Interacting Particle Markov Chain Monte Carlo - Supplementary Material

Tom Rainforth[1]* TWGR@ROBOTS.OX.AC.UK
Christian A. Naesseth[2]* CHRISTIAN.A.NAESSETH@LIU.SE
Fredrik Lindsten[3] FREDRIK.LINDSTEN@IT.UU.SE
Brooks Paige[1] BROOKS@ROBOTS.OX.AC.UK
Jan-Willem van de Meent[1] JWVDM@ROBOTS.OX.AC.UK
Arnaud Doucet[1] DOUCET@STATS.OX.AC.UK
Frank Wood[1] FWOOD@ROBOTS.OX.AC.UK

∗ equal contribution
[1] The University of Oxford, Oxford, United Kingdom
[2] Linköping University, Linköping, Sweden
[3] Uppsala University, Uppsala, Sweden

Abstract

This is the supplementary material to the main manuscript *Interacting Particle Markov Chain Monte Carlo*. We provide additional details for using all the particles and present further results on the experiments and choosing number of conditional nodes or workers.

1. Using All Particles

The Monte Carlo estimator is given by

$$\mathbb{E}[f(\mathbf{x})] \approx \frac{1}{RP}\sum_{r=1}^{R}\sum_{j=1}^{P} f(\mathbf{x}'_j[r]) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{P}\sum_{j=1}^{P} f(\mathbf{x}'_j[r]), \tag{1}$$

where we can note that $\mathbf{x}'_j[r] = \mathbf{x}_{c_j}^{b_j}$ from the internal particle system at iteration r. We can however make use of all particles to estimate expectations of interest by, for each MCMC iteration r, averaging over the sampled conditional node indices $c_{1:P}$ and corresponding particle indices $b_{1:P}$. This procedure is referred to as a Rao-Blackwellization of a statistical estimator and is (in terms of variance) never worse than the original one, and often much better. For iteration r we need to calculate the following

$$\frac{1}{P}\sum_{j=1}^{P} f(\mathbf{x}'_j[r]) = \frac{1}{P}\sum_{j=1}^{P} f(\mathbf{x}_{c_j}^{b_j}),$$

Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

where we can Rao-Blackwellize the selection of the retained particle along with each individual Gibbs update as following

$$
\begin{aligned}
\frac{1}{P}\sum_{j=1}^{P}\mathbb{E}_{c_j,b_j|\xi_{1:M},c_{1:P\setminus j}}\left[f(\mathbf{x}_{c_j}^{b_j})\right] =& \frac{1}{P}\sum_{j=1}^{P}\mathbb{E}_{c_j|\xi_{1:M},c_{1:P\setminus j}}\left[\sum_{i=1}^{N}\bar{w}_{T,c_j}^{i}f(\mathbf{x}_{c_j}^{i})\right] \\
=& \frac{1}{P}\sum_{j=1}^{P}\sum_{i=1}^{N}\mathbb{E}_{c_j|\xi_{1:M},c_{1:P\setminus j}}\left[\bar{w}_{T,c_j}^{i}f(\mathbf{x}_{c_j}^{i})\right] \\
=& \frac{1}{P}\sum_{j=1}^{P}\sum_{i=1}^{N}\sum_{m=1}^{M}\hat{\zeta}_m^j\bar{w}_{T,m}^{i}f(\mathbf{x}_m^i) \\
=& \frac{1}{P}\sum_{j=1}^{P}\sum_{m=1}^{M}\hat{\zeta}_m^j\sum_{i=1}^{N}\bar{w}_{T,m}^{i}f(\mathbf{x}_m^i) \\
=& \frac{1}{P}\sum_{m=1}^{M}\left[\left(\sum_{j=1}^{P}\hat{\zeta}_m^j\right)\cdot\left(\sum_{i=1}^{N}\bar{w}_{T,m}^{i}f(\mathbf{x}_m^i)\right)\right]
\end{aligned}
$$

where we have made use of the knowledge that the internal particle system $\{(\mathbf{x}_m^i, \bar{w}_{T,m}^i)\}$ does not change between Gibbs updates of the c_j's, whereas the $\hat{\zeta}_m^j$ do. We emphasise that this is a separate Rao-Blackwellization of each Gibbs update of the conditional node indices, such that each is conditioned upon the actual update made at $j-1$, rather than a simultaneous Rao-Blackwellization of the full batch of P updates. Though the latter also has analytic form and should theoretically be lower variance, it suffers from inherent numerical instability and so is difficult to calculate in practise. We found that empirically there was not a noticeable difference between the performance of the two procedures. Furthermore, one can always run additional Gibbs updates on the c_j's and obtain an improve estimate on the relative sample weightings if desired.

2. Choosing P

For the purposes of this study we assume, without loss of generality, that the indices for the conditional nodes are always $c_{1:P} = \{1, \dots, P\}$. Then we can show that the probability of the event that at least one conditional nodes switches with an unconditional is given by

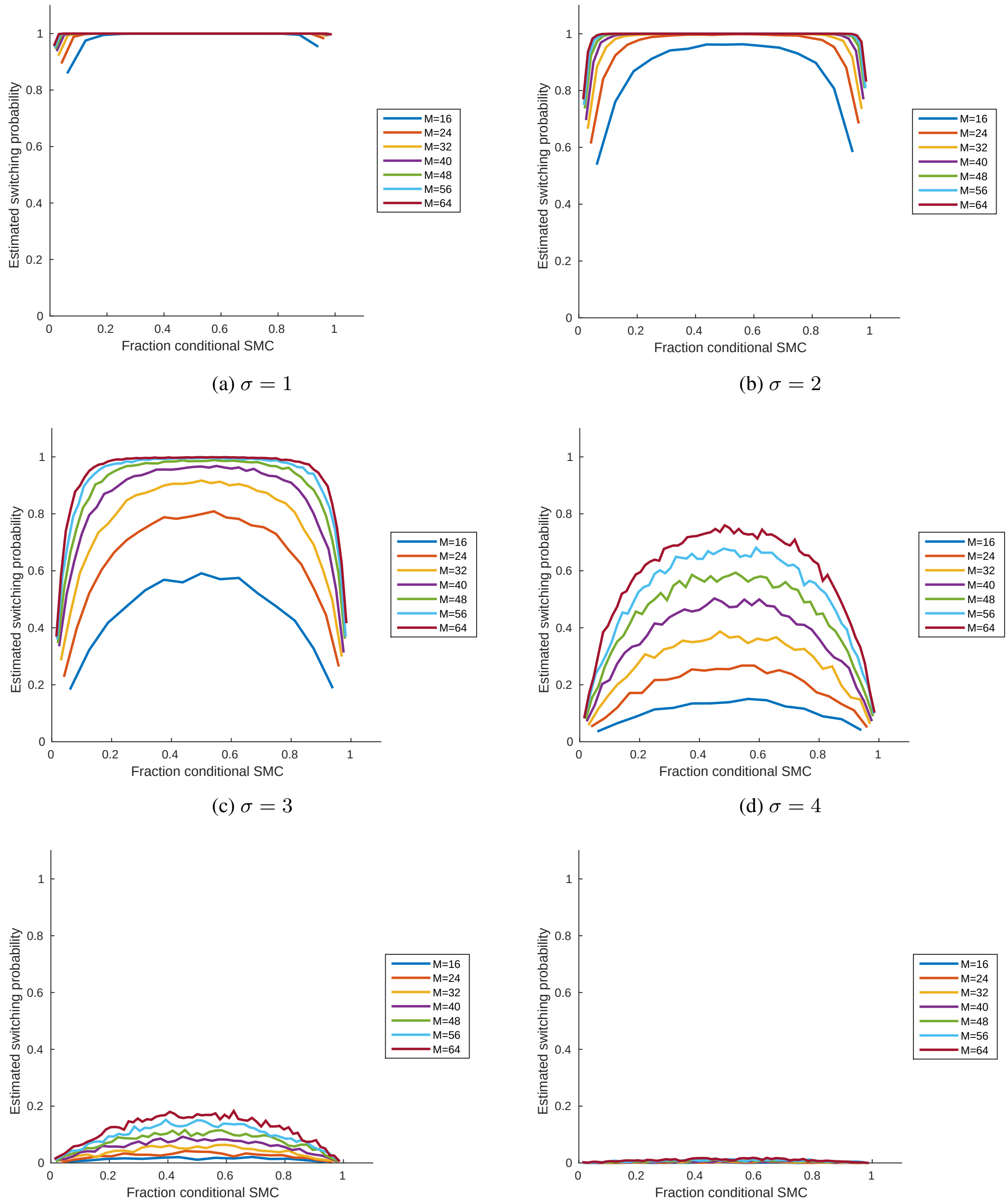
$$
\mathbb{P}(\{\text{switch}\}) = 1 - \mathbb{E}\left[\prod_{j=1}^{P}\frac{\hat{Z}_j}{\hat{Z}_j + \sum_{m=P+1}^{M}\hat{Z}_m}\right]. \tag{2}
$$

Now, there are some asymptotic (and experimental) results (Pitt et al., 2012; Bérard et al., 2014; Doucet et al., 2015) that indicate that a decent approximation for the distribution of the log of the normalisation constant estimates is Gaussian. This would mean the distributions of the conditional and unconditional normalisation constant estimates with variance σ^2 can be well-approximated as follows

$$
\log\left(\frac{\hat{Z}_j}{Z}\right) \sim \mathcal{N}(\frac{\sigma^2}{2}, \sigma^2), \quad j = 1, \dots, P, \tag{3}
$$

$$
\log\left(\frac{\hat{Z}_m}{Z}\right) \sim \mathcal{N}(-\frac{\sigma^2}{2}, \sigma^2), \quad m = P+1, \dots, M. \tag{4}
$$

A straight-forward Monte Carlo estimation of the switching probability, i.e. $\mathbb{P}(\{\text{switch}\})$, can be seen in Figure 1 for various settings of σ and M. These results seem to indicate that letting $P \approx M/2$ maximises the probability of switching.



(a) $\sigma = 1$

(b) $\sigma = 2$

(c) $\sigma = 3$

(d) $\sigma = 4$

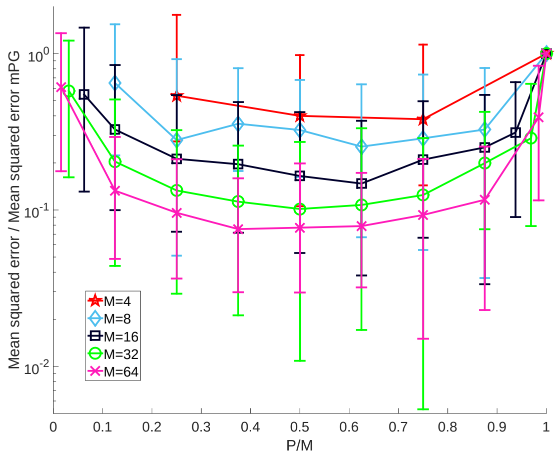
(e) $\sigma = 5$

(f) $\sigma = 6$

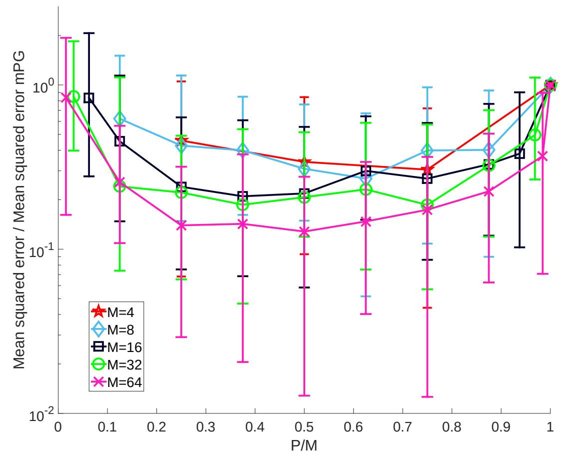
Figure 1. Estimation of switching probability for various settings of σ and M.

3. Additional Results Figures

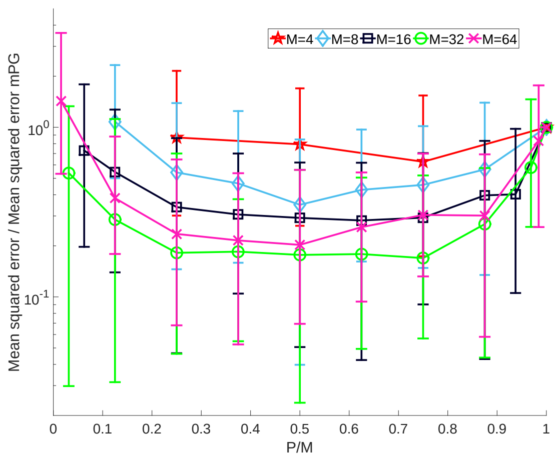
In this section we provide figures to support those in the main paper.



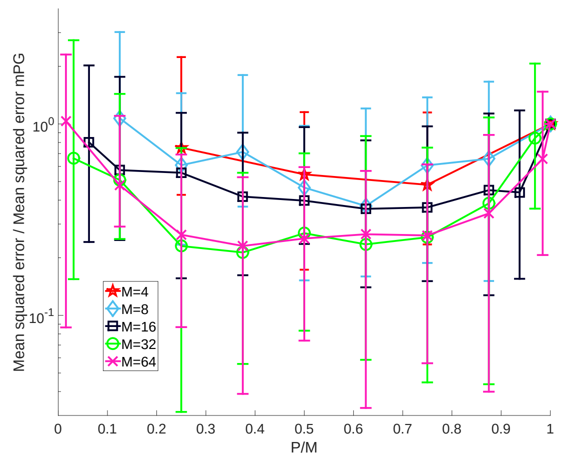
(a) Mean



(b) Standard Deviation

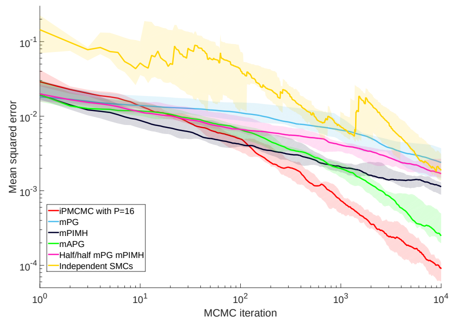


(c) Skewness

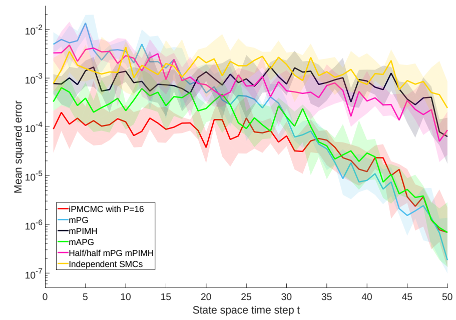


(d) Kurtosis

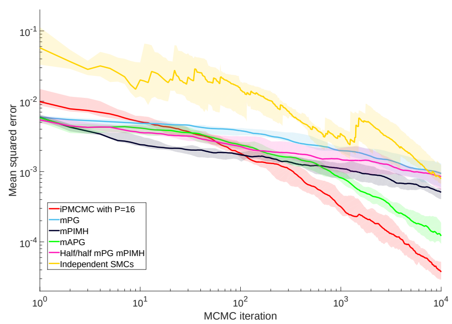
Figure 2. Median error in marginal moment estimates with different choices of P and M over 10 different synthetic datasets of the linear Gaussian state space model given in (10) after 1000 MCMC iterations. Errors are normalized by the error of a multi-start PG sampler which is a special case of iPMCMC for which $P = M$ (see Section 4). Error bars show the lower and upper quartiles for the errors. It can be seen that for all the moments then $P/M \approx 1/2$ give the best performance. For the mean and standard deviation estimates, the accuracy relative to the non-interacting distribution case $P = M$ shows a clear increase with M. This effect is also seen for the skewness and excess kurtosis estimates except for the distinction between the $M = 32$ and $M = 64$ cases. This may be because these metric are the same for the prior and the posterior such that good results for these metric might be achievable even when the samples give a poor match to the true posterior.



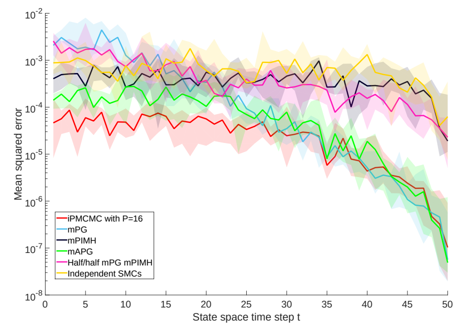
(a) Convergence in mean



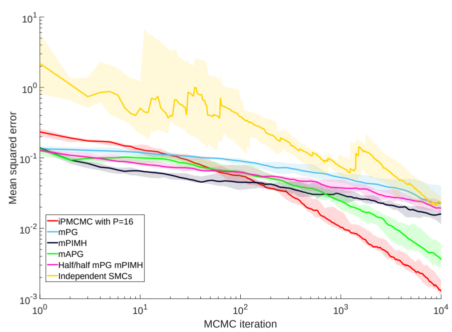
(b) Final error in mean



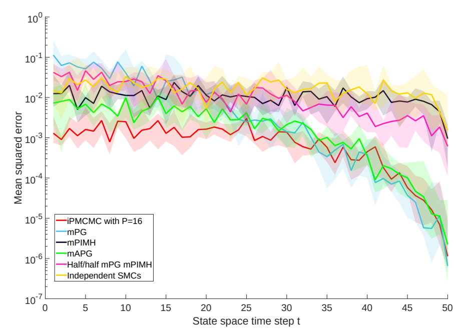
(c) Convergence in standard deviation



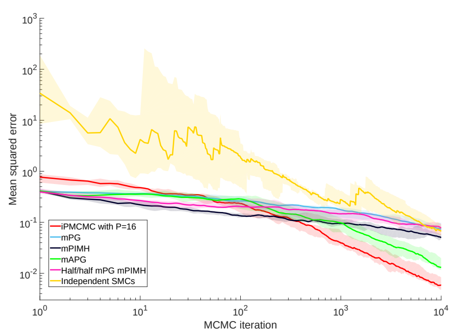
(d) Final error in standard deviation



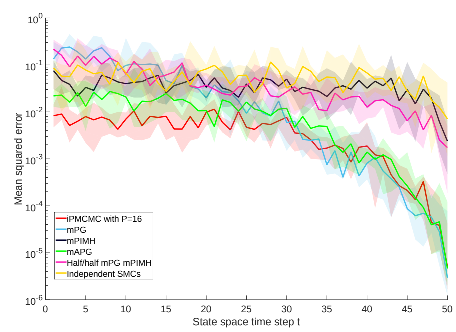
(e) Convergence in skewness



(f) Final error in skewness

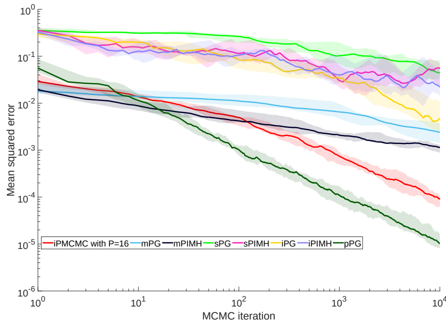


(g) Convergence in kurtosis

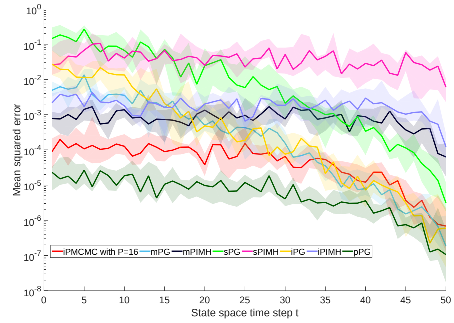


(h) Final error in kurtosis

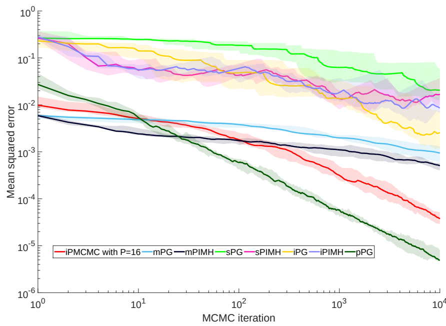
Figure 3. Mean squared error in latent variable mean, standard deviation, skewness and kurtosis averaged over all dimensions of the LGSSM as a function of MCMC iteration (left) and position in the state sequence (right) for a selection of paraellelizable SMC and PMCMC methods. See figure 3 in main paper for more details.



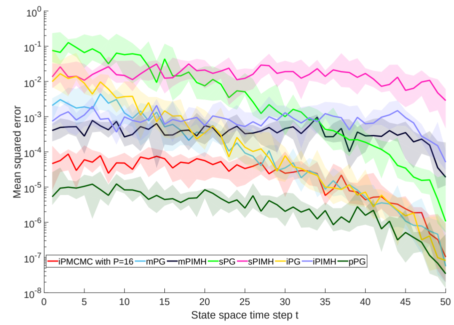
(a) Convergence in mean



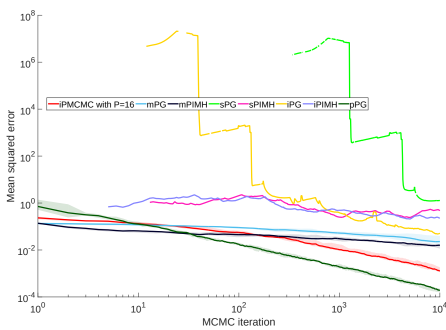
(b) Final error in mean



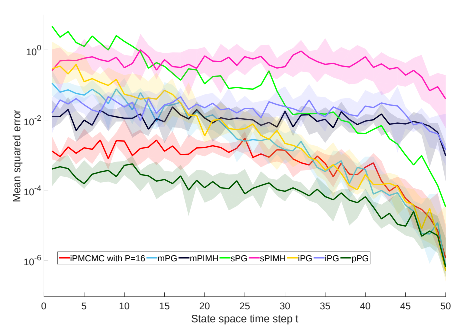
(c) Convergence in standard deviation



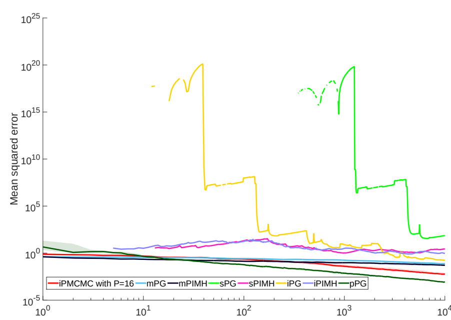
(d) Final error in standard deviation



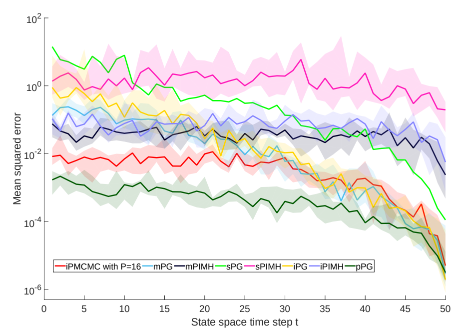
(e) Convergence in skewness



(f) Final error in skewness

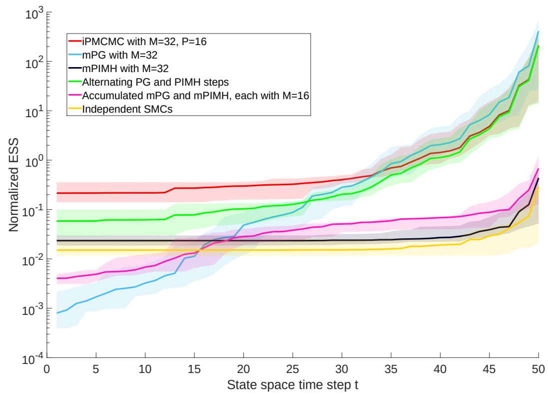


(g) Convergence in kurtosis

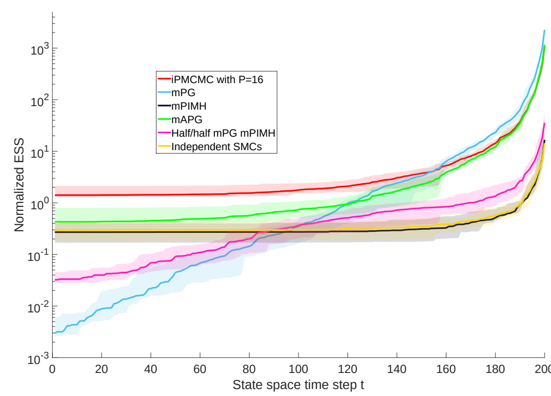


(h) Final error in kurtosis

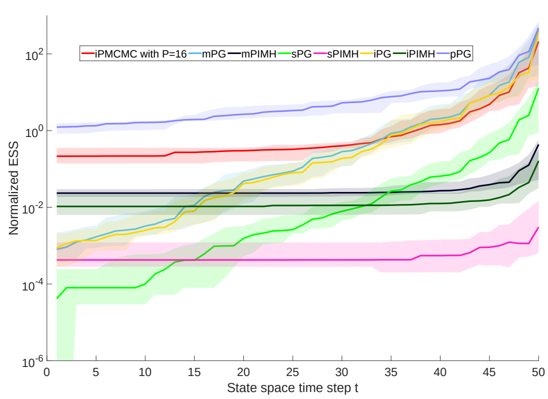
Figure 4. Mean squared error in latent variable mean, standard deviation, skewness and kurtosis averaged of all dimensions of the LGSSM as a function of MCMC iteration (left) and position in the state sequence (right) for iPMCMC, mPG, mPIMH and a number of serialized variants. Key for legends: sPG = single PG chain, sPIMH = single PIMH chain, iPG = single PG chain run 32 times longer, iPIMH = single PIMH chain run 32 times longer and pPG = single PG with 32 times more particles. For visualization purposes, the chains with extra iterations have had the number of MCMC iterations normalized by 32 so that the different methods represent equivalent total computational budget.



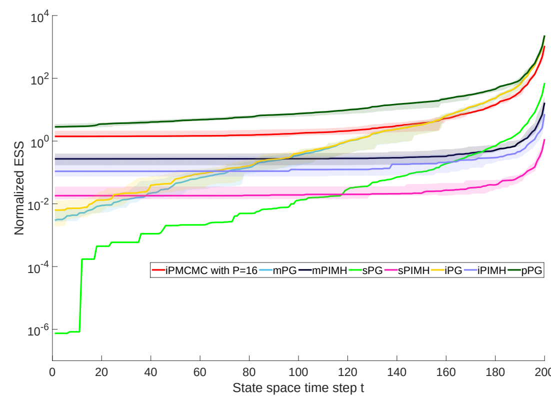
(a) ESS of distributed methods for LGSSM



(b) ESS of distributed methods for NLSSM



(c) ESS comparison to series equivalents for LGSSM



(d) ESS comparison to series equivalents for NLSSM

Figure 5. Normalized effective sample size for LGSSM (left) and NLSSM (right) for a number of distributed and series models.

References

Bérard, Jean, Del Moral, Pierre, and Doucet, Arnaud. A lognormal central limit theorem for particle approximations of normalizing constants. *Electronic Journal of Probability*, 19(94):1–28, 2014.

Doucet, Arnaud, Pitt, Michael, Deligiannidis, George, and Kohn, Robert. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, pp. asu075, 2015.

Pitt, Michael K, dos Santos Silva, Ralph, Giordani, Paolo, and Kohn, Robert. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.