BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Particle Learning for Sequential Bayesian Computation – Rejoinder

We would like to enthusiastically thank the discussants Mike Pitt, Christian Robert's multinational team, Paul Fearnhead and Dan Merl for their contributions. Hopefully our comments will make PL's scope, strengths and weaknesses clear, particularly to those readers interested in sequential parameter Bayesian computation. We would like to organize our comments into the following topics: approximating predictive densities, outliers and model misspecification, sufficient statistics, MC error accumulation, PL and MCMC and resampling. Pitt's discussion is mainly focused on PL for dynamic models (Carvalho *et al.*, 2010, Lopes and Tsay, 2010). Similarly, several of Robert *et al.* discussion are based on the mixture of Poisson distributions from Carvalho *et al.* (2009). Therefore, some readers might benefit from browsing through those papers before engaging in our comments. Fearnhead's and Merl's discussion are solely based on our chapter.

1. Approximating predictive densities. Iacobucci, Robert, Marin and Mergensen and Iacobucci, Marin and Robert suggest alternative approximations, still based on PL samples, to the predictive density. This is clearly a good idea. Examples A and B below proved some simulation evidence: PL (based on the product estimate) and MCMC (based on Chib's method) produce relatively similar results either for small or large samples. Chib's method – as it uses extra "analytical" information – might outperform the product estimator[1] in some scenarios with the well-known caveat of its potential high variability (See Example A). Neal (1999) and Polson (2007) point out that Chib's method and variants thereof can have poor MC properties which are exacerbated when MCMC convergence is prohibitively slow. The product estimate is naturally sequential and easy to implement but is potentially biased. Appealing to alternatives that exploit functional forms and/or the conditional structure of the model such as Savage-Dickey density ratio estimates or Rao-Blackwellized estimates, amongst others, is clearly preferable when available.

2. Outliers and model misspecification. One well-known fact is that all particle methods breakdown when there are outliers (misspecified models). A known drawback, also shared by all alternative particle filters, is the accumulation of MC error (for example, in the presence of sequences of outliers). We show in Carvalho *et*

[1] In a simple example of the application of Chib's method, also known as the candidate estimator (Besag, 1989), the predictive $p(y)$ is approximated by $\hat{p}(y) = p(y|\tilde{\theta})p(\tilde{\theta})/p(\tilde{\theta}_1|\tilde{\theta}_2, y)\hat{p}(\tilde{\theta}_2|y)$, where $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)$ is any value of θ, say the posterior mode or the posterior mean, and $\hat{p}(\tilde{\theta}_2|y)$ is a Monte Carlo approximation to $p(\tilde{\theta}_2|y)$, say $N^{-1}\sum_i p(\tilde{\theta}_2|\theta_{1i}, y)$, where $\theta_{11}, \ldots, \theta_{1N}$ are draws from $p(\theta_1|y)$.

al. (2010) that PL has better properties than alternative filters in the presence of outliers. Example C clearly shows in an extreme situation that even $N = 100,000$ particles will not overcome a large outlier – even though PL vastly outperforms standard alternatives.

3. Sufficient statistics. One area where we strongly disagree with Chopin, Robert and colleagues is our use of the essential state vector, Z_t. Our view is that this is key to sequential parameter Bayesian computation as it converts the sequential learning problem to a filtering problem for Z_t, i.e. find $p(Z_t|y^t)$ for $1 \leq t \leq T$. Without this extra structure, we feel that black-box sequential importance sampling algorithms and related central limit theorems are of little use in practice.

It appears that one source of confusion is that the calculation of the marginal filtering distribution $p(Z_T|y^T)$ is alligned with the full posterior smoothing problem, $p(x_1, \ldots, x_T|y^T)$. Clearly, if one solves the smoothing problem (a T dimensional joint posterior), the distribution of Z_T follows as a marginal. The converse is clearly not true – one might be able to accurately estimate the functional $p(Z_T|y^T)$ whilst having no idea about the full joint. For example, from the forward filtering PL algorithm $p(Z_1|Y^T)$ will have collapsed on one particle. We note that Carvalho *et al.* (2010) also provide a smoothing algorithm with parameter learning – extending Godsill, Doucet and West (2004) – but this is $O(N^2)$ (see discussions by Pitt and Fearnhead).

Pitt, Chopin and Robert, Chopin and Schäfer, Robert, Ryder and Chopin and Fearnhead all comment on the potential particle degeneracy of the parameter sufficient statistics. Our view is that you have to separate the concepts of degeneracy and accumulation of MC error. Now we will provide two standard examples (including the local level model of Chopin) illustrating how PL does in fact accurately learns Z_T. In example D, PL is implemented with conditional parameter sufficient statistics for a large sample size $n = 5000$ and same order of magnitude particle size $N = 1000$. Despite the very simplistic nature of the example, PL and MCMC produce fairly similar approximations. We carefully revisit the first order dynamic linear model discussed by Chopin and Schäfer in example E. It appears then that for PL to "degenerate" as the discussants suggest the time series length n will have to be many orders of magnitude larger than N. Robert *et al.* seem intent on using $N = 1000$ particles in 5000 dimensional problems and showing poor Monte Carlo performance – there really shouldn't be surprised at all with some of their findings. Addressing real problems and showing when large Monte Carlo samples are need is clearly an area for future research much in the same way that the MCMC convergence literature evolved.

One of the main criticisms running through the discussions, as well as the literature (e.g., Kantas *et al.*, 2009), is that the parameter estimation problem with sufficient statistics is equivalent to learning additive functionals of the states of the form $s_{n+1} = s_n + \phi(x_{n+1}) = \sum_{t=1}^{n} \phi(x_t, x_{t-1})$. The line of argument continues that well known limiting results, such as those in Olsson *et al.* (2008), indicate that the variance of the Monte Carlo estimates of $E(s_n|\theta, y^n)$ increases quadratically with time, since it involves approximating $p(x^n|y^n)$, the smoothing distribution. Thus, PL inherently 'degenerate', in the sense that the Monte Carlo variance will 'blow-up', and thus is unreliable. This argument appears repeatedly in the literature.

This argument is incorrect and extremely misleading for two reasons. First, what appears in the posteriors that we sample from, $p(\theta|s_n)$, are not terms like $s(x^n) = \sum_{t=1}^{n} \phi(x_t, x_{t-1})$, but rather *time-averaged* terms like $\widetilde{s}(x^n) = \sum_{t=1}^{n} \phi(x_t, x_{t-1})/n$.

This point was mentioned in the discussion by Chopin and Schäfer and, in our view, is crucial. For example, think about learning the mean α in the local level model: $y_t|x_t \sim N(\alpha + x_t, \sigma^2)$ and $x_t|x_{t-1} \sim N(x_{t-1}, \tau^2)$. Here, the posterior for α will depend on $\sum_{t=1}^n (y_t - x_t)/n = \sum_{t=1}^n y_t/n - \sum_{t=1}^n x_t/n$, and the first term is observed. More generally, the terms that appear in the posteriors are $\sum_{t=1}^n x_t/n$, $\sum_{t=1}^n x_t^2/n$, and $\sum_{t=1}^n x_t x_{t-1}/n$, all of which are time averaged.

Second, time-averaging matters. Targets like $\sum_{t=1}^n \phi(x_t, x_{t-1})/n$ do not grow for large n, at least in stationary models. Because of that, they are easier to estimate than a moving target because, for example, its variance does not increase with time (in population). Potentially, it is even easier than estimating $E(x_n|y^n)$. This can actually be seen from figures 2 and 3 in Olsson *et al.* (2008). They show the Monte Carlo error in estimating $s_2(x^n) = \sum_{t=1}^n x_t^2/n$, holding the number of particles fixed at $N = 1000$ (a very small number). It is obvious that the Monte Carlo variance decreases over time. For the local level model, we repeat these calculations in example E. Again, it is obvious the Monte Carlo variance associated with estimating $s_n = \sum_{t=1}^n x_t/n$ decreases with n (even though this model is non-stationary). See figures (c) and (d) of example E. This holds more generally, and we have verified this for a range of models and sufficient statistics. We could imagine if the model were strongly non-stationary, that time-averaging might not mitigate the error accumulation. Our conjecture is that the Monte Carlo variance decreases provided the errors in estimating the current state do not increase too rapidly. This seems to hold in common specifications.

PL parameter particles do not degenerate (as they are drawn offline if need be). Particles in PL, per se, never degenerate – we draw exactly from the mixture approximation and resampling first avoids degeneracy problems that plagued previous parameter learning attempts. This is the main advantage of PL over previous attempts where θ is part of the particle set and, after degeneration, would have to be rejuvenated (with an MCMC step).

4. Accumulation of MC error. The more interesting problem (as with MCMC convergence checks) is how MC errors accumulate in PL. General bounds, such as those provided by Chopin, seem to be of little use. Due to the simplicity of implementation, it is quite straightforward to address this via simulation. Consider the first order dynamic linear model of Chopin with $p(y_t|x_t) \sim N(x_t, \sigma^2)$, $p(x_t|x_{t-1}) \sim N(x_{t-1}, \tau^2)$ and $p(x_0) \sim N(0, C_0)$, for known variances σ^2, τ^2 and C_0. The predictive and propagation distributions needed for PL are $p(y_{t+1}|x_t) \sim N(x_t, \sigma^2 + \tau^2)$ and $(x_{t+1}|x_t, y_{t+1}) \sim N\left(Ay_{t+1} + (1-A)x_t, A\sigma^2\right)$, respectively, where $A = \tau^2/(\tau^2 + \sigma^2)$. It is instructive to analyze the MC error at the first step and then argue by induction (see e.g. Godsill et al, 2004). Here we have $p(y_1|x_0) \sim N(x_0, \sigma^2 + \tau^2)$ and $p(y_1) \sim N(0, \sigma^2 + \tau^2 + C_0)$ and $p(x_1) \sim N(0, \tau^2 + C_0)$. There is the usual relative MC error bound to approximate the marginal distribution $p^N(y_1)$ to $p(y_1)$ (functionals $\phi(x_t)$ can be analyzed in a similar fashion). We need to compare the bounds produced by PL and SIS, i.e. compare the right hand side of $Var_{PL}\left(p^N(y_1)/p(y_1)\right) \leq (Np^2(y_1))^{-1} E_{p(x_0)}[p^2(y_1|x_0)]$ to the right hand side of $Var_{SIS}\left(p^N(y_1)/p(y_1)\right) \leq (Np^2(y_1))^{-1} E_{p(x_1)}[p^2(y_1|x_1)]$, or simply study the behavior of the ratio $E_{p(x_0)}[p^2(y_1|x_0)]/E_{p(x_1)}[p^2(y_1|x_1)]$. Example F shows that, in this context, PL bounds are always smaller than SIS bounds. The only situation where PL and SIS behave similarly is when τ^2 is small relative to σ^2 and, simultaneously, C_0 is large, i.e. when the state evolution process informs very little about the observation evolution process and ones current information about where the state is

moving to is rather vague.

5. PL versus MCMC. MCMC methods have proven to be very effective in a large number of highly complex and structured frameworks, some of which studied by us in our papers and books. Our claim, mistakenly interpreted as dismissive of MCMC in the discussion by Mergensen, Iacobucci and Robert, is that PL is an attractive alternative to MCMC schemes in certain classes of models and, more importantly, MCMC is inherently non-sequential. As Pitt, one of the proponents of the APF, properly says, "the approach can clearly be used for a wide variety of existing models estimated currently by MCMC." The literature we cite in the paper include several serious applications of PL to situations other than the illustrative and pedagogical ones we decided to include. One particular example is the PL implementation for general mixture models in Carvalho *et al.* (2009).

6. Resampling schemes. Pitt, Fearnhead and Merl all suggested stratified sampling over naïve multinomial sampling. Clearly this has advantages. We support and magnify their advise and suggest that more clever resampling schemes, normalized by their computational cost, should be the norm, not the exception. This has shown to be drastically important particularly when using (partially) blind particle filters, such as the sequential importance sampling with resampling filter.

Recommendations.

- (G_0, G): MCMC schemes depend upon the not so trivial task of assessing convergence. How long should the burn-in G_0 be? (Polson, 1996). Besides, MCMC schemes produce G dependent draws.

- (T, N): PL schemes, as well as all particle filters, have to increase the number of particles N with the sample size T. Monte Carlo error is usually of the form $C_T/\sqrt{N}$, with $1/\sqrt{N}$ representing the particle filter's main strength and C_T its main weakness.

- *Propagation-resampling* schemes, such as the bootstrap filter and SIS filters, are generally outperformed by *resampling-propagation* schemes, such as AP filters and PL schemes.

- What seems, at first glance, to be a drawback of PL, i.e. the existence of several different essential vectors Z_ts for any single problem, is in fact PL's comparative advantage. The clever investigation of which essential vector to choose in a given situation can potentially lead to realistically more efficient PL schemes.

REFERENCES

Besag, J. (1989). A Candidate's formula: a curious result in Bayesian prediction. *Biometrika* **78**, 183–183.

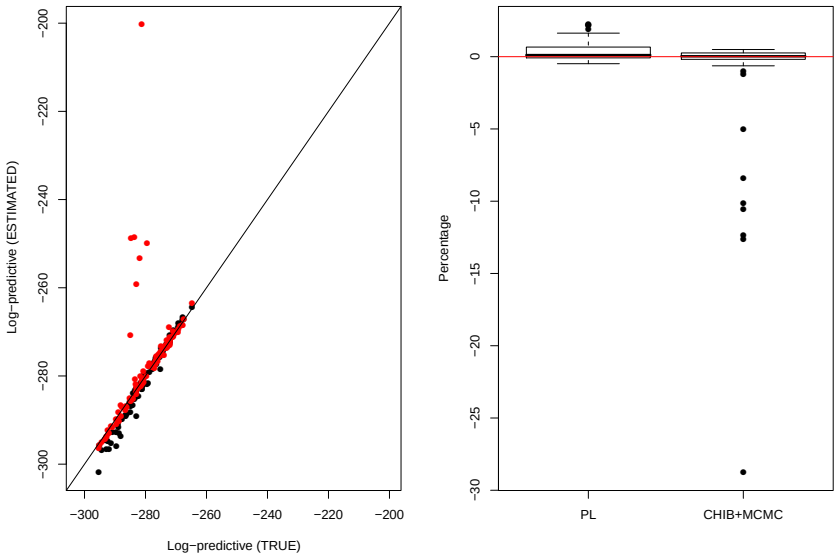
Kantas, N., Doucet, A., Singh, S. S. and Maciejowski, J. M. (2009) An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. In 15th IFAC Symposium on System Identification, SYSID 2009, 6-8 July 2009 , Saint-Malo, France.

Neal, R. M. (1999). Erroneous results in "Marginal likelihood from the Gibbs output", *Unpublished letter*. `http://www.cs.toronto.edu/ radford/ftp/chib-letter.pdf`. Department of Statistics, University of Toronto.

Olsson, J., Cappé, O., Douc, R. and Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models *Bernoulli* **14**, 155–179.

Polson, N. G. (1996). Convergence of Markov Chain Monte Carlo Algorithms. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 297–321.

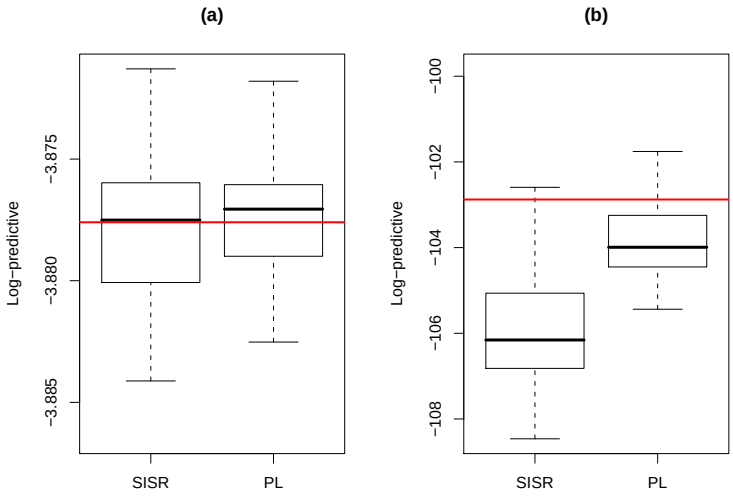
Polson, N. G. (2007). Discussion of "Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity". *Bayesian Statistics 8* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press, 401–403.



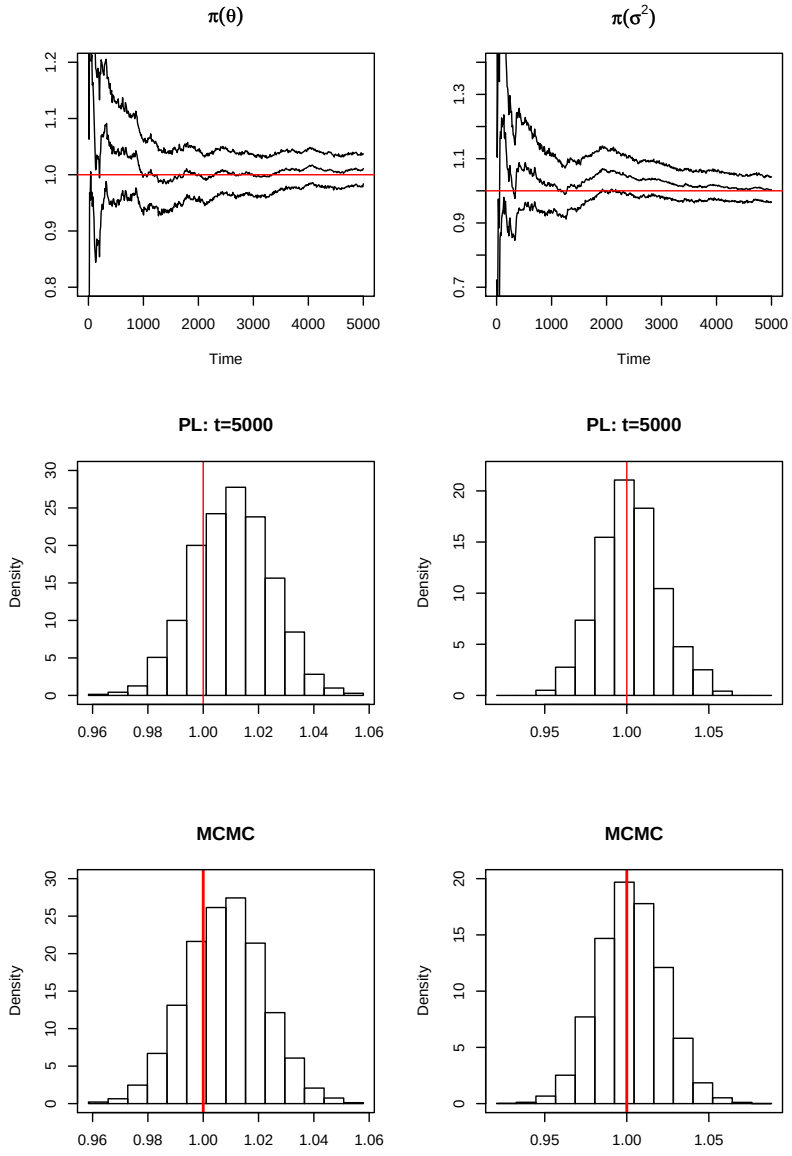
Example A: PL versus Chib's+MCMC. Comparison of PL and Chib's + MCMC when approximating $p(y)$, the predictive likelihood of a two-component mixture of Poisson distributions. For $t = 1, \ldots, n$, from $y_t \sim \alpha Poi(\gamma_1) + (1-\alpha)Poi(\gamma_2)$, where $n = 100$, $(\gamma_1, \gamma_2) = (10, 15)$ and $\alpha = 0.75$. Model is fit via PL and MCMC with prior $p(\gamma_1, \gamma_2, \alpha) = p_G(\gamma_1; 1, 0.1)\ p_G(\gamma_2; 1.5, 0.1)$, for $\gamma_1, \gamma_2 > 0$ and $\alpha \in (0, 1)$. The particle size for PL is $N = 1000$, while MCMC is run for 2000 iterations with the 2nd half kept for inference. Both MC schemes are run for each one of $S = 100$ data sets. PL seems slightly more robust than MCMC when $n = 100$, where MCMC percentage error can be as big as 30%. MCMC dominates PL when $n = 1000$, however the percentage error is below 1%.

	Mean absolute deviation					
n	20	40	60	80	100	200
PL	3.222	1.750	0.980	0.752	0.774	0.276
Chib's+PL	3.311	1.782	1.019	0.765	0.769	0.279

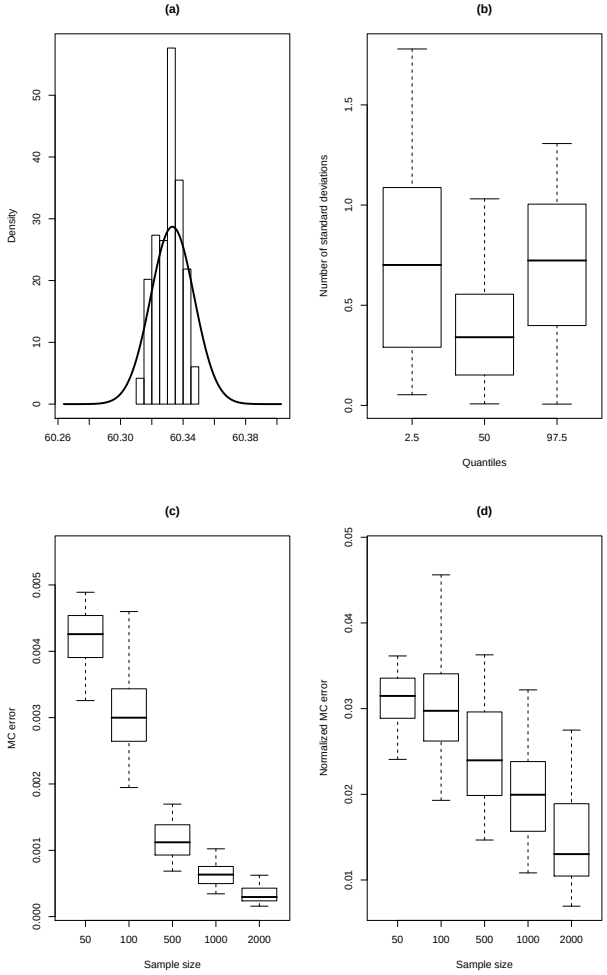
Example B: PL versus Chib's+PL. In this example we show that PL and Chib's PL produce comparable results for samples of size up to $n = 200$, which we consider large for the complexity of the model. We simulate $S = 50$ samples with n i.i.d. $N(0, 1)$ observations. The sample size n varies in $\{20, \ldots, 100, 200\}$, leading to 500 samples. For each sample we fit the simple normal model with conjugate prior for the mean and variance parameters, i.e. $y_t \sim N(\theta, \sigma^2)$ $(t = 1, \ldots, n)$, $\theta|\sigma^2 \sim N(0, \sigma^2)$ and $\sigma^2 \sim IG(10, 9)$. In this case the exact value of $p(y)$ is easily obtained since the marginal distribution of y is $t_{20}(0_n, 1.8I_n)$. We run $R = 50$ times PL, each time based on $N = 500$ particles, i.e. the same order of magnitude of the sample size. PL does not take advantage of prior conjugacy, so that during propagation θs is propagated based on resample σ^2s, which is then used to propagate σ^2s. By doing that we show that the *essential* state vector depends on both σ^2 (when propagating θ) and θ (when propagating σ^2). For any given sample size n, we compute the mean absolute error (in percentage) as $MAE(n) = \frac{100}{SR} \sum_{s=1}^{S} |\sum_{r=1}^{R} \log p_{pl}^{r}(y_s) / \log p(y_s) - R|$, where $\log p_{pl}^{r}(y_s)$ is r^{th} PL approximation to $p(y_s)$ and y_s is the s^{th} sample of size n. PL is slightly better than Chib's+PL.



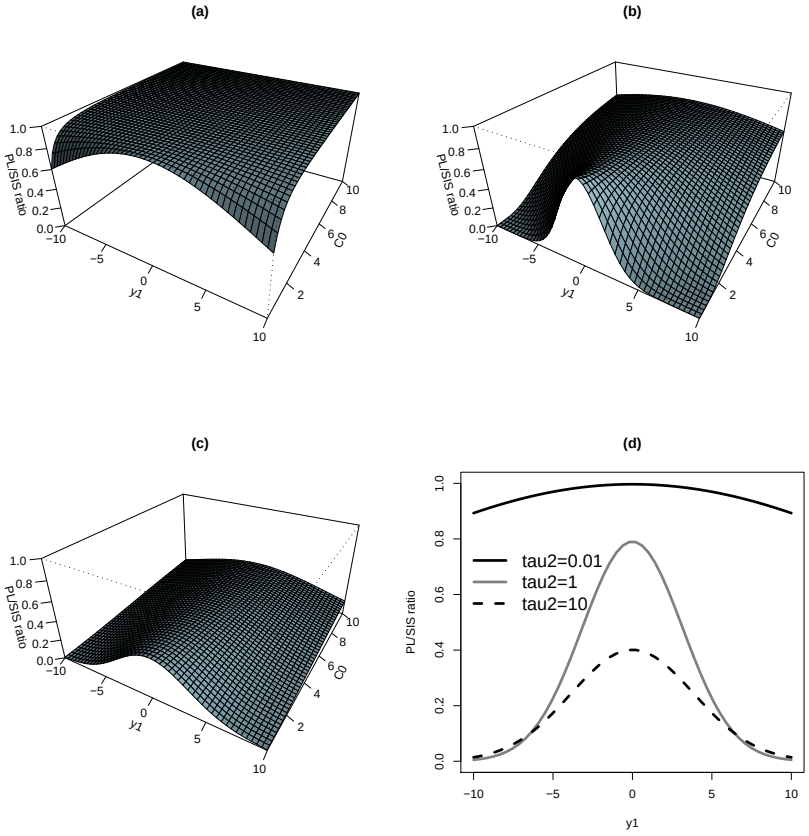
Example C: PL versus SISR. Let consider the basic local level model, i.e. $y_t|x_t \sim N(x_t, 1)$ and $x_t|x_{t-1} \sim N(x_{t-1}, 1)$, for $t = 1, \ldots, n$ and $x_0 \sim N(0, C_0)$. The MC study shows that PL has smaller MC error than SISR when approximating $\log p(y_1, y_2)$ in the presence of an outlier in the observation equation when $C_0 = 1$, $n = 2$, $y_2 = 0$ and $y_1 = 2$ (panel (a)) or $y_1 = 20$ (panel (b)).



Example D: PL versus MCMC. We simulate $n = 5000$ data points from $y_t \sim N(1, 1)$, and fit the model $y_t \sim N(\theta, \sigma^2)$ and $(\theta, \sigma^2) \sim N(m_0, C_0)IG(a_0, b_0)$, where $m_0 = 0$, $C_0 = 10$, $a_0 = 3$ and $b_0 = 2$ (relatively vague prior information). MCMC is a Gibbs sampler with full conditionals $\theta|\sigma^2, y \sim N(m_n, C_n)$ and $\sigma^2|\theta, y \sim IG(a_n, b_n)$, for $C_n = 1/(1/C_0 + n/\sigma^2)$, $m_n = C_n(m_0/C_0 + n\bar{y}/\sigma^2)$, $a_n = a_0 + n/2$ and $b_n = b_0 + \sum_{t=1}^{n}(y_t - \theta)^2/2$. The Gibbs sampler started at $\sigma^{2(0)} = 1.0$ and was run for 20,000 draws discarding the first half. PL runs from $t = 1$ to $t = n$ as follows: 1) Let $\{(m_{t-1}, C_{t-1}, a_{t-1}, b_{t-1}, \sigma^2)^{(i)}\}_{i=1}^{N}$ be the particle set at time $t - 1$, with $s_1 = 1/C_{t-1}$ and $s_2 = m_{t-1}/C_{t-1}$; 2) resample the set with weights $w_t^{(i)} \propto f_N(y_t; m_{t-1}^{(i)}, C_{t-1}^{(i)} + \sigma^{2(i)})$; 3) compute $s_1^{(i)} = \tilde{s}_1^{(i)} + 1/\tilde{\sigma}^{2(i)}$, $s_2^{(i)} = \tilde{s}_2^{(i)} + y_t/\tilde{\sigma}^{2(i)}$, $a_t = a_{t-1} + 1/2$, $C_t^{(i)} = 1/s_1^{(i)}$ and $m_t^{(i)} = C_t^{(i)} s_2^{(i)}$; 4) draw $\theta^{(i)} \sim N(m_t^{(i)}, C_t^{(i)})$; 5) compute $b_t^{(i)} = \tilde{b}_{t-1}^{(i)} + (y_t - \theta^{(i)})^2/2$; and 6) draw $\sigma^{2(i)} \sim IG(a_t^{(i)}, b_t^{(i)})$. PL results are based on $N = 1000$ particles.



Example E: Sufficient statistics. For $t = 1, \ldots, n$, let us consider the local level model where $y_t|x_t, \sigma^2 \sim N(x_t, \sigma^2)$, $x_t|x_{t-1}, \sigma^2 \sim N(x_{t-1}, \sigma^2)$, $x_0|\sigma^2 \sim N(m_0, \sigma^2)$ and $\sigma^2 \sim IG(c_0, d_0)$. It is easy to see that the joint prior of $x = (x_1, \ldots, x_n)'$ is multivariate normal with mean $\mu_0 = 1_n m_0$ and precision $\sigma^{-2}\Phi_0$, where $\Phi_{0,ij} = 0$ for all $|i - j| > 1$, $\Phi_{0,ij} = -1$ for all $|i - j| = 1$, $\Phi_{0,ii} = 2$ for all $i = 1, \ldots, n - 1$ and $\Phi_{0,nn} = 1$. Combining this (improper) prior with the normal model for $y = (y_1, \ldots, y_n)$, $y|x, \sigma^2 \sim N(x, \sigma^2 I_n)$, leads to the joint posterior of x being normal with mean $\mu_n = \Phi_n^{-1}(\Phi_0\mu_0 + y)$ and variance $\sigma^2\Phi_n^{-1}$, for $\Phi_n = \Phi_0 + I_n$. Therefore, conditional on σ^2, the posterior distribution of $s_n = \sum_{t=1}^n x_t/n = 1_n'x/n$ is normal with mean $a_n = 1_n'\mu_n/n$ and variance $\sigma^2 b_n$, where $b_n = 1_n'\Phi_n^{-1}1_n/n^2$. It is also easy to see that $\sigma^2|y \sim IG(c_n, d_n)$ where $c_n = c_0 + n/2$ and $d_n = d_0 + (y'y + \mu_0'\Phi_0\mu_0 - \mu_n'\Phi_n\mu_n)/2$, so that $s_n|y \sim t_{2c_n}(a_n, b_n d_n/c_n)$. In addition, it is easy to see that $(\sigma^2|y^t, x^t) \sim IG(c_t, d_t)$, where $y^t = (y_1, \ldots, y_t)$, $c_t = c_{t-1} + 1$ and $d_t = d_{t-1} + (y_t - x_t)^2 + (x_t - x_{t-1})^2$. In this exercise, the sample size is $n = 5000$ and particle size $N = 10000$, for $m_0 = x_0 = 0$, $c_0 = 10$, $d_0 = 9$ and $R = 50$ runs of PL. (a) Histogram approximating $p(s_n|y)$ for one of the runs. (b) Box-plots of distances (in number of standard deviations) between approximate quantiles based on the $R = 50$ histograms and the true Student's t quantiles for $p(s_n|y)$. (c) MC error measured as the standard deviation PL's estimate of $E(s_n|y^n)$ over the $R = 50$ runs and different sample sizes. (d) Same as (c) but normalized by the true value of $\sqrt{V(s_n|y^n)}$.



Example F: PL versus SIS bounds. Surface ratio $E_{p(x_0)}[p^2(y_1|x_0)]/E_{p(x_1)}[p^2(y_1|x_1)]$ for $\sigma^2 = 1$, $\tau^2 \in \{0.01, 1, 10\}$ (panels (a) through (c), respectively), $C_0 = 1$ (panel (d)), $x_0 \sim N(0, C_0)$, $(y_1|x_1) \sim N(x_1, \sigma^2)$, $(y_1|x_0) \sim N(x_0, \sigma^2 + \tau^2)$, $y_1 \sim N(0, \sigma^2 + \tau^2 + C_0)$ and $x_1 \sim N(0, \tau^2 + C_0)$. It is easy to shown that $E_{p(x_0)}[p^2(y_1|x_0)] = [2\pi(\sigma^2 + \tau^2)(2C_0 + \sigma^2 + \tau^2)]^{-1} \exp\{-y_1^2/(2C_0 + \sigma^2 + \tau^2)\}$ and $E_{p(x_1)}[p^2(y_1|x_1)] = [2\pi\sigma^2(2C_0 + \sigma^2 + 2\tau^2)]^{-1} \exp\{-y_1^2/(2C_0 + \sigma^2 + 2\tau^2)\}$.