MoM Summary

February 12, 2025

1 MoM on Importance Sampling

Example 1.1. *Consider an importance sampling with target distribution* $N(0,1)$ *and proposal distribution* $N(0, \frac{1}{1+\epsilon})$*, where* $\epsilon > 0$*. Then the weight will be*

$$w(x) \propto exp(\frac{1}{2}\epsilon x^2),$$

unbounded on $x \in \mathbf{R}$ *with bounded variance for* $\epsilon < 1$*.*

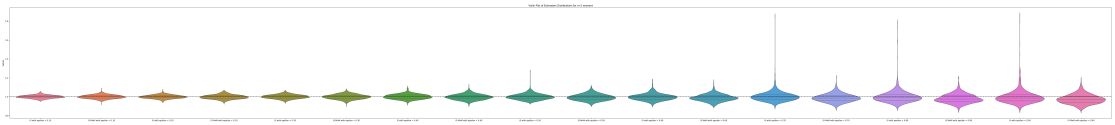

Figure 1: Plot of distribution of 1000 estimations of second moment from IS and IS-MoM.

Here, we estimate the second moment of the target distribution using 10,000 samples drawn from the proposal distributions. The x-axis represents the different values of ϵ *used in the proposal distributions. For* $\epsilon \geq 0.4$*, we observe that the MoM estimator produces distributions with thinner tails.*

(From the graph, it is evident that MoM increases the deviation of the median from the true value. Specifically, the distance between the median of the MoM-based distribution and the true value grows. However, in some cases where the IS estimator exhibits a heavy tail towards positive infinity, MoM effectively trims the tail, resulting in a leftward skew of the distribution. This trimming effect can sometimes reduce bias.)

We also evaluate the behavior of the MoM confidence interval in this example. Using 10,000 samples, we construct both the MoM confidence interval and the confidence interval derived from asymptotic variance estimation. When ϵ *is small* ($\epsilon = 0.1$)*, the two methods yield similar coverage rates, both close to the nominal*

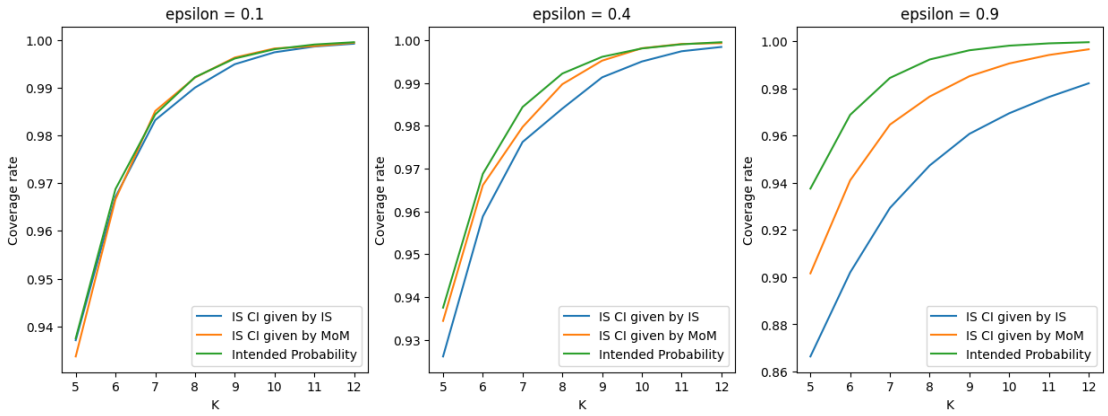

Figure 2: Plot of coverage rate against K.

level. However, as ϵ increases, the confidence interval derived from asymptotic variance exhibits a significantly smaller coverage rate compared to the MoM confidence interval, due to underestimated variance leading to poorer interval estimation. Although the MoM confidence interval also shows a reduction in coverage rate with increasing ϵ, it performs better when K is large.

It is important to note that the MoM confidence interval has greater length, and this length increases with K.

2 MoM on MCMC

Example 2.1. *In this example, we explore the application of the MoM method to estimate confidence intervals in MCMC.*

Specifically, we consider a Gibbs sampler for a multivariate normal distribution:

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right),$$

where $\rho \in [0,1]$. It is well known that the Gibbs sampler mixes quickly when the coordinates are weakly correlated (e.g., $\rho = 0.5$) but mixes slowly when the coordinates are highly correlated (e.g., $\rho = 0.99$). These two cases are used as examples in our analysis.

In this analysis, variance estimation is performed using the overlapped batch means (OBM) method [3]. The MoM confidence interval demonstrates significantly better performance, particularly in cases of slow mixing (where samples are highly correlated), where OBM typically underperforms.

We also examine an extreme case with $\rho = 0.999$, including burn-in samples in the analysis. In this scenario, the MoM confidence interval maintains strong

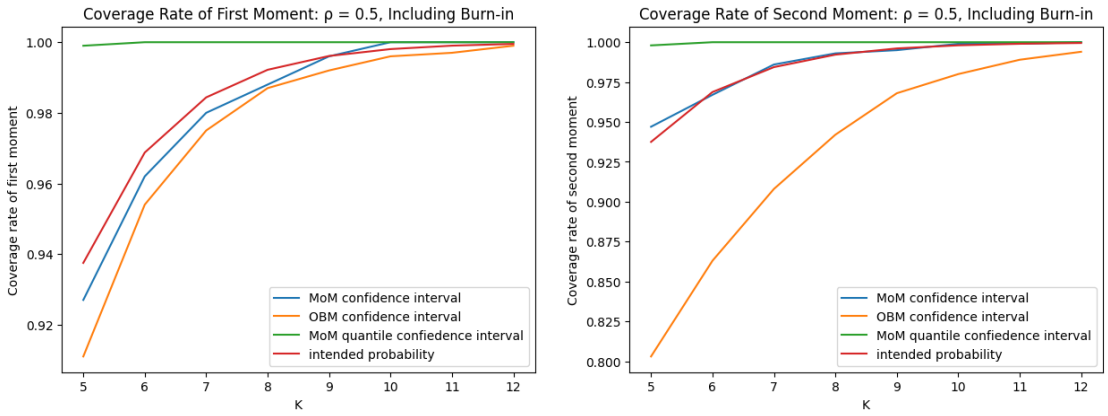

Figure 3: Plots of coverage rate of confidence interval for first moment and second moment, on the fast mixing case $\rho = 0.5$.

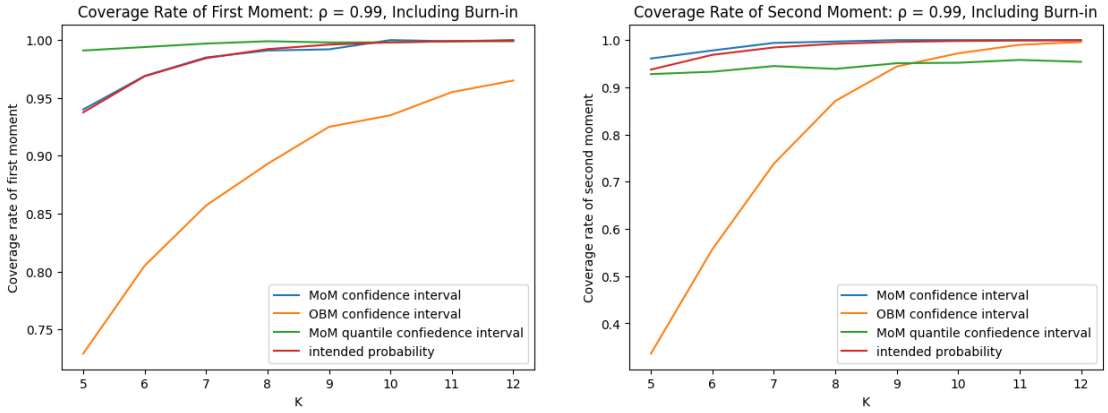

Figure 4: Plots of coverage rate of confidence interval for first moment and second moment, on the fast mixing case $\rho = 0.99$.

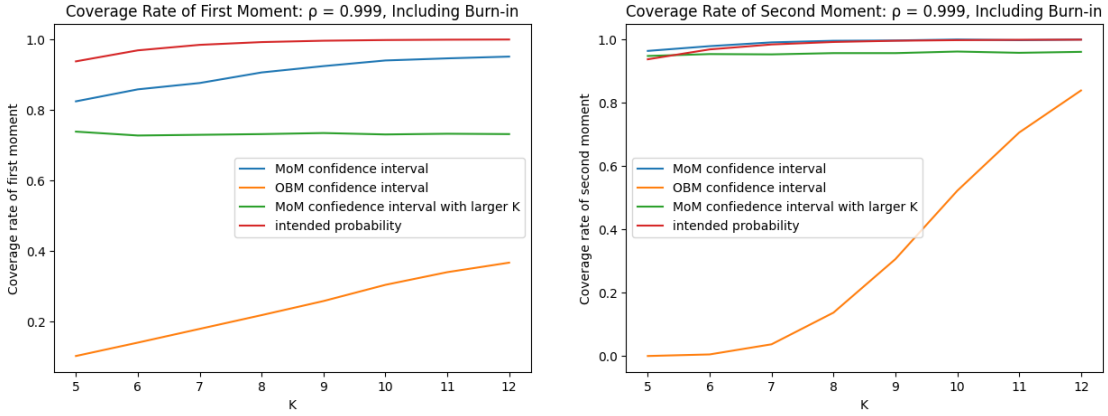

Figure 5: Coverage rate against K, with $\rho = 0.999$, burn-in included.

performance as long as the burn-in period does not constitute too large a proportion of the total sample size. This is because burn-in samples primarily affect the bounds of the interval, causing the upper bound to increase (or the lower bound to decrease). Despite this, the true value remains effectively captured within the confidence interval, although the interval length becomes larger.

Example 2.2. *This example focuses on case-deletion importance sampling, as proposed in [1]. We consider a Bayesian linear regression model with a specified prior. The full posterior is used as the proposal distribution, while the posterior with one data point removed serves as the target distribution. The posterior samples are generated using a Gibbs sampler.*

For certain data points, deletion can result in an importance weight that is unbounded but has a bounded variance. In this example, we use the stack loss data set and remove the 14th data point to illustrate this behavior.

Note that although MoM estimators exhibit better tail behavior, they cannot correct the intrinsic skewness of the IS estimators. As a result, the MoM estimator remains biased relative to the true value, which, in this case, is approximated using Gibbs sampling from the case-deletion posterior.

3 MoM on Pseudo Marginal

Consider the setup for a pseudo-marginal sampler in a Bayesian model with parameter θ and data y. The model includes a prior $v(\theta)$ and a likelihood $p(y|\theta)$, where $p(y|\theta)$ is generally assumed to be intractable.

In the pseudo-marginal MCMC framework, the intractable likelihood $p(y|\theta)$ is

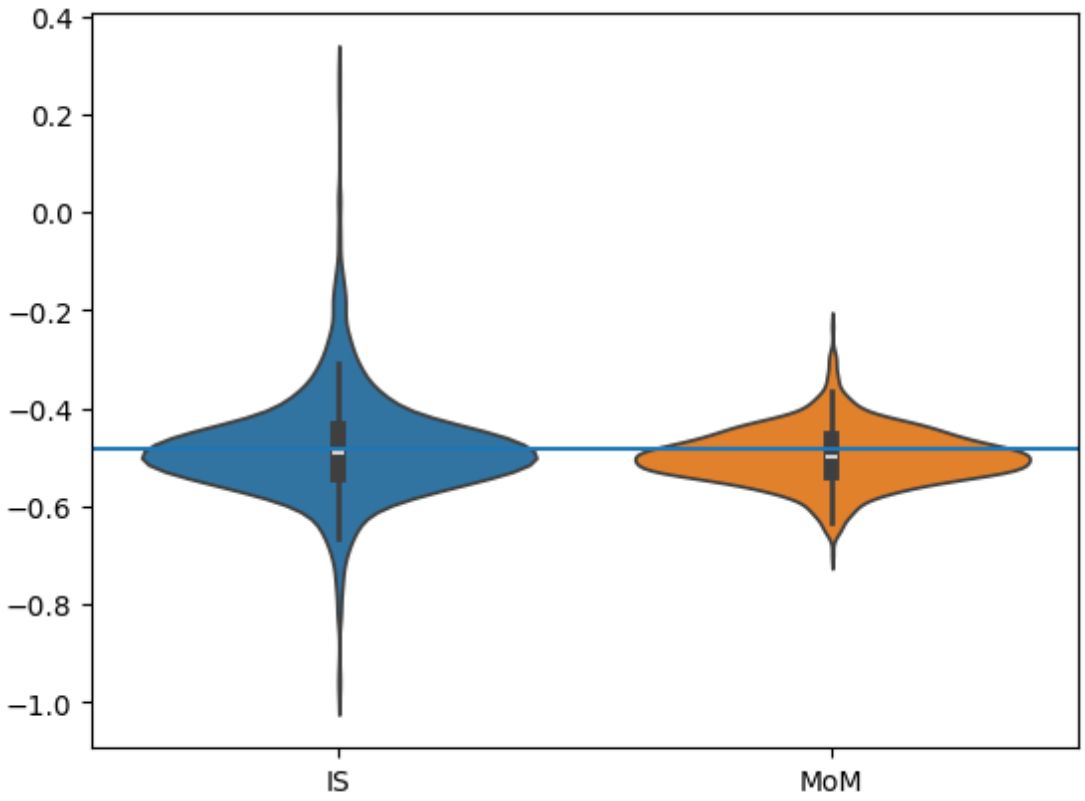

Figure 6: Plot of distribution of 1000 estimations each with 15000 samples of the last coordinate of the intercept value with the case-deleted posterior. Data point 1 deleted.

replaced by its unbiased estimator $L(\theta, Z)$. Specifically, we consider the case where

$$L(\theta, Z) = \sum_{i=1}^{N} h(z_i, \theta),$$

for some function h, with $z_i \overset{\text{i.i.d.}}{\sim} \mathbb{M}^\theta(dz)$.

In the MoM version of the pseudo-marginal method, the estimator $L(\theta, Z)$ is replaced with the corresponding MoM estimator, $MoM(\theta, Z)$.

Proposition 3.1. *Algorithm* **??** *defines the Markov kernel for a Metropolis-Hastings sampler with proposal distribution* $v(\theta)\mathbb{M}^\theta(Z)$ *and invariant distribution given by*

$$\pi_{MoM}(\theta, Z) = v(\theta)\mathbb{M}^\theta(dZ)\frac{MoM(\theta, Z)}{p(y)}.$$

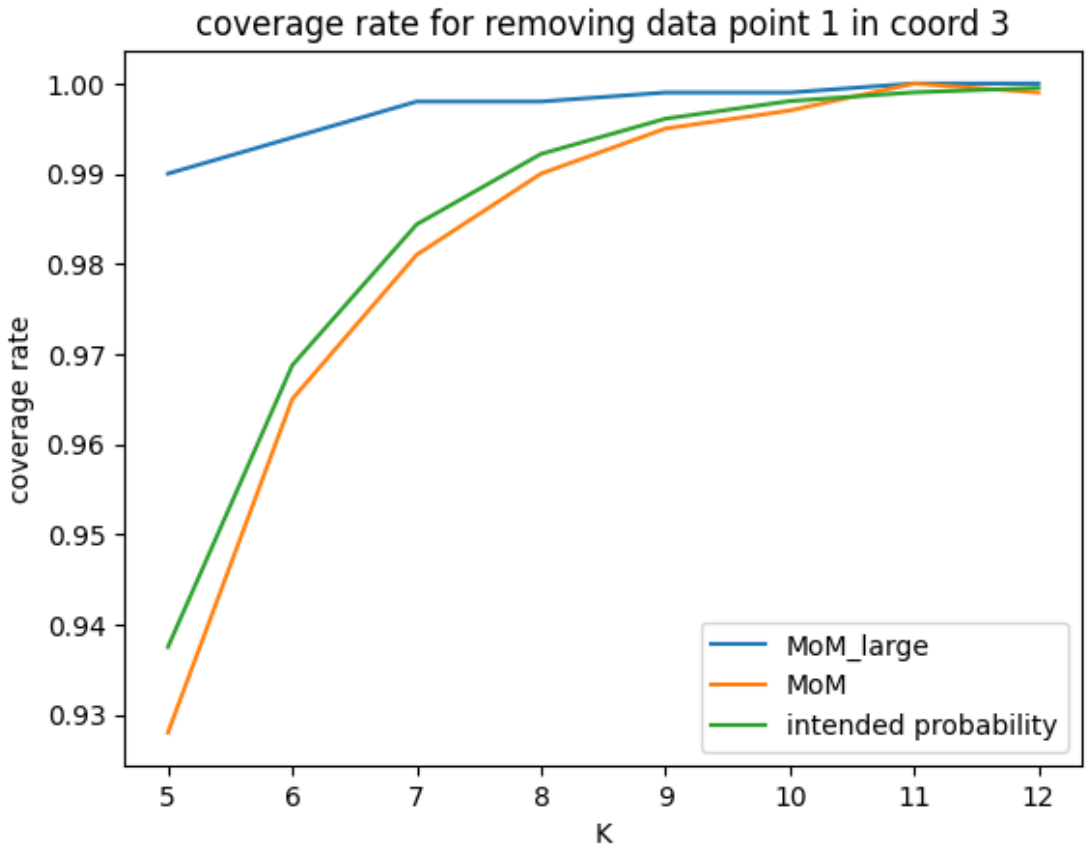

Figure 7: Plot of coverage rate against K, consisted of 1000 estimated confidence interval each with 15000 samples of the last coordinate of the intercept value with the case-deleted posterior. Data point 1 deleted.

Proof. The kernel we define can be written as

$$K(\theta, Z, d\theta, dZ) = M(\theta, d\tilde{\theta})\mathbb{M}^{\tilde{\theta}}(d\tilde{Z})r_{MoM}(\theta, Z, \theta, \tilde{Z}) + \mathbf{1}\{\theta = \tilde{\theta}, Z = \tilde{Z}\}c(\theta, Z),$$

where

$$c(\theta, Z) = 1 - \iint r_{MoM}(\theta, Z, \theta, \tilde{Z})M(\theta, d\tilde{\theta})\mathbb{M}^{\tilde{\theta}}(d\tilde{Z}),$$

represents the probability of rejection.

To verify the invariant distribution, we check the detailed balance equation. The rejection case is straightforward to verify, so we consider the acceptance case, i.e., when $\theta \neq \tilde{\theta}$ and $Z \neq \tilde{Z}$. WLOG, assume that $r_{MoM}(\theta, Z, \theta, \tilde{Z}) < 1$; then, we have $r_{MoM}(\tilde{\theta}, \tilde{Z}, \theta, Z) = 1$.

Algorithm 1 Generic MoM Pseudo Marginal Kernel

procedure INPUT(θ, Z)
  Sample $\tilde{\theta} \sim M(\theta, d\Theta)$
  Sample $\tilde{Z} \sim \mathbb{M}^{\tilde{\theta}}(d\mathbf{Z})$
  Sample $U \sim \mathcal{U}([0,1])$
  Compute

$$v = \log r_{MoM}(\theta, Z, \theta, \tilde{Z}) = \log \min \left\{ \frac{v(\tilde{\theta})\,\mathrm{MoM}(\tilde{\theta}, \tilde{Z})\, m(\theta|\tilde{\theta})}{v(\theta)\,\mathrm{MoM}(\theta, Z)\, m(\tilde{\theta}|\theta)}, 1 \right\}$$

  if $\log U \leq v$ **then**
    return $\tilde{\theta}, \tilde{Z}$
  else
    return θ, Z
  end if
end procedure

$$\begin{aligned}
&\pi(\theta, Z)K(\theta, Z, d\tilde{\theta}, d\tilde{Z}) \\
=&\frac{v(\theta)\mathbb{M}^{\theta}(dZ)\,\mathrm{MoM}(\theta, Z)}{p(y)} M(\theta, d\tilde{\theta})\mathbb{M}^{\tilde{\theta}}(d\tilde{Z}) \frac{v(\tilde{\theta})\,\mathrm{MoM}(\tilde{\theta}, \tilde{Z})\, m(\theta|\tilde{\theta})}{v(\theta)\,\mathrm{MoM}(\theta, Z)\, m(\tilde{\theta}|\theta)} \\
=&\frac{v(\tilde{\theta})\mathbb{M}^{\tilde{\theta}}(d\tilde{Z})\,\mathrm{MoM}(\tilde{\theta}, \tilde{Z})}{p(y)} M(\tilde{\theta}, d\theta)\mathbb{M}^{\theta}(dZ) \times 1.
\end{aligned}$$

□

Given the invariant distribution

$$\pi_{MoM}(\theta, Z) = v(\theta)\mathbb{M}^{\theta}(dZ)\frac{MoM(\theta, Z)}{p(y)},$$

the marginal distribution for θ is

$$\pi_{MoM}(\theta) = \frac{v(\theta)\mathbb{E}[MoM(\theta, \mathbf{Z})|\theta]}{p(y)}.$$

This is not the posterior of θ because, in general, the median is biased, i.e., $\mathbb{E}[MoM(\theta, \mathbf{Z})|\theta] \neq p(y|\theta)$.

Now, consider estimating some quantity, e.g., $\mathbb{E}[\phi(\Theta)]$, where $\phi(\Theta)$ is some function of Θ and $\Theta \sim p(\Theta|y)$. We perform importance sampling with the proposal

distribution $\pi_{MoM}(\theta, Z)$ and target distribution $\hat{\pi}(\theta, Z) = v(\theta)\mathbb{M}^{\theta}(Z)\frac{L(\theta,Z)}{p(y)}$. This gives an estimator

$$\hat{\phi} = \sum_{i=1}^{M} \frac{L(\theta_i, Z_i)}{MoM(\theta_i, Z_i)}\phi(\theta_i).$$

We now check that it is an unbiased estimator:

$$\mathbb{E}[\hat{\phi}] = \int\int L(\theta, Z)v(d\theta)\mathbb{M}^{\theta}(dZ) = \mathbb{E}[\phi(\Theta)|y].$$

4 MoM on Particle filter

References

[1] On the Variability of Case-Deletion Importance Sampling Weights in the Bayesian Linear Model on JSTOR. (n.d.). www.jstor.org. https://www.jstor.org/stable/2291464

[2] Vehtari, A., Simpson, D., Gelman, A., Yao, Y., Gabry, J. (2015, July 9). Pareto smoothed importance sampling. arXiv.org. https://arxiv.org/abs/1507.02646v9

[3] Flegal, J. M., Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. The Annals of Statistics, 38(2). https://doi.org/10.1214/09-aos735