Evaluating a Bias Compensation method for Sequential Monte Carlo(SMC)

Andreas Stasinakis

December 30, 2019

Abstract

Sequential Monte Carlo(SMC) methods are a powerful class of computational algorithms for Bayesian inference. One of their key merits is that they provide an estimate of the log marginal likelihood, which is however biased. In this project we propose a method, which uses the asymptotic distribution of the log-marginal likelihood for the bias correction. We evaluate this method empirically in order to conclude whether or not it improves the estimate compared to conventional methods. On top of this method a tuning parameter is introduced and tested as well.

Keywords: Bayesian Inference, Sequential Monte Carlo Methods, Log Likelihood estimation

1 Introduction

Before the Monte Carlo methods were discovered, Bayesian Inference was an amazing *theoritical* concept but it was hard to be implemented because, in most cases, the posterior distributions do not have a close form. Monte Carlo methods, played a major role on the practical expansion of Bayesian Inference, because it became possible to work with these posterior distributions using simulations. Monte Carlo methods are widely used in many scientific fields like Physics, Artificial intelligence, Finance etc[2].

Despite their many advantages and their exponential growth during the last decades, Monte Carlo methods are not perfect. One of the major issues that many Monte Carlo methods face, is the *Curse of Dimensionality*[6]. In particular, for high dimensional problems, the Monte Carlo estimators cannot approximate accurately the true value of the parameter, which drives to *biased* estimators. There are quite a few algorithms, that suffer from this, but in that project, only the Sequential Monte Carlo(SMC) method is discussed, using a problem defined as State Space Model in Dynamic Systems[3]. More specific, the major interest in that project is the estimation of the log likelihood of the data, $\log p(y_{1:T})$.

Nonetheless, this SMC estimator of the log likelihood is biased and in order to face this problem, a bias correction method is introduced and evaluated in this project. In the Methodology section, we first start with a brief description of the Linear Gaussian State Space Model(LG-SSM) problem, a background of how the two methods work, which parameters they use and how the evaluation is done.

In Expirements & Results section, we run simulations for the two methods for an 1000 dimensional LG-SSM problem and we compare the results. More explicitly, using several visualization and statistical measurements, we try to comment whether the proposed method reduces the bias of the SMC estimator or not.

Moreover, in the Discussion section a more subjective analysis of the results is done and a final comment on the proposed method is being given.

Finally, in the Conclusion, a summary of the project and the results is presented along with a proposal for further investigation of the bias correction method.

2 Methodology

2.1 Models

In order to compare the Bias correction(BC) method with the Sequential Monte Carlo(SMC) method, it is better to start with a simple linear problem. Therefore, in this project, the method is being tested in a simple Linear Gaussian State Space model(LG-SSM) problem[5, chapter 3.2.1]. In that way, using the properties of the Gaussian distribution, the *true value* of the log likelihood ($\log p(y_{1:T})$) can be computed using a Kalman Filter[5, chapter 3.2.2]. The obtained results could be generalized later in more complicated(non linear) problems.

The LG-SSM is being described with the three following models:

$$\begin{aligned} x_t &= 0.9x_{t-1} + e_t, && \text{(Transition Model)}, \\ y_t &= x_{t-1} + v_t, && \text{(Emission Model)}, \\ x_0 &\sim N(0, 10), && \text{(Initial Model)}, \\ \text{where } e_t, v_t &\sim N(0, 1) \end{aligned} \tag{1}$$

For both methods a number of parameters should be specified in advance. For the rest of the project the parameters below are being tested in different combinations.

T : Dimensions(steps) of the Linear Gaussian State Space Model,

N : Number of total Particles used,

M : Number of independent Particle Filters used for the BC method,

N_{BC} : Number of Particles used in each independent Particle Filter.

2.2 Sequential Monte Carlo(SMC)

A brief description of how SMC works is now being presented. The target is to estimate the log likelihood using in total N particles. The first step is to sample N particles from the initial model and define all the weights equal to $\frac{1}{N}$. After that, for each time step t, the following procedure for all the N particles is being repeated. *Resampling* particles using the weights from the previous step, *evaluating* them by computing the unnormalized weights $\widetilde{W_t^i}$ using the emission model and *propagating* them one step ahead using the transition model. Finally, the weights are being normalized and the procedure starts again. That way, only the *important* particles are being used and the rest non informative particles are being discarded. The procedure above continues for T time steps.[4, pp. 76]

For estimating the log likelihood $p(y_{1:T})$[4, pp. 54,100], we are interested in the sum of the log marginal likelihood for each time step t. More specific,

$$\log\, p(y_{1:T}) = \sum_{t=1}^{T} \left(c_t + \log \sum_{i=1}^{N} e^{v_t^i} - \text{logN}\right)$$

,

where $c_t = max\{\log \widetilde{W_t^1}, ..., \log \widetilde{W_t^N}\}$ and $v_t^i = \log \widetilde{W_t^i} - c_t$.

A drawback of this method is that the approximation for the log likelihood is biased, as it can be seen in the plot below. It can be noticed that not only a bias is introduced, but also the probability of underestimating the log likelihood is high. For this specific example, that probability is 0.9863.

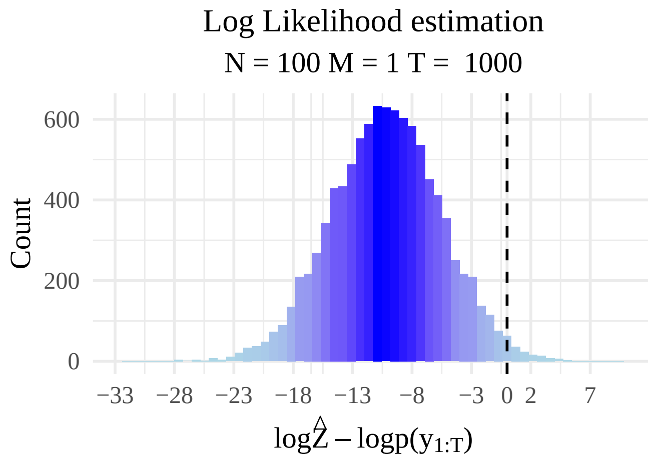

Figure 1: Histogram of 10000 samples for log likelihood estimation using SMC for N = 100. If the estimator was unbiased the mean of the distribution would have been in the dash line.

Using the Central Limit Theorem[1] for the likelihood estimator $\log\hat{Z} = \hat{\theta}$ and the true value $\log p(y_{1:T}) = \theta$, we can show that the bias and the variance of the SMC estimator have the following analytic form[4]:

$$\text{Bias} = E[\hat{\theta} - \theta] = -\frac{1}{2N}\sum_{t=0}^{T}\Big\{\int \frac{p(x_t|y_{1:T})^2}{p(x_t|y_{1:t-1})}dx - 1\Big\}$$

$$Var[\hat{\theta}] = \frac{1}{N}\sum_{t=0}^{T}\Big\{\int \frac{p(x_t|y_{1:T})^2}{p(x_t|y_{1:t-1})}dx - 1\Big\}$$

In other words, the bias of the SMC estimator can be also expressed as a function of the variance,

$$E[\hat{\theta} - \theta] = -\frac{Var[\hat{\theta}]}{2} \tag{2}$$

2.3 Bias Correction Method(BC)

The proposed Bias Correction method is mostly based on the equality (2) above. The procedure is divided into the following steps. Firstly, the tasks split in M different independent SMC, each one works as mentioned in 2.2. After that, the variance of those M independent samples of the log likelihood is being estimated. Making usage of the equality (2), we add half of the variance in each of the point estimators $\log\hat{Z}$. More specific,

$$\log\hat{Z} = \log\hat{Z} + \frac{Variance[\text{Sample of M }\log\hat{Z}]}{2} \tag{3}$$

Closing, using the Monte Carlo approximation[4, pp. 43-45], the final bias corrected point estimator is the mean of the above sample.

Moreover, all the experiments take into account the computational cost of the two methods. Therefore, the performance of them is tested under, approximately, the same computational budget. Thus for specifying different combinations of the parameters the following equality is used:

$$N = N_{BC} \cdot M, \tag{4}$$

which means that the cost for running a SMC with N particles should be (approximately) equal to the cost of running multiple(M) independent particle filters consist of N_{BC} particles each.

2.4 Bias Correction using a tuning parameter

In this project a variation of the BC method(equation 3) is also being proposed. A tuning parameter $\gamma \in [0, 1]$ is being added to control the variance of the BC estimator. Therefore the following H estimator is being tested:

$$H(N_{BC}, M, \gamma) = \sum_{i=1}^{M} \log(\hat{Z}_i) + \gamma \cdot \frac{\text{Variance}(\log \hat{Z_{1:M}})}{2}, \tag{5}$$

where $Z_{1:M}$ is a sample of M estimations for the log likelihood using N_{BC} particles each.

2.5 Evaluation

Both methods depend on the number of total particles N. Therefore for the rest of the project they are represented as a function of M and N_{BC} under the condition $N = M \cdot N_{BC}$. Thus, $BC(M, N_{BC})$ and $SMC(N) = SMC(M \cdot N_{BC})$.

2.5.1 Visualization Performance Criteria

- Histograms of the difference between the estimated log likelihood ($\log \hat{Z}$) and the true value ($\log p(y_{1:T})$) are being presented for both methods using the same computational budget(i.e $SMC(M \cdot N_{BC})$ and $BC(M, N_{BC})$). In this way, a better picture about the distribution of the two estimators under the same computational budget is being given.
- Histograms of the BC and a parallel SMC are also plotted. The parallel SMC is defined as a combination of M particle filters with N_{BC} particles each. Therefore it is the same procedure as the BC but without the correction part(i.e $SMC(M, N_{BC})$ and $BC(M, N_{BC})$).
- Histograms of the BC and SMC are being presented, but not for the same computational budget. Specifically, the BC is being tested for M particle filters of N_{BC} particles each, but only N_{BC} in total are used for the Standard method(i.e compare $SMC(N_{BC})$ with $BC(M, N_{BC})$).

2.5.2 Numeric Performance Criteria

Moreover, the three following performance criteria for the estimator $\log \hat{Z} = \hat{\theta}$ and the true value $\log p(y_{1:T}) = \theta$ of the log likelihood are being computed.

- *Bias:* $\text{Bias}(\hat{\theta}, \theta) = E(\hat{\theta}) - \theta$,
- *Variance* : $\text{Var}(\hat{\theta})$,
- *Root Mean Square Error:* $\text{RMSE} = \sqrt{Var(\hat{\theta}) + \text{Bias}(\hat{\theta}, \theta)^2}$.

3 Expirements & Results

3.1 BC vs SMC

The key component of the proposed BC method is the CLT(equalities 2 and 3). For this reason the number of N_{bc} for each individual particle filter should be chosen in order for the CLT to be activated. To achieve that, a sufficient enough number of particles is needed. In this project the number of particles used for the BC method, among others which have been tested, is $N_{BC} = 100$. The simulations have been ran in different combinations of M and N_{BC}. Only the important results are presented here(see Appendix A for different simulations). As can be seen in figure 2, for that amount of particles in each independent particle filter, the sampling distribution looks approximately normal. That allows us to start the experiments for a total number of particles $N = 1000$, $M = 10$ and $N_{BC} = 100$. Another simulation for $N = 2000$ has been also tested, but not presented here(see Appendix B).

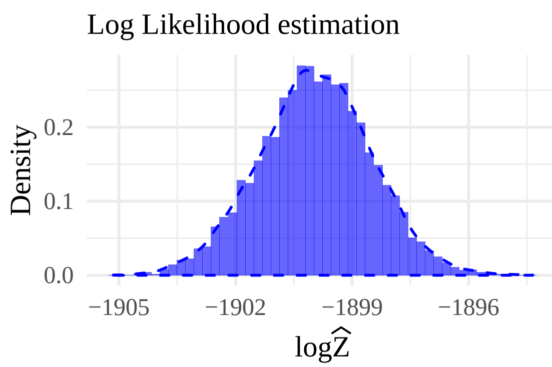

Figure 2: Density of a 10000 sample for the Log likelihood estimation for $M = 10$ and $N_{BC} = 100$.

The results in Table 1 show a reduction of the bias between the SMC and the BC method from -1.07 to approximately 0.655. The variance of the BC estimator was increased by 1200% though, which leads to a tripling of the $RMSE$. Moreover, for the parallel $SMC(10, 100)$ only the variance was reduced in comparison to the $SMC(100)$.

Table 1: Performance Criteria using M = 10, $N_{BC} = 100$ and N = 1000 for Standard SMC, BC and parallel SMC.

	SMC(10X100)	BC(10,100)	SMC(10,100)	SMC(100)
Bias	-1.070409	0.6552289	-10.115433	-10.15214
Variance	2.130812	27.8038373	2.104916	21.43957
RMSE	1.810076	5.3132271	10.218938	11.15811

From the visualization results, it can be seen that the BC estimator has a skewed to the right distribution with high variance in contrast to all the variations of the SMC, for which their distributions are approximately normal even if it leads to a bias estimation.

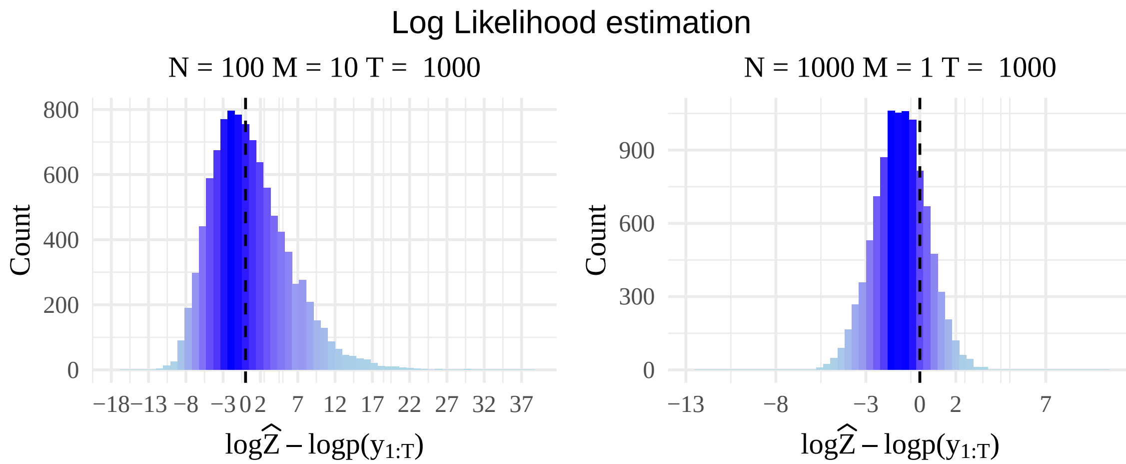

Figure 3: Graph of 10000 samples for BC(left) and SMC(right) for total number of particles N equal to 1000. That experiment matches the computational cost of both methods, given that the BC is done in a single GPU.

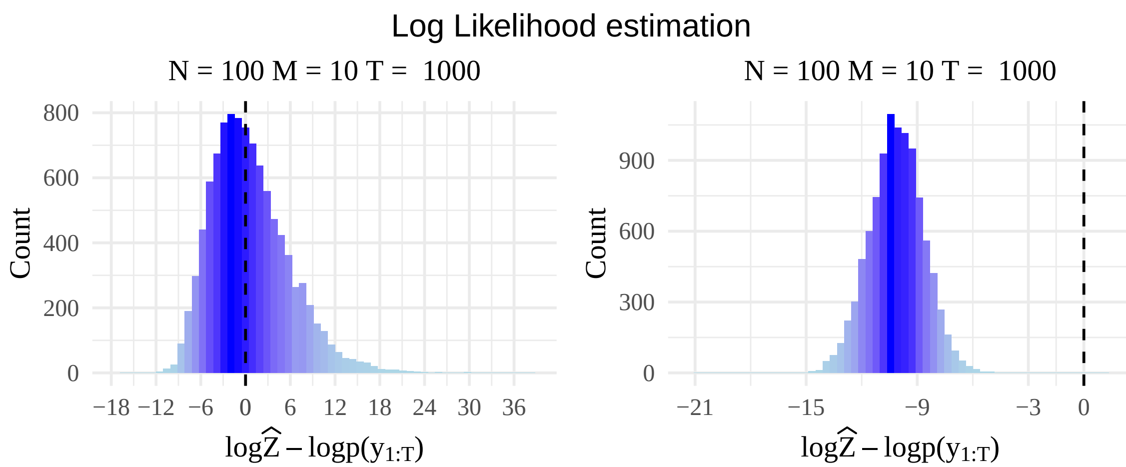

Figure 4: Graph of 10000 samples for BC(left) and parallel Standard SMC(right) using 1000 particles in total. In that case we match the computational cost given that both methods are parallelized in M independent GPU.

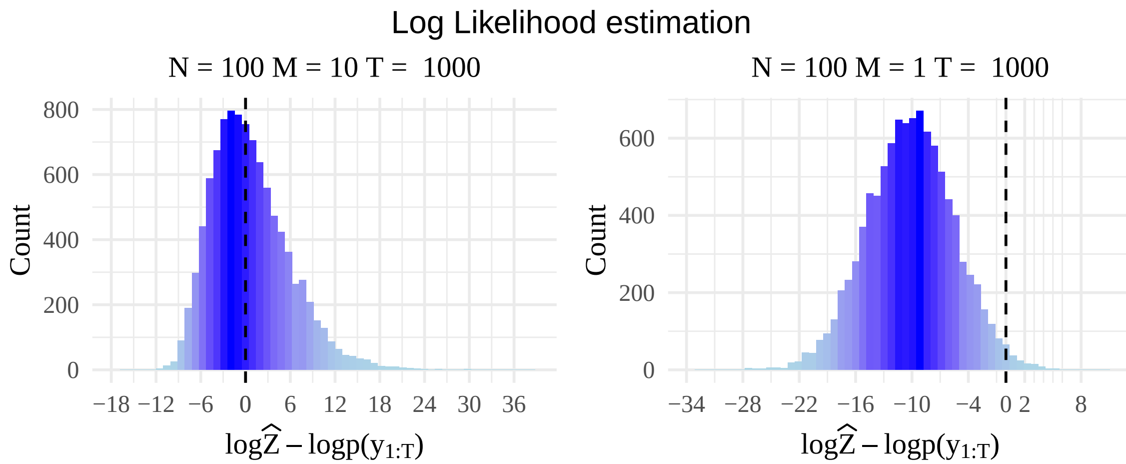

Figure 5: Graph of 10000 samples for BC(left) and Standard SMC(right) using 1000 particles and 100 particles in total respectively. Therefore the total computational cost is the same, only if we assume, that the parallelization procedure for the BC method has no extra cost.

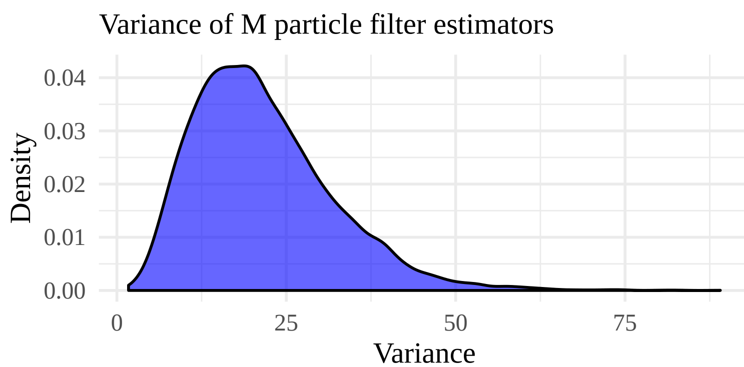

Figure 6: Density of the variance. It is the variance of the sample, which consists of M independent SMC estimators. Half of this value is later being used for the bias correction method.

3.2 BC vs SMC using a tuning parameter.

As discussed in the results above, the BC leads to a reduction of the bias of the SMC estimator, but that comes with the cost of an increase in the variance.

Using the same computational cost as before, we now try to adjust the variance of the estimator using a tuning parameter γ(equality 5). As we can see from Figure 7, while γ is increasing, the variance is increasing as well. On the other hand, the MSE and the absolute value of the bias are decreasing until a point, when they start increasing again. The value of γ, which minimizes the MSE is 0.77 and the value which minimizes the bias is 0.94.

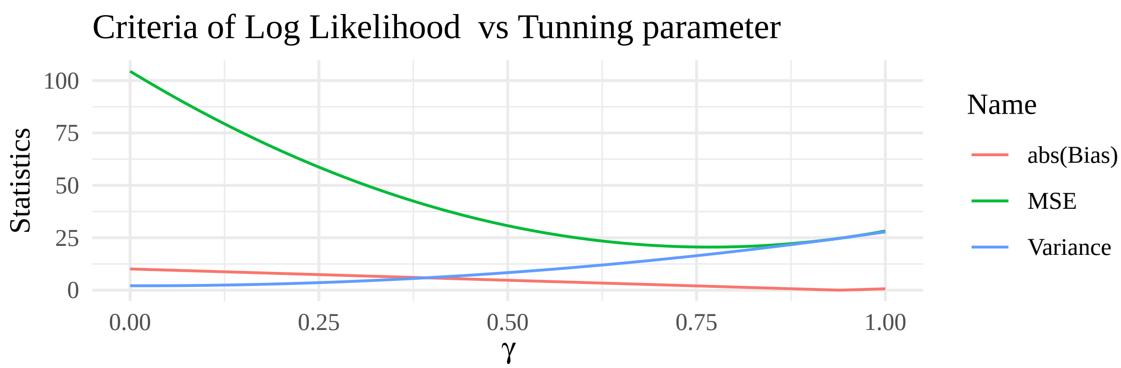

Figure 7: Plot of MSE, variance and absolute value of bias for different values of the tuning parameter between 0 and 1.

Table 2 shows that for $\gamma = 0.94$ a reduction in terms of bias has been accomplished, but the variance has only been reduced by approximately 11% compared to the Normal BC(i.e $\gamma = 1$). On the other hand, for $\gamma = 0.77$, the variance has been reduced by 38%, but it is still high(17.23) and the absolute value of the bias is even higher than the SMC(1.82 and 1.07 respectively).

Moreover, the histograms of the two different values of γ show, that despite the reduction of the variance, the distribution of the Log Likelihood estimator remains skewed to the right.

Table 2: Performance Criteria for $N = 1000$ of different methods. The first line is the optimal tuning parameter for minimizing the MSE, while the second line is the optimal parameter for minimizing the absolute bias. The last three lines have been already discussed in previous paragraphs but they are included in terms of comparison. For $\gamma = 0$ no variance is added to the estimator. Therefore it is the normal SMC. For $\gamma = 1$ all the proportion of the sample variance is added, consequently the BC method is used.

	γ	MSE	Variance	|Bias|
Min MSE	0.77	20.549867	17.231821	1.8220234
Min Bias	0.94	24.775060	24.777456	0.0089892
$BC(M, N_{BC})$	1.00	28.230382	27.803837	0.6552289
$SMC(M, N_{BC})$	0.00	104.426694	2.104916	10.1154332
$SMC(M \cdot N_{BC})$	0.00	3.276375	2.130812	1.0704093

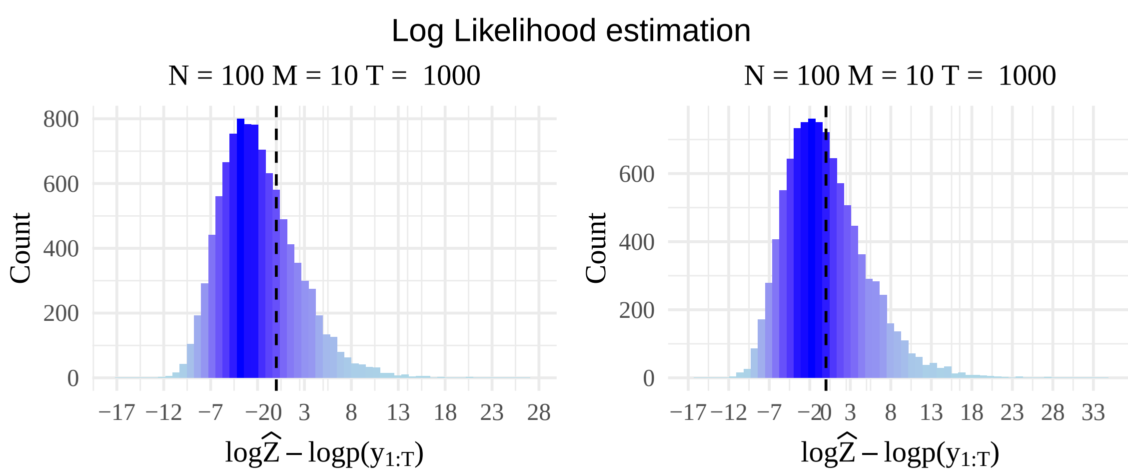

Figure 8: Histograms of log likelihood estimation for two different values of tuning parameter. On the left we use the value, which minimizes the MSE(0.77) and on the right the value, which minimizes the bias(0.94).

4 Discussion

In this project a method for reducing the bias of the SMC estimator is introduced. From Table 1 it is clear, that a reduction of the bias has been accomplished. Unfortunately, the variance of the BC estimator has a tremendous increase from 2.13 to 27.8. In order to look deeper into that result, we also plot(Figure 6) the variance of each sample, which consists of $M = 10$ point estimators using $N_{BC} = 100$ each. That is the variance, which is later added to each SMC estimator for the BC method. One might say that the distribution of the variance has a large variance too. Therefore, when we simulate a sample of M log likelihood estimators, the variance of that sample is high and it leads to a high variance in the BC estimator.

The results of Table 1 are more clearly represented on Figure 3,4 and 5, on which a better understanding of the BC can be given. More specific, in Figure 3 the mass of the distribution of the BC is closer to 0 than the SMC method. Additionally, the distribution of the former is skewed to the right and its variance is much larger than the variance of the latter.

For the Figure 4 we can compare the two methods in terms of computational cost, assuming that we can parallelize the M particle filters for both of them. The parallel SMC has captured the shape of a normal distribution with low variance but it is clear that the estimation is inaccurate. This is expected because we are just averaging over the standard SMC. Therefore only a reduction in the variance has been accomplished.

A comparison between $BC(10, 100)$ and $SMC(100)$ is presented in Figure 5. In that case, we assume that parallel procedure is done *without* any additional cost. The difference between the bias of the two estimators is large. For the same number of particles(100) on each PF, the BC performs more efficient than the SMC. One of the advantages of the BC is, that the independent particle filters can be parallelized in contrast to the standard SMC in which all simulations have to be done in the same GPU/computer node.

In all the results in section 3.1 we came across with a drawback of the BC method. That is the large variance of the estimator. In order to counter this problem, we add a tuning parameter γ, which allows us to control the *Bias - Variance trade off*. The first attempt was to find the optimal γ in terms of MSE but during the experiments we realized, that this leads to a less accurate estimation of the log likelihood. An explanation for that could be, that the MSE is a combination of the variance and the bias of the estimator. Unfortunately, the scale of the variance is much higher than the scale of the bias. That means that while we are trying to reduce the variance by minimizing the MSE, we may have an increase in the bias of the estimator.

From the Table 2 we can say, that using a tuning parameter both the bias and the variance of the BC estimator are being reduced. For $\gamma = 0.94$ we have a bias, which is really close to zero(0.009) but only a small reduction of the variance is achieved. In contrast, minimizing the MSE, increased the bias of the estimator without any significant reduce of the variance.

Therefore adding a tuning parameter does not give us an *optimal solution* in terms of the *Bias - Variance trade off*. An impressive reduction of the bias has been accomplished but it did not manage to solve the problem of the large variance.

5 Conclusion

In conclusion, in this project a new method for reducing the bias of the *SMC* estimator using the variance of the sample is proposed and tested in a simple *LG-SSM*. The experiments show, that the method reduces the bias of the estimator but at the same time the variance of the new estimator is large. In order to counter this, a tuning parameter, which controls the *Bias - Variance trade off*, is added. The updated version of the BC estimator led to a greater decrease of the bias but without any significant effect to the variance of the estimator. However, given the reduction of the bias, it may be interesting for future exploration, to combine the BC method with one of the known variance reduction methods.

References

[1] Nicolas Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Annals of Statistics*, 2004.

[2] A E Gelfand. Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics - W. Walter R. Gilks, Sylvia Richardson, David J. Spiegelhalter - Google Books. *Markov chain Monte Carlo in practice*, 1995.

[3] Bo Li, Thomas Bengtsson, and Peter Bickel. Curse-of-dimensionality revisited: Collapse of importance sampling in very large scale systems. pages 1–18, 2008.

[4] Thomas B Schön and Fredrik Lindsten. Learning of dynamical systems. Particle filters and Markov chain methods. 2017.

[5] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. Probabilistic robotics (intelligent robotics and autonomous agents series). *Intelligent robotics and autonomous agents, The MIT ...*, 2005.

[6] R. H. Wilcox. Adaptive control processes—A guided tour, by Richard Bellman, Princeton University Press, Princeton, New Jersey, 1961, 255 pp. *Naval Research Logistics Quarterly*, 1961.

Appendix A

Bias Correction for different combinations of M and N_{BC} for $N = 1000$.

In order to choose the optimal M, several combinations have been tested. As it can be seen in Table 3, for a fixed N, while M is increasing, the bias is increasing as well. This can be explained from the fact that for higher value of M, the number of particles N_{BC} is decreasing, therefore the CLT is not activated yet.

Table 3: Performance Criteria for M = 10,15,20,25 only the BC vs the Standard SMC using 1000 particles in total

	SMC(10X100)	BC(10,100)	BC(15,67)	BC(20,50)	BC(25,40)
Bias	-1.070409	0.6552289	1.452707	2.797775	4.363312
Variance	2.130812	27.8038373	41.181889	58.032788	74.347084
RMSE	1.810076	5.3132271	6.579371	8.115080	9.663237

Appendix B

Bias Correction for N = 2000.

A higher number of total particles has been also tested(N = 2000) for $M = 5$ and $N_{BC} = 400$. While the $N \to \infty$, the SMC estimator approaches the true value of the parameter. The probability that the standard SMC method underestimates the log likelihood is now 0.7024. Despite that, the estimation of the BC for the bias is again more accurate and also the variance has been reduced in compare to the experiment for N = 1000.

Table 4: Performance Criteria for M = 5 between BC,SMC and parallelized SMC

	SMC(2000)	BC(5,400)	SMC(5,400)
Bias	-0.5369281	0.0614668	-2.595236
Variance	1.0682514	4.5231477	1.056489
RMSE	1.1646615	2.1275511	2.791349

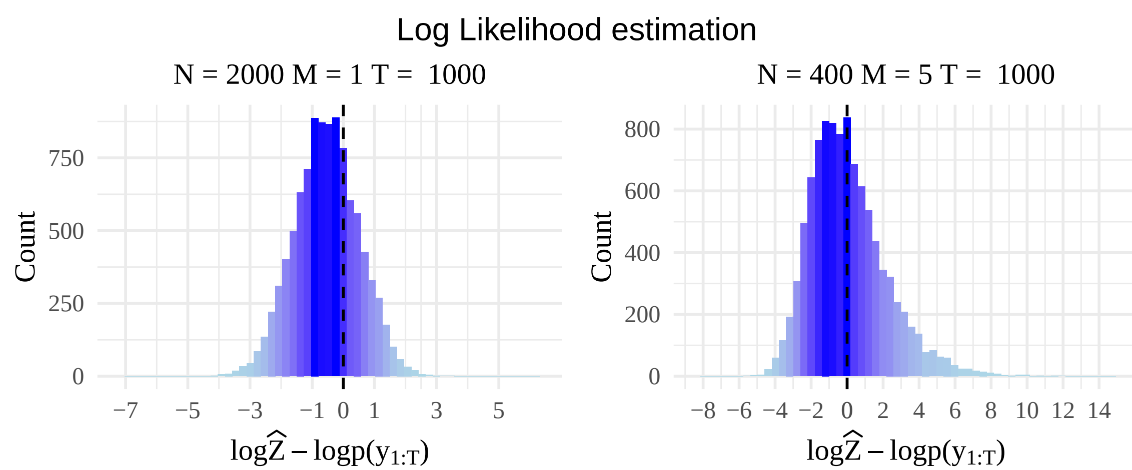

Figure 9: Histograms of the standard SMC(left top) for 2000 particles, BC(right top) for M = 5

Bias Correction using tuning parameter for N = 2000.

Table 5: Performance Criteria for M = 5, N = 400. The first line is the optimal tuning parameter for minimizing the MSE, while the second line is the optimal parameter for minimizing the bias. The last three lines have been already discussed in previous paragraphs, but they are included for comparison.

	γ	MSE	Variance	|Bias|
Min MSE	0.65	3.263006	2.509174	0.8683790
Min Bias	0.98	4.384467	4.384836	0.0083328
$BC(M, N_{BC})$	1.00	4.526474	4.523148	0.0614668
$SMC(M, N_{BC})$	0.00	7.791631	1.056489	2.5952356
$SMC(M \cdot N_{BC})$	0.00	1.356436	1.068251	0.5369281

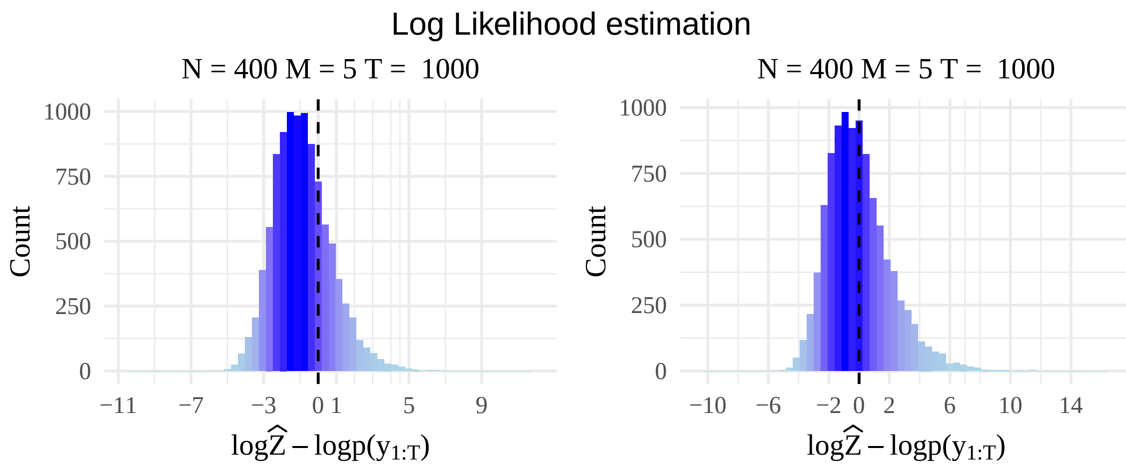

Figure 10: Histograms of different values of tuning parameter for M = 5 and 400 particles for each PF. On the left is the value which minimized the MSE while on the right the value which minimizes the bias.