MCMC-Based Bayesian Change Point Detection: a Comparison between the Gibbs and Metropolis-within-Gibbs Samplers

Naitong Chen

March 15, 2021

1 Introduction

Change point detection solves the class of problems where one would like to identify times when abrupt changes in the underlying data generating process of a time series occurs. In practice, this setting is widely seen in areas such as stock analysis and climate studies [CG97; ENJ04]. In the literature, a wide variety of approaches, both frequentist and Bayesian, have been developed to efficiently solve the change point problems [KFE12; CGS92]. In the Bayesian approaches, the problem is often framed as estimating the posterior distribution of the change point locations. The common advantages of the Bayesian approach to statistical inference apply in the change point setting. Compared to the point estimates of the change point locations, the posterior distribution of the change point locations of a given sequence provides a more flexible and holistic understanding of the time series. For example, having access to the approximate posterior distribution of the change point locations allows us to estimate the probability that a change has occured at some given time.

One of the earliest Bayesian approaches to the change point problem is developed in [CGS92] through Markov Chain Monte Carlo (MCMC). Particularly, a Gibbs sampler is used to obtain samples for estimating the posterior distribution of the change point locations. It is worth noting that, although only the detection of a single change point is discussed in [CGS92], there is a natural extension to the multiple change point setting when the number of change points is known. Specifically, instead of sampling a single change point from the distribution of change points conditioned on all other parameters in the model, a set of change points can be sampled from the conditional distribution over all possible configurations of the change point locations. It is then obvious that the number of possible configurations of the change point locations grows linearly in the length of the time series and exponentially in the number of change points to be detected. As a result, the computational complexity of the Gibbs sampler becomes exponentially more expensive.

The scalability issue described above is present in virtually all change point detection methods. In fact, many of the more recent developments in MCMC-based approaches can be viewed as trying to bypass this computational bottleneck either through clever reformulation of the problem [Ste94; LL01] or the use of Metropolis-Hastings (MH) and potentially well-designed proposal distributions [Gre95; AL08]. The approach in [AL08] is one of the easiest and most intuitive attempts at circumventing the computational challenge posed in [CGS92]. Instead of computing the full conditional distribution of the change point locations, a uniform proposal over all possible configurations of change point locations is accepted/rejected based on the MH ratio. However, while the expensive full conditional distribution no longer needs to be evaluated, exploring this potentially enormous space of change point locations using a uniform proposal may not be the most efficient, thus posing a different challenge.

In this report, we compare the Bayesian change point detection problems proposed in [CGS92] and [AL08]. Using a piecewise constant model, we would like to explore, under different settings, whether there is a trade-off between evaluating the expensive full conditional distribution and searching over a potentially large space using a cheaper sampler that risks low acceptance rates. Through a number of experiments, we find that although the approach in [CGS92] is more favourable in all models that are tested, the approach in [AL08] offers very competitive performance. However, we note that only a uniform proposal distribution is used.

By designing more informative proposal distribution, as already been explored in [BF+18], proposal-based MCMC methods applied to the change point problems would be a very attractive choice when the number of possible change point configurations is too large to be handled by a Gibbs sampler.

2 Problem Setup

Below, we start by specifying the piecewise constant model and outlining the procedures of the two MCMC methods before presenting the discussing the experimental results.

2.1 Pairwise Constant Model

Consider a sequence of random variables $(Y_1, \cdots, Y_n)$, where $\forall i \in \{1, \cdots, n\}, Y_i \in \mathbb{R}$, such that

$$Y_i \sim \begin{cases} N(\mu_1, 1) & 1 \leq i \leq r_1 \\ N(\mu_2, 1) & r_1 < i \leq r_2 \\ \quad \vdots & \\ N(\mu_{k+1}, 1) & r_k < i \leq n \end{cases},$$

where k is known, $\boldsymbol{r} = (r_1, \cdots, r_k)$ denotes the set of integer valued change point locations, and $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_{k+1})$ denotes the underlying means of each segment. Note that the jth segment is defined to be $\boldsymbol{Y}_{r_{j-1}+1:r_j} = (Y_{r_{j-1}+1}, \cdots, Y_{r_j})$, with $r_0 = 0, r_{k+1} = n$ by convention. In addition, we assume that $\boldsymbol{r}$ and $\boldsymbol{\mu}$ are independent and that the random variables are i.i.d. within each segment.

Given a sequence of realized observations $\boldsymbol{y} = (y_1, \cdots, y_n)$ from the above data generating process, its likelihood is defined by

$$L(\boldsymbol{r}, \boldsymbol{\mu} \mid \boldsymbol{y}) = \prod_{i=1}^{k+1} \prod_{j=r_{i-1}+1}^{r_i} f(\mu_i; y_j).$$

Note $f(\mu_i; y_j)$ is the likelihood of μ_i being the mean of a normal distribution with variance 1 given the observation y_j.

Then once a set of prior distributions are defined on the parameters $\boldsymbol{r}$ and $\boldsymbol{\mu}$, the goal of the Bayesian change point detection methods is to estimate the posterior distributions $p(\boldsymbol{r} \mid \boldsymbol{y})$ and $p(\boldsymbol{\mu} \mid \boldsymbol{y})$. We focus on the posterior distribution of the change point locations $p(\boldsymbol{r} \mid \boldsymbol{y})$.

For simplicity, we assume the priors on the segment means to be

$$\mu_i \sim N(m, 1), \forall i = 1, \cdots, k+1, \text{and } m \in \mathbb{R}.$$

As a prior on the change point locations, we assume that $\boldsymbol{r}$ follows a discrete uniform distribution over all possible configurations of the change point locations. Specifically, for any gien possible configuration $\boldsymbol{r}_c$,

$$p_{\boldsymbol{r}}(\boldsymbol{r}_c) = \frac{1}{\binom{n-1}{k}}.$$

Note that by our specification, it is not possible for the last observation in the sequence to be a change point.

2.2 MCMC Simulation Procedures

Under the above specification, to obtain a set of samples for $\boldsymbol{r}$ and $\boldsymbol{\mu}$, given $\left(\boldsymbol{r}^i, \boldsymbol{\mu}^i\right)$, the Gibbs sampler described in [CGS92] generates the next sample $\left(\boldsymbol{r}^{i+1}, \boldsymbol{\mu}^{i+1}\right)$ through the following steps:

1. generate $\boldsymbol{r}^{i+1}$ from the probability mass function

$$p\left(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{\mu^i}\right) = \frac{L\left(\boldsymbol{r}, \boldsymbol{\mu}^i \mid \boldsymbol{y}\right)}{\sum_{j=1}^{\binom{n-1}{k}} L\left(\boldsymbol{r}^j, \boldsymbol{\mu}^i \mid \boldsymbol{y}\right)},$$

   where $\boldsymbol{r}^j$ in the denominator denotes the jth configuration from the permutation of all possible configurations;

2. generate $\boldsymbol{\mu}^{i+1}$ by

$$\mu_j^{i+1} \sim N\left(\frac{m + \sum_{l=r_{j-1}+1}^{r_j^{i+1}} y_l}{r_j^{i+1} - r_{j-1}^{i+1}}, \left(1 + r_j - r_{j-1}\right)^{-1}\right), \forall j \in \{1, \cdots, k+1\}. \tag{1}$$

Similarly, given $\left(\boldsymbol{r}^i, \boldsymbol{\mu}^i\right)$, the Metropolis-within-Gibbs (MWG) sampler described in [AL08] generates the next sample $\left(\boldsymbol{r}^{i+1}, \boldsymbol{\mu}^{i+1}\right)$ through the following steps:

1. generate $\boldsymbol{r}'$ from the discrete uniform prior $p_{\boldsymbol{r}}$;
2. accept $\boldsymbol{r}'$ as $\boldsymbol{r}^{i+1}$ with probability $\min\left(1, \beta\left(\boldsymbol{\mu}^i, \boldsymbol{r}^i, \boldsymbol{r}'\right)\right)$, where

$$\beta\left(\boldsymbol{\mu}^i, \boldsymbol{r}^i, \boldsymbol{r}'\right) = \frac{L\left(\boldsymbol{r}', \boldsymbol{\mu}^i \mid \boldsymbol{y}\right)}{L\left(\boldsymbol{r}^i, \boldsymbol{\mu}^i \mid \boldsymbol{y}\right)};$$

3. otherwise set $\boldsymbol{r}^{i+1} = \boldsymbol{r}^i$;
4. generate μ^{i+1} following the last step of the Gibbs sampler.

Despite the use of conjugate priors when sampling the means of each segment, the trade-off between evaluating the expensive full conditional distribution of the change point locations and searching over a potentially large space using a cheaper sampler that risks low acceptance rates is still present. Specifically, the cost of computing $p\left(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{\mu}^i\right)$ from the Gibbs sampler grows linearly in the length of the time series and exponentially in the number of change points to be detected. The search space size of all configurations of change point locations from the MWG sampler also grows in the same manner.

Before presenting the experiment results, it is worth noting that the posterior distribution

$$\begin{aligned}
p(\boldsymbol{r} \mid \boldsymbol{y}) &= \int_{\mu_1} \cdots \int_{\mu_{k+1}} p(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{r}) p_{\boldsymbol{r}}(\boldsymbol{r}) p(\mu_1) \cdots p(\mu_{k+1}) d\mu_1 \cdots d\mu_{k+1} \\
&= p_{\boldsymbol{r}}(\boldsymbol{r}) \left(\int_{\mu_1} p(\boldsymbol{y}_{1:r_1} \mid \mu_1) p(\mu_1) d\mu_1\right) \cdots \left(\int_{\mu_{k+1}} p(\boldsymbol{y}_{r_k+1:n} \mid \mu_{k+1}) p(\mu_{k+1}) d\mu_{k+1}\right) \\
&\coloneqq p_{\boldsymbol{r}}(\boldsymbol{r}) Q(1, r_1, \mu_1) \cdots Q(r_k + 1, n, \mu_{k+1}).
\end{aligned}$$

can be fully evaluated. Note that $p(\mu_i)$ denotes the prior distribution of the ith segment mean, and

$$p(\boldsymbol{y}_{r_{i-1}+1:r_i} \mid \mu_i) = \prod_{j=r_{i-1}+1}^{r_i} f(\mu_i; y_j)$$

denotes the likelihood of μ_i given all observations from the ith segment of the sequence.

This posterior distribution can be evaluated using Baye's rule because the $Q(\cdot)$'s are the normalizing constants of the known posterior distributions of the segment means defined in Eq. (1). This allows us to more directly compare the quality of the samples generated from the two MCMC methods in the next section.

3 Experiment

4 Discussion

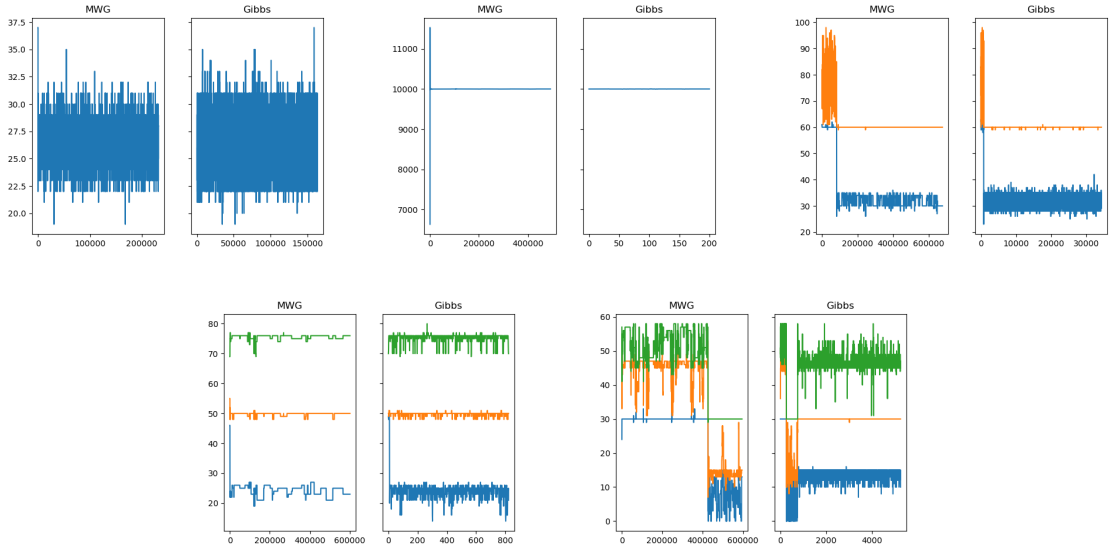

Figure 1: trace without burnin

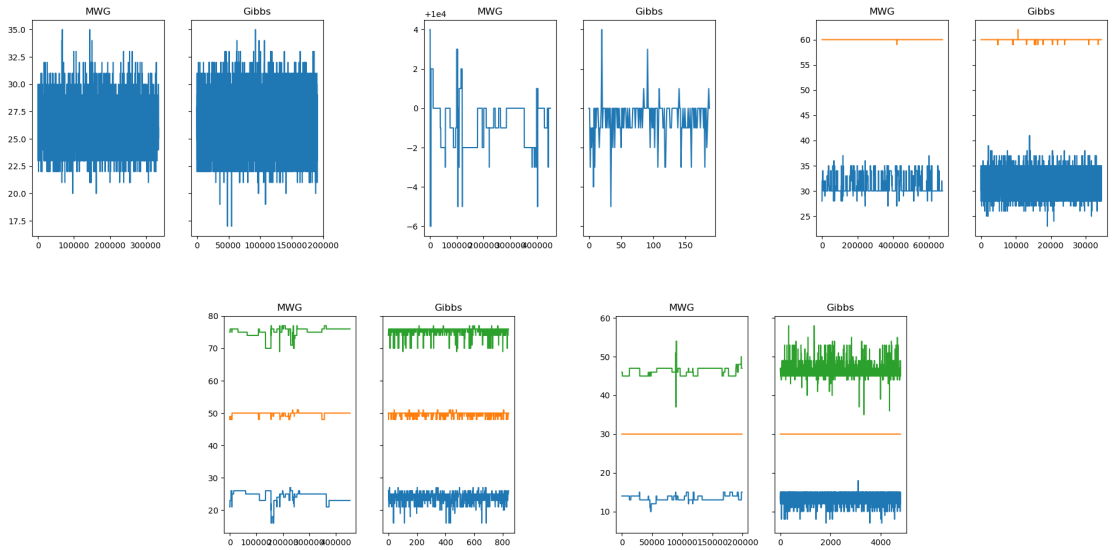

Figure 2: trace after burnin

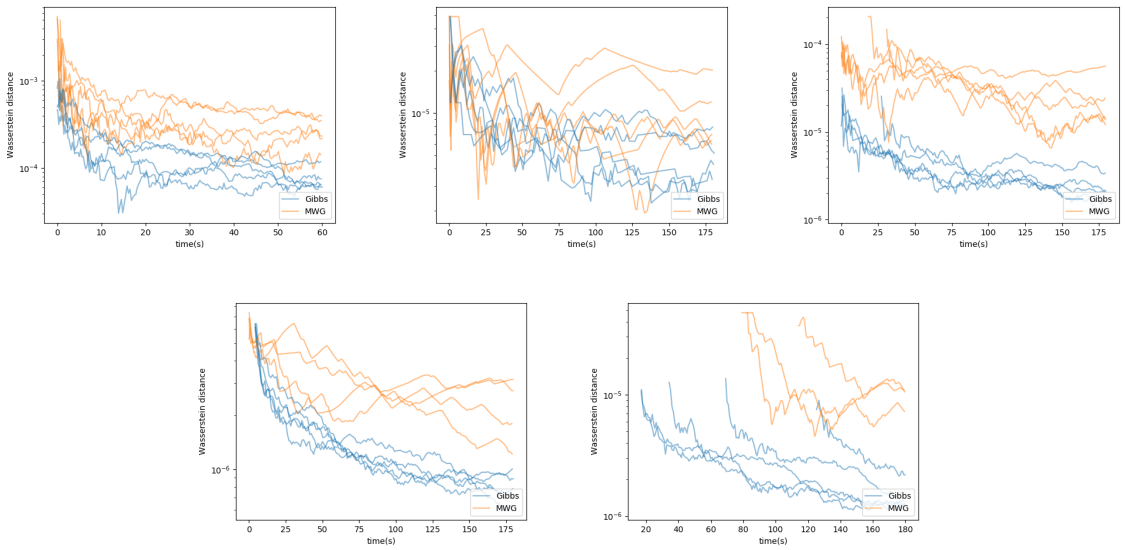

Figure 3: wasserstein

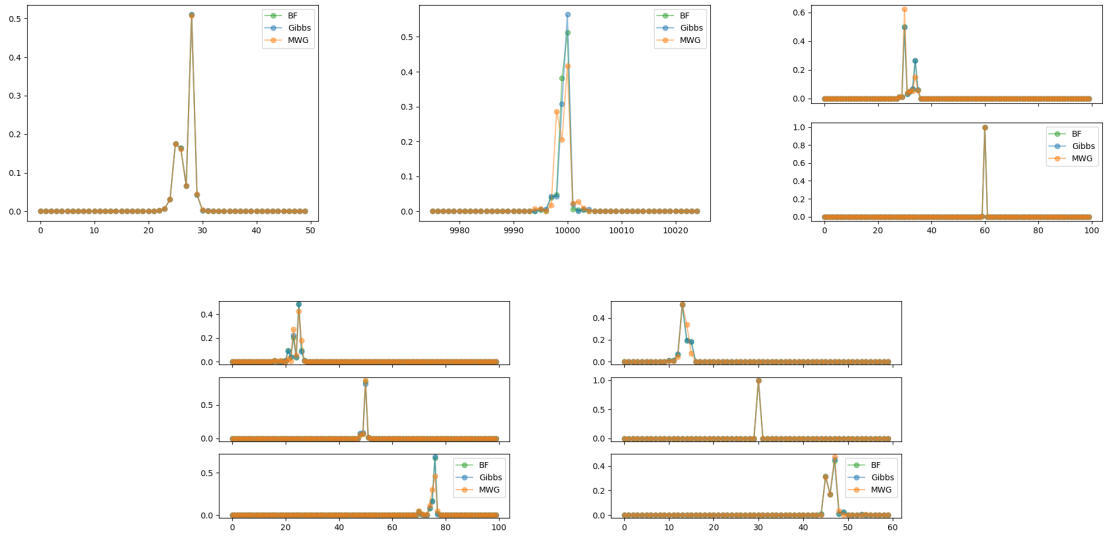

Figure 4: marginal

A Supplementary Plots

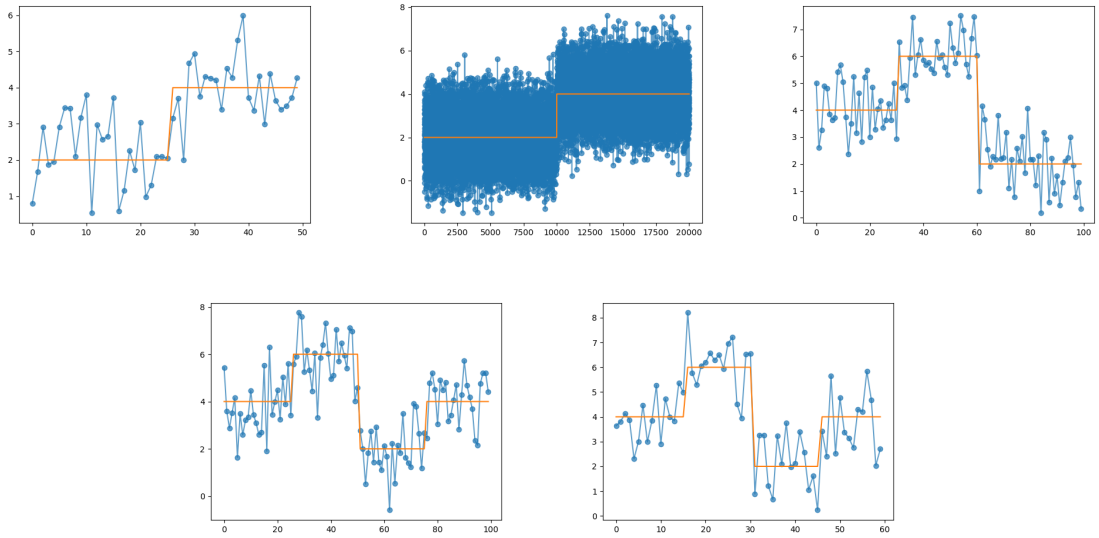

Figure 5: sequence

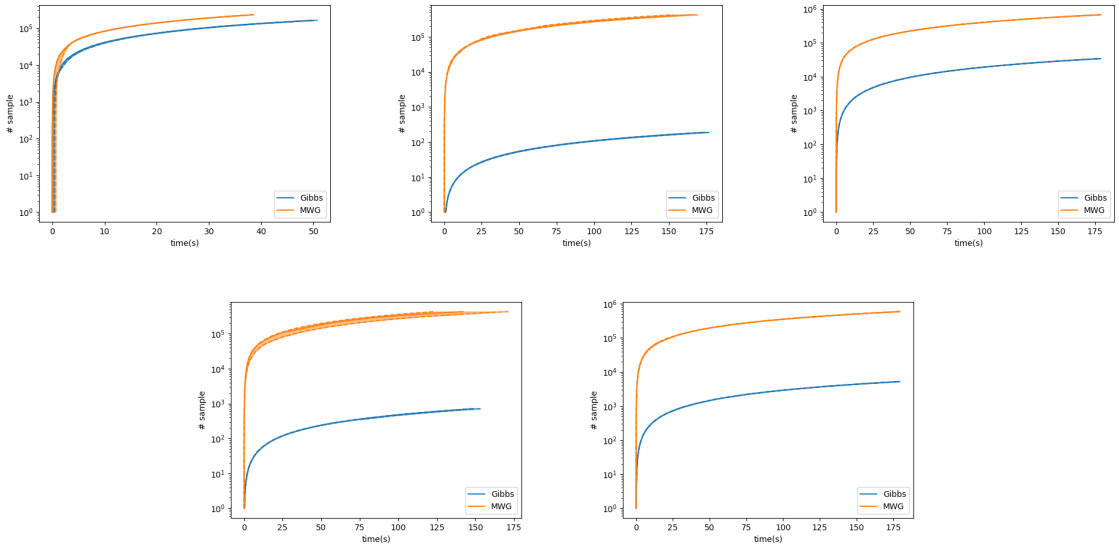

Figure 6: sample size

References

[AL08] J. Antoch and D. Legát. "Application of MCMC to change point detection". In: *Applications of Mathematics* 53.4 (2008), pp. 281–296.

[BF+18] A. Benson, N. Friel, et al. "Adaptive MCMC for multiple changepoint analysis with applications to large datasets". In: *Electronic Journal of Statistics* 12.2 (2018), pp. 3365–3396.

[CGS92] B. P. Carlin, A. E. Gelfand, and A. F. Smith. "Hierarchical Bayesian analysis of changepoint problems". In: *Journal of the royal statistical society: series C (applied statistics)* 41.2 (1992), pp. 389–405.

[CG97] J. Chen and A. K. Gupta. "Testing and locating variance changepoints with application to stock prices". In: *Journal of the American Statistical association* 92.438 (1997), pp. 739–747.

[ENJ04] J. B. Elsner, X. Niu, and T. H. Jagger. "Detecting shifts in hurricane rates using a Markov chain Monte Carlo approach". In: *Journal of climate* 17.13 (2004), pp. 2652–2666.

[Gre95] P. J. Green. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82.4 (1995), pp. 711–732.

[KFE12] R. Killick, P. Fearnhead, and I. A. Eckley. "Optimal detection of changepoints with a linear computational cost". In: *Journal of the American Statistical Association* 107.500 (2012), pp. 1590–1598.

[LL01] M. Lavielle and E Lebarbier. "An application of MCMC methods for the multiple change-points problem". In: *Signal processing* 81.1 (2001), pp. 39–53.

[Ste94] D. Stephens. "Bayesian retrospective multiple-changepoint identification". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 43.1 (1994), pp. 159–178.

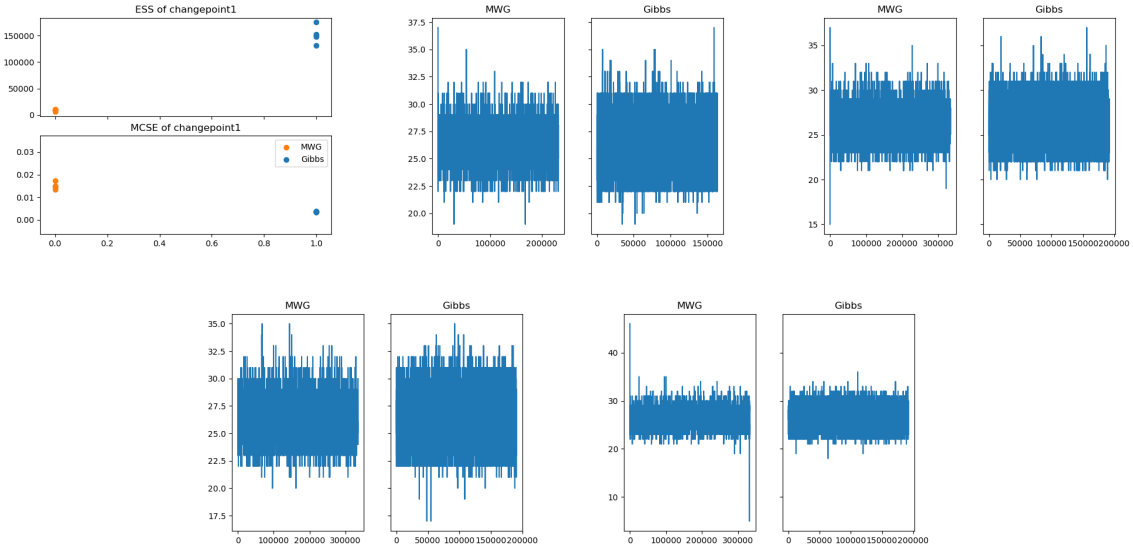

Figure 7: model1: trace of entire sequence

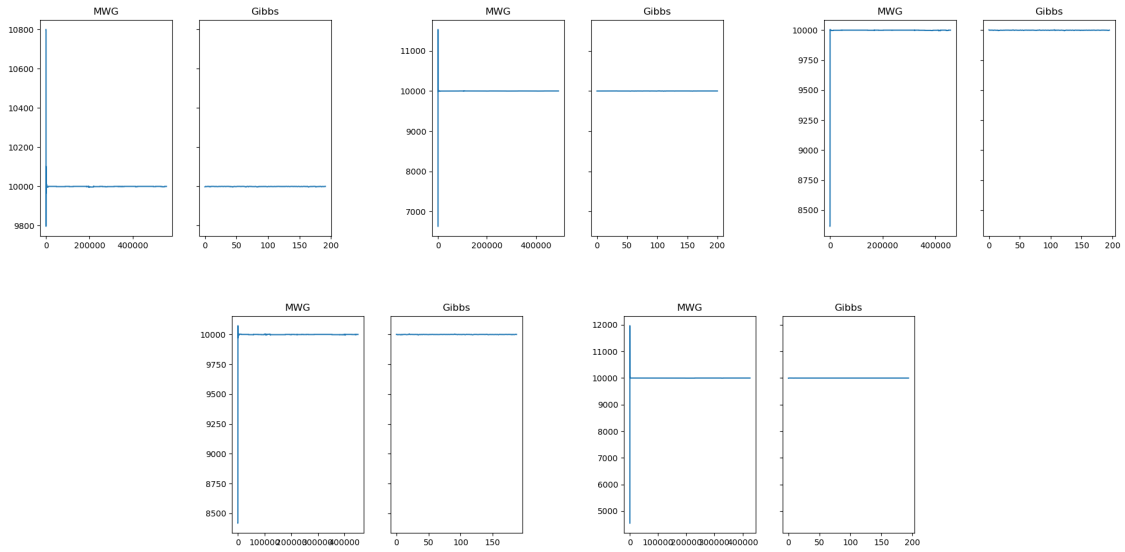

Figure 8: model2: trace of entire sequence

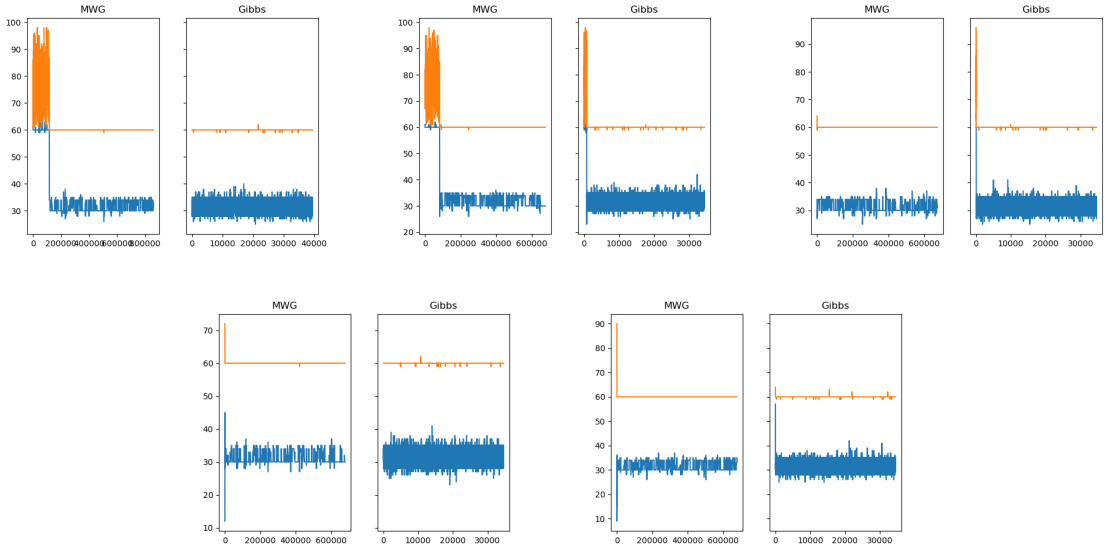

Figure 9: model3: trace of entire sequence

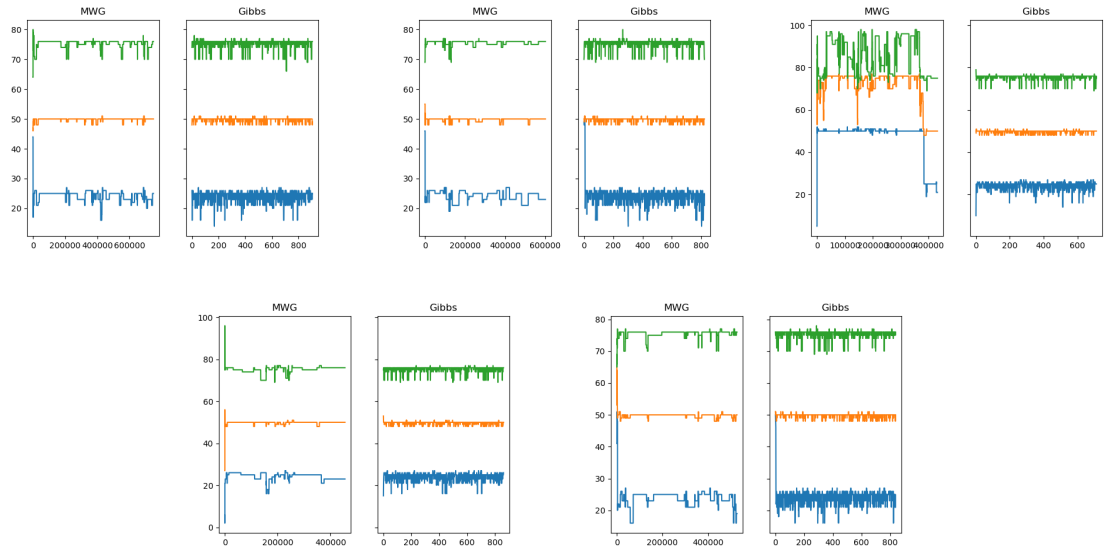

Figure 10: model4: trace of entire sequence

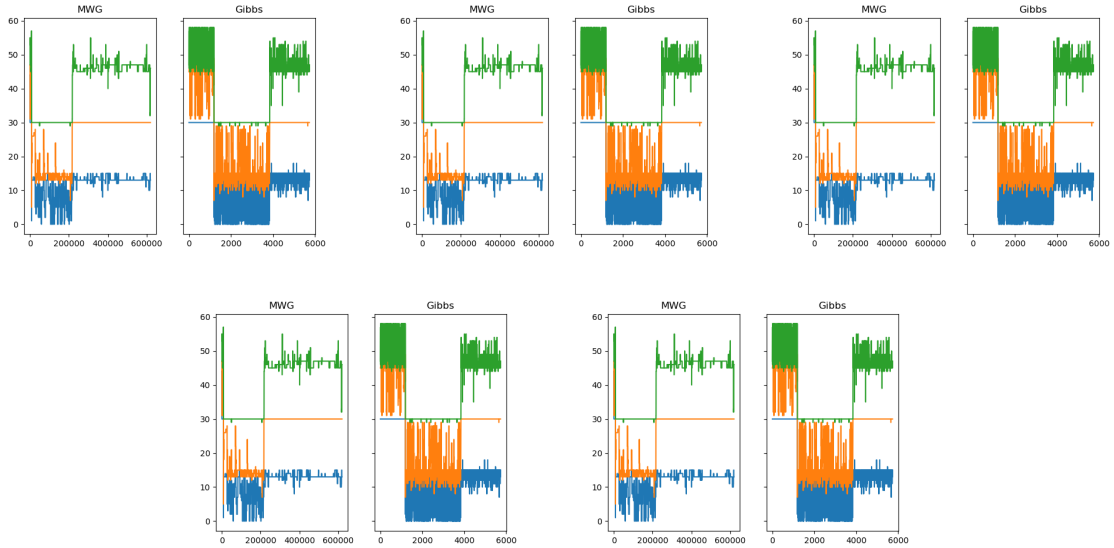

Figure 11: model5: trace of entire sequence

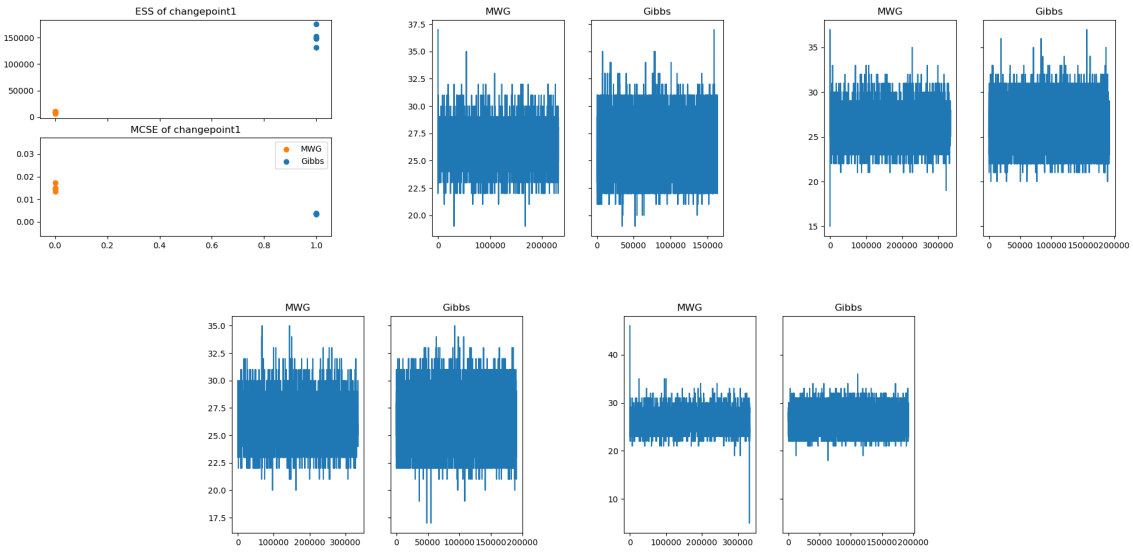

Figure 12: model1: trace after burn in

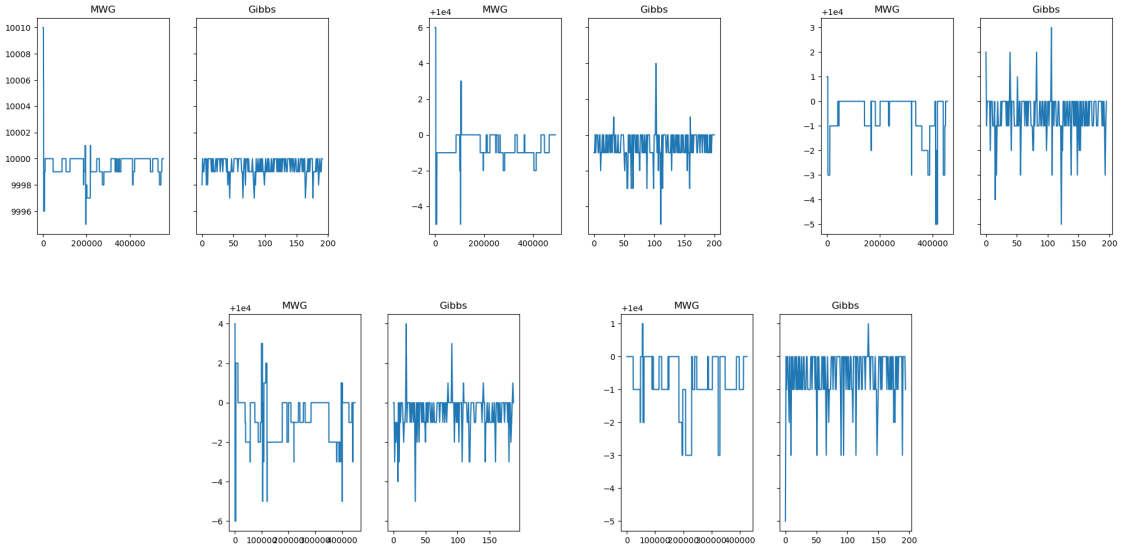

Figure 13: model2: trace after burn in

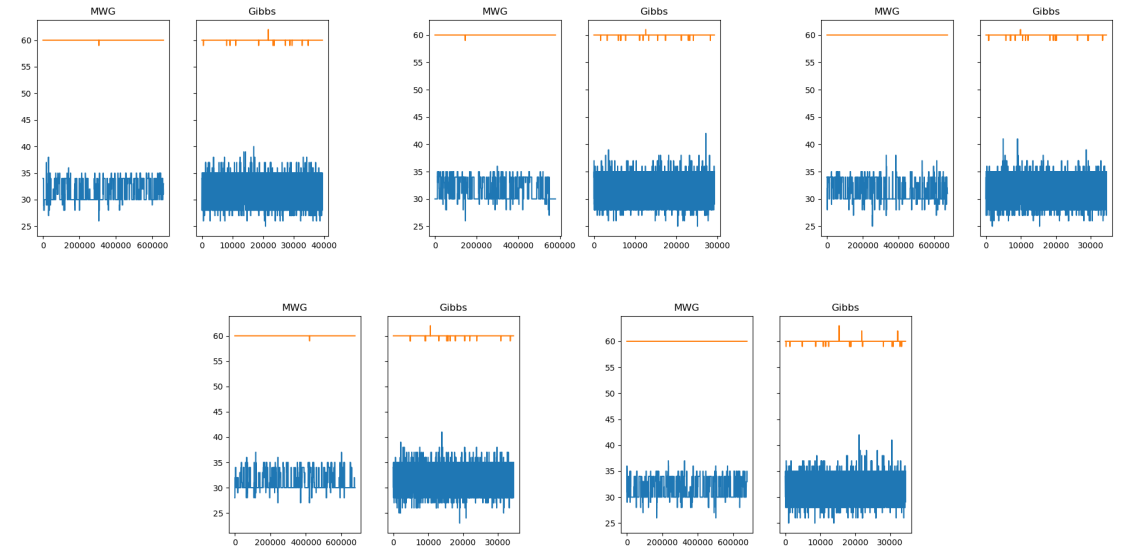

Figure 14: model3: trace after burn in

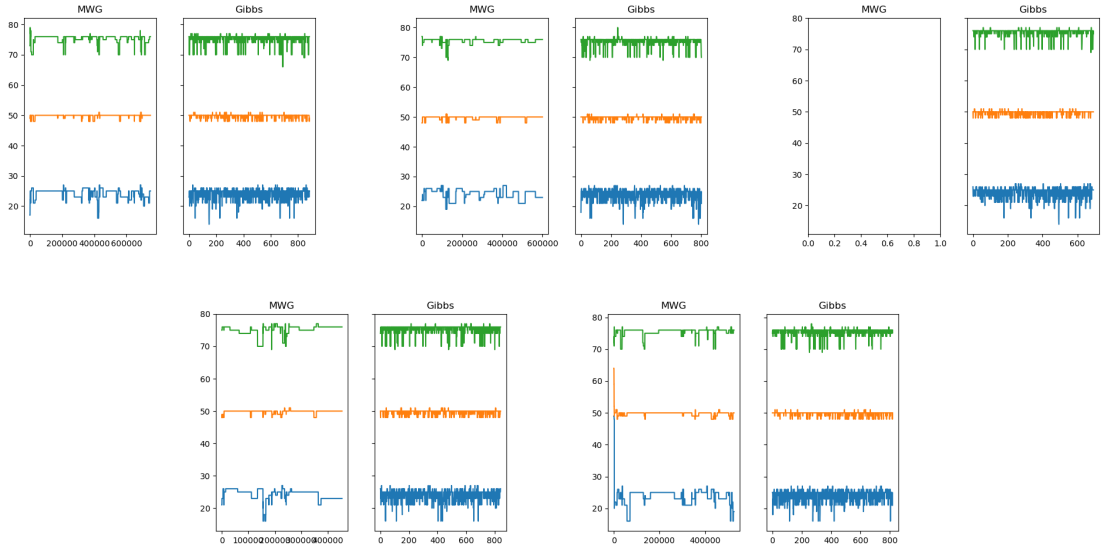

Figure 15: model4: trace after burn in

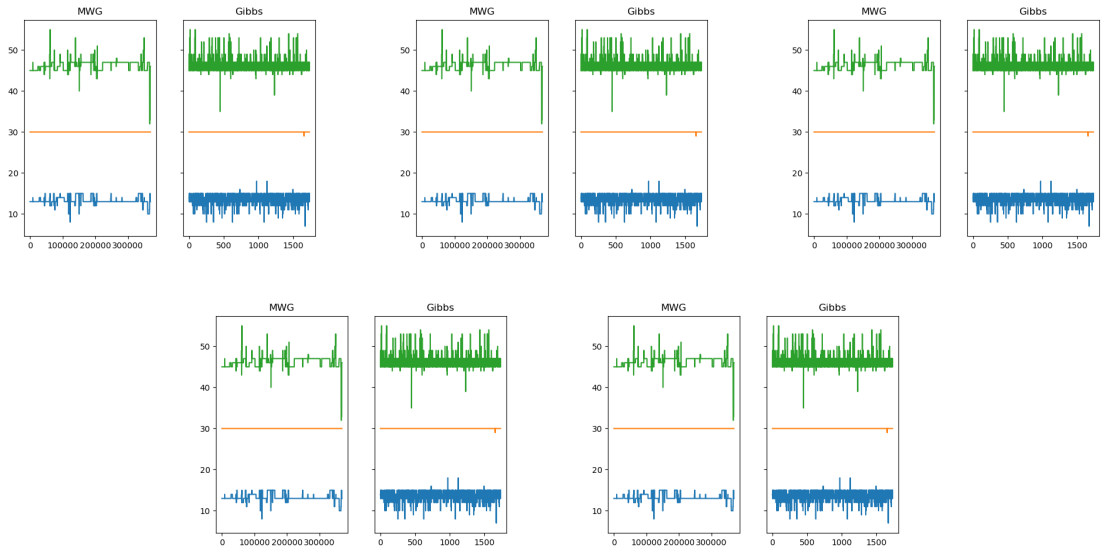

Figure 16: model5: trace after burn in

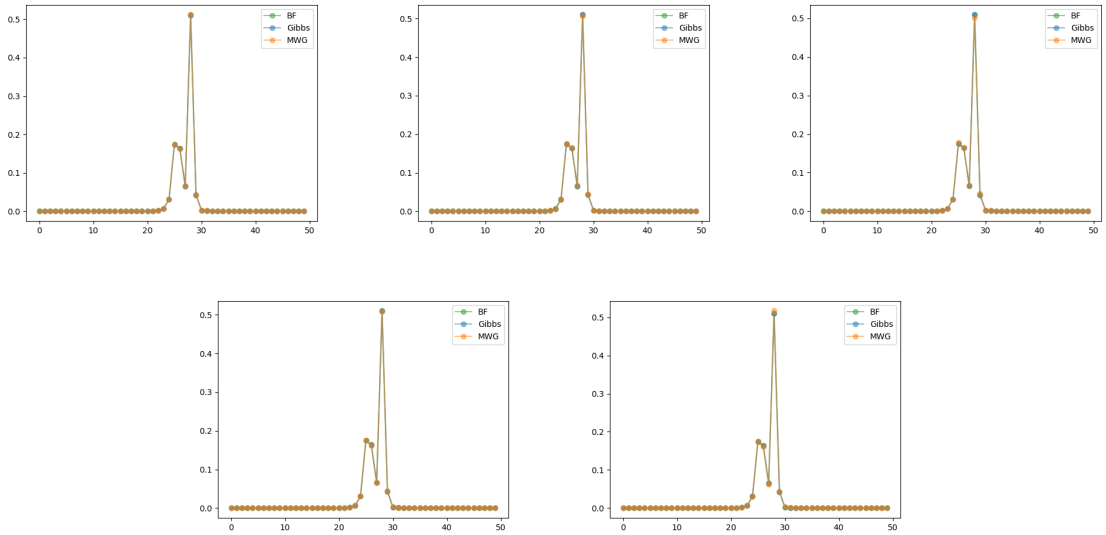

Figure 17: model1: posterior approximation

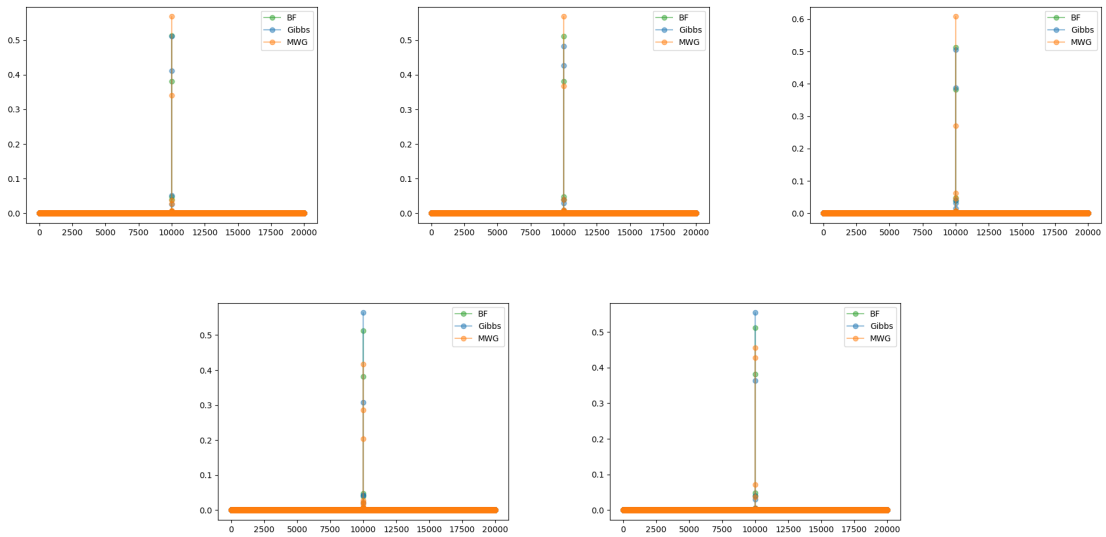

Figure 18: model2: posterior approximation

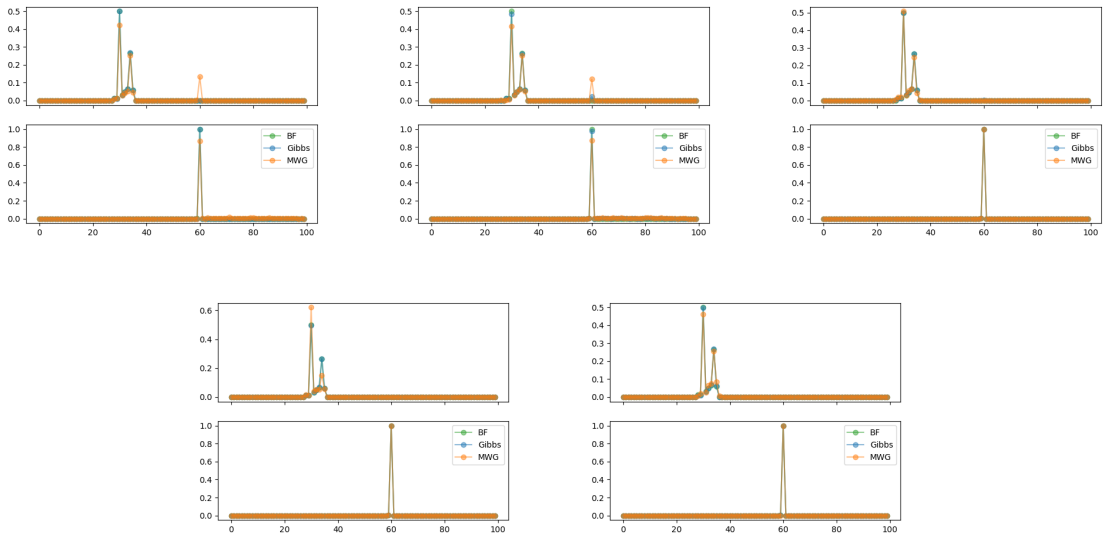

Figure 19: model3: posterior approximation

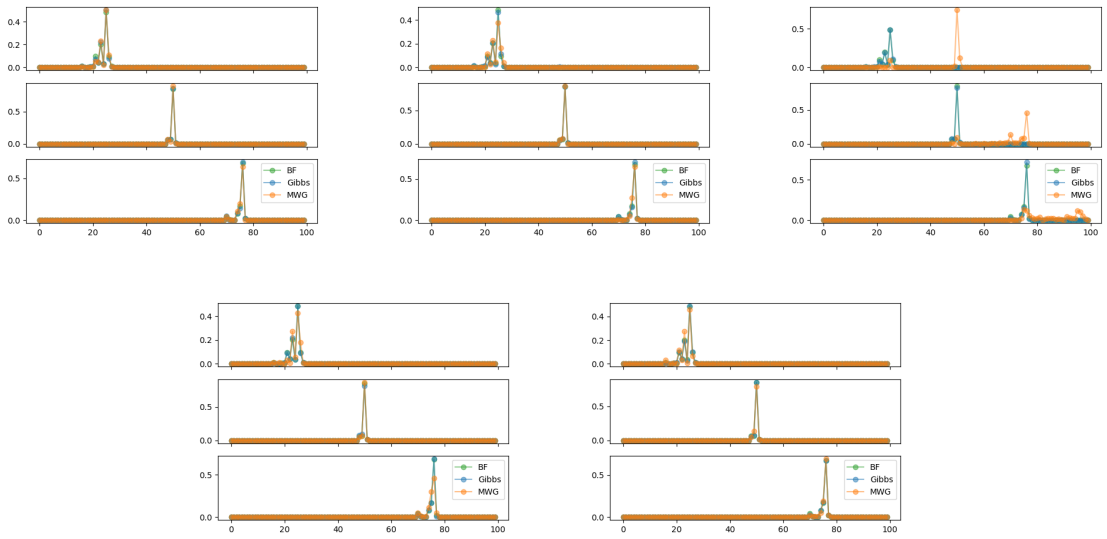

Figure 20: model4: posterior approximation

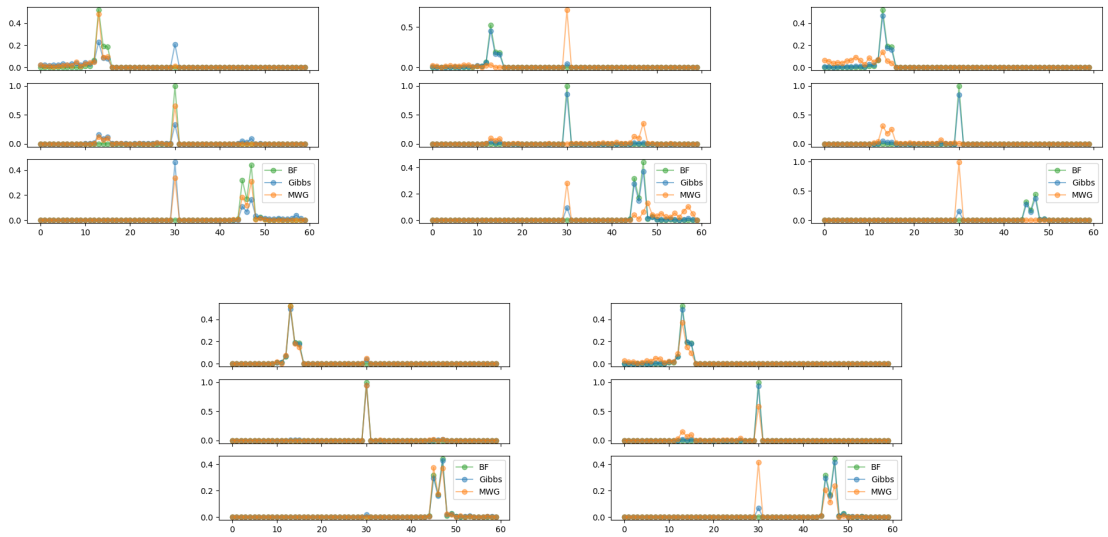

Figure 21: model5: posterior approximation

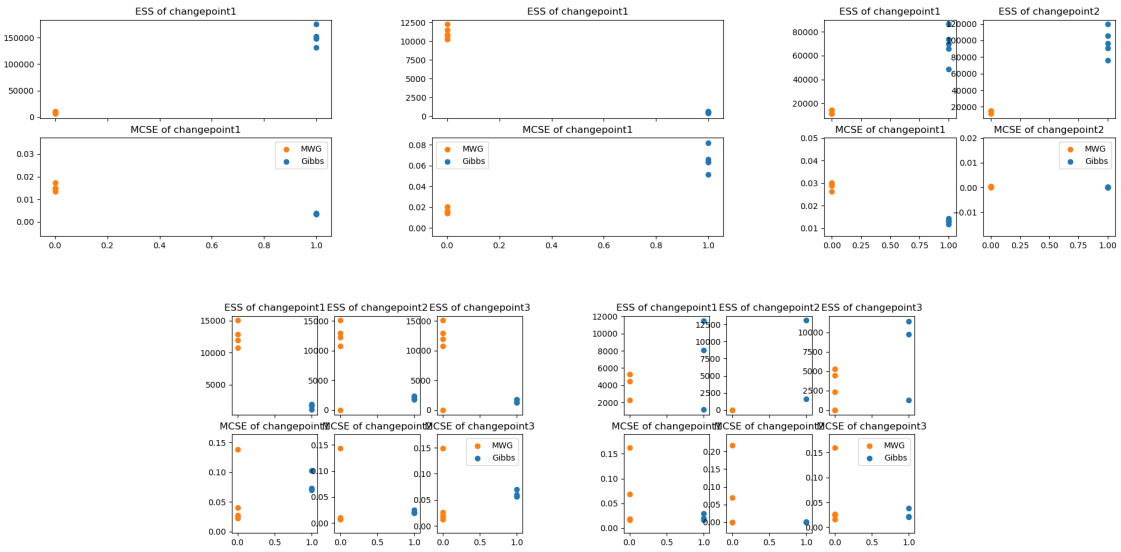

Figure 22: ess and se