants, the state estimate is the mean value of the posterior distribution, whereas for the MHE, it is the mode of the approximated posterior.

3.2.1 CSTR example

The MHE-smoothing based SMC algorithm is applied to a continuously stirred tank reactor (CSTR) under poor initial guess, which was also studied in (Chen et al., 2004). The states C and T are modeled by the following equations:

$$\begin{aligned}
\frac{dC}{dt} &= \frac{q}{V}\,(C_0 - C) - k\,C\,e^{-E_A/T}\ , \\
\frac{dT}{dt} &= \frac{q}{V}\,(T_0 - T) - \frac{\Delta H}{\rho\,C_p}\,k\,C\,e^{-E_A/T} - \frac{U\,A}{\rho\,C_p\,V}(T - T_c)\ ,
\end{aligned}$$

where C is concentration, T is temperature and others are known model parameters, whose values are listed in Table 3.2. Discritization of the above differential equations plus a linear measurement equation give the following state space model:

$$\begin{aligned}
x_k &= \left[\ C_k \quad T_k\ \right]^T \\
&= \begin{bmatrix} (1 - \frac{\Delta t\ q}{V} - \Delta t\ k\ e^{-E_A/T_{k-1}}) & 0 \\ -\frac{\Delta t\ \Delta H\ k\ e^{-E_A/T_{k-1}}}{\rho\ C_p} & (1 - \frac{\Delta t\ q}{V} - \frac{\Delta t\ U\ A}{\rho\ C_p\ V}) \end{bmatrix} x_{k-1} + \\
&\quad \begin{bmatrix} \frac{\Delta t\ q\ C_0}{V} \\ \frac{\Delta t\ q\ T_0}{V} + \frac{\Delta t\ U\ A\ T_c}{\rho\ C_p\ V} \end{bmatrix} + \omega_k\ , \\
y_k &= x_k + \nu_k\ ,
\end{aligned}$$

Param.	Value	Unit	Param.	Value	Unit
q	100	L/min	ΔH	$-50,000$	J/mol
V	100	L	r	1000	g/L
C_0	1.0	mol/L	C_p	0.239	J/g K
k	7.2×10^{10}	1/min	U	5000	J/cm^2 min K
E_A	8750	K	T_0	350	K
A	10	cm^3	T_c	305	K

Table 3.2: CSTR model parameters (Henson and Seborg, 1997).

The operating conditions are listed in Table 3.2. The system noise at the scale of the normalized state variables are $p(\omega) \sim \mathcal{N}(0, Q)$, where $Q = \sigma_\omega^2 I_2, \sigma_\omega^2 = 2.5 \times 10^{-7}$, and the measurement noises are $p(\nu) \sim \mathcal{N}(0, R)$, where $R = \sigma_\nu^2 I_2, \sigma_\nu^2 = 0.0025$. I_m is a $m \times m$ identity matrix. The initial condition is assumed to be

$$x_1 = \begin{bmatrix} 0.5 \\ 3.5 \end{bmatrix}.$$

However, a poor choice of the prior is given as $\mathcal{N}(\mu_1, P_1)$, where $P_1 = \sigma_v^2 \cdot I_2$ and

$$\mu_1 = \begin{bmatrix} 2.5 \\ 3.7 \end{bmatrix}.$$

Furthermore, $K = 400$ measurements are available for the model, and system states need to be estimated at each of these 400 time points.

In this case study, we show that EKF smoothing has relatively similar performance as MHE smoothing in terms that both have close values of $\mu_{1|\ell}$ and they tend to be closer to the true value of x_1 as ℓ increases. We also illustrate the prior-checking process as to how the smoothing length is decided by comparing observed predictive density value to threshold values.

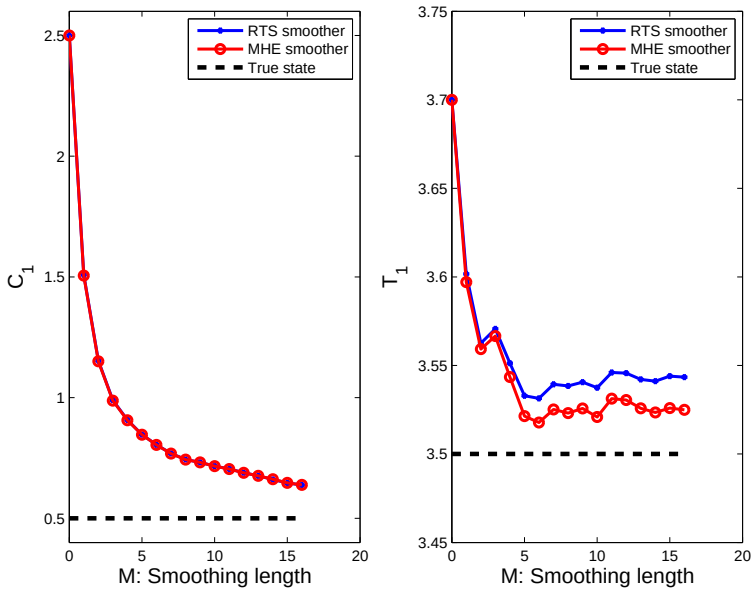
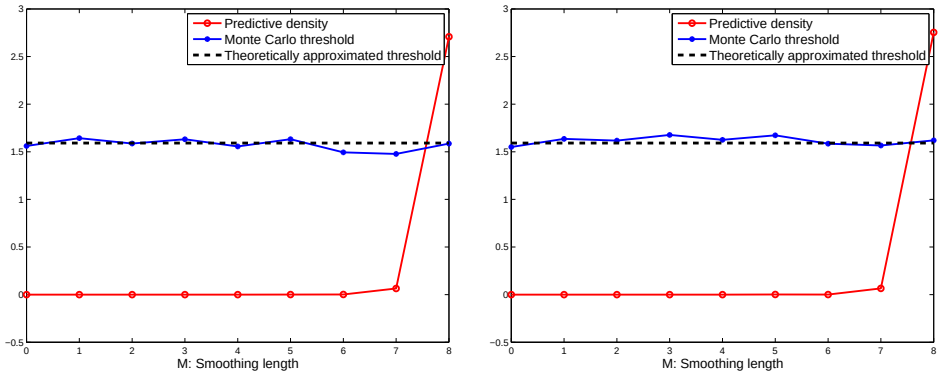


Figure 3.2: Smoothing estimate of initial value

Figure 3.2 shows, in a typical run, the smoothed estimate $\mu_{1|\ell}$ for the initial state under different smoothing length $\ell = 0, 1, \ldots, 16$. Note that $\mu_{1|0}$ denotes the given prior mean. The dashed line represents the true initial state value while dotted solid line draws the smoothed estimates for RTS smoothing and circled solid line is for MHE smoothing. In this case study, MHE and EKF have similar performance for the state C_1 while MHE smoother approaches faster and is closer to the true value of state T_1 than RTS smoother. With the increase of smoothing length, *i.e.*, the number of observations used, the smoothed estimates show a trend in getting closer and closer to the true values. However, the amount of gain in terms of the decrease in estimation error $|x_1 - \mu_{1|\ell}|$ gets smaller and smaller.



(a) MHE Smoothing (b) RTS Smoothing

Figure 3.3: Determine the smoothing length by comparing with threshold value

Figure 3.3 demonstrates the process as to how the smoothing length is determined by comparing the predictive density values to their respective threshold values. Circled line represents the predictive density values with increasing smoothing lengths and solid line is for their corresponding threshold values. In more details, predictive density value γ_1 is found to be smaller than its threshold value at time point 1 at the beginning of SMC estimation. MHE or RTS smoother is used to obtain a smoothed prior mean $\mu_{1|\ell}$ for $\ell = 1, 2, \ldots$, which are shown in Figure 3.2. The SMC sampling restarts with the new prior mean $\mu_{1|\ell}$ and recalculates the new predictive density value, denoted by $\gamma_{1|\ell}$, at time 1. If $\gamma_{1|\ell}$ is still smaller than its threshold value, smoothing procedure continues to find $\mu_{1|\ell+1}$ by using one more observation. Figure 3.3 shows that both MHE smoother and RTS smoother use 8 observations to find a satisfying prior mean, that is, the predictive density value is larger than its threshold value after 8 observations are used for smoothing. MHE smoother and RTS smoother have similar performance as seen in this figure.

Theoretically approximated threshold values are also drawn as dashed line in Figure 3.3 by assuming the prior is Gaussian, whereas it is actually a truncated Gaussian due to constrained state space. These approximated theoretical values are given by

$$\frac{1}{2\pi|R+P_1|^{1/2}}e^{\{-0.5\chi^2(1-\alpha,df)\}}. \tag{3.10}$$

In this study, α is set to be 0.05 and the degree of freedom, df, is 2. Note that these values are very close to the Monte Carlo estimate, and that they fall on a straight line as a function of the smoothing length, since the threshold does not depend on the value of smoothed value of the prior mean $\mu_{1|\ell}$ for any ℓ.

The smoothing length also varies from one simulation to another. In this study we run the CSTR model 100 times with the same initial state value, however, with stochastic realizations of the state values and measurements beyond the initial time point. The resulting 100 empirical smoothing lengths are summarized in Figure 3.4 for both MHE and RTS smoother. The distribution of smoothing length is very similar for both smoothing methods, and both vary from 7 to 16.

At this point, it is worth noting that EBSIR (Chen et al., 2004) can also be applied for this model, since the observation equation admits a closed form of $[x_1|y_1]$. However, it is limited to smoothing length of one, since $p(x_1|y_{1:\ell})$ is not known in closed form when $\ell > 1$. In contrast, MHE smoothing can be implemented for more observations.

3.2.2 McKeithan Network

The McKeithan reaction network was discussed in McKeithan (1995), Chaves and Sontag (2002). It exhibits strong nonlinearities in both the state and the measurement

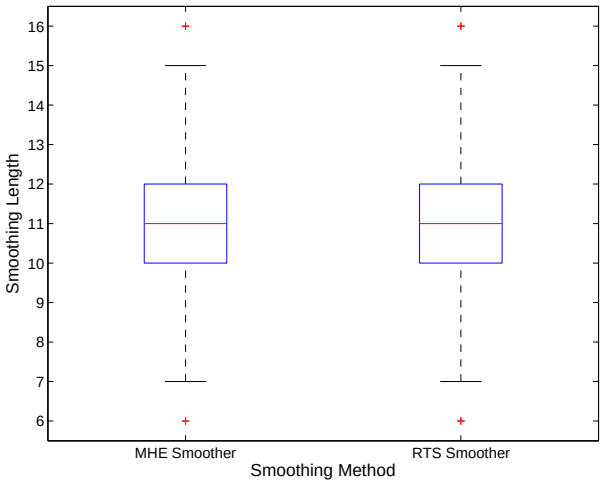


Figure 3.4: Empirical distribution of observed smoothing length over 100 simulations

equations as shown below:

$$\begin{aligned}
\dot{A} &= -k_1 AB + k_3 C + k_4 D, \\
\dot{B} &= -k_1 AB + k_3 C + k_4 D, \\
\dot{C} &= k_2 AB - (k_3 + \beta_3) C, \\
\dot{D} &= \beta_3 C - k_4 D, \\
y &= [AB^2 \quad AD]^T,
\end{aligned}$$

The corresponding discrete time state space representation is as follows:

$$\begin{aligned}
x_k &= \begin{bmatrix} A_k & B_k & C_k & D_k \end{bmatrix}^T \\
&= \Delta t \begin{bmatrix} 1/\Delta t & -kA_{k-1} & k_3 & k_4 \\ -kB_{k-1} & 1/\Delta t & k_3 & k_4 \\ kB_{k-1} & 0 & 1/\Delta t - k_3 - \beta_3 & 0 \\ 0 & 0 & \beta_3 & 1/\Delta t - k_4 \end{bmatrix} x_{k-1} \\
&\quad + \omega_{k-1}, \\
y_k &= \begin{bmatrix} A_{k-1} B_{k-1}^2 & A_{k-1} D_{k-1} \end{bmatrix}^T + \nu_k,
\end{aligned}$$

where $k_1 = k_2 = 6$, $k_3 = 0.5$, $k_4 = 7$, $\beta_3 = 1$, and$\Delta t = 0.01$. The initial value is $x_1 = (1, 3, 3, 2)^T$, while its prior is set to $\mathcal{N}(\mu, \sigma^2 I_4)$,where $\mu = \begin{bmatrix} 5 & 4 & 8 & 7 \end{bmatrix}'$, $\sigma^2 = 0.5$. The system noises are iid Gaussian with covariance matrix $Q = 10^{-4} I_4$. The measurement noises are also iid Gaussian with covariance matrix $R = I_2$.

In this case study, we show that smoothing via RTS does not perform as well as via MHE. Increasing smoothing length does not help RTS find a closer smoothed estimate to the true value. Prior checking is compared between predictive density value and effective sample size, both of which have similar performance. Based on the newly obtained smoothed prior, mean square error of SMC estimation is smaller than that of RTS smoothing.

Figure 3.5 shows the smoothed estimate for each component of the initial state under different smoothing lengths in a typical run of the model simulation. MHE and RTS smoothing estimate are both drawn for a comparison. The dashed lines represent the true values of the initial state. It can be seen that, the RTS smoother, shown in solid line, looks promising only for the state component A_1 and D_1, but does not yield any better estimate for the other two components than their original initial guess. MHE smoother, in dotted line, obtains much accurate estimates and moves quickly to the true value when only one observation y_1 is used. Adding more observations does not seem to make meaningful contribution.

Figure 3.6(a) depicts the observed predictive density and its threshold value at each time point without prior smoothing. The predictive density values are drawn in dotted line, while threshold values are drawn in solid line. As seen in Figure 3.6(a), nearly all the predictive density values are below the threshold values, provides a strong evidence indicating the incompatibility between the specified prior and the

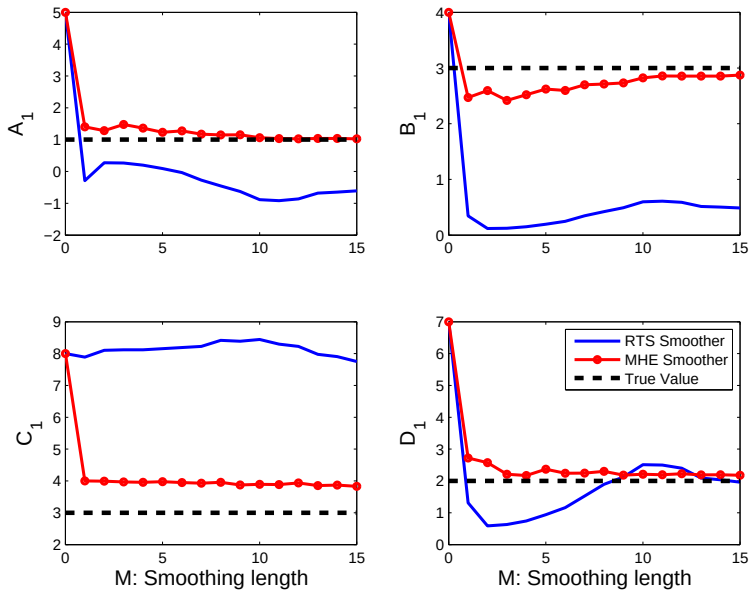
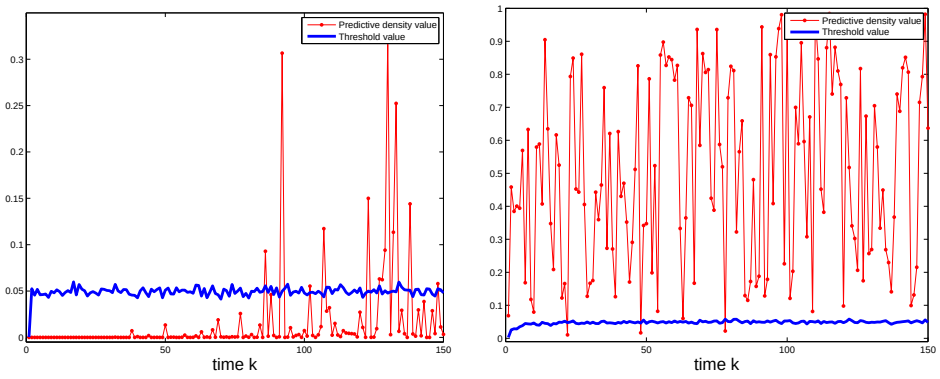


Figure 3.5: Smoothed estimate of the initial state at different smoothing lengths.

observed data. After MHE smoothing for the poor prior, the observed predictive density and the threshold value are also shown in Figure 3.6(b) above the same legend. It is seen that there are only 3 out of 150 data below their threshold values. This is reasonable since the threshold corresponds to a 0.05 level.

Figure 3.7 demonstrates the inability of RTS smoothing to improve upon the initial poor prior. It shows that, after RTS smoothing is performed with some observations, the observed predictive density value is above the threshold values corresponding to only a few smoothing lengths at the beginning, but drops below the threshold quickly for most of the remaining smoothing lengths. The algorithm in Table 3.1 suggests that one needs to use several observations for smoothing. However, as shown in Figure 3.5 the smoothed estimate is not getting any closer to the true value. Therefore, model checking will once again fail later, if RTS smoother was used.



(a) No smoothing (b) MHE smoothing

Figure 3.6: Observed predictive density value (solid line) and threshold value at 5% level (dashed line).

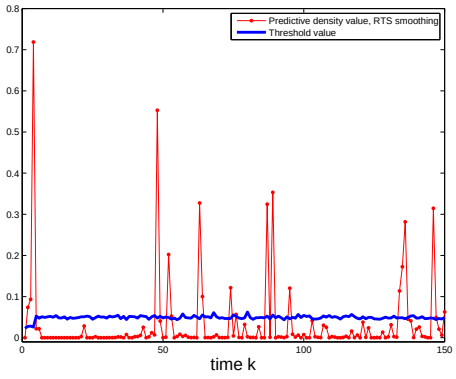
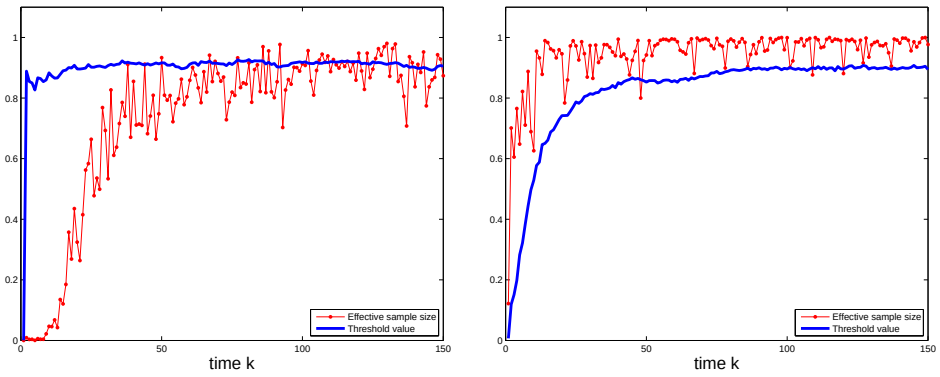


Figure 3.7: Values of the Observed predictive density (solid line) and threshold at 5% level (dashed line) for the RTS smoother.

Prior checking with the effective sample size $\tilde{N}_k$ is also performed for this model. The same set of simulated data as in Figure 3.2.2 is used. The dotted line in Figure 3.8(a) represents the $\tilde{N}_k$ values at each time point k under the original prior setting, and the solid line is their corresponding threshold values. The threshold value for effective sample size is obtained in the same way as for predictive density values. That is, by treating ESS as a random variable before the current observation is available, we find its empirical CDF and locate the lower $20-$quantile value as the threshold value at current time. Similar to Figure 3.6(a), most of the $\tilde{N}_k$ values are below the threshold line. Figure 3.8(b) presents similar information after choosing a compatible prior based on MHE smoothing according to the algorithm in Table 3.1. It can be seen that nearly all the $\tilde{N}_k$ values are larger than their corresponding threshold values. Thus, one could possibly use ESS based criterion for prior checking, but the amount of computational effort for this criterion would be more than that for the predictive value criterion, since the former does variance calculation and the latter computes only mean value.

In order to compare the posterior performance of the MHE and RTS smoothed priors as well as the plain SMC without any smoothing, we use a smoothing length of 1 for both smoothers for the sake of simplicity as well as to make a fair comparison. Figure 3.9 presents the MSE for the estimation of each of the four states over 100 simulation runs as a measure of performance of these methods. Here, RTS-SMC denotes SMC with RTS smoothing; and MHE-SMC denotes SMC with MHE smoothing, whereas SMC denotes that only the SMC estimate was used without any smoothing. Clearly, MHE-SMC has the least mean squared error.



(a) No smoothing (b) MHE smoothing

Figure 3.8: Effective Sample Size values observed for No smoothing and MHE smoothing.

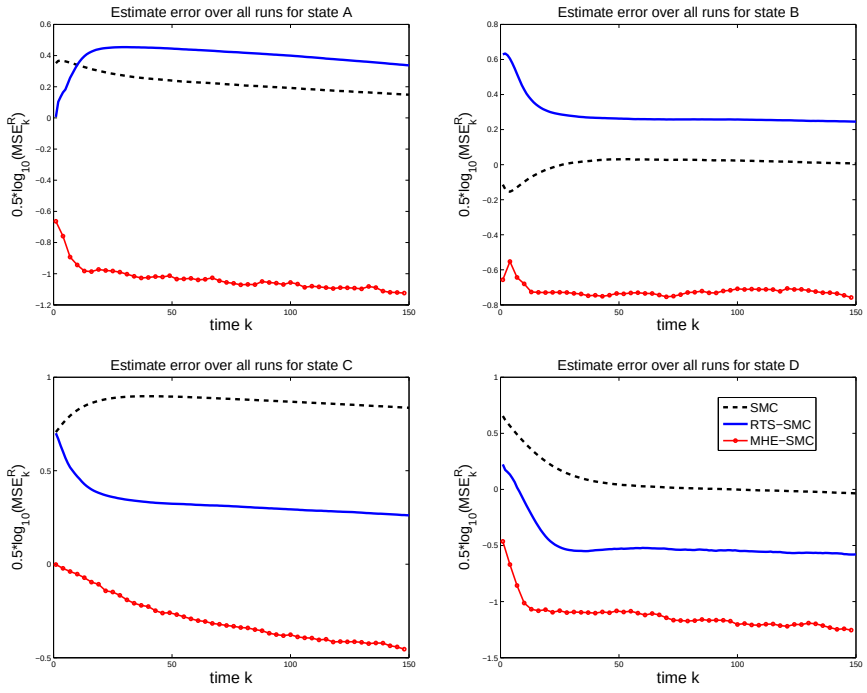


Figure 3.9: Estimation Errors for pure SMC, MHE-SMC and RTS-SMC methods

3.3 Example of an Unsuccessful Smoothed Prior: Constrained Batch Reactor

When the measurements model outputs provides information about a subspace of state space, it is fairly well understood that the prior information could have decisive impact on the state estimates. Specifically for state space models, prior gives the starting point for the initial state and a direction to begin estimation. If the given prior is poor, but possibly compatible with the observations that provide information about a subspace, smoothing might not help at all. An example is illustrated below.

The constrained batch reactor (CBR) model was discussed in Haseltine and Rawlings (2005) with the following specifications:

$$\begin{aligned} x_k &= \begin{bmatrix} A_k & B_k & C_k \end{bmatrix}^T \\ &= x_{k-1} + \Delta t \lambda \gamma + \omega_k \\ y_k &= \begin{bmatrix} -1 & 1 & 1 \end{bmatrix} x_k + \nu_k, \end{aligned} \tag{3.11}$$

with the stoichiometric matrix

$$\lambda = \begin{bmatrix} -1 & 1 & 1 \\ 0 & -2 & 1 \end{bmatrix},$$

the reaction rates

$$\gamma = \begin{bmatrix} s_1 A_{k-1} - s_2 B_{k-1} C_{k-1} \\ s_3 B_{k-1}^2 - s_4 C_{k-1} \end{bmatrix},$$

with the vector of constants s given by

$$s = \begin{bmatrix} 0.5 & 0.4 & 0.2 & 0.1 \end{bmatrix}.$$

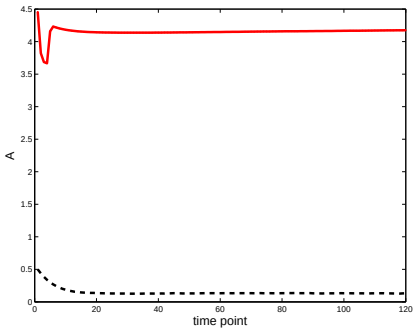
The system noise ω_k follows $\mathcal{N}(0, \sigma_\omega^2 I_3)$ where $\sigma_\omega^2 = 0.001^2$, and measurement noise is $\mathcal{N}(0, 0.1^2)$. Furthermore, the true value of initial state x_1 used to generate simulated states ans measurements is $\begin{bmatrix} 0.5 & 0.05 & 0 \end{bmatrix}^T$. All three states in this model

are constrained to be nonnegative. One of the discussed priors in Haseltine and Rawlings (2005) is $x_1 \sim \mathcal{N}(\mu, \sigma^2 I_3)$, with $\mu = \begin{bmatrix} 4 & 0 & 4 \end{bmatrix}^T$ and $\sigma^2 = 0.5^2$.

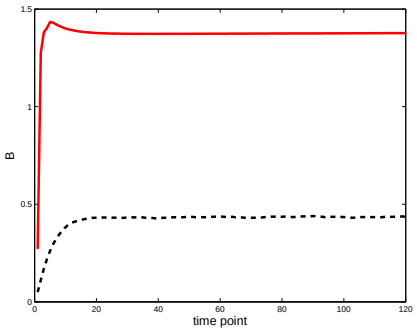
Clearly, the state vector is three dimensional, while the measurement is just a linear combination of these three. Thus, only a linear subspace of state space is observed with independent noise. Furthermore, the prior mean is expected to be compatible with the observation y_1. As a result, the predictive density value testing usually passes at time 1.

First, we apply SMC directly to the model with the given prior, and the estimation result is shown in Fig. 3.10. It can be seen that the SMC estimates do not get closer to the true values at all. MHE smoothing is then applied to the initial state with increasing smoothing lengths. The smoothed result, shown in Fig. 3.11, does not give satisfactory result for any length up to 15 and still diverges thereafter. Finally, the predictive density value is larger than the threshold values with a few observations, showing that smoothed estimates are compatible with the observations even though, in fact the estimated state values are very far from the true values.

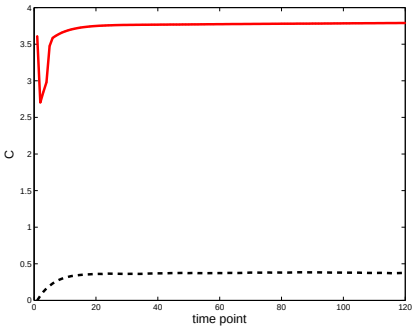
The McKeithan network model in 3.2.2 may seem to contradict the above analysis, since this network has a 4-dimensional state vector and a 2-dimensional observation, but MHE smoothing with only a few observations seems to improve the prior estimate, and the MSE is quite small. However, a careful examination of the singular value decomposition (SVD) of its simulated system states indicates that the corresponding state space seems to have a smaller effective dimension. We simulated the McKeithan network model 100 times, thus generating 100 sets of simulated ordered singular values. The log of these singular values are presented in Fig. 3.12. Typically, 2 smallest values are very near zero while the other two are larger. This seems to



(a) State A SMC Estimation

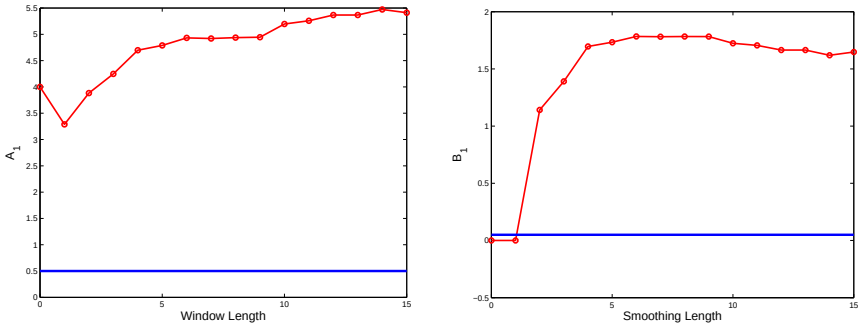


(b) State B SMC Estimation

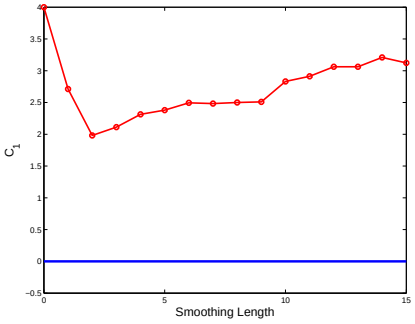


(c) State C SMC Estimation

Figure 3.10: SMC estimation of CBR model without smoothing. Solid line is SMC result and dashed line is the true value.



(a) A_1 MHE Smoothing (b) B_1 MHE Smoothing



(c) C_1 MHE Smoothing

Figure 3.11: MHE smoothing for the initial state of the CBR model. Circled line represents the smoothed result while solid line is the true value for states A_1, B_1 and C_1 respectively.

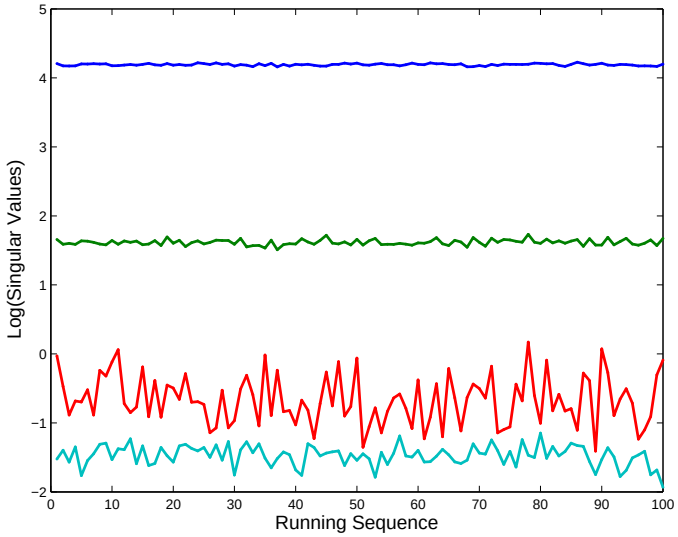


Figure 3.12: Log singular values for the McKeithan network model for 100 runs.

indicate that the states in the McKeithan network effectively live in a two dimensional space, while the other two dimensions may be considered as noise. Thus a sequence of two dimensional measurement may be quite informative about a sequence of four dimensional state space.

3.4 Conclusions

This chapter suggests predictive density as a tool to detect a poor prior in Sequential Monte Carlo sampling. A novel combination of MHE smoothing with SMC filtering is proposed to obtain a compatible empirical prior. The smoothing is only applied to the first few time points, after which regular SMC takes over. The case studies have shown that this strategy dramatically improves the performance of SMC, when the initial state is poorly specified. This work shows how two seemingly different approaches - SMC, suitable for online, recursive state estimation, and MHE,

that provides point estimates in a non-recursive fashion - can be combined to help produce more accurate state estimates in a commonly encountered practical situation. It is also worth noticing that for some models whose observation space has a lower dimension than the state space, smoothing might not find a reliable estimate of initial values. Thus caution is needed in this situation. We continue to investigate various properties of the predictive density, as well as the proposed MHE smoothed SMC estimation for a variety of dynamic systems.

CHAPTER 4

CONSTRAINED SMC ESTIMATION

It is quite common to encounter nonlinear and non-Gaussian dynamic processes with constraints in practice, the so called constrained estimation problems. For example, concentration level is inherently nonnegative. Ordered constraints specify the relative order for some elements of the state vector. Some examples of ordered constraints for static models are given in Gelfand et al. (1992).

The regular SMC sampling procedure described in Chapter 1 does not consider constraints on the states. In this chapter, a practical acceptance/rejection approach is reported to extend the SMC sampling to the constrained state estimation. This procedure keeps replacing the set of particles that fail to meet the constraints with newly generated "well behaved" samples until all the particles satisfy the requirements. Similar idea is discussed in Gelfand et al. (1992) for constrained parameter estimation with Gibbs sampler in static problems. Chen (2004) proposed the acceptance/rejection scheme for dynamic model estimation through the SMC sampling. In this chapter, we demonstrate its accuracy and fast estimation capability for a more complicated model via some meaningful evaluation criteria. Furthermore, it is also

shown that the constrained SMC algorithm shares the same theoretical properties as the unconstrained one under certain conditions. [1]

4.1 Background

Constrained estimation has been a challenging problem for dynamic models. The simplest linear Gaussian model (1.2) is still analytically tractable with interval constraints for the state variables. However, for more complicated constraints, nonlinear or non-Gaussian models, methods for solving such problems either rely on crude approximations to the model or resort to complex numerical algorithms.

Extended Kalman Filter is a convenient method to do Bayesian inference under nonlinear dynamic models with the assumption of additive Gaussian noises. Its straightforward extension to simple constrained estimation problem is based on a truncated Gaussian distribution of the system states (Haseltine and Rawlings, 2005). EKF is usually the fastest method for estimating dynamic models due to its recursive and closed-form computing. However, its reliability and accuracy are not as good as other methods introduced below; see the examples in Section 4.3 for an illustration.

Moving Horizon Estimation has a built-in capability of handling the type of constraints that limit system states within a convex set. By assuming that system noises and measurement noises are mutually independent, additive Gaussian or are approximated by Gaussian random variables, an objective function is formulated such that its least-squares solution can be found by constrained quadratic programming on special moving windows, as described in Chapter 3. The objective function for the window at the first time point is listed in Equation (3.5) for the purpose of prior

[1] A brief summary of this chapter is in Lang et al. (2007)

smoothing. The objective functions for subsequent windows are similar to the very first one, except for arrival cost, which summarizes the predictive distribution of the state at the start of the respective window (Tenny, 2002). Haseltine and Rawlings (2005) shows that MHE approach outperforms EKF, and constrained EKF, in terms of smaller estimation error for a variety of constrained estimation problems. However, three significant shortcomings of MHE still exist. The first most obvious one is that MHE approach is time-consuming in practice because of its non-recursive formulation and intensive numerical operations. Its performance also suffers due to the fact that the prior at the beginning of each moving window is assumed to (multivariate) Gaussian distribution. Figure 4.1 shows prior distributions at a few time points in a simulation of the McKeithan network model discussed in Section 4.3.2. It is evident that these distributions do not have symmetric bell shapes, and in fact, some look like truncated, unimodal or multimodal. As a result, approximation by Gaussian, or any fixed-shape distribution, tends to reduce accuracy of MHE. Finally, MHE cannot handle ordered or other complex constraints on the states, unless one uses specially designed optimization algorithms.

4.2 Constrained SMC

Compared to EKF or MHE, simulation methods like SMC are extremely flexible to handle constraints. The particles that fail to pass the constraints are invalid samples and can be readily replaced by another round of sampling until all the samples are compatible with the constraints. Literally speaking, many more types of constraints can be processed by this acceptance/rejection scheme, *e.g.,* ordered constraints.

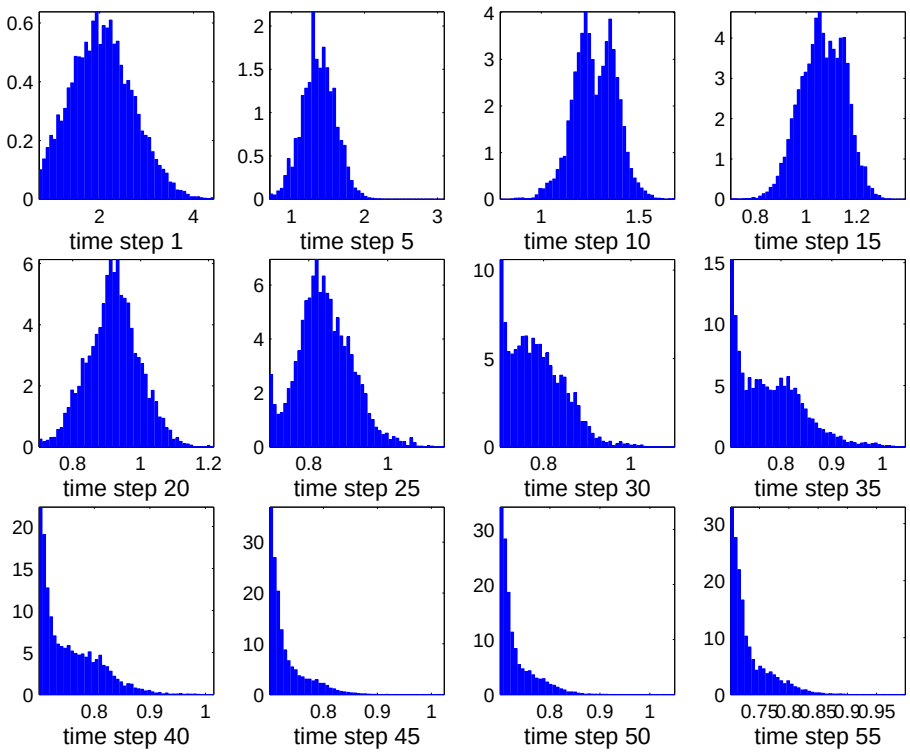


Figure 4.1: Evolution of the prior of McKeithan reaction network. (Inside each subfigure, y-axis is density value obtained from histogram.)

4.2.1 Acceptance/Rejection Algorithm

In fact, the acceptance/rejection operation guarantees that the underlying distribution of the states is truly represented by the generated samples. Therefore constrained SMC shares the same asymptotic property as stated in Theorem 1 with the regular SMC, if the conditions therein are satisfied. The following proposition is just a restatement of that result in the context of constrained estimation problem.

Proposition 4.2.1. *Under the conditions in Theorem 1, for all k and all $y_{1:k}$ in a possibly constrained dynamic model, the empirical distribution $\hat{p}^N(x_k|y_{1:k})$ obtained from the constrained SMC sampling has the following asymptotic property:*

$$\left\| \hat{p}^N(x_k|y_{1:k}) - p(x_k|y_{1:k}) \right\|_1 \to 0$$

in probability as $N \to \infty$.

Theoretically, setting the invalidated samples' weights to 0 is equivalent to the repeated acceptance/rejection steps. However, it is better to have more diversity among the samples than to have zero-weighted particles and discard them for sure in the resampling step that will eventually follow.

4.2.2 Constrained SMC algorithm

The proposed approach extends existing SMC algorithms to ensure satisfaction of inequality constraints. Equality constraints may be imposed by including them in the state or measurement equations (Ungarala and Bakshi, 2001). This approach, represented by the pseudo-code in Table 4.1, extends previous work on unconstrained estimation (Chen et al., 2004).

- FOR times $k = 1, 2, 3, \ldots$
  - — FOR $i = 1, 2, 3, \ldots, N$
    - – **Draw samples** $\tilde{x}_k^{(i)} \sim q(\cdot|x_{1:k-1}^{(i)}, y_{1:k})$ **until** $\tilde{x}_k^{(i)} \in \mathbf{S}_k$
    - – Compute weight $\tilde{w}_k^{(i)}$
  - — END FOR
  - — Normalize $\tilde{w}_k^{(i)}$ to find $w_k^{(i)}$
- END FOR

Table 4.1: Algorithm for Estimation by Constrained SMC.

The additional step in our acceptance/rejection procedure is in the sampling step of the existing SMC algorithm, therefore, only the sampling procedure is updated here. Other steps are described in Chapter 1. In Table 4.1, $\mathbf{S}_k$ denotes the set of states satisfying the constraints at time k. A complete pseudo code description of this algorithm is seen in Lang et al. (2007) where prior importance function is used. The modified step shown in bold in Table 4.1 reinforces the constraints for all the generated samples. Note that, in most cases, the initial samples $\{x_1^{(i)}, i = 1, \ldots, N\}$ also need to satisfy some constraints. For example, in Figure 4.1, the first sub-figure shows that the samples are from a truncated Gaussian distribution. Therefore, drawing samples from prior distributions needs another acceptance/rejection step. The steps shown in bold face in Table 4.1 may require a larger number of samples than unconstrained estimation, but as shown by the illustrative examples in Section 4.3, the computational complexity still remains reasonable and better than MHE. Usually, the number of rejected samples is not large, mainly because almost all prior samples already satisfy the constraints.

4.3 Case Studies

The performance of constrained SMC is compared with EKF and MHE using the same two models as in last chapter: a continuously stirred tank reactor (CSTR), and a constrained McKeithan reaction network. Performance evaluations are made on computing time, the overall mean-squared error (MSE), and mean-squared error averaged across realizations, MSE_k^R, in Equation (3.9). MSE is defined as

$$\text{MSE} = \frac{1}{L}\sum_{k=1}^{L}\text{MSE}_k^R,$$

where L is the number of simulation times.

Simulations are run on a 2.0 GHz CPU with 512MB RAM personal computer. MHE is run under GNU/Octave with a special package for computational efficiency due to Tenny (2002). The constrained SMC is run under Matlab with no particular design for efficiency. Allowing for uncertainty, $L = 100$ sets of simulated data are generated to test each method. Performance evaluation is thus a summary of these 100 sets of estimation results.

4.3.1 Constrained Adiabatic CSTR

We use the same operating conditions as listed in Table 3.2 and the same system settings as in Section 3.2.1. There are 400 measurements in each realization. A non-negative constraint is enforced on the concentration C_k at all time points.

As shown in Table 4.2 and Figure 4.2, the MSE values show that SMC is the most accurate one with performance slightly better than MHE and both are much better than EKF. However, computation time used by SMC is consistently less than that used by MHE. Of course, EKF requires the least computing time since it is based on

	Concentration ($\times 10^{-5}$)	Normalized Temperature ($\times 10^{-4}$)	CPU time ($\times 10^{-1}$)	Parameters
EKF	19.47 ± 2.5	7.53 ± 0.76	0.01 ± 0.00	
MHE	3.54 ± 1.8	1.14 ± 0.42	0.88 ± 0.01	$h = 2$
	3.49 ± 1.8	1.14 ± 0.41	1.55 ± 0.02	$h = 5$
	3.46 ± 1.8	1.14 ± 0.41	2.40 ± 0.03	$h = 10$
SMC	3.29 ± 2.0	1.09 ± 0.41	0.06 ± 0.00	$N = 1000$
	3.25 ± 1.9	1.07 ± 0.41	0.11 ± 0.01	$N = 2000$

Table 4.2: Performance Comparison under the Constrained CSTR Model.

the closed-form solution in (A.1) and its most time consuming operation, the matrix inversion, is trivial in this case.

Figure 4.3 confirms the estimation performance for the SMC sampling. As a matter of fact, EKF performs quite well during the beginning 100 points and then experiences an abrupt increase in errors. A typical run of the simulation presented in Figure 4.4 indicates that the CSTR enters rapid state changes at that time and the concentration level drops nearly to zero, the critical value for the constraint. The EKF's poor performance after an abrupt change in the system states reflects the coarse first-order approximation to the nonlinear state transition equation in this situation. In contrast, MHE uses the smoothing approach for estimating the arrival cost for each window. For this nonlinear problem, the smoother does not reduce to EKF, but utilizes information about the measurements in each window to estimate a more accurate arrival cost at the beginning of each window than the arrival cost obtained via filtering (Tenny, 2002). For SMC, the dynamic process is well approximated by the particles that satisfy the constraint.

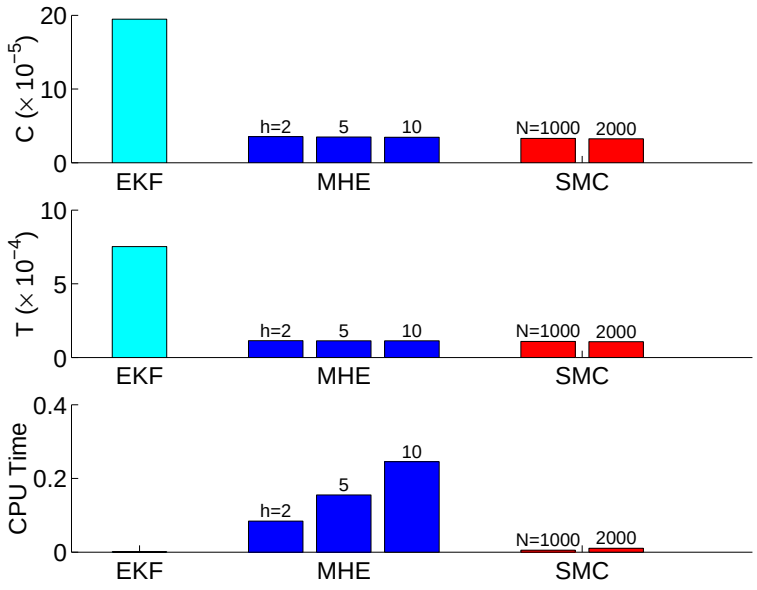


Figure 4.2: MSE and CPU Time Comparison for the Constrained CSTR Model.

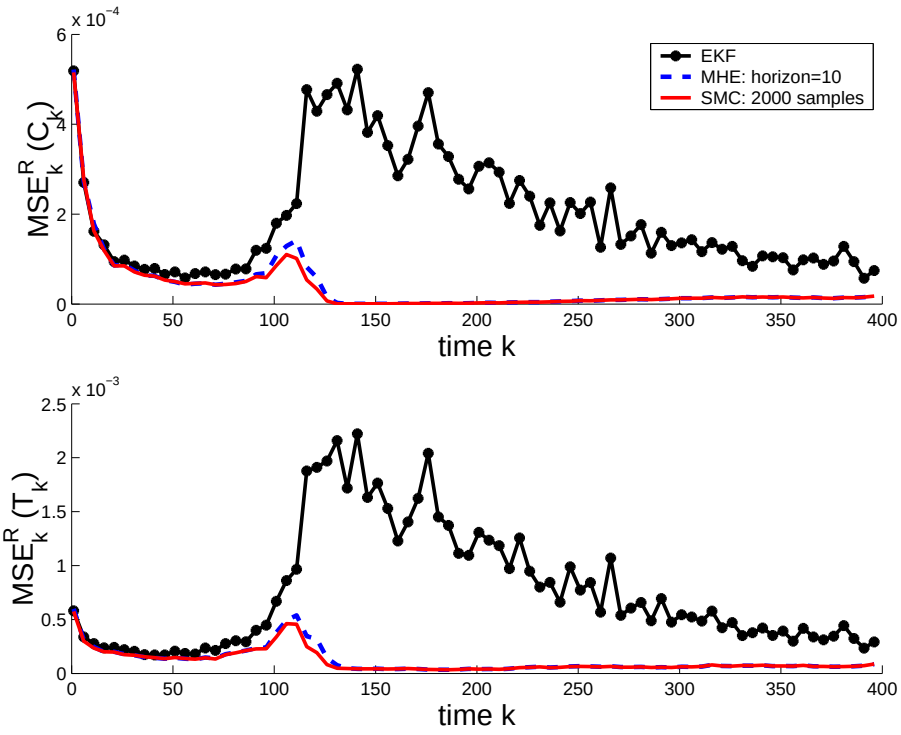


Figure 4.3: MSE_k^R Comparison for the Constrained CSTR Model.

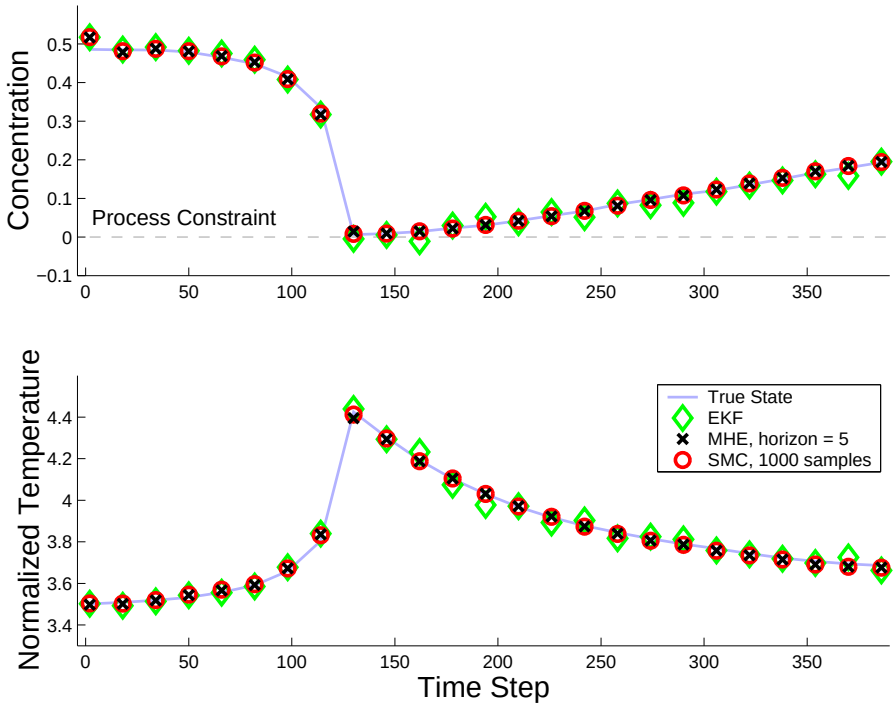


Figure 4.4: Estimation Result of a Typical Realization from the Constrained CSTR Model.

	A ($\times 10^{-3}$)	B ($\times 10^{-2}$)	C ($\times 10^{-2}$)	D ($\times 10^{-3}$)	CPU time ($\times 10^{-1}$)	Parameter
EKF	6.4 ± 4.0	3.6 ± 3.8	25.8 ± 10.3	5.1 ± 2.0	0.01 ± 0.00	
MHE	3.1 ± 1.7	2.3 ± 1.5	5.3 ± 3.5	1.8 ± 0.8	1.08 ± 0.02	$h = 2$
	2.9 ± 1.4	2.3 ± 1.3	4.7 ± 3.0	1.7 ± 0.8	1.69 ± 0.04	$h = 5$
	2.8 ± 1.3	2.2 ± 1.1	4.4 ± 2.8	1.7 ± 0.8	2.60 ± 0.07	$h = 10$
SMC	2.4 ± 1.1	2.0 ± 1.1	3.6 ± 2.7	1.4 ± 0.6	0.08 ± 0.01	$N = 1000$
	2.4 ± 1.2	1.9 ± 1.0	3.6 ± 2.6	1.4 ± 0.6	0.15 ± 0.01	$N = 2000$

Table 4.3: MSE and CPU Time Comparison for the Constrained McKeithan Network.

4.3.2 Constrained McKeithan Network

The McKeithan model as discussed in Section 3.2.2 is also used here to demonstrate better performance of the constrained SMC than the MHE and EKF methods. The initial value is set to $x_1 = (1, 3, 3, 2)^T$, and the constraint is set to D_k such that $D_k > 0.7$ for all k. The initial prior distribution is set to be $N(x_1, 0.5I_4)$. There are 1000 measurements.

The MSE and CPU-times listed in Table 4.3 and Figure 4.5 demonstrate that the constrained SMC for general state-space models performs better than the MHE and the EKF. Once again, the EKF is the fastest method because of the approximated closed-form solutions. However, it is also the worst one in terms of overall estimation error. Note that the constrained SMC has the smallest estimation error of the three methods compared here. The SMC performs better than the MHE in this case, mainly because this system is highly non-linear and MHE uses Gaussian distribution to approximate arrival cost at the beginning of each moving window.

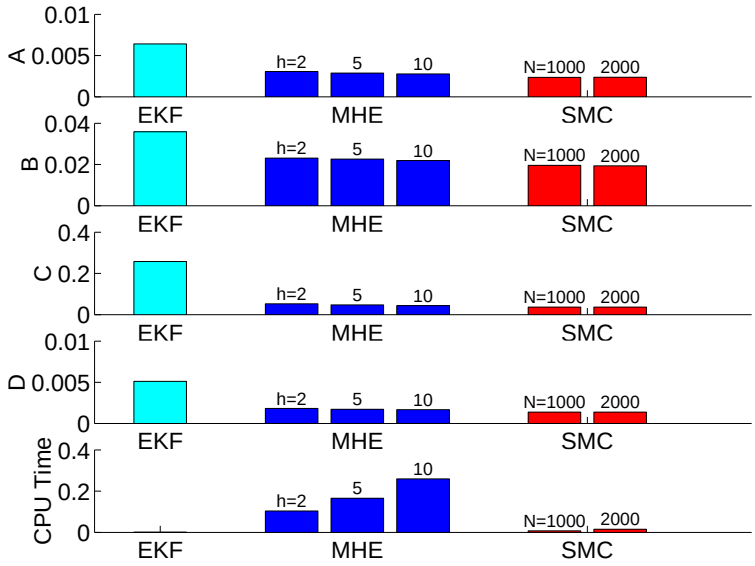


Figure 4.5: MSE and CPU Time Comparison for the Constrained McKeithan Network.

Figure 4.6 shows MSE_k^R, the mean-squared error averaged across 100 realizations. It can be seen that the EKF, shown in dotted line, tends to have larger and larger error values for state components A and B. This indicates a divergence trend for the EKF. Once again, the constrained SMC shown in solid line is slightly better than MHE, which is drawn in a dashed line. However, the CPU time of SMC is tremendously less than that of MHE. Similar conclusion is also backed up by a typical run of these three methods shown in Figure 4.7.

4.4 Conclusions

A practical approach is introduced to extend the SMC sampling to constrained dynamic models for Bayesian inference. The new algorithm, constrained SMC, enforces constraints through extra acceptance/rejection procedures to ensure that all

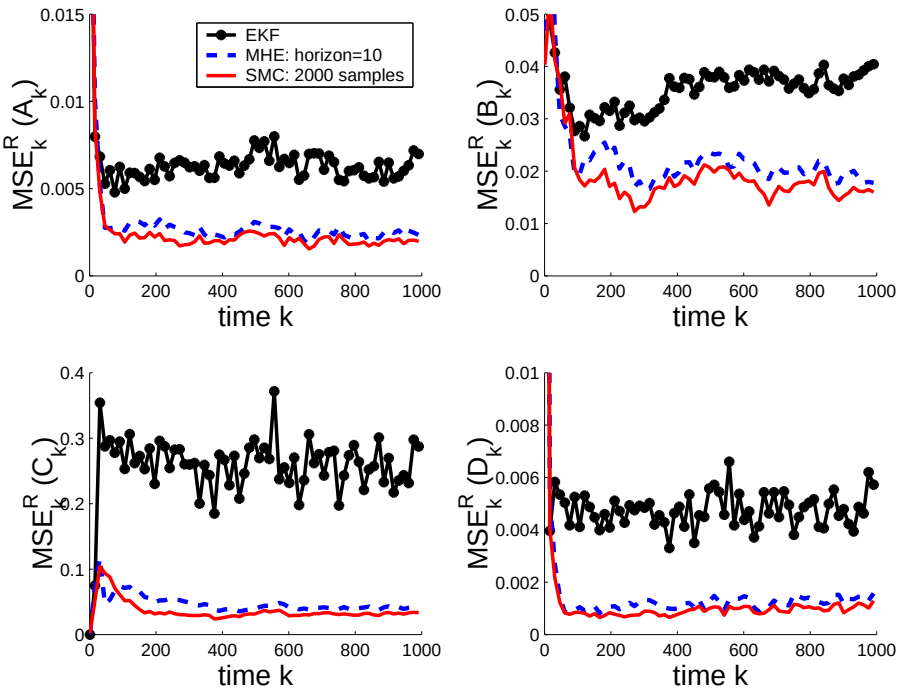


Figure 4.6: MSE_k^R Comparison for the Constrained McKeithan Network.

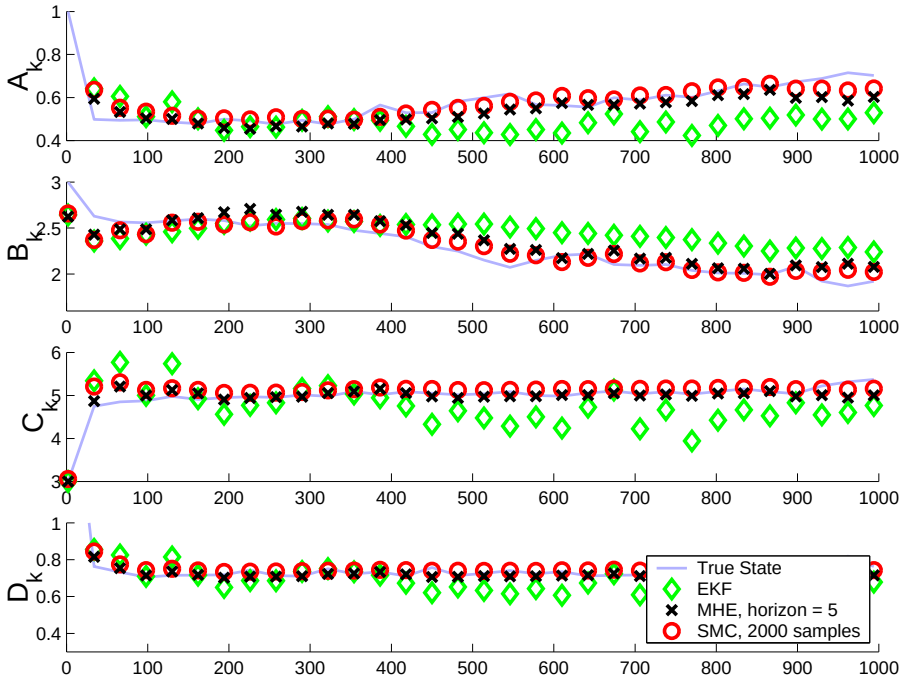


Figure 4.7: Estimation Result of a Typical Realization from the Constrained McKeithanNet.

the generated samples represent the target distribution. The proposed algorithm is also verified to possess the theoretical properties of unconstrained SMC. In addition, two dynamic models are used to demonstrate its superior performance over EKF in terms of less estimation error. Its estimation is also shown to be at least as good as a popular powerful numerical algorithm, Moving Horizon Estimation, which has a built-in capability of handling constraints. Since MHE assumes additive Gaussian noises and approximates prior by Gaussian distribution such that point estimate of states is obtained by constrained quadratic programming, constrained SMC can be applied to more dynamic models with non-Gaussian noises and is expected to give more information by providing the posterior distribution. Furthermore, constrained SMC runs much faster than MHE.

CHAPTER 5

POPULATION PHARMACOKINETICS MODELING

A novel application of SMC for estimation in Bayesian population Pharmacokinetic (PK) model, which is essentially a static model with both population and individual-specific parameters, is introduced in this chapter. individual-specific parameters. SMC, as discussed in previous chapters, is a powerful simulation method to perform Bayesian inference for general dynamic models. Many successful applications of SMC have been reported in various research areas. Application of regular SMC to the population PK model offers some advantages, but also creates some concerns, that must be addressed.

Compared to the popular Bayesian simulation method, Markov Chain Monte Carlo, SMC's sequential recursive updating scheme makes it a better choice for fast estimation. However, regular SMC suffers from particle impoverishment with static parameters, such as the population parameters in PK models. To handle this concern, an iteration of particle moving with Gibbs sampling is used to diversify particles. The performance of SMC with particle rejuvenation is well demonstrated in the case study.

The rest of the chapter is organized as follows. An introduction to PK modeling and parameter estimation are given in Section 5.1, followed by the specific SMC

sampling algorithm with particle moving in the context of population PK modeling. In the end, a one-compartment PK model, used for modeling real data obtained from Cadralazine study, is given to demonstrate the performance of the suggested SMC approach.

5.1 Introduction to Population pharmacokinetics Modeling

Pharmacokinetics is the study of physiological process, or the way how a drug is handled in vivo, within a human or an animal body as a system. Such processes include absorption, distribution, metabolism, and elimination, which are known as ADME. A classical approach to investigating ADME is to represent a body as a series of compartments, leading to differential equations of the kinetics, and possibly their analytical solutions can be obtained. The resulting model is then fitted to a series of repeatedly measured concentrations of a drug during a limited time (Gibaldi and Perrier, 1982). Consider the example of the one-compartment PK model that is used in the case-study section. The time course of drug concentrations over time is given by:

$$C_i(t) = \frac{d}{\alpha_i} \exp\left(-\frac{\beta_i}{\alpha_i} t\right), \tag{5.1}$$

where $C_i(t)$ is the modeled drug concentration of the subject i at time t; d is the initial dose of drug; α_i and β_i are, respectively, volume distribution and clearance rate for the i^{th} subject. The parameters α_i and β_i are unknown, and are assumed to be governed by the ADME processes. Conditional on the modeled concentration level $C_i(t_{ij})$, the measured concentration y_{ij} of the subject i at time t_{ij} is usually assumed to be either normal,

$$y_{ij} \sim \mathcal{N}\left(C_i(t_{ij}), \tau^{-1}\right), \tag{5.2}$$

or log-normal,

$$\log y_{ij} \sim \mathcal{N}\left(\log C_i(t_{ij}), \tau^{-1}\right). \tag{5.3}$$

In either case, estimating $\theta_i = (\alpha_i, \beta_i)$ for each i, the so-called individual-specific parameters, is of central interest in the PK modeling for determining an appropriate dose for an individual. The precision τ in Equation (5.2) or (5.3) is known as population parameter, since it is the property for all the individuals in the study. Note that, the problem setting in (5.2) or (5.3) essentially means that, conditional on population parameters, estimation of individual parameter θ_i depends on the data from subject i only, *i.e.*, $y_{i1}, \ldots, y_{in_i}$.

Population pharmacokinetics, as its name suggests, primarily focuses on population parameters, *i.e.*, variability or distribution of the individual-specific parameters, across the whole target population. Such consideration plays an important role in policy making. For example, it would be difficult to recommend an appropriate level of drug dose if its effect and safety has large variability under a targeted population. Official guideline for population PK modeling is available online at the web site http://www.fda.gov/cder/Guidance/1852fnl.pdf. Clearly, once an appropriate probability distribution for individual-specific parameters is specified, say, multivariate normal, conditional on the underlying population parameters, one has a hierarchical model for the drug-concentration measurements involving the individual-specific and population parameters. Now one can use the measurements to estimate the population parameters. In general, we denote population parameter by φ, and assume that the individual-specific parameters $[\theta_i|\varphi]$ are *i.i.d.*, following some specified distribution; see e.g., Equation (5.4b).

Bayesian method fits seamlessly into population PK modeling in that its hierarchical framework is exactly the structure between the individual-specific and population parameters. Following the convention of Wakefield et al. (1994), a general Bayesian framework of the population PK model, the so-called three-stage model, is defined as:

$$[y_{ij}|\theta_i, \tau] \sim \mathcal{N}(C_i(t_{ij}), \tau^{-1}), i = 1, \ldots, n; j = 1, \ldots, n_i, \tag{5.4a}$$

$$[\theta_i|\mu, \Omega] \sim \mathcal{N}(\mu, \Omega^{-1}), \tag{5.4b}$$

$$[\tau] \sim \text{Gamma}(a, b), \tag{5.4c}$$

$$[\mu] \sim \mathcal{N}(\eta, V^{-1}), \tag{5.4d}$$

$$[\Omega] \sim \text{Wishart}(U, \rho). \tag{5.4e}$$

Note that, the population parameters now are $\varphi = (\mu, \Omega, \tau)$. The problem of interest is the posterior distribution of φ given all the measurements. The values of the prior hyperparameters $a, b, \eta, U, V,$ and ρ are assumed to be given as part of the prior information. In practice, their values are usually selected to make the prior information as non-informative as possible to ease the concern of prior's effect on the estimation result.

Markov Chain Monte Carlo is a powerful simulation scheme to solve complex Bayesian integration problems (Geman and Geman, 1984). It constructs a Markov chain such that the chain's stationary distribution is the desired joint posterior distribution of the underlying random variables (Tierney, 1994, Robert and Casella, 1998). Extensive literature is available on the applications of MCMC to a variety of practical problems; see, e.g., Gilks et al. (1996). In particular, its application to population PK modeling is discussed in Wakefield et al. (1994), Møuller and Rosner (1997), Lunn

et al. (2002) and others. A windows software package, WinBUGS (Spiegelhalter et al., 2000) facilitate Gibbs sampling for MCMC. In addition, PKBugs (Lunn et al., 1999), an add-on for WinBUGS, is specialized for the population PK model estimation. The reliability of MCMC relies heavily on its convergence, or the mixing capability of the chain (Atherya et al., 1996), and a good sampler is expected to have a rapid mixing time. However, diagnostics for mixing/convergence are not yet satisfying (Brooks and Gelman, 1998).

The following sections discuss a novel application of SMC to the population PK model, which is essentially a static model since its individual-specific and population parameters do not have the dynamic feature. A similar idea of using SMC in static models was proposed in Chopin (2002). However, the model discussed in Chopin (2002) is only a two-stage Bayesian hierarchical model consisting of only "population parameters.". As a result, the methodology in Chopin (2002) cannot be applied straightforwardly to the population PK model. In this chapter, we use an augmented estimation space, $(\theta_1, \ldots, \theta_n, \varphi)$. Then the marginal posterior distribution of φ is automatically obtained from the samples of the augmented state space generated by SMC estimation. Furthermore, an extra step of Gibbs sampling for the population parameters is introduced at each step of SMC iteration to avoid some difficulties which could result from the random-walk moving scheme suggested in Chopin (2002). Storvik (2002) also introduced a Bayesian parameter estimation scheme, however, to the dynamic models.

5.2 Preliminary SMC PK Model Estimation

As discussed in the previous chapters, SMC is predominantly used in state-space models for Bayesian analysis, see, e.,g., Doucet et al. (2001). Population PK models do not have dynamically evolving states as state-space models. Furthermore, measurements from different subjects are usually presented all in a batch. However, we propose to process the subject-wise measurements sequentially, since each subject brings in independent individual-specific parameters from the underlying population.

For the simplicity of discussion, denote $y_k = \{y_{k1}, \ldots, y_{kn_k}\}$, *i.e.*, all the n_k measured data of subject k. Also denote $y_{1:k} = \{y_1, \ldots, y_k\}$, that is, all the available measurements from subject 1 to subject k. Similarly, $\theta_{1:k} = \{\theta_1, \ldots, \theta_k\}$. The joint posterior distribution of $\theta_{1:k}$ and φ is

$$\begin{aligned} & p(\theta_{1:k}, \varphi | y_{1:k}) \\ \propto & p(\theta_{1:k}, \varphi | y_{1:k-1}) p(y_k | \theta_{1:k}, \varphi, y_{1:k-1}) \\ = & p(\theta_{1:k-1}, \varphi | y_{1:k-1}) p(\theta_k | \theta_{1:k-1}, \varphi, y_{1:k-1}) p(y_k | \theta_k, \varphi) \\ = & p(\theta_{1:k-1}, \varphi | y_{1:k-1}) p(\theta_k | \varphi) p(y_k | \theta_k, \varphi). \end{aligned} \tag{5.5}$$

By the model specification in (5.4), the likelihood value $p(y_k|\theta_k)$ is computed as

$$p(y_k | \theta_k, \varphi) = \prod_{j=1}^{n_k} p(y_{kj} | \theta_k, \varphi). \tag{5.6}$$

And $p(\theta_k|\varphi) = p(\theta_k|\mu, \Omega)$ is the specified prior in equation (5.4b);

From Equation (5.5), it is clear that the posterior distribution of $\theta_{1:k}$ and φ given all the measurements up to subject k is in a recursive form with regard to index k. As introduced in Chapter 1, the importance function $\pi_k(\theta_{1:k}, \varphi)$ is chosen such that $\pi_k(\theta_{1:k}, \varphi) = \pi_{k-1}(\theta_{1:k-1}, \varphi) p(\theta_k|\varphi)$, with $\pi_0 = p(\varphi)$, the given prior on the population parameter φ as a starting point. We assume that, after the data from

Sample $\tau^{(\ell)}, \ell = 1, \ldots, N$ using (5.4c).
Sample $\Omega^{(\ell)}, \ell = 1, \ldots, N$ using (5.4e).
Sample $\mu^{(\ell)}, \ell = 1, \ldots, N$ using (5.4d).
Set all weights $w_0^{(\ell)} = 1/N$.
FOR subject $k = 1, \ldots, n$
— Sample $\theta_k^{(\ell)}$ using (5.4b) for each $\mu^{(\ell)}, \tau^{(\ell)}; \ell = 1, \ldots, N$
— Calculate weight $w_k^{(\ell)}$ for each $\ell = 1, \ldots, N$ with Equation (5.8)
— Resampling with replacement such that
Prob{particle ℓ is selected}$= w_k^{(\ell)}, \ell = 1, \ldots, N$.
END FOR

Table 5.1: A preliminary SMC algorithm for population PK model.

subject $(k-1)$ have been processed, we have N particles containing sample-weight pairs $\{\theta_1^{(\ell)}, \ldots, \theta_{k-1}^{(\ell)}, \varphi^{(\ell)}\}, \ell = 1, \ldots, N$ with normalized weights $\{w_{k-1}^{(1)}, \ldots, w_{k-1}^{(N)}\}$. In order to process the measurements y_k for subject k with this form of importance function, the previous samples are augmented by $\{\theta_k^{(\ell)}\}, \ell = 1, \ldots, N$, which are randomly sampled from $\pi_\theta(\theta_k; \varphi)$. Then the new unnormalized weights are obtained by (1.4) as

$$\begin{aligned} \tilde{w}_k^{(\ell)} &= \frac{p(\theta_1^{(\ell)}, \ldots, \theta_k^{(\ell)}, \varphi^{(\ell)} | y_{1:k})}{\pi_k(\theta_1^{(\ell)}, \ldots, \theta_{k-1}^{(\ell)}, \varphi^{(\ell)})} \\ &= w_{k-1}^{(\ell)} p(y_k | \theta_k^{(\ell)}, \tau^{(\ell)}). \end{aligned} \tag{5.7}$$

The new weights are then normalized through (1.6).

Particularly, when resampling is done for each index k, weight updating reduces to

$$w_k^{(\ell)} = \frac{p(y_k | \theta_k^{(\ell)}, \tau^{(\ell)})}{\sum_{\ell=1}^N p(y_k | \theta_k^{(\ell)}, \tau^{(\ell)})}. \tag{5.8}$$

A preliminary SMC algorithm for population PK model is listed in Table 5.1. However, this regular version of SMC sampling does not work well for the population PK model. An operational approach will be discussed in the next section.

As can be seen in the Algorithm in Table 5.1, the particles for the population parameters are only generated from their prior distribution at the beginning of the SMC estimation. After that, there is no opportunity for them to evolve into new values (locations). Thus, population parameter particles have less and less distinct values following each stage of sampling/resampling. As a result, the underlying posterior distribution of population parameters is approximated by only a few delta functions. Such approximation does not provide useful information on distribution of the population parameters, and this algorithm needs to be modified.

5.3 SMC PK Model Estimation with Particle Moving

In this section, we introduce particle moving under the convenient conjugate prior settings in Equation (5.4). Essentially, it is an extra one-iteration of Gibbs sampling for all the population parameters following particle resampling inside each SMC iteration. Note that resampling is performed every time for an easy implementation of Gibbs sampling, and after that the particles still represent the underlying posterior distribution with equal weights, therefore, only one-step Gibbs sampling is needed. On the other hand, MCMC needs many iterations of Gibbs sampling since it begins with arbitrary initial state. Section 1.2.3 discussed other moving methods designed for general static models.

Let us denote $U_0 = U, V_0 = V, \eta_0 = \eta, a_0 = a,$ and $b_0 = b$. Define the following variables recursively:

$$U_k = U_{k-1} + (\theta_k - \mu)'(\theta_k - \mu), \tag{5.9a}$$

$$V_k = V_{k-1} + \Omega, \tag{5.9b}$$

$$\eta_k = V_k^{-1}(V_{k-1}\eta_{k-1} + \Omega\theta_k), \tag{5.9c}$$

$$a_k = a_{k-1} + \frac{1}{2}n_k, \tag{5.9d}$$

$$b_k = \left(b_{k-1}^{-1} + \sum_{j=1}^{n_k}(y_{kj} - C_k(t_{kj}))^2\right)^{-1}. \tag{5.9e}$$

Given the posterior particles for all the individual-specific and population parameters after the $(k-1)th$ subject's measurements have been processed, all the available measurements on the current individual k as well as the individual-specific parameters for the subject k, along with the updated weights, the full conditional distributions for each of the population parameters are given by:

$$[\mu|y_{1:k}, \theta_{1:k}, \Omega, \tau] \sim \mathcal{N}(\eta_k, V_k^{-1}) \tag{5.10a}$$

$$[\Omega|y_{1:k}, \theta_{1:k}, \tau, \mu] \sim \text{Wishart}(U_k, k + \rho) \tag{5.10b}$$

$$[\tau|y_{1:k}, \theta_{1:k}, \mu, \Omega] \sim \text{Gamma}\,(a_k, b_k) \tag{5.10c}$$

Note that since the population PK model is three stage hierarchical model, given θ_k, $y_1, y_2, \ldots, y_{n_k}$ and μ, Ω are conditionally independent. Therefore, the full conditional of μ, Ω doesn't involve the measurements on the kth subject. Furthermore, it is also worth noting that, given the particles after the resampling in the $(k-1)th$ updating, the kth updating cycle was initiated by sampling $\theta_k^{(\ell)}$ from the full conditional of θ_k given all others variables, at stage $(k-1)$. Thus the population parameters particle moving via the one step of the Gibbs sampler completes the full cycles of sampling

Sample $\tau^{(\ell)}, \ell = 1, \ldots, N$ using (5.4c).
Sample $\Omega^{(\ell)}, \ell = 1, \ldots, N$ using (5.4e).
Sample $\mu^{(\ell)}, \ell = 1, \ldots, N$ using (5.4d).
Set all weights $w_0^{(\ell)} = 1/N$.
FOR times $k = 1, \ldots, n$
— Sample $\theta_k^{(\ell)}$ using (5.4b) for each $\{\mu^{(\ell)}, \tau^{(\ell)}, \Omega^{(\ell)}\}; \ell = 1, \ldots, N$
— Calculate weight $w_k^{(\ell)}$ for each ℓ in (5.8)
— Resample with replacement such that
Prob{particle ℓ is selected}$= w_k^{(\ell)}, \ell = 1, \ldots, N$.
— For each $\ell = 1, \ldots, N$
— Draw μ conditional on $\{y_{1:k}, \theta^{(\ell)}, \Omega^{(\ell)}, \tau^{(\ell)}\}$ using Equation (5.10a)
— Replace $\mu^{(\ell)} = \mu$, the new sample
— Draw Ω conditional on $\{y_{1:k}, \theta^{(\ell)}, \mu^{(\ell)}, \tau^{(\ell)}\}$ using Equation (5.10b)
— Replace $\Omega^{(\ell)} = \Omega$, the new sample
— Draw τ conditional on $\{y_{1:k}, \theta^{(\ell)}, \Omega^{(\ell)}, \mu^{(\ell)}\}$ using Equation (5.10c)
— Replace $\tau^{(\ell)} = \tau$, the new sample
— END FOR
END FOR

Table 5.2: An operational SMC algorithm with Gibbs sampling for population PK model.

from the full conditionals. In other words, for each subject k, individual-specific parameters are drawn/predicted from the SMC based posterior after the measurements from subject $(k-1)$have been processed. Then the augmented particles, which contain the population parameters, are resampled by their weights. After resampling, an iteration of Gibbs-sampling is done to draw new μ, Ω, τ for each particle based on the above full conditionals. Finally the old values of μ, Ω, τ are replaced by the newly generated ones. This completes the Gibbs sampler iteration for the subject k. The new algorithm is listed in Table 5.2.

5.4 Prior Specifications and Simulation

Prior settings for SMC are important. Bayesian PK modeling usually specifies non-informative prior, for example, small precision V in (5.4d). A non-informative prior tends to have a minimal effect on the posterior distribution. But SMC simulation often begins with sampling from the prior, and small precision or large variance tends to draw extreme values of random variables, which are either unrealistic in the sense that its weights tend to be very small, or impose difficulties in numerical computations, especially for the log-normal distribution. At the resampling step, extreme particles are very likely to be replaced by other particles, because they have small weight. Thus, computing resources are not well utilized. As a result, the prior should not be as "flat" as what MCMC recommends.

MCMC simulation uses a long chain with a burn-in period and thinning-out. The value of the burn-in period depends on how fast the chain's realizations mix. Thus its choice is on a case-by-case basis. After the burn-in period, one of every m random variables are selected from the remaining chain to provide less correlated samples. In the following case study, we chose to run 6000 iterations of Gibbs sampler in MCMC simulation, ignore the first 1000 samples as burn-in period and select 1 out of every 5 samples in the remaining ones to represent the posterior distribution. The SMC also used 1000 particles.

5.5 Cadralazine PK Model Study

Cadralazine data is analyzed in many papers (Racine et al., 1986, Wakefield et al., 1994). There are 10 cardiac failure subjects in this study. $6-8$ blood samples were taken for each subject following a single 30mg intravenous bolus dose of cadralazine

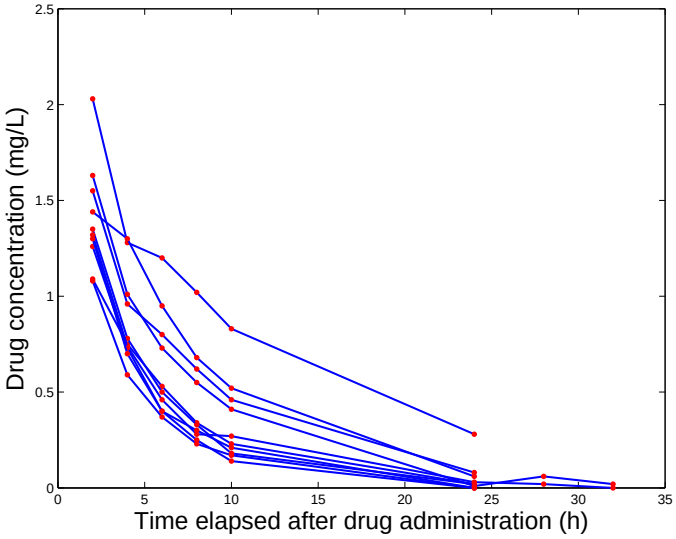


Figure 5.1: Plasma concentrations for 10 subjects in the cadralazine study. Each line represents a series of data measured at different time for the same subject.

(Lunn et al., 1999). The actual data, available in Wakefield et al. (1994), is shown in Figure 5.1, where each solid line represents a subject.

5.5.1 Modeling of Cadralazine Data

The cadralazine data is modeled as a one-compartment model with first-order elimination process. Since it is intravenous bolus administered, the full dose of the drug reaches systemic circulation instantly. The drug concentration is described by (5.1) and is assumed to follow a normal distribution as specified in (5.2).

5.5.2 SMC Estimation of Cadralazine Data

In this study, set $a = 0.01, b = 100$ for the gamma prior distribution of τ, we set $\eta = [1.065 \quad 2.708]'$ (Lunn et al., 1999), $V = 0.1I_2, U = 0.5I_2, \rho = 2$. SMC simulation is done with 1000 particles and MCMC has an iteration of 6000 times

(Lunn et al., 1999) with a burn-in period of 1000. After thinning out of MCMC chain, both methods have the same number of samples.

Figure 5.2 shows the posterior distribution of μ given by SMC (top) and MCMC (bottom) for a typical run of the MCMC and SMC based posterior distributions. Both methods provide similar results based on 100 runs of Monte Carlo simulations for this data. Figure 5.3 shows box plot of the simulated posterior means of α (left) and β (right) obtained in each of 100 runs of SMC and MCMC. It can be seen that the SMC and MCMC estimates are close to each other, however, MCMC estimates have smaller Monte Carlo variance. Figure 5.4 shows the standard deviation of posterior α (left) and β (right) for all runs of SMC and MCMC. We see again that MCMC has smaller standard deviation than SMC in general. Figure 5.5 shows the CPU time needed by SMC and MCMC respectively. On the average, SMC uses 13.8s while MCMC uses 35.7s per run. Thus SMC and MCMC provide similar estimates, but MCMC takes on the average 2.5 times the computational resources than SMC.

5.6 Conclusion

In this chapter we successfully applied sequential Monte Carlo to the population PK models for Bayesian inference. It is much faster than the the popular Bayesian simulation method, MCMC. The example used is a simple one-compartment model. However, the application of SMC to two-, or three-compartment models is an analogous extension in a straightforward manner. Further complexity to PK models can be added by drug administration method or dosing frequency. These complexities do not put any additional difficulties for SMC sampling, they simply tend to make

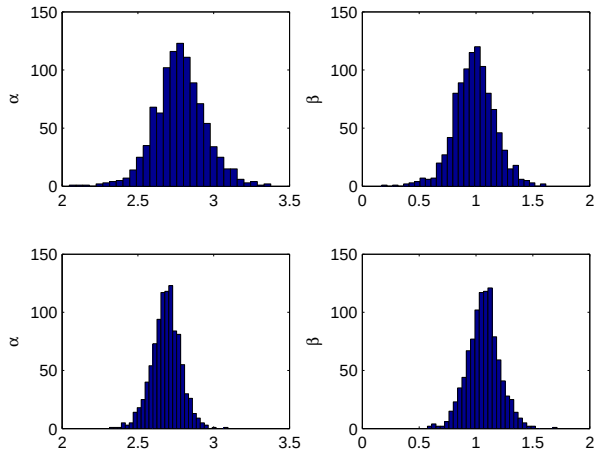


Figure 5.2: Posterior distribution of α and β of drug cadralazine given by SMC (top) and by MCMC (bottom).

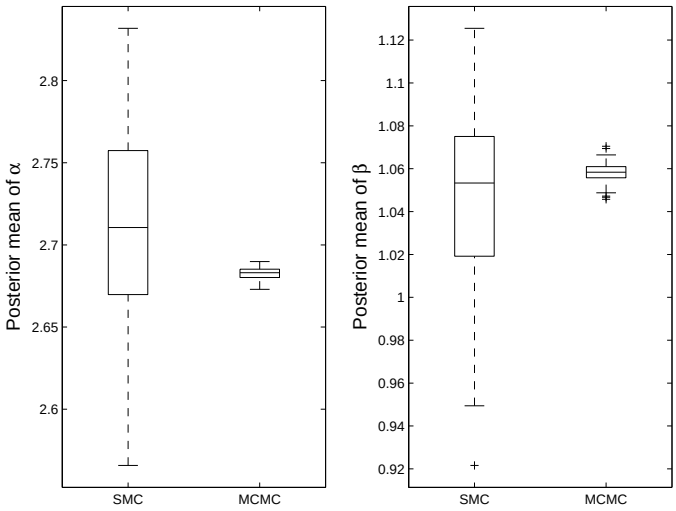


Figure 5.3: Box plot of posterior mean of α (left) and β (right) for 100 runs of SMC and MCMC.

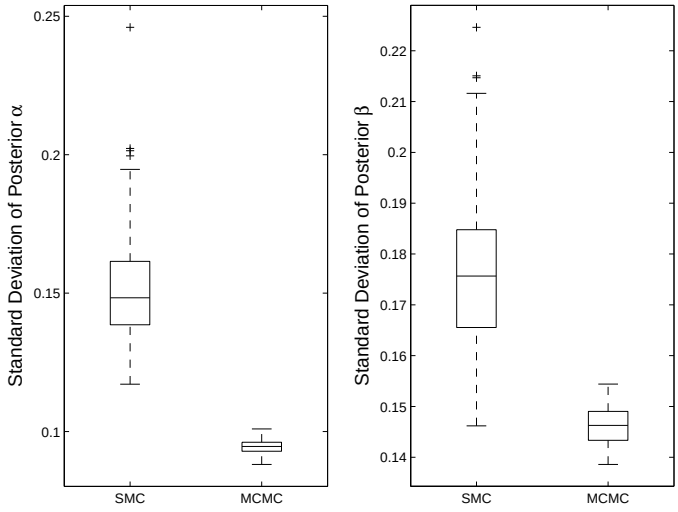


Figure 5.4: Box plot of standard deviation of α (left) and β (right) for 100 runs of SMC and MCMC.

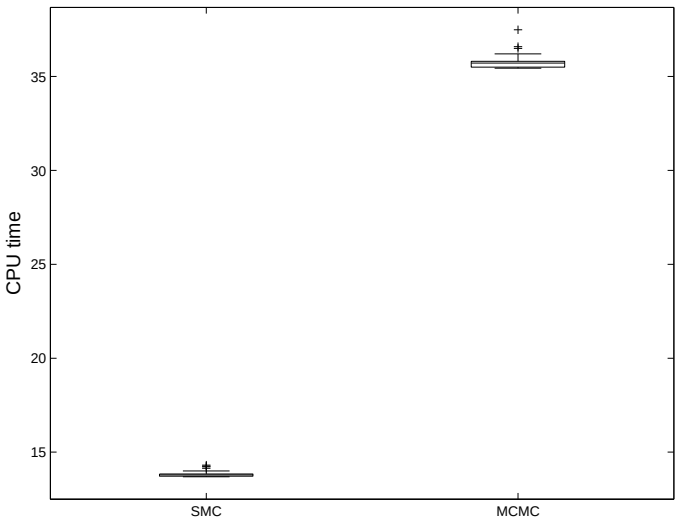


Figure 5.5: Box plot of CPU times for 100 runs of SMC (left) and MCMC (right).

likelihood calculation more complex. Future research would involve pharmacokinetic and pharmacodynamic (PK-PD) modeling.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

Theoretical study and practical improvements on sequential Monte Carlo are presented in this dissertation. By using expectation-normalized importance weights, asymptotic properties of effective sample size are investigated under different resampling control schemes, as the number of particles goes to infinity. In addition, optimal importance function is proved to be superior to the prior importance function in terms of larger effective sample size.

The enhancements on SMC performance make it robust under possibly incompatible prior and applicable to constrained estimation. As a simulation method, SMC is adversely affected by an incompatible prior that conflicts with the observed data. Predictive density value is recommended to detect an incompatible prior and furthermore, to find an appropriate smoothing length using a numerical optimization approach, *moving horizon estimation*, to find a smoothed compatible prior. For the constrained estimation problem, we introduce an extra acceptance/rejection step to reinforce the generated samples to represent the desired distribution under constraints. Several case studies are reported in the dissertation. The SMC sampling is shown to be easily applied to nonlinear non-Gaussian dynamic models with accurate estimation performance and fast enough computing speed.

A novel application of SMC is also presented for the population pharmacokinetic model estimation, which is essentially a static model. One case study involving one compartment population PK model demonstrates that SMC can provide posterior marginal distributions similar to those from the popular Bayesian simulation method based on Markov Chain Monte Carlo Gibbs Samplers, with a comparatively much smaller execution/running time. Thus SMC can be considered to be a better alternative for Bayesian estimation for population PK modeling problems.

Future work on improving performance of sequential Monte Carlo involves complex dynamic models with non-Gaussian noise and larger dimension of state space than that of observations. As demonstrated in Chapter 3, a seemingly compatible prior could actually be an opposite one because the proposed predictive density value fails to detect that. Another challenging direction, which is not addressed in the dissertation, is the application or theoretical study of SMC on the models with closed-loop feed back control. Extensive research has been done on estimation in this case, however, control with SMC is an area that needs attention. How to apply this powerful simulation tool in the control area is necessarily an interdisciplinary joint effort.

APPENDIX A

KALMAN FILTER AND RTS SMOOTHER

A.1 Kalman Filter

For linear Gaussian model with a Gaussian prior, the optimal estimates are given by Kalman Filter, which is listed below:

$$\begin{aligned} \mu_{k|k-1} &= F\mu_{k-1|k-1}, \\ P_{k|k-1} &= FP_{k-1|k-1}F^T + Q, \\ K_k &= P_{k|k-1}H^T(HP_{k|k-1}H^T + R)^{-1}, \\ \mu_{k|k} &= \mu_{k|k-1} + K_k(y_k - H\mu_{k|k-1}), \\ P_{k|k} &= P_{k|k-1} - K_kHP_{k|k-1}. \end{aligned} \tag{A.1}$$

Parameters in the above equations are defined in 1.2.

A.2 Extended Kalman Filter

Extended Kalman Filter (EKF) is obtained when the first-order Taylor series expansion is used to approximate both equations in (1.1) and thus to F and H for the above Kalman Filter.

A.3 RTS Smoother

Rauch et al. (1965) presented the well known smoother, which gives an analytic solution to the smoothed estimate of system ststes for linear Gaussian models. Similar as the EKF, it can be also applied to nonlinear Gaussian models with Taylor series expansion to both equations. The optimal smoothing results are recursively given below as $j = k, k-1, \ldots$:

$$\begin{aligned} \mu_{j-1|k} &= \mu_{j-1|j-1} + C_{j-1}(\mu_{j|k} - \mu_{j|j-1}), \\ P_{j-1|k} &= P_{j-1|j-1} + C_j(P_{j|k} - P^{-1}_{j|j-1})C_j^T, \end{aligned} \tag{A.2}$$

where

$$C_j = \Sigma_{j|j} F^T P^{-1}_{j+1|j}. \tag{A.3}$$

Basically, the above RTS smoother makes forward KF/EKF estimation as time evolves, saves intermittent data and then performs smoothing with j beginning from k and changing backward to a time point ℓ. If one is interest only in the smoothed state at a particular time, a one-pass recursive updating is preferred as seen in Rauch (1963).

$$\begin{aligned} \mu_{\ell|k} &= \mu_{\ell|k-1} + \left(\prod_{i=\ell}^{k-1} C_i\right) K_k(y_k - H\mu_{k|k-1}), \\ P_{\ell|k} &= P_{\ell|k-1} - \left(\prod_{i=\ell}^{k-1} C_i\right) K_k H P_{k|k-1} \left(\prod_{i=\ell}^{k-1} C_i\right)', \end{aligned} \tag{A.4}$$

where C_i is defined in Equation (A.3).

APPENDIX B

PREDICTIVE DENSITY AFTER SMOOTHING

Linear Gaussian model is assumed. RTS smoother, or equivalently, MHE smoother, finds the smoothed initial state estimate $\mu_{1|k}, k = 1, \ldots,$, which is then used to replace the prior mean μ_1. Let $\gamma_{\ell|k}$ denote the new predictive density value at time point ℓ when $\mu_{1|k}$ is used as the new prior mean. Intuitively, it is expected that $\gamma_{1|k} > \gamma_1$. In this section, it will be proved that $\gamma_{1|1} > \gamma_1$ and on average $\log \gamma_{1|k}$ is larger than $\log \gamma_1$. Denote $\Psi_{\ell|k-1} = \mathrm{Var}\{y_\ell | y_{1:k-1}\}$, and let $\Psi_{\ell|0} = \Psi_\ell$.

Lemma B.0.1. *$\gamma_{1|1} \geq \gamma_1$ for any y_1 and μ_1 with equality holding if and only if $y_1 = H\mu_1$.*

Before proving the above lemma, two other supporting lemmas are given below.

Denote $A > 0$ if A is positive definite (matrix); Further denote $A > B$ if $A - B$ is positive definite. Note $A > B > 0$ is equivalent to all the three conditions are met: $A > 0, B > 0,$ and $A > B$. Only real symmetric matrices are considered.

Lemma B.0.2. *If $A > B > 0$, then $B^{-1} > A^{-1} > 0$.*

Proof. The proof can be done either through a direct application of formula (1.18), or through a very interesting statistical computation (Rao, 2006). □

Lemma B.0.3. *If $A, B > 0$, then*

$(A+B)^{-1}B(A+B)^{-1}B(A+B)^{-1} < (A+B)^{-1}$.

Proof.

$$\begin{aligned}
&(A+B)^{-1}B(A+B)^{-1}B(A+B)^{-1} \\
=&\left(I - A(A+B)^{-1}\right)^T (A+B)^{-1} \left(I - A(A+B)^{-1}\right) \\
=&(A+B)^{-1} - 2(A+B)^{-1}A(A+B)^{-1} + (A+B)^{-1}A(A+B)^{-1}A(A+B)^{-1} \\
=&(A+B)^{-1} - (A+B)^{-1}A(A+B)^{-1} - (A+B)^{-1}A\left[A^{-1} - (A+B)^{-1}\right]A(A+B)^{-1} \\
<&(A+B)^{-1},
\end{aligned}$$

Since $(A+B)^{-1}A(A+B)^{-1} > 0$, $[A^{-1} - (A+B)^{-1}] > 0$ (by lemma B.0.2, and then $(A+B)^{-1}A\left[A^{-1} - (A+B)^{-1}\right]A(A+B)^{-1} > 0$. □

Proof. Lemma B.0.1 is proved below.

We first derive the expression of $\gamma_{1|1}$, the new predictive density value when only y_1 is used for smoothing to obtain $\mu_{1|1}$.

$$\begin{aligned}
y_1 - H\mu_{1|1} &= y_1 - H\left[\mu_{1|0} + K_1\left(y_1 - H\mu_{1|0}\right)\right] \\
&= (I - HK_1)(y_1 - H\mu_1) \\
&= \left[I - HP_1H^T(HP_1H^T + R)^{-1}\right](y_1 - H\mu_1) \\
&= R(HP_1H^T + R)^{-1}(y_1 - H\mu_1)
\end{aligned} \tag{B.1}$$

Then,

$$\begin{aligned}
&(y_1 - H\mu_{1|1})^T\Psi_{1|0}^{-1}(y_1 - H\mu_{1|1}) \\
=&(y_1 - H\mu_1)^T(HP_1H^T + R)^{-1}R(HP_1H^T + R)^{-1}R(HP_1H^T + R)^{-1}(y_1 - H\mu_1)
\end{aligned}$$

By lemma B.0.3,

$$\begin{aligned}
\leq&(y_1 - H\mu_1)^T(HP_1H^T + R)^{-1}(y_1 - H\mu_1) \\
=&(y_1 - H\mu_1)^T\Psi_{1|0}^{-1}(y_1 - H\mu_1).
\end{aligned} \tag{B.2}$$

Now compare the new predictive density value to the original one.

$$
\begin{aligned}
& 2\left(\log(\gamma_{1|1}) - \log(\gamma_1)\right) \\
=& (y_1 - H\mu_1)^T \Psi_{1|0}^{-1}(y_1 - H\mu_1) - (y_1 - H\mu_{1|1})^T \Psi_{1|0}^{-1}(y_1 - H\mu_{1|1}) \qquad \text{(B.3)} \\
\geq& 0.
\end{aligned}
$$

Therefore, $\gamma_{1|1} \geq \gamma_1$ for any y_1 and μ_1 with equality holding if and only if $y_1 = H\mu_1$. □

APPENDIX C

EXPECTATION-NORMALIZED EFFECTIVE SAMPLE SIZE IN LINEAR GAUSSIAN MODELS

In this appendix, observation y_k is assumed to be a random variable before it is available at time k. Then $\bar{N}_k$, as defined a function of y_k in Equation 2.2, is also a random variable, and its distribution is investigated for the linear Gaussian model. The definition of $\bar{N}_k$ in 2.2 is repeated below:

$$\bar{N}_k = \frac{p^2(y_k|y_{1:k-1})}{\mathbb{E}^{x_k|y_{1:k-1}}\left\{p^2(y_k|x_k)\right\}}.$$

We start by simplifying the denominator $\mathbb{E}^{x_k|y_{1:k-1}}\left\{p^2(y_k|x_k)\right\}$.

$$\begin{aligned} p^2(y_k|x_k) &= \frac{1}{(2\pi)^{d_y}|R|}\exp\left\{-(y_k-x_k)'R^{-1}(y_k-x_k)\right\}, \\ &\propto \frac{1}{(2\pi)^{d_y/2}|R/2|^{1/2}}\exp\left\{-\frac{1}{2}(y_k-x_k)'(R/2)^{-1}(y_k-x_k)\right\}, \\ &= \phi(y_k;\ x_k, \frac{R}{2}), \end{aligned} \tag{C.1}$$

where $\phi(x;\mu,\sigma^2)$ denotes the Gaussian density function of x with mean μ and variance σ^2. Therefore, $p^2(y_k|x_k)$ can be regarded as a Gaussian density function $p_*(y_k|x_k) = \phi(y_k; x_k, \frac{R}{2})$ up to some known constant, which does not depend on y_k. Compared to $p(y_k|x_k)$, the new function $p_*(y_k|x_k)$ has the same Gaussian mean but one half of the old covariance. Equivalently we can think that the covariance matrix of measurement

noise is reduced by half at time k. In addition, it is known that

$$\begin{aligned}\mathbb{E}^{x_k|y_{1:k-1}}\{p(y_k|x_k)\} &= p(y_k|y_{1:k-1}),\\ &= \phi(y_k;\ H\mu_{k|k-1}, HP_{k|k-1}H' + R).\end{aligned}$$

Then $\mathbb{E}^{x_k|y_{1:k-1}}\{p_*(y_k|x_k)\}$ has the same form as $p(y_k|y_{1:k-1})$ except that R needs to be replaced by its half to get the correct result

$$\mathbb{E}^{x_k|y_{1:k-1}}\{p_*(y_k|x_k)\} = \phi(y_k;\ H\mu_{k|k-1}, HP_{k|k-1}H' + R/2). \tag{C.2}$$

Then it is obtained immediately that

$$\begin{aligned}\bar{N}_k &= \frac{p^2(y_k|y_{1:k-1})}{\mathbb{E}^{x_k|y_{1:k-1}}\{p^2(y_k|x_k)\}},\\ &\propto \frac{\phi(y_k;\ H\mu_{k|k-1}, HP_{k|k-1}H'/2 + R/2)}{\phi(y_k;\ H\mu_{k|k-1}, HP_{k|k-1}H' + R/2)},\\ &\propto \phi(y_k;\ H\mu_{k|k-1}, \Delta_{k|k-1}),\end{aligned} \tag{C.3}$$

where

$$\begin{aligned}\Delta_{k|k-1} &= \left[\left(\frac{1}{2}HP_{k|k-1}H' + \frac{1}{2}R\right)^{-1} - \left(HP_{k|k-1}H' + \frac{1}{2}R\right)^{-1}\right]^{-1},\\ &= \left[HP_{k|k-1}H' + R\right]\left[1 + \frac{1}{2}\left(HP_{k|k-1}H\right)^{-1}R\right].\end{aligned} \tag{C.4}$$

Note that the proportional constant is time dependent.

BIBLIOGRAPHY

Anderson, B. D. and Moore, J. B. (1979). *Optimal Filtering*. Prentice-Hall, New Jersey.

Atherya, K. B., Doss, H., and Sethuraman, J. (1996). On the convergence of the markov chain simulation method. *The Annals of Statistics*, 24:69–100.

Bayarri, M. and Berger, J. (1997). Measures of surprise in bayesian analysis.

Bayarri, M. J. and Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95(452):1127–1142.

Berzuini, C. and Gilks, W. (2001). Resample-move filtering with cross-model jumps. In Doucet, A., de Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*, chapter 6, pages 117–138. Springer, New York.

Box, G. E. P. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, 143(4):383–430.

Brooks, S. P. and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992). A monte carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, 87(418):493–500.

Carter, C. K. and Kohn, R. (1996). Markov chain monte carlo in conditionally gaussian state space models. *Biometrika*, 83(3):589–601.

Chaves, M. and Sontag, E. D. (2002). State-estimator for chemical reaction networks of feinberg-horn-jackson zero deficiency type. *European Journal of Control*, 8(4):343–359.

Chen, W., Bakshi, B. R., Goel, P. K., and Ungarala, S. (2004). Bayesian estimation of unconstrained nonlinear dynamic systems via sequential monte carlo sampling. *Industrial & Engineering Chemistry Research*, 43(14):4012–4025.

Chen, W.-s. (2004). *Bayesian estimation by sequential Monte Carlo sampling for nonlinear dynamic systems*. PhD thesis, the Ohio State University.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539–552.

Chopin, N. (2004). Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *Annals of Statistics*, 32(6):2385–2411.

Crisan, D. (2001). Particle filters—a theoretical perspective. In Doucet, A., de Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*, chapter 2, pages 17–41. Springer, New York.

Crisan, D. and Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3):736–746.

Del Moral, P. and Miclo, L. (2000). Branching and interacting particles systems. approximations of feynman-kac formulae with applications to non-linear filtering. In *Séminaire de Probabilitités XXXIV. Lecture Notes in Math*, pages 1–145. Springer, Berlin.

Doucet, A., de Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10:197–208.

Evans, M. and Moshonov, H. (2006). Checking for prior-data confict. *Bayesian Analysis*, 1(4):893–914.

Gelfand, A. E., Smith, A. F. M., and Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, 87(418):523–532.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Geweke, J. (1988). Antithetic acceleration of monte carlo integration in bayesian inference. *Journal of Econometrics*, 38:73–90.

Geweke, J. and Tanizaki, H. (2001). Bayesian estimation of state-space model using the metropolishastings algorithm within gibbs sampling. *Computational Statistics and Data Analysis*, 37(2):151–170.

Gibaldi, M. and Perrier, D. (1982). *Pharmacokinetics*. Dekker, New York.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, London.

Goel, P. K., Lang, L., and Bakshi, B. R. (2006). Sequential Monte Carlo in bayesian inference for dynamic models: an overview. In Upadhyay, S. K., Singh, U., and Dey, D. K., editors, *Bayesian Statistics and Its Applications*. Anamaya Publishers, New Delhi, India.

Good, I. J. (1956). The surprise index for the multivariate normal distribution. *The Annals of Mathematical Statistics*, 27(4):1130–1135.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 29(1):83–100.

Haseltine, E. L. and Rawlings, J. B. (2005). Critical evaluation of extended kalman filtering and moving-horizon estimation. *Industrial & Engineering Chemistry Research*, 44(8):2451–2460.

Henson, M. A. and Seborg, D. E. (1997). *Nonlinear Process Control*. Prentice Hall PTR, Upper Saddle River, NJ.

Hürzeler, M. and Kunsch, H. R. (1998). Monte carlo approximations for general state-space models. *Journal of Computational and Graphical Statistics*, 7(2):175–193.

Isard, M. and Blake, A. (1998). Condensation—conditional density propagation for visual tracking. *Int. J. Comput. Vision*, 29(1):5–28.

Julier, S. J. and Uhlmann, J. K. (1995). A new approach for filtering nonlinear systems. In *Proceedings of the American Control Conference*, pages 1628–1623.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *Journal of American Statistical Association*, 82(400):1032–1041.

Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation of the gaussian-sum smoother. *Ann. Inst. Statist. Math.*, 46(4):605–623.

Kitagawa, G. (1996). Monte carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25.

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288.

Künsch, H. R. (2005). Recursive monte carlo filters: Algorithms and theoretical analysis. *Annals of Statistics*, 33(5):1983–2021.

Lang, L., Chen, W., Bakshi, B. R., Goel, P. K., and Ungarala, S. (2007). Bayesian estimation of constrained nonlinear dynamic systems via sequential monte carlo sampling. *Automatica*, 43(9):1615–1622.

Lang, L., Goel, P. K., and Bakshi, B. R. (2006). A smoothing based method to improve performance of sequential Monte Carlo estimation under poor prior. In *CPC7 Chemical Process Control, Jan. 8-13, 2006*, Lake Louise, Alberta, Canada.

Liu, J. S. (1996). Metropolized independent sampling with comparison to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119.

Liu, J. S. and Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044.

Lunn, D. J., Best, N., Thomas, A., Wakefield, J., and Spiegelhalter, D. (2002). Bayesian analysis of population pk/pd models: General concepts and software. *Journal of Pharmacokinetics and Pharmacodynamics*, 29(3):271–307.

Lunn, D. J., Wakefield, J., Thomas, A., Best, N., and Spiegelhalter, D. (1999). *PKBugs User Guide Version* 1.1. Imperial College of Science, Technology and Medicine.

McKeithan, T. W. (1995). Kinetic proofreading in t-cell receptor signal transduction. *Proceedings of the National Academy of Sciences*, 92:5042–5046.

Møuller, P. and Rosner, G. L. (1997). A bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, 92:1279–1292.

Naylor, J. C. and Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, 31:214–225.

Racine, A., Grieve, A. P., Fluhler, H., and Smith, A. F. M. (1986). Bayesian methods in practice: Experiences in the pharmaceutical industry (with discussion). *Applied Statistics*, 35(2):93–150.

Rao, C. R. (2006). Statistical proofs of some matrix theorems. *International Statistical Review*, 74(2):169–185.

Rao, C. V. and Rawlings, J. B. (2000). *Nonlinear Moving Horizon State Estimation*, pages 45–70. Nonlinear Model Predictive Control. Birkhauser, Basel, Switzerland.

Rauch, H. E. (1963). Solutions to the linear smoothing problem. *IEEE Transactions on Automatic Control*, 8(4):371–372.

Rauch, H. E., Tung, F., and Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450.

Robert, C. P. and Casella, G. (1998). *Monte Carlo Statistical Methods*. Springer, New York.

Robertson, D. G., Lee, J. H., and Rawlings, J. B. (1996). A moving Horizon-Based approach for Least-Squares estimation. *AIChE Journal*, 42(8):2209–2224.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 12(4):1151–1172.

Rubin, D. B. (1987). Using the sir algorithm to simulate posterior distributions. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics 3*, pages 395–402. Oxford University Press.

Spiegelhalter, D., Thomas, A., and Best, N. (2000). *WinBUGS Version* 1.3 *User Manual*. Medical Research Council Biostatistics Unit, Cambridge.

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing*, 50(2):281–289.

Tenny, M. J. (2002). *Computational Strategies for Nonlinear Model Predictive Control*. PhD thesis, University of Wisconsin-Madison.

Tierney, L. (1994). Markov chains for exploring posterior distributions with discussion. *The Annals of Statistics*, 22:1701–1762.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86.

Ungarala, S. and Bakshi, B. R. (2001). Multiscale bayesian estimation and data rectification. In Petrosian, A. A. and Meyer, F. G., editors, *Wavelets in Signal and Image Analysis*, pages 69–110. Kluwer Academic Publishers, Dodrecht, The Netherlands.

van der Merwe, R., de Freitas, J. F. G., Doucet, A., and Wan, E. A. (2000). The unscented particle filter. Technical Report CUED/F-INFENG/TR 380. Technical report, Cambridge University Engineering Department, Cambridge, England.

Wakefield, J. C., Smith, A. F. M., Racine-Poon, A., and Gelfand, A. E. (1994). Bayesian analysis of linear and non-linear population models by using gibbs sampler. *Applied Statistics*, 43(1):201–221.

Weaver, W. (1948). Probability, rarity, interest, and surprise. *The Scientific Monthly*, 67(6):390–392.