ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2018)
80, *Part* 5, *pp.* 839–870

The correlated pseudomarginal method

George Deligiannidis and Arnaud Doucet

University of Oxford, UK

and Michael K. Pitt

King's College London, UK

[Received July 2017. Revised May 2018]

Summary. The pseudomarginal algorithm is a Metropolis–Hastings-type scheme which samples asymptotically from a target probability density when we can only estimate unbiasedly an unnormalized version of it. In a Bayesian context, it is a state of the art posterior simulation technique when the likelihood function is intractable but can be estimated unbiasedly by using Monte Carlo samples. However, for the performance of this scheme not to degrade as the number T of data points increases, it is typically necessary for the number N of Monte Carlo samples to be proportional to T to control the relative variance of the likelihood ratio estimator appearing in the acceptance probability of this algorithm. The correlated pseudomarginal method is a modification of the pseudomarginal method using a likelihood ratio estimator computed by using two correlated likelihood estimators. For random-effects models, we show under regularity conditions that the parameters of this scheme can be selected such that the relative variance of this likelihood ratio estimator is controlled when N increases sublinearly with T and we provide guidelines on how to optimize the algorithm on the basis of a non-standard weak convergence analysis. The efficiency of computations for Bayesian inference relative to the pseudomarginal method empirically increases with T and exceeds two orders of magnitude in some examples.

Keywords: Asymptotic posterior normality; Correlated random numbers; Intractable likelihood; Metropolis–Hastings algorithm; Particle filter; Random-effects model; Weak convergence

1. Introduction

Consider a Bayesian model where the likelihood of the observations y is denoted by $p(y \mid \theta)$ and the prior for the parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ admits a density $p(\theta)$ with respect to Lebesgue measure $\mathrm{d}\theta$. Then the posterior density of interest is $\pi(\theta) \propto p(y \mid \theta)p(\theta)$. We slightly abuse the notation by using the same symbols for distributions and densities.

A standard approach to compute expectations with respect to $\pi(\theta)$ is to use the Metropolis–Hastings (MH) algorithm to generate an ergodic Markov chain of invariant density $\pi(\theta)$. Given the current state θ of the Markov chain, one samples a candidate θ' which is accepted with a probability which depends in part on the likelihood ratio $p(y \mid \theta')/p(y \mid \theta)$. For many latent variable models, the likelihood is intractable and it is thus impossible to implement the MH algorithm. In this context, Markov chain Monte Carlo schemes targeting the joint posterior distribution of the parameter and latent variables are often inefficient as the parameter and latent variables can be strongly correlated under the posterior, or cannot even be used if only forward simulation of the latent variables is feasible; see, for example, Ionides *et al.* (2006), Johndrow *et al.* (2016) and Andrieu *et al.* (2010), section 2.3, for a detailed discussion.

Address for correspondence: M. K. Pitt, Department of Mathematics, King's College London, Strand, London, WC2R 2LS, UK.
E-mail: michael.pitt@kcl.ac.uk

 1369–7412/18/80839

Contrary to these approaches, the pseudomarginal algorithm directly mimics the MH scheme targeting the marginal $\pi(\theta)$ by substituting an estimator of the likelihood ratio $p(y|\theta')/p(y|\theta)$ for the true likelihood ratio in the MH acceptance probability (Lin *et al.*, 2000; Beaumont, 2003; Andrieu and Roberts, 2009). This estimator is obtained by computing a non-negative unbiased estimator of $p(y \mid \theta')$ and dividing it by the estimator of $p(y \mid \theta)$ computed when θ was accepted. This simple yet powerful idea has become popular as it is often possible to obtain a non-negative unbiased estimator of intractable likelihoods and it provides state of the art performance in many scenarios; see, for example, Andrieu *et al.* (2010) and Flury and Shephard (2011). Qualitative convergence results for this procedure have been obtained by Andrieu and Roberts (2009) and Andrieu and Vihola (2015).

Assuming that the likelihood estimator is evaluated by using importance sampling or particle filters for state space models with N particles, it has also been shown under various assumptions by Pitt *et al.* (2012), Doucet *et al.* (2015) and Sherlock *et al.* (2015) that N should be selected such that the variance of the log-likelihood ratio estimator should take a value between 1.0 and 3.0 in regions of high probability mass to minimize the computational resources that are necessary to achieve a prespecified asymptotic variance for a particular pseudomarginal average. As the number T of data $y = (y_1, \ldots, y_T)$ increases, this implies that N should increase linearly with T (Bérard *et al.* (2014), theorem 1) and the computational cost of the pseudomarginal algorithm is thus of order T^2 at each iteration. This can be prohibitive for large data sets.

The reason for this is that the pseudomarginal algorithm is based on an estimator of $p(y \mid \theta')/p(y \mid \theta)$ that is obtained by dividing estimators of $p(y \mid \theta)$ and $p(y \mid \theta')$ which are independent given θ and θ'. However, when one is interested in estimating a ratio, using positively correlated estimators of the numerator and denominator typically provides a lower variance ratio estimator than if these estimators were independent; see, for example, Koop (1972). This is exploited by the proposed correlated pseudomarginal method which correlates these estimators by correlating the auxiliary random variates that are used to obtain them. Two implementations of this generic idea are detailed. We show how to correlate importance sampling estimators for random-effects models and particle filter estimators for state space models by using the Hilbert sort procedure that was proposed by Gerber and Chopin (2015).

We study in detail the large sample properties of the correlated pseudomarginal scheme for random-effects models. In this scenario, the log-likelihood ratio estimator based on our correlation scheme satisfies a conditional central limit theorem (CLT) whenever N grows to ∞ sublinearly with T and the Euclidean distance between θ and θ' is of order $1/\sqrt{T}$. When the posterior concentrates towards a Gaussian density of standard deviation $1/\sqrt{T}$, this CLT can be used to show that a space-rescaled version of the correlated pseudomarginal chain converges weakly to a discrete time Markov chain on the parameter space. The integrated auto-correlation time of the weak limit is not impacted by how fast N goes to ∞ with T. However, the lower this growth rate is, the more correlated the auxiliary variables need to be to control the variance of this estimator. We provide results suggesting that N needs to grow at least at rate $\sqrt{T}$ for the integrated auto-correlation time of the original correlated pseudomarginal chain to remain finite as $T \to \infty$. We use these results to provide practical guidelines on how to optimize the performance of the algorithm for large data sets which are validated experimentally. In our numerical examples on random-effects models and state space models, the correlated pseudomarginal method always outperforms the pseudomarginal method and the improvement increases with T from 20 to 50 times when T is a few hundred to more than 100 times when T is a few thousand.

The rest of the paper is organized as follows. In Section 2, we introduce the correlated pseudomarginal algorithm and detail its implementation for random-effects and state space models. In Section 3, we present various CLTs for the log-likelihood estimator and log-likelihood ratio

estimators that are used by the pseudomarginal and correlated pseudomarginal methods. In Section 4, we exploit these results to analyse and optimize the correlated pseudomarginal kernel in the large sample regime. We demonstrate experimentally the efficiency of this methodology in Section 5 and discuss various potential extensions in Section 6. All the proofs are given in the on-line supplementary material. The numerical results have been generated by using Ox version 4.0 (Doornik, 2007). The computer code to replicate the experiments is available on line from `https://github.com/mikepitt1969/correlated`.

2. Metropolis–Hastings and correlated pseudomarginal schemes

2.1. Metropolis–Hastings algorithm

The transition kernel Q_{MH} of the MH algorithm targeting $\pi(\theta)$ by using a proposal distribution $q(\theta, \mathrm{d}\theta') = q(\theta, \theta')\mathrm{d}\theta'$ is given by

$$Q_{\mathrm{MH}}(\theta, \mathrm{d}\theta') = q(\theta, \mathrm{d}\theta')\alpha_{\mathrm{MH}}(\theta, \theta') + \{1 - \varrho_{\mathrm{MH}}(\theta)\}\delta_{\theta}(\mathrm{d}\theta'), \tag{1}$$

where

$$r_{\mathrm{MH}}(\theta, \theta') = \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')} = \frac{p(y \mid \theta')\, p(\theta')q(\theta', \theta)}{p(y \mid \theta)\, p(\theta)q(\theta, \theta')}, \tag{2}$$

and

$$\begin{aligned} \alpha_{\mathrm{MH}}(\theta, \theta') &= \min\{1, r_{\mathrm{MH}}(\theta, \theta')\}, \\ \varrho_{\mathrm{MH}}(\theta) &= \int q(\theta, \mathrm{d}\theta')\alpha_{\mathrm{MH}}(\theta, \theta'). \end{aligned} \tag{3}$$

Implementing this MH scheme requires being able to evaluate the likelihood ratio $p(y \mid \theta')/p(y \mid \theta)$ appearing in the expression of $r_{\mathrm{MH}}(\theta, \theta')$. When it is not possible to evaluate this ratio exactly, this MH algorithm cannot be implemented.

2.2. The correlated pseudomarginal algorithm

Assume that $\hat{p}(y \mid \theta, U)$ is a non-negative unbiased estimator of the intractable likelihood $p(y \mid \theta)$ when $U \sim m$. Here U corresponds to the $\mathcal{U}$-valued auxiliary random variables that are used to obtain the estimator. We assume that $m(\mathrm{d}u) = m(u)\mathrm{d}u$ and introduce the joint density $\bar{\pi}(\theta, u)$ on $\Theta \times \mathcal{U}$, where

$$\bar{\pi}(\theta, u) = \pi(\theta)\, m(u)\, \hat{p}(y \mid \theta, u)/p(y|\theta). \tag{4}$$

As $\hat{p}(y \mid \theta, U)$ is unbiased, $\bar{\pi}(\theta, u)$ admits $\pi(\theta)$ as the marginal density. The correlated pseudomarginal algorithm is an MH scheme targeting (4) with proposal density $q(\theta, \mathrm{d}\theta')K(u, \mathrm{d}u')$ where K admits an m-reversible Markov transition density, i.e.

$$m(u)K(u, u') = m(u')K(u', u). \tag{5}$$

This yields the acceptance probability

$$\alpha_Q\{(\theta, u), (\theta', u')\} = \min\left\{1, r_{\mathrm{MH}}(\theta, \theta')\frac{\hat{p}(y \mid \theta', u')/p(y \mid \theta')}{\hat{p}(y \mid \theta, u)/p(y \mid \theta)}\right\}. \tag{6}$$

The correlated pseudomarginal algorithm admits $\bar{\pi}(\theta, u)$ as an invariant density by construction and its transition kernel Q is given by

$$Q\{(\theta, u), (\mathrm{d}\theta', \mathrm{d}u')\} = q(\theta, \mathrm{d}\theta')K(u, \mathrm{d}u')\alpha_Q\{(\theta, u), (\theta', u')\} + \{1 - \varrho_Q(\theta, u)\}\delta_{(\theta, u)}(\mathrm{d}\theta', \mathrm{d}u'), \tag{7}$$

Table 1. Algorithm 1: correlated pseudomarginal algorithm

1, sample $\theta' \sim q(\theta, \cdot)$
2, sample $\varepsilon \sim \mathcal{N}(\mathbf{0}_M, I_M)$ and set $U' = \rho U + \sqrt{(1-\rho^2)}\varepsilon$
3, compute the estimator $\hat{p}(y \mid \theta', U')$ of $p(y \mid \theta')$
4, with probability

$$\alpha_Q\{(\theta, U), (\theta', U')\} = \min\left\{1, \frac{\hat{p}(y \mid \theta', U')}{\hat{p}(y \mid \theta, U)} \frac{p(\theta')}{p(\theta)} \frac{q(\theta', \theta)}{q(\theta, \theta')}\right\}$$

output (θ', U'); otherwise, output (θ, U)

where $1 - \varrho_Q(\theta, u)$ is the corresponding rejection probability. For $K(u, u') = m(u')$, we recover the pseudomarginal scheme. Data-informed proposals such as the preconditioned Crank-Nicolson Langevin proposal of Cotter *et al.* (2013) and its extensions proposed by Titsias and Papaspiliopoulos (2018) could also be used to update the auxiliary random variates at the cost of more complex acceptance probabilities.

Let $\varphi(z; \mu, \Sigma)$ be the multivariate normal density of argument z, mean μ and covariance matrix Σ and let $X \sim \mathcal{N}(\mu, \Sigma)$ denote a sample from this distribution. Henceforth, we focus on the case where the likelihood estimator is computed by using $M \geqslant 1$ standard normal random variables and the corresponding Crank-Nicolson proposal (Cotter *et al.*, 2013) is used. Hence we have

$$\begin{aligned} m(u) &= \varphi(u; \mathbf{0}_M, I_M), \\ K_\rho(u, u') &= \varphi\{u'; \rho u, (1-\rho^2) I_M\}, \end{aligned} \tag{8}$$

where $\rho \in (-1, 1)$, $\mathbf{0}_M$ is the $M \times 1$ vector with 0-entries and I_M the $M \times M$ identity matrix. It is straightforward to check that K_ρ is m reversible. There is no loss of generality to select m as a normal density since inversion techniques can be used to form any random variable of interest. (For example, in Section 2.3.2, it is necessary to generate uniform random variates and these may be constructed as $\Phi(u_i)$ where u_i is a scalar element of u and Φ the cumulative distribution function of the standard normal distribution.)

The selection of m as a normal distribution and K_ρ as a proposal is advantageous because K_ρ can be interpreted as a discretized Ornstein–Uhlenbeck process. This is key in establishing the main theoretical result of Section 3 whose proof is simplified by the use of Itô's lemma and Stein's lemma. This allows us to provide useful guidelines on how to optimize the parameters of the correlated pseudomarginal. Moreover, K_ρ is cheap to simulate from and admits a single interpretable parameter.

Algorithm 1 in Table 1 summarizes how to simulate from $Q\{(\theta, U), \cdot\}$. Contrary to the pseudomarginal method corresponding to $\rho = 0$, we need to store the vector u instead of $\hat{p}(y \mid \theta, u)$ to implement the algorithm when $\rho \neq 0$. In the applications that are considered, this overhead is mild.

The rationale behind the correlated pseudomarginal scheme is that if $(\theta, u) \mapsto \hat{p}(y \mid \theta, u)$ is a sufficiently regular function and (θ, U) and (θ', U') are sufficiently 'close' then we expect the ratio estimator $\hat{p}(y \mid \theta', U')/\hat{p}(y \mid \theta, U)$ to have small relative variance and therefore to mimic the 'exact' MH scheme Q_{MH} better. In many situations, the posterior $\pi(\theta)$ will be approximately normal for large data sets with covariance scaling like $1/\sqrt{T}$, so an appropriately scaled MH random walk or auto-regressive proposal $q(\theta, \mathrm{d}\theta')$ will ensure that θ and θ' are close. We explain in Section 3 how ρ can be selected as a function of T to ensure that U and U' are sufficiently close that the log-likelihood ratio estimator $\log\{\hat{p}(y \mid \theta', U')/\hat{p}(y \mid \theta, U)\}$ satisfies a conditional

CLT at stationarity. As explained in Section 1, properties of this estimator and in particular its asymptotic distribution and variance at stationarity are critical to our analysis of the correlated pseudomarginal scheme in the large sample regime that is detailed in Section 4.

2.3. Application to latent variable models

2.3.1. Random-effects models

Consider the model

$$X_t \overset{\text{IID}}{\sim} f_\theta(\cdot), \qquad Y_t | X_t \sim g_\theta(\cdot | X_t), \tag{9}$$

where $\{X_t; t \geqslant 1\}$ are $\mathbb{R}^k$-valued latent variables, $\{Y_t; t \geqslant 1\}$ are $\mathbf{Y}$-valued observations, $\mathbf{Y}$ being a topological space, and $f_\theta(\cdot)$ and $g_\theta(\cdot|x)$ are densities with respect to reference Borel measures. For any $i < j$, let $i:j = \{i, i+1, \ldots, j\}$. For a realization $Y_{1:T} = y_{1:T}$, the likelihood satisfies

$$p(y_{1:T}|\theta) = \prod_{t=1}^{T} p(y_t|\theta), \tag{10}$$

with

$$p(y_t|\theta) = \int g_\theta(y_t|x_t) f_\theta(x_t) \, \mathrm{d}x_t. \tag{11}$$

If the T integrals appearing in expression (10) are intractable, we can estimate them by using importance sampling to obtain the following unbiased likelihood estimator

$$\hat{p}(y_{1:T} \mid \theta, U) = \prod_{t=1}^{T} \left\{ \frac{1}{N} \sum_{i=1}^{N} \omega(y_t, X_{t,i}; \theta) \right\}, \tag{12}$$

where the importance weight $\omega(y, U_{t,i}; \theta)$ is given by

$$\omega(y_t, U_{t,i}; \theta) = \frac{g_\theta(y_t \mid X_{t,i}) \, f_\theta(X_{t,i})}{q_\theta(X_{t,i} \mid y_t)}, \tag{13}$$

assuming that there is a deterministic map $\Xi_t : \mathbb{R}^p \times \Theta \to \mathbb{R}^k$ such that $X_{t,i} = \Xi_t(U_{t,i}; \theta) \sim q_\theta(\cdot|y_t)$ for $U_{t,i} \sim \mathcal{N}(\mathbf{0}_p, I_p)$. Let U be the column vector consisting of all the components of $U_{t,i}$ for $t \in 1:T$ and $i \in 1:N$. It is clear that $U \sim \mathcal{N}(\mathbf{0}_M, I_M)$ where $M = TNp$.

2.3.2. State space models

Consider a generalization of model (9)–(10) where the latent variables $\{X_t; t \geqslant 1\}$ now arise from a homogeneous $\mathbb{R}^k$-valued Markov process of initial density ν_θ and Markov transition density f_θ with respect to Lebesgue measure, i.e., for $t \geqslant 1$,

$$X_1 \sim \nu_\theta, \qquad X_{t+1}|X_t \sim f_\theta(\cdot|X_t), \qquad Y_t|X_t \sim g_\theta(\cdot|X_t). \tag{14}$$

For a realization $Y_{1:T} = y_{1:T}$, the likelihood satisfies the predictive decomposition

$$p(y_{1:T} \mid \theta) = p(y_1 \mid \theta) \prod_{t=2}^{T} p(y_t \mid y_{1:t-1}, \theta), \tag{15}$$

with

$$p(y_t \mid y_{1:t-1}, \theta) = \int g_\theta(y_t \mid x_t) p_\theta(x_t \mid y_{1:t-1}) \, \mathrm{d}x_t, \tag{16}$$

where $p_\theta(x_1 \mid y_{1:0}) = \nu_\theta(x_1)$ and $p_\theta(x_t \mid y_{1:t-1})$ denotes the posterior density of X_t given $Y_{1:t-1} =$

$y_{1:t-1}$ for $t \geqslant 2$. Importance sampling estimators of the likelihood have relative variance typically increasing exponentially with T so the likelihood is usually estimated by using particle filters instead.

Particle filters propagate N random samples, termed particles, over time by using a sequence of resampling steps and importance sampling steps using the importance densities $q_\theta(x_1 \mid y_1)$ at time 1 and $q_\theta(x_t \mid y_t, x_{t-1})$ at times $t \geqslant 2$. Let $\Xi_1 : \mathbb{R}^p \times \Theta \to \mathbb{R}^k$ and $\Xi_t : \mathbb{R}^k \times \mathbb{R}^p \times \Theta \to \mathbb{R}^k$ for $t \geqslant 2$ be deterministic maps such that $X_1 = \Xi_1(V; \theta) \sim q_\theta(\cdot \mid y_1)$ and $X_t = \Xi_t(x_{t-1}, V; \theta) \sim q_\theta(\cdot \mid y_t, x_{t-1})$ for $t \geqslant 2$ if $V \sim \mathcal{N}(\mathbf{0}_p, I_p)$. We also propose to use normal random variables to obtain the uniform random variables that are necessary to sample the categorical distributions appearing in the resampling steps. By using these representations, we obtain an unbiased estimator $\hat{p}(y_t \mid \theta, U)$ of $p(y_t \mid \theta)$ where U follows a multivariate normal distribution (Del Moral, 2004). When this estimator is used within a pseudomarginal scheme, the resulting algorithm is known as the particle marginal MH algorithm (Andrieu *et al.*, 2010). However, if this likelihood estimator is used in the correlated pseudomarginal context, the likelihood ratio estimator $\hat{p}(y_{1:T} \mid \theta', u') / \hat{p}(y_{1:T} \mid \theta, u)$ can significantly deviate from 1 even when (θ, u) is close to (θ', u') and the true likelihood is continuous at θ. This is because the resampling steps introduce discontinuities in the particles that are selected when θ and u are modified, even slightly (Malik and Pitt, 2011).

To reduce the variability of this likelihood ratio estimator, we use a resampling scheme based on the Hilbert sort procedure that was introduced by Gerber and Chopin (2015). This procedure is based on the Hilbert space filling curve which is a continuous fractal map $H : [0, 1] \to [0, 1]^k$ whose image is $[0, 1]^k$. It admits a pseudoinverse $h : [0, 1]^k \to [0, 1]$, i.e. $H \circ h(x) = x$ for all $x \in [0, 1]^k$. For most points x and x' that are close in $[0, 1]^k$, their images $h(x)$ and $h(x')$ tend to be close. This property can be used to build a 'sorted' resampling procedure which will ensure that when the parameter or auxiliary variables change only slightly the particles that are selected remain close. Practically, this resampling procedure proceeds as follows:

(a) the $\mathbb{R}^k$-valued particles are projected in the hypercube $[0, 1]^k$ by using a bijection $\varkappa : \mathbb{R}^k \to [0, 1]^k$;
(b) the resulting $[0, 1]^k$-valued particles are projected on $[0, 1]$ by using the pseudoinverse h;
(c) these projected $[0, 1]$-valued particles are sorted;
(d) the systematic resampling scheme proposed by Carpenter *et al.* (1999) is used on the sorted points.

Introduce the importance weights $\omega_1(x_1; \theta) = \nu_\theta(x_1)\, g_\theta(y_1 \mid x_1) / q_\theta(x_1 \mid y_1)$ and $\omega_t(x_{t-1}, x_t; \theta) = f_\theta(x_t \mid x_{t-1})\, g_\theta(y_t \mid x_t) / q_\theta(x_t \mid y_t, x_{t-1})$ for $t \geqslant 2$. The only difference between the resulting particle filter summarized by algorithm 2 in Table 2 and the algorithm of Gerber and Chopin (2015) is that we use normal random variates instead of randomized quasi-Monte-Carlo points in $[0, 1]^p$. For the mapping $\varkappa$, we adopt the logistic transform that was used in Gerber and Chopin (2015).

If we denote by U the column vector composed of the components of $(U_{1,1}, \ldots, U_{T,N}, U_1^R, \ldots, U_{T-1}^R)$, then $U \sim \mathcal{N}(\mathbf{0}_M, I_M)$ where $M = TNp + T - 1$. The corresponding unbiased likelihood estimator is given by

$$\hat{p}(y_{1:T} \mid \theta, U) = \left\{ \frac{1}{N} \sum_{i=1}^{N} \omega_1(X_{1,i}; \theta) \right\} \prod_{t=2}^{T} \left\{ \frac{1}{N} \sum_{i=1}^{N} \omega_t(X_{t-1, \sigma_{t-1}(A_{t-1,i})}, X_{t,i}; \theta) \right\}. \tag{17}$$

We can now use this estimator within the correlated pseudomarginal scheme. Many valid alternatives and generalizations of this scheme are possible as discussed in Section 6. For example, we found that introducing an additional Hilbert sort step (Table 2) after resampling can slightly improve performance without affecting the scaling properties.

Table 2. Algorithm 2: particle filter using Hilbert sort

1, sample $U_{1,i} \sim \mathcal{N}(\mathbf{0}_p, I_p)$ and set $X_{1,i} = \Xi_1(U_{1,i}; \theta)$ for $i \in 1:N$
2, for $t = 1, \ldots, T-1$:

(a) find the permutation σ_t such that $h \circ \varkappa\,(X_{t,\sigma_t(1)}) \leqslant \ldots \leqslant h \circ \varkappa\,(X_{t,\sigma_t(N)})$ if $k \geqslant 2$, or $X_{t,\sigma_t(1)} \leqslant \ldots \leqslant X_{t,\sigma_t(N)}$ if $k = 1$;
(b) sample $U_t^R \sim \mathcal{N}(0,1)$, set $\bar{U}_{t,i} = (i-1)/N + \Phi(U_t^R)/N$ for $i \in 1:N$;
(c) sample $A_{t,i} \sim F_t^{-1}(\bar{U}_{t,i})$ for $i \in 1:N$ where F_t^{-1} is the generalized inverse distribution function of the categorical distribution with weights $\{\omega_1(X_{1,\sigma_1(i)}; \theta); i \in 1:N\}$ if $t = 1$ and $\{\omega_t(X_{t-1,\sigma_{t-1}(A_{t-1,\sigma_t(i)})}, X_{t,\sigma_t(i)}; \theta); i \in 1:N\}$ for $t \geqslant 2$;
(d) sample $U_{t+1,i} \sim \mathcal{N}(\mathbf{0}_p, I_p)$ and set $X_{t+1,i} = \Xi_{t+1}(X_{t,\sigma_t(A_{t,i})}, U_{t+1,i}; \theta)$ for $i \in 1:N$

2.4. Discussion

Ideas related to the correlated pseudomarginal scheme have previously been proposed: Lee and Holmes (2010) suggested combining pseudomarginal steps with updates where only θ is updated while U is held fixed, but this scheme scales poorly with T as it still uses pseudomarginal steps. Andrieu *et al.* (2012) proposed combining pseudomarginal steps with steps where θ is held fixed and correlation between $\hat{p}(y \mid \theta, U)$ and $\hat{p}(y \mid \theta, U')$ is introduced by sampling U' by using an m-reversible Markov kernel K. However, the crucial selection of K was not discussed. It was independently proposed by Dahlin *et al.* (2015) to use the correlation scheme (8) but the guidelines for the correlation parameter ρ therein do not ensure that the variance of the log-likelihood ratio estimator is controlled as T increases. This work also relies on a standard particle filter.

As the density m of U is independent of θ, it might be argued that a Gibbs algorithm sampling alternately from the full conditional densities $\bar{\pi}(\theta|u)$ and $\bar{\pi}(u|\theta)$ of $\bar{\pi}(\theta, u)$ could mix well. Related ideas have been explored in Papaspiliopoulos *et al.* (2007). Such a Gibbs strategy is usually not implementable in the applications that are considered here. Particle Gibbs samplers have been proposed to mimic this strategy but their computational complexity is of order T^2N per iteration for state space models when using such a parameterization (Lindsten *et al.* (2014), section 6.2). Thus they are not competitive with the pseudomarginal algorithm whose cost is of order T^2 per iteration. An alternative approach for updating U given θ, which has been proposed by Murray and Graham (2016), is to use elliptical slice sampling. However, in this context, no guidelines for the selection of N have been proposed. Experimentally, this method is not competitive with an appropriately tuned correlated pseudomarginal scheme when the same value of N is used for both methods. We observed that elliptical slice sampling is attempting many moves on the ellipse which are not on the support of the slice, thus requiring multiple expensive evaluations of the simulated likelihood for each sample.

3. Asymptotics of the log-likelihood ratio estimators

To understand the quantitative properties of the correlated pseudomarginal scheme, it is key to establish the statistical properties of the likelihood ratio estimator appearing in its acceptance probability (6). For the random-effects models that were introduced in Section 2.3.1, we establish conditional CLTs for the log-likelihood estimator (12) and the corresponding log-likelihood ratio estimators used by the pseudomarginal and the correlated pseudomarginal algorithms when $N \to \infty$ and $T \to \infty$. Here N will be a deterministic function of T denoted by N_T. We show that these estimators exhibit very different behaviours, underlining the benefits of correlated pseudomarginal over pseudomarginal schemes.

Consider a sequence of random variables $\{M^T; T \geqslant 1\}$ defined on a probability space $(\Omega, \mathcal{G}, P)$ and a sequence of sub-σ-algebras $\{\mathcal{G}^T; T \geqslant 1\}$ and write $\rightarrow_{\mathrm{P}}$ to denote convergence in probability. We also write $M^T|\mathcal{G}^T \Rightarrow \lambda$ if $M \sim \lambda$ and $\mathbb{E}[f(M^T)|\mathcal{G}^T] \rightarrow_{\mathrm{P}} \mathbb{E}[f(M)]$ as $T \rightarrow \infty$ for any bounded continuous function f.

Henceforth, we shall make the assumption that $Y_t \sim^{\mathrm{IID}} \mu$ and write $\mathcal{Y}^T$ for the σ-field that is spanned by $Y_{1:T}$. When additionally $U \sim m$, we denote the associated probability measure, expectation and variance by $\mathbb{P}$, $\mathbb{E}$ and $\mathbb{V}$. As our limit theorems consider the asymptotic regime where $T \rightarrow \infty$ and $N_T \rightarrow \infty$, we should write m_T and π_T instead of m and π and similarly U^T, U_t^T and $U_{t,i}^T$ instead of U, U_t and $U_{t,i}$. The probability space is defined precisely in section A.1 of the on-line supplementary material. For notational simplicity we do not emphasize here this dependence on T but it should be kept in mind that we are dealing with triangular arrays of random variables. We can write unambiguously $\mathbb{E}[\psi(Y_1, U_{1,1}; \theta)]$ rather than $\mathbb{E}[\psi(Y_1, U_{1,1}^T; \theta)]$ as $U_{1,1}^T \sim \mathcal{N}(\mathbf{0}_p, I_p)$ under $\mathbb{P}$ for any $T \geqslant 1$.

3.1. Asymptotic distribution of the log-likelihood error

Let $\gamma(y_1; \theta)^2 = \mathbb{V}\{\varpi(y_1, U_{1,1}; \theta)\}$ be the conditional variance given $Y_1 = y_1$ and $\gamma(\theta)^2 = \mathbb{V}\{\varpi(Y_1, U_{1,1}; \theta)\} = \mathbb{E}[\gamma(Y_1; \theta)^2]$ the unconditional variance of the normalized importance weight

$$\varpi(Y_t, U_{1,1}; \theta) = \frac{\omega(Y_t, U_{1,1}; \theta)}{p(Y_t \mid \theta)}, \tag{18}$$

where $\omega(Y_t, U_{1,1}; \theta)$ is defined in equation (13).

Our first result establishes conditional CLTs for the log-likelihood error

$$Z_T(\theta) = \log\{\hat{p}(Y_{1:T} \mid \theta, U)\} - \log\{p(Y_{1:T} \mid \theta)\}, \tag{19}$$

when U arises from the proposal m or from the equilibrium distribution

$$\bar{\pi}(u|\theta) = \frac{\bar{\pi}(\theta, u)}{\pi(\theta)} = \prod_{t=1}^{T} \frac{\hat{p}(Y_t \mid \theta, u_t)}{p(Y_t \mid \theta)} \varphi(u_t; \mathbf{0}_{pN_T}, I_{pN_T}), \tag{20}$$

with $\bar{\pi}(\theta, u)$ as defined in equation (4).

Theorem 1. Let $N_T = \lceil \beta T^{\alpha} \rceil$ with $\frac{1}{3} < \alpha \leqslant 1$, $\beta > 0$ and $Y_t \sim^{\mathrm{IID}} \mu$.

(a) If $\mathbb{E}[\varpi(Y, U_{1,1}; \theta)^8] < \infty$ and $U \sim m$ then

$$T^{(\alpha-1)/2} Z_T(\theta) + \tfrac{1}{2} T^{(1-\alpha)/2} \beta^{-1} \gamma(\theta)^2 | \mathcal{Y}^T \Rightarrow \mathcal{N}\{0, \beta^{-1}\gamma(\theta)^2\}. \tag{21}$$

(b) If $\mathbb{E}[\varpi(Y_1, U_{1,1}; \theta)^9] + \mathbb{E}[\gamma(Y_1; \theta)^4] < \infty$ and $U \sim \bar{\pi}(\cdot|\theta)$ then

$$T^{(\alpha-1)/2} Z_T(\theta) - \tfrac{1}{2} T^{(1-\alpha)/2} \beta^{-1} \gamma(\theta)^2 | \mathcal{Y}^T \Rightarrow \mathcal{N}\{0, \beta^{-1}\gamma(\theta)^2\}. \tag{22}$$

Remark 1. To establish results (21) and (22), for $\frac{1}{2} < \alpha \leqslant 1$, the conditions $\mathbb{E}[\varpi(Y_1, U_{1,1}; \theta)^4] < \infty$ and $\mathbb{E}[\varpi(Y_1, U_{1,1}; \theta)^5] < \infty$ respectively are sufficient.

For particle filters, a CLT for $Z_T(\theta)$ of the form (21) has already been established for the case $\alpha = 1$ in Bérard *et al.* (2014), when using multinomial resampling under strong mixing assumptions. We conjecture that both result (21) and result (22) hold under weaker assumptions

for $\frac{1}{3} < \alpha < 1$ and the Hilbert sort resampling scheme. However, it is very technically challenging to establish this result. In the simpler scenario where one uses systematic resampling, such a CLT has not yet been established. Some of the technical problems which arise when attempting to carry out such an analysis are detailed in Gentil and Rémillard (2008).

Result (21) suggests that, for large T under the proposal, $Z_T(\theta)$ is approximately normal with mean $-\beta^{-1}T^{1-\alpha}\gamma(\theta)^2/2$ and variance $\beta^{-1}T^{1-\alpha}\gamma(\theta)^2$. Result (22) suggests that at equilibrium $Z_T(\theta)$ is approximately normal with the same variance but opposite mean.

3.2. Asymptotic distribution of the log-likelihood ratio error

Assume that we are at state (θ, U) and propose (θ', U') using $\theta' \sim q(\theta, \cdot)$ and $U' \sim m$ as in the pseudomarginal algorithm or $\theta' \sim q(\theta, \cdot)$ and $U' \sim K_\rho(U, \cdot)$ as in the correlated pseudomarginal algorithm. In both cases, the acceptance ratio (6) depends on the log-likelihood ratio error

$$R_T(\theta, \theta') = \log\left\{\frac{\hat{p}(Y_{1:T} \mid \theta', U')}{\hat{p}(Y_{1:T} \mid \theta, U)}\right\} - \log\left\{\frac{p(Y_{1:T} \mid \theta')}{p(Y_{1:T} \mid \theta)}\right\}. \tag{23}$$

We examine here the limiting distribution of $R_T(\theta, \theta + \xi/\sqrt{T})$ for fixed θ and ξ, the rationale being that the posterior typically concentrates at rate $1/\sqrt{T}$ when T increases. Thus a correctly scaled random-walk proposal for an MH algorithm will be of the form $\theta' = \theta + \xi/\sqrt{T}$ where the distribution of ξ is independent of T.

For the pseudomarginal algorithm, we have the following conditional CLT.

Theorem 2. Let θ and ξ be fixed. Assume that $\vartheta \mapsto \varpi(y_1, u_{1,1}; \vartheta)$ and $\vartheta \mapsto \mathbb{E}[\varpi(Y_1, U_{1,1}; \vartheta)^9]$ are continuous at $\vartheta = \theta$ for any $(y_1, u_{1,1}) \in \mathbf{Y} \times \mathbb{R}^p$, $\vartheta \mapsto \gamma(\vartheta)$ is continuously differentiable at $\vartheta = \theta$ and $\mathbb{E}[\varpi(Y_1, U_{1,1}; \vartheta)^9] + \mathbb{E}[\gamma(Y_1; \theta)^4] < \infty$. For $N_T = \lceil \beta T^\alpha \rceil$ with $\frac{1}{3} < \alpha \leqslant 1$, $\beta > 0$, $Y_t \sim^{\mathrm{IID}} \mu$, $U \sim \bar{\pi}(\cdot|\theta)$ and $U' \sim m$ where U and U' are independent, we have

$$T^{(\alpha-1)/2} R_T(\theta, \theta + \xi/\sqrt{T}) + T^{(1-\alpha)/2}\beta^{-1}\gamma(\theta)^2 \mid \mathcal{Y}^T \Rightarrow \mathcal{N}\{0, 2\beta^{-1}\gamma(\theta)^2\}. \tag{24}$$

This result shows that the log-likelihood ratio error in the pseudomarginal case can have only a limiting variance of order 1 if N_T is proportional to T. The log-likelihood ratio estimator that is used by the correlated pseudomarginal exhibits a markedly different behaviour if we consider the Crank-Nicolson proposal (8), $U' \sim K_{\rho_T}(U, \cdot)$, with

$$\rho_T = \exp\left(-\psi \frac{N_T}{T}\right), \tag{25}$$

for some $\psi > 0$. Denote by $\mathcal{F}^T$ the σ-field that is spanned by $\{Y_t; t \in 1:T\}$ and $\{U_{t,i}; t \in 1:T, i \in 1:N\}$. We also denote the Euclidean norm by $\|\cdot\|$ and write $\nabla_u f = (\partial_{u^1} f, \ldots, \partial_{u^p} f)'$ for a real-valued function $f : \mathbb{R}^p \to \mathbb{R}$ where $u = (u^1, \ldots, u^p)$.

Theorem 3. Let θ and ξ be fixed. Let $Y_t \sim^{\mathrm{IID}} \mu$, $U \sim \bar{\pi}(\cdot|\theta)$ and $U' \sim K_{\rho_T}(U, \cdot)$ where ρ_T is given by equation (25). Under assumptions 1–6 in section A.5 of the on-line supplementary material, if $N_T \to \infty$ as $T \to \infty$ with $N_T/T \to 0$, we have

$$R_T(\theta, \theta + \xi/\sqrt{T}) \big| \mathcal{F}^T \Rightarrow \mathcal{N}\{-\kappa(\theta)^2/2, \kappa(\theta)^2\}, \tag{26}$$

where

$$\kappa(\theta)^2 = 2\psi\, \mathbb{E}\{\|\nabla_u \varpi(Y_1, U_{1,1}; \theta)\|^2\}. \tag{27}$$

Assumptions 1–6 in the supplementary material are differentiability and integrability assumptions on $\varpi(y, u; \theta)$ with respect to y, u and θ. This result states that the limiting variance of the log-likelihood ratio for the correlated pseudomarginal scheme at equilibrium is of order 1 when N_T grows sublinearly with T, although it will typically grow exponentially with p, the dimension of $U_{1,1}$. Moreover, the distribution of the log-likelihood ratio error is asymptotically independent of U, suggesting that the correlated pseudomarginal chain is less prone to sticking than the pseudomarginal chain at stationarity.

This conditional CLT has not been established for particle filters. For univariate state space models, i.e. $k = 1$, we have observed experimentally on various stationary state space models that a similar conditional CLT appears to hold. For multivariate state space models, the CLT appears to hold only conditionally on $\mathcal{Y}^T$ when N_T grows at least at rate $T^{k/(k+1)}$; see Section 5.

4. Analysis and optimization

4.1. Weak convergence in the large sample regime

The use of weak convergence techniques to analyse and optimize Markov chain Monte Carlo schemes was pioneered by Roberts *et al.* (1997) and has found numerous applications ever since; see, for example Sherlock *et al.* (2015) for a recent application to the pseudomarginal method. The high level idea behind this approach is to identify an appropriate asymptotic regime under which a component of the original Markov chain, rescaled appropriately, converges to a limiting process which is simpler to analyse and optimize. To the best of our knowledge, all previous contributions have considered the asymptotic regime where $d \to \infty$, d being the parameter dimension, while T is fixed. In these scenarios, under time rescaling, the limiting Markov process is usually a diffusion. We analyse here the correlated pseudomarginal scheme under the standard large sample regime of asymptotic statistics where d is fixed and $T \to \infty$. In this context, after space rescaling, the parameter component of the correlated pseudomarginal chain, targeting the posterior $\pi_T(\theta)$ that is associated with the observations $Y_{1:T}$, converges towards a discrete time Markov chain. Our analysis assumes that the statistical model is sufficiently regular to ensure that $\{\pi_T(\theta); T \geqslant 1\}$ can be approximated by normal densities which concentrate. Here $\pi_T(\theta)$ is interpreted as the density of a $\mathcal{Y}^T$-measurable random probability measure; see, for example, Berti *et al.* (2006) and Crauel (2003) for a formal definition. We write $\to_{\mathbb{P}^Y}$ to denote convergence in probability with respect to the law of $\{Y_t; t \geqslant 1\}$.

Assumption 1. There is a $d \times d$ positive definite matrix $\bar{\Sigma}$, a parameter value $\bar{\theta} \in \mathbb{R}^d$ and an $\mathbb{R}^d$-valued random sequence $\{\hat{\theta}_T; T \geqslant 1\}$, $\hat{\theta}_T$ being $\mathcal{Y}^T$ measurable, such that as $T \to \infty$

$$\int |\pi_T(\theta) - \varphi(\theta; \hat{\theta}_T, \bar{\Sigma}/T)| \, \mathrm{d}\theta \to_{\mathbb{P}^Y} 0, \qquad \hat{\theta}_T \to_{\mathbb{P}^Y} \bar{\theta}.$$

This assumption will be satisfied if a Berstein–von Mises theorem holds; see van der Vaart (2000), section 10.2, for sufficient conditions.

Consider the stationary correlated pseudomarginal chain $\{(\vartheta_n^T, \mathbf{U}_n^T); n \geqslant 0\}$ with proposal $q_T(\theta, \theta')$ targeting the random measure $\bar{\pi}_T(\mathrm{d}\theta, \mathrm{d}u) = \pi_T(\mathrm{d}\theta)\bar{\pi}_T(\mathrm{d}u|\theta)$ associated with the observations $Y_{1:T}$. By rescaling the parameter component of the correlated pseudomarginal chain using $\tilde{\vartheta}_n^T := \sqrt{T}(\vartheta_n^T - \hat{\theta}_T)$, we obtain the stationary Markov chain $\{(\tilde{\vartheta}_n^T, \mathbf{U}_n^T); n \geqslant 0\}$ with initial distribution $(\tilde{\vartheta}_0^T, \mathbf{U}_0^T) \sim \tilde{\pi}_T$ where

$$\left.\begin{aligned} \tilde{\pi}_T(\tilde{\theta}, u) &= \tilde{\pi}_T(\tilde{\theta})\tilde{\pi}_T(u|\tilde{\theta}), \\ \tilde{\pi}_T(\tilde{\theta}) &= \pi_T(\hat{\theta}_T + \tilde{\theta}/\sqrt{T})/\sqrt{T}, \\ \tilde{\pi}_T(u|\tilde{\theta}) &= \bar{\pi}_T(u|\hat{\theta}_T + \tilde{\theta}/\sqrt{T}), \end{aligned}\right\} \tag{28}$$

and the associated proposal density for the parameter becomes

$$\tilde{q}_T(\tilde{\theta}, \tilde{\theta}') = \frac{q_T(\hat{\theta}_T + \tilde{\theta}/\sqrt{T}, \hat{\theta}_T + \tilde{\theta}'/\sqrt{T})}{\sqrt{T}}. \tag{29}$$

We shall assume here that we use a random-walk proposal that is scaled appropriately.

Assumption 2. The proposal density is of the form

$$q_T(\theta, \theta') = \sqrt{T}\upsilon\{\sqrt{T}(\theta' - \theta)\}, \tag{30}$$

where υ is a probability density on $\mathbb{R}^d$, i.e. $\theta' \sim q_T(\theta, \cdot)$ when $\theta' = \theta + \xi/\sqrt{T}$ with $\xi \sim \upsilon$.

Finally, we assume that a uniform version of the CLT of theorem 3 holds in a neighbourhood of $\bar{\theta}$, where $\bar{\theta}$ is specified in assumption 1. We denote by $d_{\mathrm{BL}}(\mu, \nu)$ the bounded Lipschitz metric between two probability measures μ and ν; see, for example, van der Vaart (2000), page 332, or section A.9 of the on-line supplementary material.

Assumption 3. There is a neighbourhood $N(\bar{\theta})$ of $\bar{\theta}$ such that the log-likelihood ratio error that is considered in theorem 3 with $\xi \sim \upsilon(\cdot)$ satisfies as $T \to \infty$

$$\sup_{\theta \in N(\bar{\theta})} \mathbb{E}(d_{\mathrm{BL}}[\mathrm{law}\{R_T(\theta, \theta + \xi/\sqrt{T}) | \mathcal{F}^T\}, \mathcal{N}\{-\kappa(\theta)^2/2, \kappa(\theta)^2\}] | \mathcal{Y}^T) \to_{\mathbb{P}^Y} 0.$$

In assumption 3, the expectation is with respect to Y_t, U and U' distributed as in theorem 3. For the random-effects model of Section 2.3.1, we prove that assumption 3 holds under regularity conditions that are given in section A.6 of the on-line supplementary material.

Under assumption 2, the proposal that is defined in equation (29) satisfies $\tilde{q}_T(\tilde{\theta}, \tilde{\theta}') = \upsilon(\tilde{\theta}' - \tilde{\theta}) := \tilde{q}(\tilde{\theta}, \tilde{\theta}')$. In this case, the corresponding transition kernel of the rescaled correlated pseudomarginal chain is given by

$$Q_T\{(\tilde{\theta}, u), (\mathrm{d}\tilde{\theta}', \mathrm{d}u')\} = \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') K_{\rho_T}(u, \mathrm{d}u') \alpha_{Q_T}\{(\tilde{\theta}, u), (\tilde{\theta}', u')\} + \{1 - \varrho_{Q_T}(\tilde{\theta}, u)\} \delta_{(\tilde{\theta}, u)}(\mathrm{d}\tilde{\theta}', \mathrm{d}u') \tag{31}$$

with acceptance probability

$$\alpha_{Q_T}\{(\tilde{\theta}, u), (\tilde{\theta}', u')\} = \min\left\{1, \frac{\tilde{\pi}_T(\tilde{\theta}', u')\, \tilde{q}(\tilde{\theta}', \tilde{\theta}) K_{\rho_T}(u', u)}{\tilde{\pi}_T(\tilde{\theta}, u)\, \tilde{q}(\tilde{\theta}, \tilde{\theta}') K_{\rho_T}(u, u')}\right\},$$

and corresponding rejection probability $1 - \varrho_{Q_T}(\tilde{\theta}, u)$. The kernel Q_T is assumed to be $\mathcal{Y}^T$ measurable. Let $\Theta_T = \{\tilde{\vartheta}_n^T; n \geqslant 0\}$ denote the non-Markov stationary space-rescaled parameter sequence arising from the correlated pseudomarginal chain. The following result shows that the sequences $\{\Theta_T; T \geqslant 1\}$ converge weakly as $T \to \infty$ to a stationary Markov chain corresponding to the penalty method—an 'ideal' Monte Carlo technique which cannot be practically implemented (Ceperley and Dewing, 1999; Nicholls *et al.*, 2012).

Theorem 4. If assumptions 1–3 hold and $\vartheta \mapsto \kappa(\vartheta)$ is locally Lipschitz at $\vartheta = \bar{\theta}$ then the random probability measures on $(\mathbb{R}^d)^\infty$ given by the laws of $\{\Theta_T; T \geqslant 1\}$ converge weakly in probability $\mathbb{P}^Y$ as $T \to \infty$ to the law of a stationary Markov chain $\{\tilde{\vartheta}_n; n \geqslant 0\}$ defined by $\tilde{\vartheta}_0 \sim \mathcal{N}(0, \bar{\Sigma})$ and $\tilde{\vartheta}_n \sim P(\tilde{\vartheta}_{n-1}, \cdot)$ for $n \geqslant 1$ with

$$P(\tilde{\theta}, \mathrm{d}\tilde{\theta}') = \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \alpha_P(\tilde{\theta}, \tilde{\theta}') + \{1 - \varrho_P(\tilde{\theta})\}\, \delta_{\tilde{\theta}}(\mathrm{d}\tilde{\theta}'), \tag{32}$$

and

$$\alpha_P(\tilde{\theta}, \tilde{\theta}') = \int \varphi(\mathrm{d}r; -\kappa^2/2, \kappa^2) \min\left\{1, \frac{\varphi(\tilde{\theta}'; 0, \bar{\Sigma})\, \tilde{q}(\tilde{\theta}', \tilde{\theta})}{\varphi(\tilde{\theta}; 0, \bar{\Sigma})\, \tilde{q}(\tilde{\theta}, \tilde{\theta}')} \exp(r)\right\},$$

$1 - \varrho_P(\tilde{\theta})$ being the corresponding rejection probability and $\kappa := \kappa(\bar{\theta})$.

The consequence of this result is that, as $T \to \infty$, only the asymptotic distribution of the log-likelihood ratio error at the central parameter value $\bar{\theta}$ impacts the acceptance probability of the limiting chain. For large T and a proposal of the form that is specified in assumption 2, we thus expect some of the quantitative properties of the correlated pseudomarginal kernel Q, where we now omit T from the notation, to be captured by the Markov kernel

$$\hat{Q}(\theta, \mathrm{d}\theta') = q(\theta, \mathrm{d}\theta')\alpha_{\hat{Q}}(\theta, \theta') + \{1 - \varrho_{\hat{Q}}(\theta)\}\delta_\theta(\mathrm{d}\theta'), \tag{33}$$

with

$$\alpha_{\hat{Q}}(\theta, \theta') = \int \varphi(\mathrm{d}r; -\kappa^2/2, \kappa^2) \min\{1, r_{\mathrm{MH}}(\theta, \theta') \exp(r)\},$$

where $1 - \varrho_{\hat{Q}}(\theta)$ is the corresponding rejection probability and r_{MH} is defined in equation (2). We have obtained equation (33) by using the change of variables $\theta = \hat{\theta}_T + \tilde{\theta}/\sqrt{T}$ and substituting the true target for its normal approximation in equation (32), hence removing a level of approximation.

4.2. A bounding Markov chain

We analyse here the stationary Markov chain with transition kernel $\hat{Q}$ arising from our weak convergence analysis. To state our results, we need the following notation. For any real-valued measurable function h, probability measure μ and Markov kernel K on a measurable space $(E, \mathcal{E})$, we write $\mu(h) = \int_E h(x)\mu(\mathrm{d}x)$, $Kh(x) = \int_E K(x, \mathrm{d}x')h(x')$ and

$$K^n h(x) = \int_E \int_E K^{n-1}(x, \mathrm{d}z) K(z, \mathrm{d}x')h(x')$$

for $n \geqslant 2$ with $K^1 = K$. We also introduce the Hilbert space $L^2(\mu) = \{h : E \to \mathbb{R} \mid \mu(h^2) < \infty\}$ equipped with the inner product $\langle g, h\rangle_\mu = \int_E g(x)h(x)\mu(\mathrm{d}x)$. For any $h \in L^2(\mu)$, the autocorrelation at lag $n \geqslant 0$ is $\phi_n(h, K) = \langle \bar{h}, K^n h\rangle_\mu / \mu(\bar{h}^2)$ where $\bar{h} = h - \mu(h)$. The integrated autocorrelation time that is associated with a function h under a Markov kernel K is given by $\mathrm{IF}(h, K) = 1 + 2\Sigma_{n=1}^{\infty} \phi_n(h, K)$ and will be referred to subsequently as the *inefficiency*. For $\mu(\mathrm{d}x) = \mu(\mathrm{d}x_1, \mathrm{d}x_2)$, we shall slightly abuse the notation and write $\mathrm{IF}(h, K)$ instead of $\mathrm{IF}(g, K)$ when $g(x_1, x_2) = h(x_1)$ or $g(x_1, x_2) = h(x_2)$. When estimating $\mu(h)$, $n\,\mathrm{IF}(h, K)$ samples from a stationary Markov chain of μ-invariant transition kernel K are necessary to obtain an estimator of approximately the same precision as an average of n independent draws from μ; see, for example, Geyer (1992).

We provide an upper bound on $\mathrm{IF}(h, \hat{Q})$ which we exploit to provide guidelines on how to optimize the performance of the correlated pseudomarginal scheme in Section 4.4. The inefficiency $\mathrm{IF}(h, \hat{Q})$ is difficult to work with but we give an upper bound that depends only on $\mathrm{IF}(h, Q_{\mathrm{MH}})$ and κ. To proceed, we introduce an auxiliary Markov kernel Q^* given by

$$Q^*(\theta, \mathrm{d}\theta') = \varrho_{\mathrm{U}}(\kappa) Q_{\mathrm{MH}}(\theta, \mathrm{d}\theta') + \{1 - \varrho_{\mathrm{U}}(\kappa)\}\delta_\theta(\mathrm{d}\theta'), \tag{34}$$

where Q_{MH} is defined in equation (1) and

$$\varrho_{\mathrm{U}}(\kappa) = \int \varphi(\mathrm{d}r; -\kappa^2/2, \kappa^2) \min\{1, \exp(r)\} = 2\Phi(-\kappa/2). \tag{35}$$

We denote by $\bar{\varrho}_{Q^*}(\kappa)$ and $\bar{\varrho}_{\hat{Q}}(\kappa)$ the average acceptance probability of Q^* and $\hat{Q}$ respectively, at stationarity. The kernel Q^* is a 'lazy' version of Q_{MH} which satisfies the following properties.

Proposition 1. The kernel Q^* is π reversible and $\mathrm{IF}(h, \hat{Q}) \leqslant \mathrm{IF}(h, Q^*)$ for any $h \in L^2(\pi)$, where

$$\mathrm{IF}(h, Q^*) = \{1 + \mathrm{IF}(h, Q_{\mathrm{MH}})\}/\varrho_{\mathrm{U}}(\kappa) - 1, \tag{36}$$

with equality when $\varrho_{\mathrm{MH}}(\theta) = 1$ for all $\theta \in \Theta$, and

$$\bar{\varrho}_{Q^*}(\kappa) = \varrho_{\mathrm{U}}(\kappa)\pi(\varrho_{\mathrm{MH}}) \leqslant \bar{\varrho}_{\hat{Q}}(\kappa). \tag{37}$$

Moreover, Q^* is geometrically ergodic if Q_{MH} is geometrically ergodic.

For any π- or $\bar{\pi}$-invariant Markov kernel K, we define the relative inefficiency $\mathrm{RIF}(h, K)$ and the auxiliary relative computing time $\mathrm{ARCT}(h, K)$ with respect to the MH kernel Q_{MH} using the exact likelihood by

$$\begin{aligned} \mathrm{RIF}(h, K) &:= \frac{\mathrm{IF}(h, K)}{\mathrm{IF}(h, Q_{\mathrm{MH}})}, \\ \mathrm{ARCT}(h, K) &:= \sqrt{\left\{ \frac{\mathrm{RIF}(h, K)}{\kappa^2 \varrho_{\mathrm{U}}(\kappa)} \right\}}. \end{aligned} \tag{38}$$

We next minimize $\mathrm{ARCT}(h, Q^*)$, which is an upper bound on $\mathrm{ARCT}(h, \hat{Q})$, with respect to κ—this quantity is a component of the function that we need to minimize to optimize the performance of the correlated pseudolikelihood algorithm; see Section 4.4.

Proposition 2. The following results hold.

(a) If $\mathrm{IF}(h, Q_{\mathrm{MH}}) = 1$, then

$$\mathrm{RIF}(h, Q^*) = \{2 - \varrho_{\mathrm{U}}(\kappa)\}/\varrho_{\mathrm{U}}(\kappa),$$

and $\mathrm{ARCT}(h, Q^*)$ is minimized at $\kappa = 1.35$, at which point $\varrho_{\mathrm{U}}(\kappa) = 0.50$, $\mathrm{RIF}(h, Q^*) = 2.99$ and $\mathrm{ARCT}(h, Q^*) = 1.81$.

(b) As $\mathrm{IF}(h, Q_{\mathrm{MH}}) \to \infty$,

$$\mathrm{RIF}(h, Q^*) = 1/\varrho_{\mathrm{U}}(\kappa),$$

and $\mathrm{ARCT}(h, Q^*)$ is minimized at $\kappa = 1.50$, at which point $\varrho_{\mathrm{U}}(\kappa) = 0.43$, $\mathrm{RIF}(h, Q^*) = 2.20$ and $\mathrm{ARCT}(h, Q^*) = 1.47$.

(c) $\mathrm{RIF}(h, Q^*)$ and $\mathrm{ARCT}(h, Q^*)$ are decreasing functions of $\mathrm{IF}(h, Q_{\mathrm{MH}})$. The minimizing argument rises monotonically from 1.35 to 1.50 as $\mathrm{IF}(h, Q_{\mathrm{MH}})$ increases from 1 to ∞.

Fig. 1 displays $\varrho_{\mathrm{U}}(\kappa)$, $\mathrm{RIF}(h, Q^*)$ and $\mathrm{ARCT}(h, Q^*)$ against κ. The two scenarios that are displayed are for $\mathrm{IF}(h, Q_{\mathrm{MH}}) = 1$, corresponding to the 'perfect' proposal case where $q(\theta, \theta') = \pi(\theta')$, and for the limiting case where $\mathrm{IF}(h, Q_{\mathrm{MH}}) \to \infty$. These correspond to parts (a) and (b) of proposition 2. From Fig. 1, it is also clear that $\mathrm{ARCT}(h, Q^*)$, for both scenarios, is fairly flat as a function of κ. The function only approximately doubles relative to the minimum at $\kappa = 1$ or $\kappa = 4$.

4.3. A lower bound on the integrated auto-correlation time

We stress here that theorem 4 does not imply that the inefficiency of the correlated pseudomarginal scheme converges, as $T \to \infty$, to the inefficiency of the limiting chain that is identified

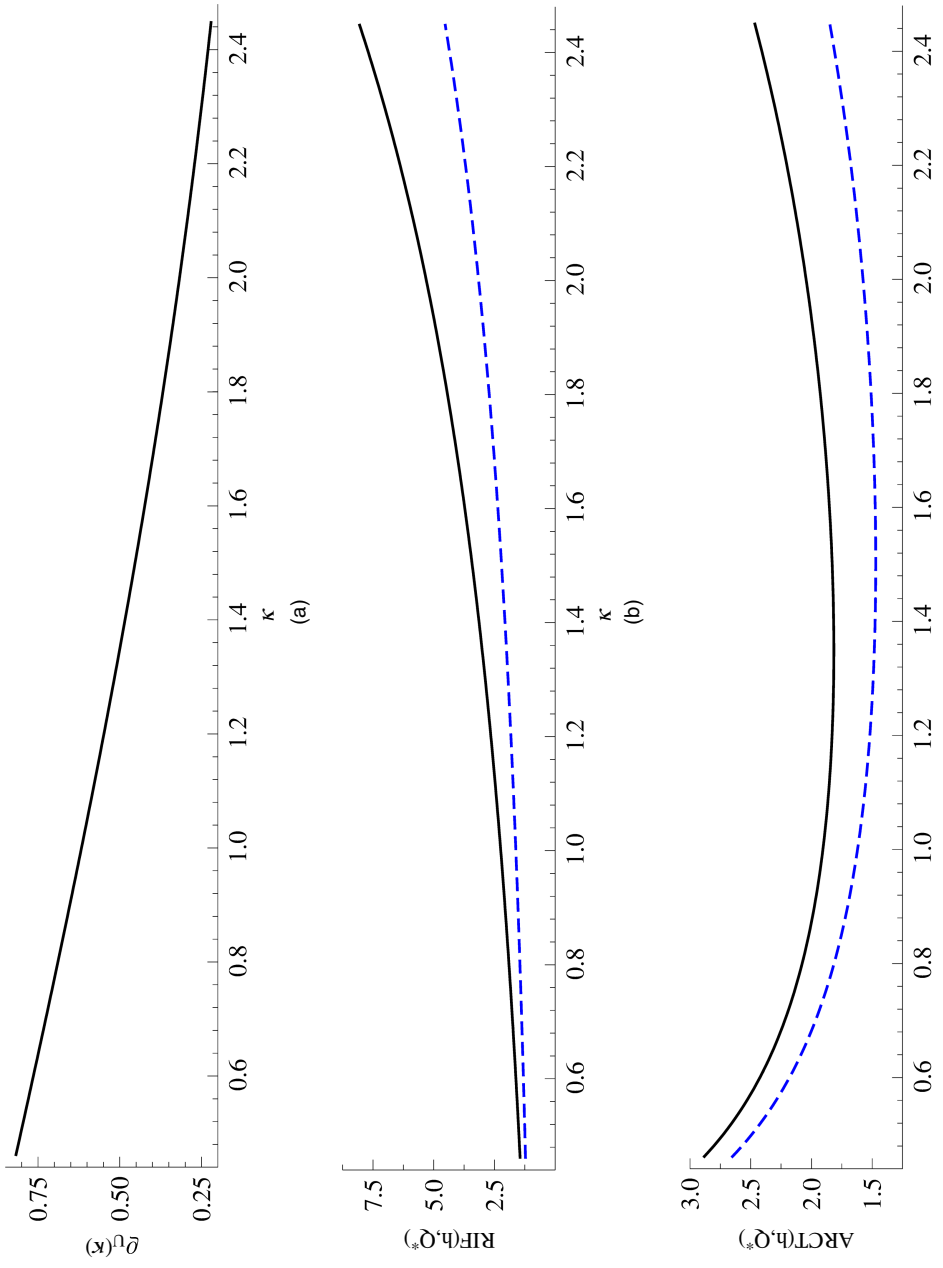

Fig. 1. Illustrations of proposition 2: (a) acceptance probability $\varrho_U(\kappa)$ against κ; (b) relative inefficiency $\mathrm{RIF}(h, Q^*)$ against κ (——, $\mathrm{IF}(h, Q_{\mathrm{MH}}) = 1$; – – –, $\mathrm{IF}(h, Q_{\mathrm{MH}}) \to \infty$); (c) auxiliary relative computing time $\mathrm{ARCT}(h, Q^*)$ against κ (——, $\mathrm{IF}(h, Q_{\mathrm{MH}})=1$; – – –, $\mathrm{IF}(h, Q_{\mathrm{MH}}) \to \infty$)

therein. In fact, whereas theorem 4 holds whenever $N_T \to \infty$ and $N_T = o(T)$ as $T \to \infty$, our next result suggests that N_T must grow at least as fast as $\sqrt{T}$ for the inefficiency of the correlated pseudomarginal scheme to remain bounded. To simplify the presentation in this section, we assume further on that $d = 1$.

In the correlated pseudomarginal context, the sequence of auxiliary variables $\{\mathbf{U}_n; n \geqslant 0\}$ evolves at a much slower scale than $\{\vartheta_n; n \geqslant 0\}$ as it is driven by the proposal K_{ρ_T}, where ρ_T is given by equation (25). When N_T grows too slowly with T, we expect and observe empirically that the inefficiency $\mathrm{IF}(h, Q_T)$, for any function h, is of the same order as the inefficiency of $\{\mathbb{E}[h(\vartheta_n)|\mathbf{U}_n]; n \geqslant 0\}$. Moreover, under regularity conditions (see for example Doucet *et al.* (2013), lemma 2), we have for large T

$$\mathbb{E}[h(\vartheta_n)|\mathbf{U}_n] = h(\hat{\theta}_T) + \frac{\bar{\Sigma}}{2T}\nabla_{\vartheta,\vartheta} h(\hat{\theta}_T) + \frac{\bar{\Sigma}}{T}\nabla_{\vartheta} h(\hat{\theta}_T)\Psi(\hat{\theta}_T, \mathbf{U}_n) + O_{\mathbb{P}}(T^{-2}), \tag{39}$$

where

$$\Psi(\hat{\theta}_T, U) = \nabla_{\vartheta} \log\{\hat{p}(Y_{1:T} \mid \hat{\theta}_T, U)/p(Y_{1:T} \mid \hat{\theta}_T)\} \tag{40}$$

is the error in the simulated score at $\hat{\theta}_T$ and will be referred to as the score error. As a first step, we obtain a lower bound on $\mathrm{IF}(\Psi, Q_T)$.

Proposition 3. Under regularity conditions given in section A.10 of the on-line supplementary material, there is a constant $C > 0$ such that $\mathrm{IF}(\Psi, Q_T) \geqslant C\mathbb{V}_{\bar{\pi}_T}(\Psi)$ $\mathbb{P}^Y$—almost surely.

It follows from calculations that are similar to those in section A.11 in the on-line supplementary material (see also Lindsten and Doucet (2016), proposition 3) that under regularity conditions there exists $A > 0$ such that $\mathbb{V}_{\bar{\pi}_T}(\Psi) \sim AT/N$ $\mathbb{P}^Y$—almost surely. By combining equation (39) and proposition 3, we thus expect the inefficiency of $\{\mathbb{E}[h(\vartheta_n)|\mathbf{U}_n]; n \geqslant 0\}$ to be lower bounded by a term of order

$$\frac{\mathrm{IF}(\Psi, Q_T)\mathbb{V}_{\bar{\pi}_T}(\Psi/T)}{\mathbb{V}_{\bar{\pi}_T}(h)} \geqslant B\frac{T}{N_T}\frac{T^{1-\alpha}}{T^2}T = BT^{1-2\alpha}$$

for $N_T = \lceil \beta T^{\alpha} \rceil$, some constant $B > 0$ and T sufficiently large. This result suggests that a necessary condition for $\mathrm{IF}(h, Q_T)$ to remain finite as $T \to \infty$ is to have N_T growing at least at rate $\sqrt{T}$. This is validated by the experimental results of Section 5 which also suggest that this rate is sufficient.

4.4. Optimization

We provide a heuristic to select the parameters of the correlated pseudomarginal scheme to optimize its performance which is validated by experimental results in Section 5. Again, we set $d = 1$ for simplicity. For a test function $h : \Theta \to \mathbb{R}$, we want to minimize

$$\mathrm{CT}(h, Q_T) = N_T \ \mathrm{IF}(h, Q_T), \tag{41}$$

where the factor N_T arises from the fact that the computational cost of the likelihood estimator is proportional to N_T for random-effects models. The results of Section 4.3 suggest that we should choose the number of Monte Carlo samples to scale as $N_T = \beta T^{1/2}$ so that $\rho_T = \exp(-\psi\beta T^{-1/2})$. It remains to determine ψ and β.

To evaluate equation (41), we first decompose the functional of interest evaluated at the parameter at the nth iteration as

$$h(\vartheta_n) = f(\mathbf{U}_n) + g(\vartheta_n, \mathbf{U}_n),$$

where

$$f(\mathbf{U}) := \mathbb{E}_{\bar{\pi}_T}[h(\vartheta)|\mathbf{U}],$$
$$g(\vartheta, \mathbf{U}) := h(\vartheta) - \mathbb{E}_{\bar{\pi}_T}\{h(\vartheta)|\mathbf{U}\}. \tag{42}$$

It is easy to check that

$$\mathbb{V}_{\bar{\pi}_T}(h)\,\mathrm{IF}(h, Q_T) \leqslant 2\,\mathbb{V}_{\bar{\pi}_T}(f)\,\mathrm{IF}(f, Q_T) + 2\,\mathbb{V}_{\bar{\pi}_T}(g)\,\mathrm{IF}(g, Q_T).$$

Assumption 1 combined with mild regularity assumptions on h and integrability conditions shows that $\mathbb{V}_{\bar{\pi}_T}\{h(\vartheta_n)\} \approx \bar{\Sigma}_h/T$, where $\bar{\Sigma}_h = |h'(\bar{\theta})|^2\bar{\Sigma}$. Since $f(\mathbf{U}_n)$ and $g(\vartheta_n, \mathbf{U}_n)$ are clearly uncorrelated, it follows that $\mathbb{V}_{\bar{\pi}_T}(h) = \mathbb{V}_{\bar{\pi}_T}(f) + \mathbb{V}_{\bar{\pi}_T}(g)$. From equation (39) we have $\mathbb{V}_{\bar{\pi}_T}(f) \approx \bar{\Sigma}^2\mathbb{V}_{\bar{\pi}_T}(\Psi/T) \approx \bar{\Sigma}_f/(TN_T)$; therefore

$$\mathbb{V}_{\bar{\pi}_T}(g) \approx \frac{\bar{\Sigma}_h}{T} - \frac{\bar{\Sigma}_f}{TN_T} \approx \frac{\bar{\Sigma}_h}{T}.$$

Using the reasoning of Section 4.3 and the calculations above we obtain

$$\begin{aligned}\mathrm{IF}(h, Q_T) &\leqslant \frac{2}{\bar{\Sigma}_h}\{\mathbb{V}_{\bar{\pi}_T}(\sqrt{T}f)\mathrm{IF}(f, Q_T) + \mathbb{V}_{\bar{\pi}_T}(\sqrt{T}g)\mathrm{IF}(g, Q_T)\} \\ &\approx \frac{2}{\bar{\Sigma}_h}\left\{\frac{\bar{\Sigma}_f}{N_T}\mathrm{IF}(\Psi, Q_T) + \bar{\Sigma}_h\mathrm{IF}(g, Q_T)\right\}. \end{aligned} \tag{43}$$

Proposition 3 states that $\mathrm{IF}(\Psi, Q_T)$ is of order at least T/N_T in probability as $T \to \infty$. Numerical results suggest that in fact we have $\mathrm{IF}(\Psi, Q_T) \approx A/\{\delta_T\varrho_{\mathrm{U}}(\kappa)\}$ where $\delta_T = \psi N_T/T = -\log(\rho_T)$ as illustrated in Section 5.1, Fig. 5. Hence, by substituting this expression of $\mathrm{IF}(\Psi, Q_T)$ in approximation (43), it follows that

$$\mathrm{IF}(h, Q_T) \lesssim \frac{2}{\bar{\Sigma}_h}\left\{\frac{\bar{\Sigma}_f}{N_T}\frac{A}{\delta_T\varrho_{\mathrm{U}}(\kappa)} + \bar{\Sigma}_h\,\mathrm{IF}(g, Q_T)\right\},$$

where the symbol '$\lesssim$' means that an approximation has been used. It can also be observed empirically from Fig. 4, described in Section 5.1, that the auto-correlations of $g(\vartheta_n, \mathbf{U}_n)$ decay exponentially, at a rate that is independent of T. We expect that, at least approximately, we have $\mathrm{IF}(g, Q_T) \approx \mathrm{IF}(h, \hat{Q}_T)$ in probability. Therefore overall, for some constant $B > 0$, we have that

$$\mathrm{IF}(h, Q_T) \lesssim 2\left\{\frac{B}{\varrho_{\mathrm{U}}(\kappa)\delta_T N_T} + \mathrm{IF}(h, \hat{Q}_T)\right\}. \tag{44}$$

We are interested in optimizing $\mathrm{CT}(h, Q_T) = N_T\,\mathrm{IF}(h, Q_T)$ with respect to ψ and β where we recall from equation (27) that $\delta_T = \psi N_T/T = \psi\beta/\sqrt{T} = \kappa^2\beta/(\gamma^2\sqrt{T})$ as $\kappa^2 = \psi\gamma^2$. Therefore

$$\mathrm{CT}(h, Q_T) \lesssim 2T^{1/2}\left\{\frac{C}{\beta\varrho_{\mathrm{U}}(\kappa)\kappa^2} + \beta\,\mathrm{IF}(h, \hat{Q}_T)\right\}, \tag{45}$$

where $C = B\gamma^2$, and the upper bound on $\mathrm{CT}(h, Q_T)$ is minimized at

$$\beta^* = \sqrt{\left\{\frac{C}{\varrho_{\mathrm{U}}(\kappa)\kappa^2\,\mathrm{IF}(h, \hat{Q}_T)}\right\}}.$$

By plugging β^* in the right-hand side of expression (45), we obtain by proposition 1 that

$$\mathrm{CT}(h, Q_T) \lesssim 4\sqrt{\{C\ \mathrm{IF}(h, Q_{\mathrm{MH}})T\}}\,\mathrm{ARCT}(h, \hat{Q}_T) \lesssim 4\sqrt{\{C\ \mathrm{IF}(h, Q_{\mathrm{MH}})T\}}\,\mathrm{ARCT}(h, Q_T^*) \quad (46)$$

where ARCT was introduced in expression (38). In practice we minimize $\mathrm{ARCT}(h, Q_T^*)$ with respect to κ, following proposition 2. The minimizer $\hat{\kappa}$ is a function of $\mathrm{IF}(h, Q_{\mathrm{MH}})$ which varies only slightly as $\mathrm{IF}(h, Q_{\mathrm{MH}})$ varies from 1 to ∞ as observed in Fig. 1. Consequently, we propose the following procedure to optimize the performance of the correlated pseudomarginal. Let T be fixed and sufficiently large for the asymptotic assumptions to hold approximately. First, we choose a candidate value for N and determine $\hat{\psi}$ such that the standard deviation of the log-likelihood ratio estimator around the mode of the posterior, estimated through a preliminary run, satisfies $\hat{\kappa} \approx 1.4$. Second, fixing ψ at $\hat{\psi}$, we evaluate for several values of β the computation time $\mathrm{CT}(h, Q_T)$ which we assume is of the form of the upper bound (45), i.e.

$$\mathrm{CT}(h, Q_T) = C_0/\beta + C_1\beta, \quad (47)$$

with κ and T kept constant; see Fig. 6 in Section 5.1 for empirical results. This function is minimized for $\beta = \sqrt{(C_0/C_1)}$. Practically we evaluate $\mathrm{CT}(h, Q_T)$ on only a subset of the data. We then estimate through regression the constants C_0 and C_1 by $\hat{C}_0$ and $\hat{C}_1$ which in turn provide the following estimate of β:

$$\hat{\beta} = \sqrt{(\hat{C}_0/\hat{C}_1)}. \quad (48)$$

We examine in Section 5.1 the assumptions that were made here, illustrate this procedure and demonstrate its robustness.

5. Applications

5.1. Random-effects model

We illustrate the performance of the pseudomarginal and correlated pseudomarginal schemes on a simple Gaussian random-effects model where

$$\begin{aligned} X_t &\overset{\mathrm{IID}}{\sim} \mathcal{N}(\theta, 1), \\ Y_t | X_t &\sim \mathcal{N}(X_t, 1). \end{aligned} \quad (49)$$

We are interested in estimating θ (which has a true value of 0.5) to which we assign a zero-mean Gaussian prior with large variance. In this scenario, the likelihood is known as $Y_t \sim \mathcal{N}(\theta, 2)$. This enables detailed experimental analysis of the log-likelihood error and the log-likelihood ratio error. This also enables us to implement the MH algorithm with the true likelihood. The same normal random-walk proposal is used for all three schemes (MH, pseudomarginal and correlated pseudomarginal) and the following unbiased estimator of the likelihood is used for the pseudomarginal and correlated pseudomarginal schemes:

$$\begin{aligned} \hat{p}(y_{1:T} \mid \theta, U) &= \prod_{t=1}^{T} \hat{p}(y_t \mid \theta, U_t), \\ \hat{p}(y_t \mid \theta, U_t) &= \frac{1}{N} \sum_{i=1}^{N} \varphi(y_t; \theta + U_{t,i}, 1), \qquad U_{t,i} \overset{\mathrm{IID}}{\sim} \mathcal{N}(0, 1). \end{aligned} \quad (50)$$

The inefficiency is estimated for all three schemes for $h(\theta) = \theta$ using $1 + 2\Sigma_{n=1}^{L} \hat{\phi}_n$ where $\hat{\phi}_n$ is the estimated correlation for θ at lag n and L is a suitable cut-off value. We use the notation

$Z = \log\{\hat{p}(y_{1:T} \mid \theta, U)/p(y_{1:T} \mid \theta)\}$ and $W = \log\{\hat{p}(y_{1:T} \mid \theta', U')/p(y_{1:T} \mid \theta')\}$ where $\theta' \sim q(\theta, \cdot)$ and $U' \sim K_\rho(U, \cdot)$ and write $R = W - Z$ for $R_T(\theta, \theta')$ defined in equation (23).

As discussed in Section 4, for large data sets, the relative inefficiency $\mathrm{RIF} = \mathrm{IF}/\mathrm{IF}_{\mathrm{MH}}$ and associated relative computing time $\mathrm{RCT} = N\,\mathrm{RIF}$ of the correlated pseudomarginal scheme depend on the standard deviation κ of R at stationarity and the correlation parameter ρ. To validate experimentally the results of Section 3, we first analyse the case where $T = 8192$ in more detail. We run the correlated pseudomarginal algorithm by using a random-walk proposal for $N = 80$ and $\rho = 0.9963$, so that $\kappa = 1.145$. The draws of W and Z at equilibrium, together with R, are displayed in Fig. 2. The draws of Z are approximately distributed according to $\mathcal{N}(\sigma^2/2, \sigma^2)$ (Fig. 2(c)), where the variance σ^2 is high. The draws of R appear uncorrelated (in unreported tests) and their histogram is indistinguishable from the expected theoretical distribution $\mathcal{N}(-\kappa^2/2, \kappa^2)$ established in theorem 3 (Fig. 2(d)). This is in agreement with theorem 1, equation (22), the posterior of θ being concentrated. The resulting draws and correlogram (Figs 2(b) and 2(f)) of θ demonstrate low persistence.

For the pseudomarginal scheme, it is necessary to take $N = 5000$ samples to ensure that the variance of Z evaluated at a central value $\hat{\theta}$ is approximately 1 (Doucet *et al.*, 2015). We next validate experimentally the theoretical results of Section 4 by investigating the performance of the correlated pseudomarginal algorithm for this data set, varying N, and thus also $\kappa^2 = \mathbb{V}(R)$, while keeping $\rho = 0.9963$. Fig. 3 displays the values of RIF and RCT against κ as well as the marginal acceptance probabilities, showing that RCT is approximately minimized around $\kappa = 1.6$ close to the minimizing argument of $\mathrm{ARCT}(h, Q_T^*)$ that was established in proposition 2 which satisfies expression (46). Figs 3(c) and 3(d) show that $\log(\kappa^2)$ decreases linearly with $\log(N)$ as expected (Fig. 3(d)) and that the marginal probability of acceptance in the correlated pseudomarginal scheme is close to the asymptotic lower bound (Fig. 3(c)) given by expression (37). From these experimental results, it is clear that, for all values of N that were considered, the gains of the correlated pseudomarginal scheme over the pseudomarginal method in terms of RCT are very significant. The optimal value of N for the correlated pseudomarginal scheme is 35 ($\kappa = 1.6$) which gives $\mathrm{RCT} = 61$ against a value of $\mathrm{RCT} = 14100$ for the pseudomarginal scheme. Consequently, the pseudomarginal method would take more than 200 times as long in computation time to produce an estimate of the posterior mean of θ of the same accuracy.

We next investigate the performance of the correlated pseudomarginal method when T and $N = \beta\sqrt{T}$ vary while ψ, or equivalently ρ, is scaled such that κ is approximately constant. The results are recorded in Table 3. They suggest that the scaling $N = \beta\sqrt{T}$ is successful as $\mathrm{IF}_{\mathrm{CPM}}$ appears to stabilize whereas the scaling $N = \beta T$ is necessary for $\mathrm{IF}_{\mathrm{PM}}$ to stabilize. Experimental results that are not reported here confirm that, if N grows at a slower rate than $\sqrt{T}$, then $\mathrm{IF}_{\mathrm{CPM}}$ increases without bound with T.

We now justify empirically some of the assumptions that were made in Section 4 to guide the selection of the parameters ψ and β. First, we show that the correlated pseudomarginal process can be thought of as a combination of two different processes: a 'slow' moving component $f(\mathbf{U}_n) \approx \hat{f}(\mathbf{U}_n) = \hat{\theta}_T + \bar{\Sigma} T^{-1} \Psi(\hat{\theta}_T, \mathbf{U}_n)$, the modified score error associated with the score error $\Psi(\hat{\theta}_T, \mathbf{U}_n)$ defined in equation (40) and a 'fast' component $g(\vartheta_n, \mathbf{U}_n) = \vartheta_n - f(\mathbf{U}_n) \approx \hat{g}(\vartheta_n, \mathbf{U}_n) = \vartheta_n - \hat{f}(\mathbf{U}_n)$. We display these components for a correlated pseudomarginal run and the associated correlograms in Fig. 4 for fixed κ. We also illustrate in Fig. 5 that $\mathrm{IF}(\Psi, Q_T) \approx A/\{\delta_T \varrho_{\mathrm{U}}(\kappa)\}$ where $\delta_T = \psi N_T/T = -\log(\rho_T)$. The optimization scheme that was developed in Section 4.4 essentially selects β such that the asymptotic variances of both the slow and the fast components $\hat{f}(\mathbf{U}_n)$ and $\hat{g}(\vartheta_n, \mathbf{U}_n)$ are of the same order.

To apply the optimization procedure, we first run the algorithm for $N = 20$ and tune ψ to obtain $\hat{\kappa} \approx 1.4$. For the resulting value $\hat{\psi}$, we then evaluate $\mathrm{CT}_{\mathrm{CPM}} = N\,\mathrm{IF}_{\mathrm{CPM}}$ for various values

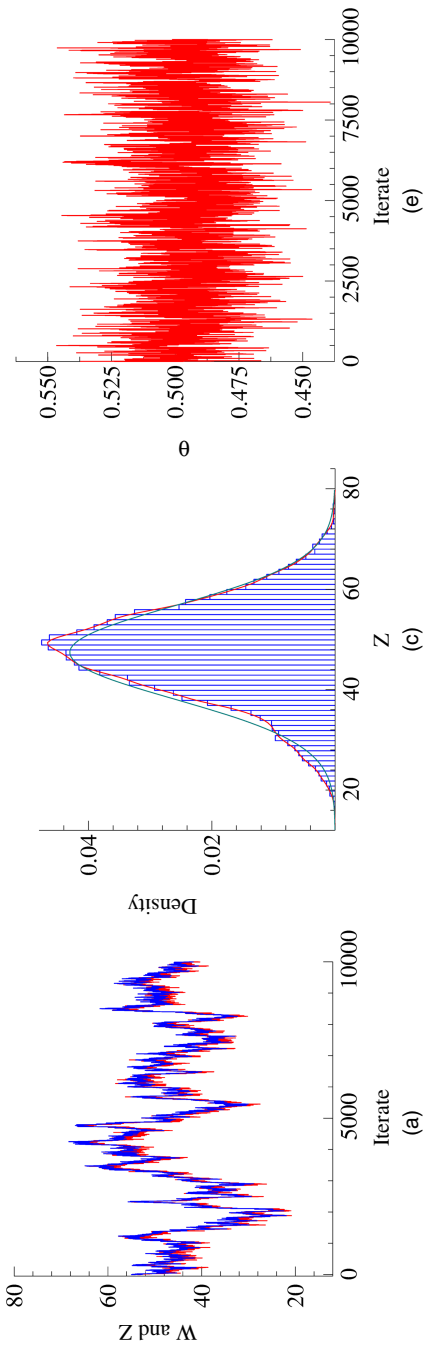

Fig. 2. Random-effects model using the correlated pseudomarginal algorithm ($T = 8192$, $N = 80$ and $\rho = 0.9963$): (a) first 10000 iterations of W (——) and Z (——) and (b) the difference R; histograms of (c) Z and (d) R and the theoretical Gaussian densities; (e) draws of θ and (f) the corresponding correlogram

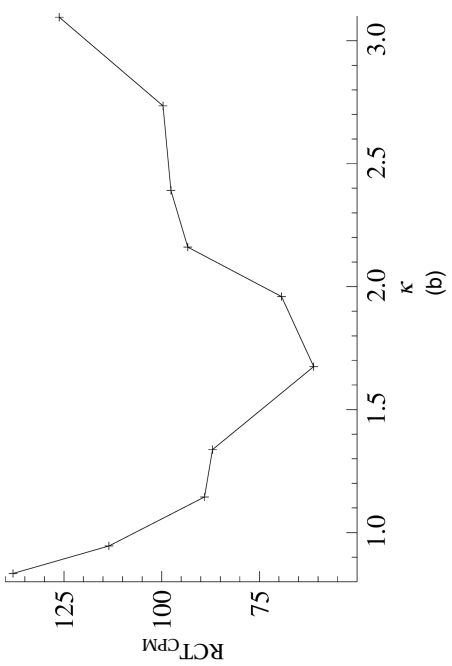

Fig. 3. Random-effects model using the correlated pseudomarginal algorithm ($T = 8192$, ρ fixed and various N): (a) $\mathrm{RIF}_{\mathrm{CPM}}$ (+) and RIF_{Q^*} for $\mathrm{IF}(h, Q_{\mathrm{MH}}) = 1$ (+) and $\mathrm{IF}(h, Q_{\mathrm{MH}}) = \infty$ (□) against κ (see proposition 2); (b) $\mathrm{RCT}_{\mathrm{CPM}}$ against κ; (c) acceptance probability of the correlated pseudomarginal and the theoretical lower bound, of equation (37), against κ; (d) $\log(\kappa^2)$ against $\log(N)$

Table 3. Random-effects model†

T	N	ρ	κ^2	$\bar{\varrho}_{\mathrm{MH}}$	IF_{MH}	$\bar{\varrho}_{\mathrm{CPM}}$	IF_{CPM}	RIF_{CPM}
1024	19	0.9894	2.0	0.71	10.71	0.48	43.26	4.04
2048	28	0.9925	1.9	0.69	8.21	0.49	38.50	4.61
4096	39	0.9947	1.7	0.72	11.75	0.51	21.01	1.79
8192	56	0.9962	1.8	0.81	15.61	0.50	24.25	1.55
16384	79	0.9974	1.8	0.70	9.37	0.50	20.05	2.14

†Inefficiency and acceptance probabilities $\bar{\varrho}_{\mathrm{MH}}$ and $\bar{\varrho}_{\mathrm{CPM}}$ for the MH and correlated pseudomarginal algorithms, $N = \beta\sqrt{T}$ and ρ selected such that κ^2 is approximately constant.

of β and perform a regression based on equations (47) and (48). Practically, we use only a subset of the data to perform this optimization to speed up computation. The results are fairly insensitive to the size of this subset as illustrated in Fig. 6 and suggest selecting β around 0.25.

5.2. *Heston stochastic volatility model*

We investigate here the empirical performance of the correlated pseudomarginal algorithm on the Heston model (Heston, 1993; Chopin and Gerber, 2017), which is a popular stochastic volatility model with leverage which is a partially observed diffusion model. The logarithm of the observed price $P(t)$ evolves according to

$$\begin{aligned} \mathrm{d}\log P(t) &= \sigma(t)\mathrm{d}B(t), \\ \mathrm{d}\sigma^2(t) &= \upsilon\{\mu - \sigma^2(t)\}\mathrm{d}t + \omega\sigma(t)\mathrm{d}W(t), \end{aligned}$$

where $\sigma(t)$ is a stationary latent spot stochastic volatility process such that $\sigma^2(t) \sim \mathcal{G}(\alpha, \beta)$ where $\mathcal{G}(\alpha, \beta)$ is the gamma distribution of shape $\alpha = 2\mu\upsilon/\omega^2$ and rate $\beta = 2\upsilon/\omega^2$. The Brownian motions $B(t)$ and $W(t)$ are correlated with $\chi = \mathrm{corr}\{B(t), W(t)\}$. The returns $Y_s = \log\{P(\tau_s)\} - \log\{P(\tau_{s-1})\}$ are observed at equally spaced times $\tau_0 < \ldots < \tau_T$, where $\Delta = \tau_s - \tau_{s-1}$ for all $s = 1, \ldots, T$. Conditionally on the volatility $\sigma^2(t)$ and driving process $W(t)$, we have

$$Y_s \sim \mathcal{N}\{\chi\gamma_s; (1-\chi^2)\sigma_s^{2*}\}, \tag{51}$$

$$\begin{aligned} \sigma_s^{2*} &= \int_{\tau_{s-1}}^{\tau_s} \sigma^2(t)\,\mathrm{d}t, \\ \gamma_s &= \int_{\tau_{s-1}}^{\tau_s} \sigma(t)\,\mathrm{d}W(t). \end{aligned} \tag{52}$$

To perform inference, we first reparameterize the model in terms of $x(t) = \log\{\sigma^2(t)\}$. We apply Itô's lemma to $x(t)$ and discretize the resulting diffusion by using an Euler scheme. We write $x_i^s = x(\tau_s + \epsilon i)$, where $\epsilon = \Delta/I$ for $i = 0, \ldots, I$ so that $x_I^s = x_0^{s+1}$. The evolution of these latent variables is given by

$$x_{i+1}^s = x_i^s + \epsilon\left[\upsilon\{\mu\exp(-x_i^s) - 1\} - \frac{\omega^2}{2}\exp(-x_i^s)\right] + \sqrt{\epsilon}\omega\exp\left(-\frac{x_i^s}{2}\right)\eta_i,$$

where $\eta_i \sim^{\mathrm{IID}} \mathcal{N}(0,1)$ for $i = 0, \ldots, I-1$. Under the Euler scheme, the returns satisfy

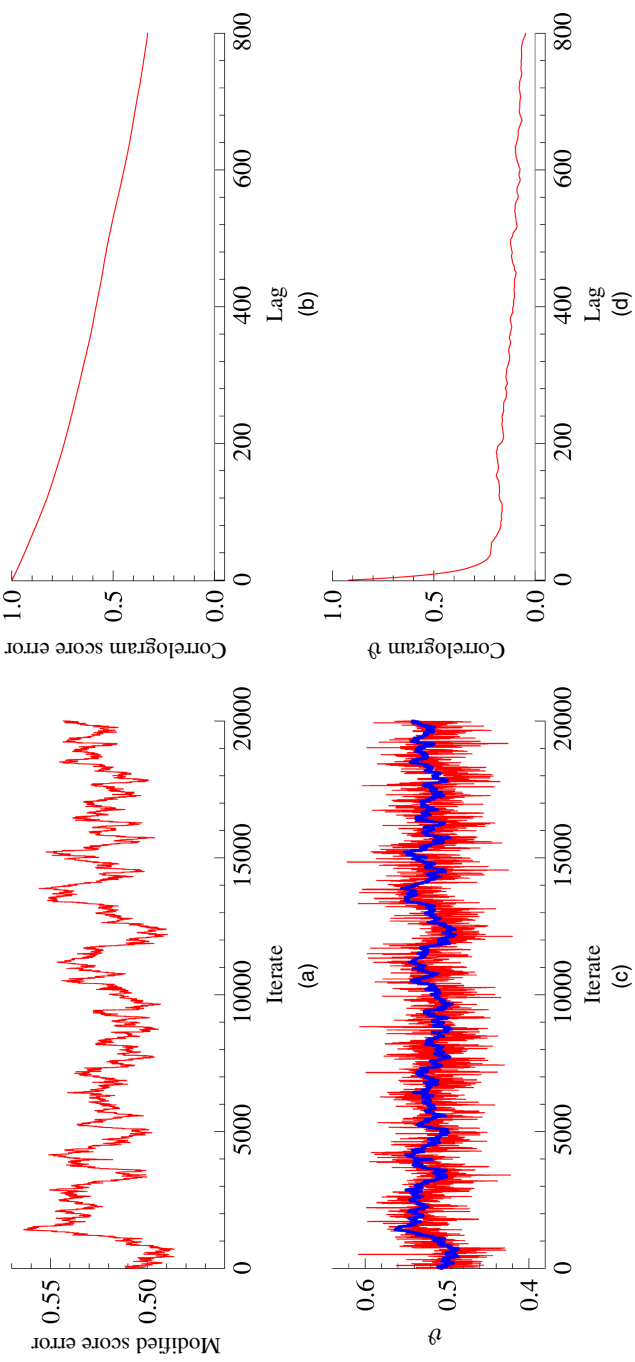

Fig. 4. Random-effects model using the correlated pseudomarginal algorithm ($T = 2560$, $\beta = 0.12$, $N = 6$ and $\rho = 0.9977$): (a) modified score error $\hat{f}(\mathbf{U}_n)$ and (b) its correlogram; (c) parameter ϑ_n (——) and modified score error (——); (d) correlogram ϑ_n; (e) residual $\hat{g}(\vartheta_n, \mathbf{U}_n)$ and (f) correlogram

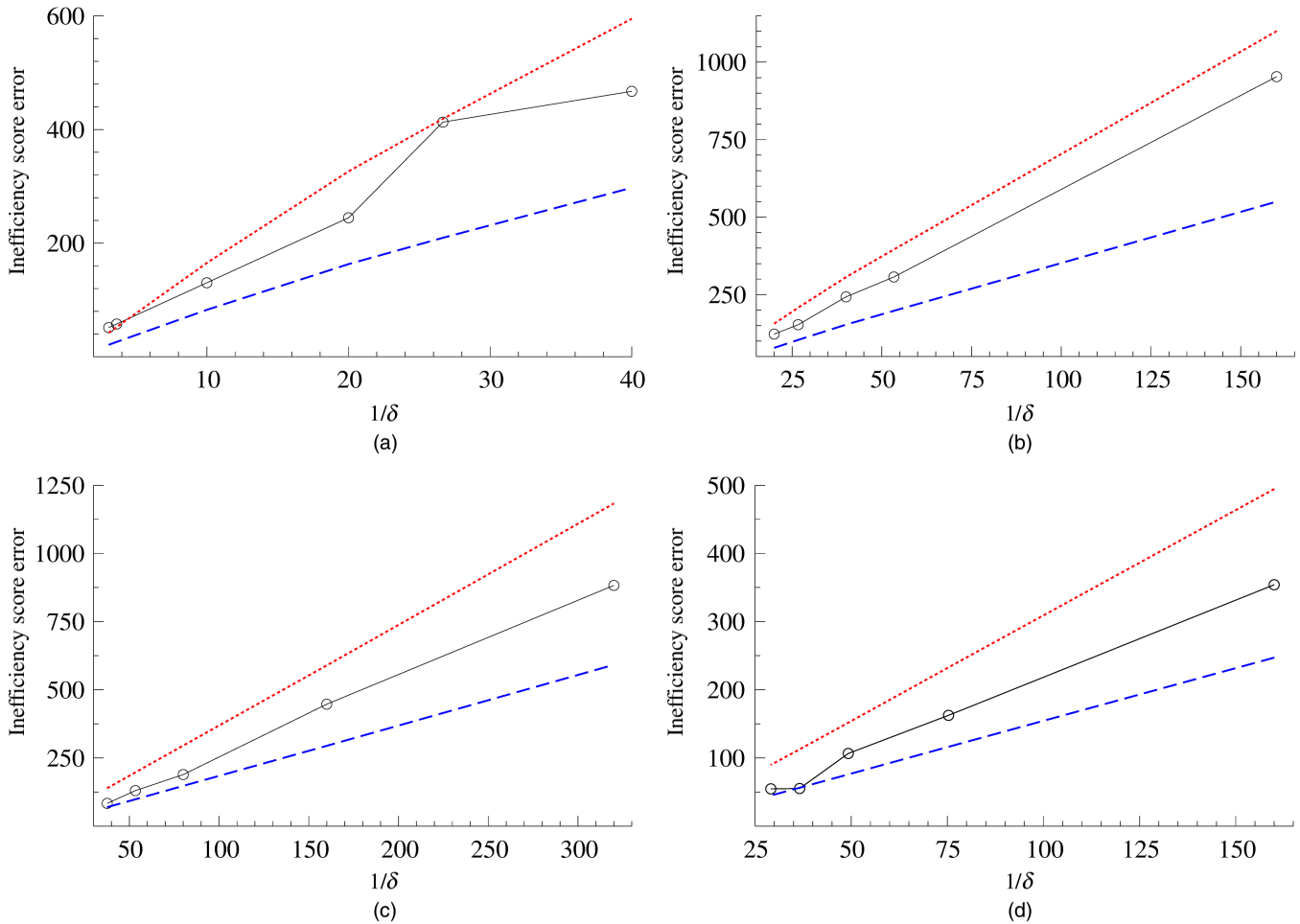

Fig. 5. Random-effects model using the correlated pseudomarginal algorithm ($T = 320$)—inefficiency of the score error (———) plotted against $1/\delta$ for four values of κ^2 = (a) 9.5, (b) 4.9, (c) 1.42, (d) 0.75 and upper bound $2/(\delta\bar{\varrho}_{\text{CPM}})$ (·····) and lower bound $1/(\delta\bar{\varrho}_{\text{CPM}})$ (– – –) with corresponding acceptance probabilities $\bar{\varrho}_{\text{CPM}}$ (a) 0.12, (b) 0.27, (c) 0.54 and (d) 0.65

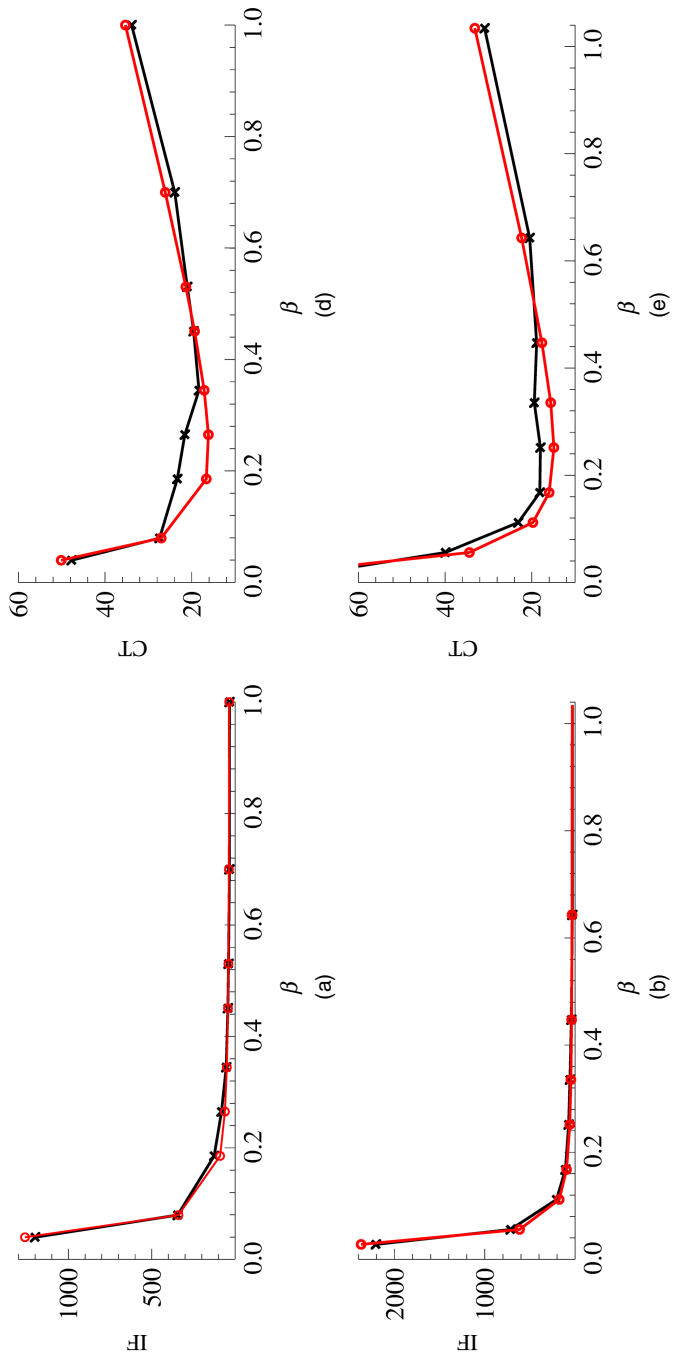

Fig. 6. Random-effects model, (a)–(c) IF and (d)–(f) CT as a function of β (——, regression fit based on the estimated CT): (a), (d) $T = 2560$; (b), (e) $T = 1280$; (c), (f) $T = 320$

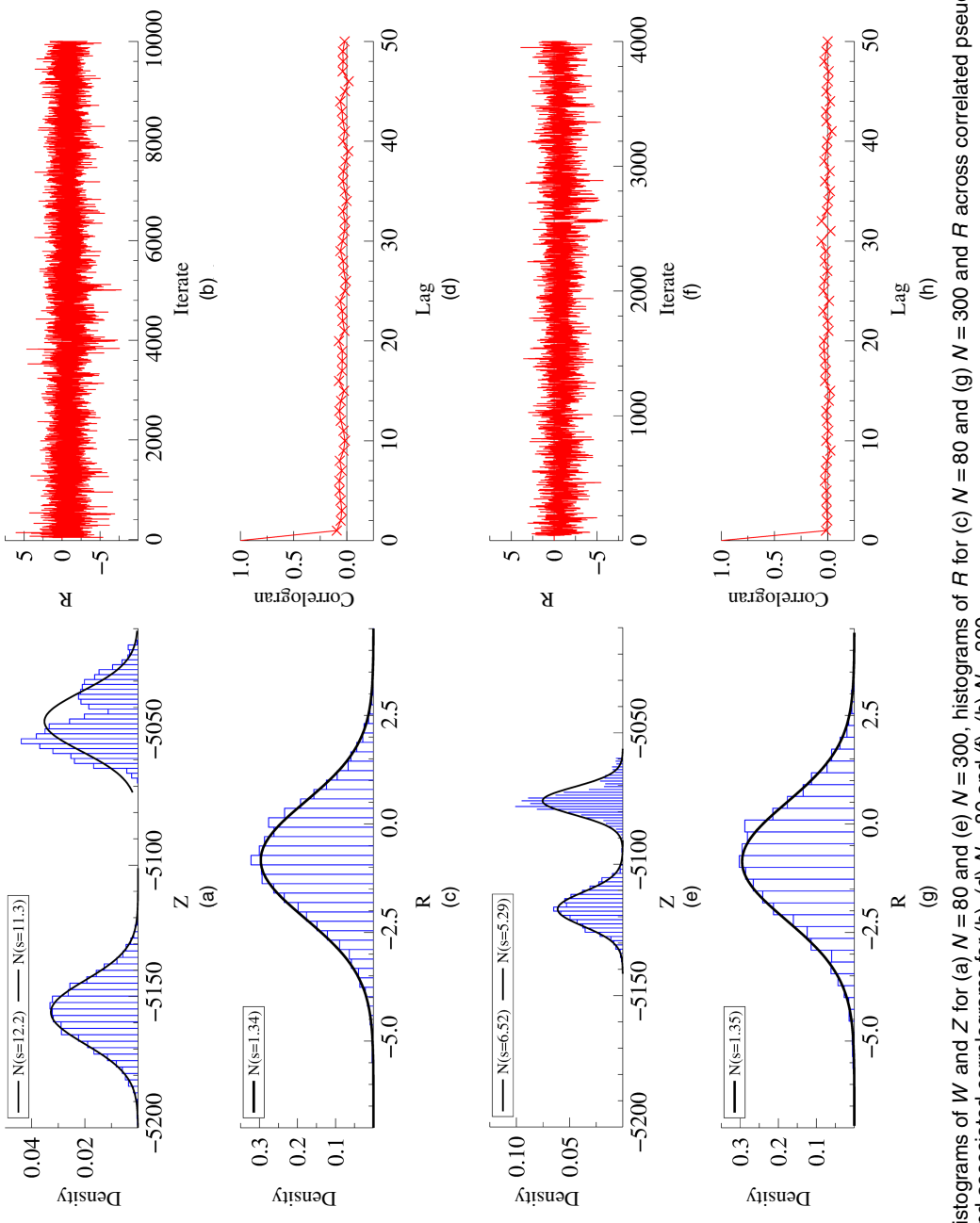

Fig. 7. Histograms of W and Z for (a) $N = 80$ and (e) $N = 300$, histograms of R for (c) $N = 80$ and (g) $N = 300$ and R across correlated pseudomarginal iterations and associated correlograms for (b), (d) $N = 80$ and (f), (h) $N = 300$

Table 4. Heston model: posterior means and standard deviations over 10000 iterations†

N	μ	ϕ	ω	χ	ρ	$\bar{\varrho}_{\mathrm{CPM}}$
$\mathbb{E}(\theta)$ $(SD(\theta))$						
80	1.258 (0.098)	0.981 (0.0027)	0.142 (0.0099)	−0.676 (0.027)	0.9975	
150	1.253 (0.098)	0.981 (0.0028)	0.142 (0.0105)	−0.672 (0.034)	0.9953	
300	1.255 (0.099)	0.981 (0.0028)	0.142 (0.0110)	−0.671 (0.032)	0.9907	
$CT(\theta)$						
80	9995	12555	13571	33794		0.276
150	19691	20256	17931	32588		0.272
300	32970	30432	35103	35505		0.281

†The computing time CT = IF × N for the correlated pseudomarginal scheme for $N = \beta\sqrt{T}$ and ρ selected such that $\kappa \approx 1.4$ at $\hat{\theta}$ and acceptance probability $\bar{\varrho}_{\mathrm{CPM}}$.

$$Y_s \sim \mathcal{N}\{\chi\hat{\gamma}_s; (1-\chi^2)\hat{\sigma}_s^{2*}\}, \tag{53}$$

$$\begin{aligned} \hat{\sigma}_s^{2*} &= \epsilon \sum_{i=1}^{I} \exp(x_i^s), \\ \hat{\gamma}_s &= \sqrt{\epsilon}\sum_{i=1}^{I} \exp(x_i^s/2)\eta_i, \end{aligned} \tag{54}$$

where $\hat{\sigma}_s^{2*}$ and $\hat{\gamma}_s$ are the Euler approximations of expressions (52). We are interested in inferring $\theta = (\mu, \upsilon, \omega, \chi)$ given $T = 4000$ daily returns $y_{1:T}$ from the Standard & Poors 500 index from August 15th, 1990, to July 3rd, 2006. We use here $I = 10$. Although the state is scalar, it is very difficult to perform inference by using standard Markov chain Mante Carlo techniques as this involves $TI = 40000$ highly correlated latent variates.

We first run the correlated pseudomarginal scheme keeping the parameter fixed at the posterior mean $\hat{\theta}$, estimated from a full correlated pseudomarginal run, and only updating the auxiliary variables. We display the histograms of $Z = \log\{\hat{p}(y_{1:T} \mid \hat{\theta}, U)\}$, $W = \log\{\hat{p}(y_{1:T} \mid \hat{\theta}, U')\}$ and $R = \log\{\hat{p}(y_{1:T} \mid \hat{\theta}, U')/\hat{p}(y_{1:T} \mid \hat{\theta}, U)\}$ in Fig. 7 for $N = 80$ and $N = 300$ using the parameters given in Table 4. We observe that R is approximately distributed according to $\mathcal{N}(-\kappa^2/2, \kappa^2)$ for $\kappa = 1.35$ in both cases. Additionally the sequence of estimates is almost uncorrelated across correlated pseudomarginal iterations.

Using $N = 300$, we first select $\psi = 0.125$ to achieve $\kappa = 1.4$ at $\hat{\theta}$. We then run the correlated pseudomarginal scheme by using a random-walk proposal for other values of N, $N = \beta\sqrt{T}$, and compute CT = IF × N. These results are summarized in Table 4. The posterior estimates are in very close agreement across the various values of N. In unreported results, we observe empirically that the dependence of CT on β for parameters $(\mu, \phi := \exp(-\upsilon), \omega, \chi)$ matches equation (47) which can be optimized, suggesting an optimal value of N around 70–80. As in the random-effects scenario, we observe on data sets of increasing length that the scaling $N = \beta\sqrt{T}$ is successful as $\mathrm{IF}_{\mathrm{CPM}}$ appears to stabilize. In this context, the pseudomarginal method is extremely expensive computationally as we need approximately $N = 20000$ to obtain a standard deviation of Z around 1 (Doucet *et al.*, 2015), our implementation taking 7 min per iteration to run on a standard desktop computer. In terms of CT, the correlated pseudomarginal scheme is approximately 100 times more efficient than the pseudomarginal scheme.

5.3. Linear Gaussian state space model

We examine empirically the performance of the correlated pseudomarginal method for multivariate state space models by using the particle filter with Hilbert sort described in algorithm 2 (Table 2) and compare it with the pseudomarginal method. Attention is restricted to a linear Gaussian state space model which allows exact calculation of the likelihood and of the log-likelihood error $Z_T(\theta, U) = \log\{\hat{p}(Y_{1:T} \mid \theta, U)/p(Y_{1:T} \mid \theta)\}$. Similar empirical results for nonlinear non-Gaussian state space models were observed.

We consider the model that was discussed in Guarniero *et al.* (2017) and Jacob *et al.* (2016) where $\{X_t; t \geqslant 1\}$ and $\{Y_t; t \geqslant 1\}$ are $\mathbb{R}^k$ valued with

$$X_1 \sim \mathcal{N}(\mathbf{0}, I_k), \qquad X_{t+1} = A_\theta X_t + V_{t+1}, \qquad Y_t = X_t + W_t, \tag{55}$$

where $V_t \sim^{\text{IID}} \mathcal{N}(\mathbf{0}_k, I_k)$, $W_t \sim^{\text{IID}} \mathcal{N}(\mathbf{0}_k, I_k)$ and $A_\theta^{i,j} = \theta^{|i-j|+1}$.

We use the transition density of $\{X_t; t \geqslant 1\}$ as proposal density within the particle filter. We investigate the variance of the error in the log-likelihood estimator $Z = \log\{\hat{p}(y_{1:T}|\theta, U)/p(y_{1:T}|\theta)\}$ by running the correlated pseudomarginal procedure holding the parameter fixed and equal to its true value $\theta = 0.4$. Next, we investigate the variance of the error in the log-likelihood ratio estimator $R = \log\{\hat{p}(y_{1:T}|\theta', U')/p(y_{1:T}|\theta')\} - Z$ where $U' \sim K_\rho(U, \cdot)$ is the proposal when $\theta' = \theta$. This is performed for various values of T, with $N = \lceil \beta T^\alpha \rceil$ and $\rho = \exp(-\psi N/T)$ for $k \in \{2, 3, 4\}$.

We shall now discuss the choice of α for state space models. In sharp contrast with random-effects models, we found empirically that there are dimension-dependent limitations to the realized correlation that can be achieved through the particle filter with Hilbert sort. In particular we found that, because of resampling, the realized correlation is limited by $\min\{1 - c_1 N^{-1/k}, 1 - c_2\delta\}$ for some constants c_1 and c_2, unless we set δ extremely small. Since the inefficiency tends to increase if we set δ too small, we balance the two terms by choosing $\delta = N^{-1/k}$, thus setting $\alpha = k/(k+1)$ for the following examples.

We run the correlated pseudomarginal chain for 1000 iterations, recording $\kappa^2 = \mathbb{V}(R)$ and $\sigma^2 = \mathbb{V}(Z)$. The values of β and ψ have been chosen so that they result in a particular target value of κ^2 as will be evident from the following tables. The asymptotic acceptance probability of the correlated pseudomarginal scheme is thus in this case given by $\varrho_{\text{CPM}}(\kappa) := \varrho_{\text{U}}(\kappa) = 2\Phi(-\kappa/2)$ whereas it is $\varrho_{\text{PM}}(\sigma) = 2\Phi(-\sigma/\sqrt{2})$ for the pseudomarginal scheme (Doucet *et al.*, 2015).

The results for $k = 2$ are reported in Table 5, where the two eigenvalues of A_θ are 0.56 and 0.24. The scaling rule proposed results in values of κ^2 which are approximately constant, remaining at values that are close to 2 for $T \geqslant 1600$. The implied acceptance probability of the

Table 5. Linear state space model: results for $k = 2$ for varying T†

T	N	$\delta = -log(\rho)$	κ^2	σ^2	$\varrho_{\text{CPM}}(\kappa)$	$\varrho_{\text{PM}}(\sigma)$
100	18	0.0216	2.59	16.3	0.42	0.004
400	46	0.0138	2.71	20.5	0.41	0.0013
1600	116	0.0087	2.01	34.1	0.48	3.6×10^{-5}
6400	294	0.0055	2.07	49.7	0.47	6.0×10^{-7}
25600	742	0.0034	1.97	105.9	0.48	3.4×10^{-13}

†State dimension $k = 2$ with $\beta = 0.854$, $\psi = 0.12$ and $\alpha = \frac{2}{3}$.

Table 6. Linear state space model: results for $k = 3$ for varying T†

T	N	$\delta = -log(\rho)$	κ^2	σ^2	$\varrho_{\mathrm{CPM}}(\kappa)$	$\varrho_{\mathrm{PM}}(\sigma)$
100	49	0.0205	3.15	13.7	0.37	0.0089
400	140	0.0147	2.97	16.6	0.39	0.0039
1600	397	0.0104	3.44	26.7	0.35	0.00025
6400	1124	0.0074	3.03	34.1	0.38	3.66×10^{-5}
25600	3181	0.0052	2.69	49.4	0.41	6.74×10^{-7}

†State dimension $k = 3$ with $\beta = 1.57$, $\psi = 0.042$ and $\alpha = \frac{3}{4}$.

CPM scheme $\varrho_{\mathrm{CPM}}(\kappa)$ therefore settles at a value that is close to 0.5. By contrast, the marginal variance σ^2 increases at the expected rate $T^{1-\alpha}$ and accordingly the acceptance probability of the corresponding pseudomarginal scheme, $\varrho_{\mathrm{PM}}(\sigma)$, is very low even for $T = 100$. Similar results are found for the case $k = 3$, which are reported in Table 6, where the eigenvalues of A_θ are (0.6605, 0.3360, 0.2035), resulting in a model with moderately high persistence. In this case we set $\alpha = \frac{3}{4}$. Although less dramatic, the implied gain of the correlated pseudomarginal method over the pseudomarginal method is substantial even for $T = 100$ and increases with T.

The full correlated pseudomarginal procedure is now implemented for $T = 400$ and $T = 6400$ when $k = 2$ and $k = 3$ by using the parameters of Tables 5 and 6. An auto-regressive proposal is employed for θ which is based on the posterior mode and the second derivative at this point (Tran *et al.*, 2016b).

The results for $k = 3$ and $T = 6400$ are shown in Fig. 8. The mixing for θ is fairly rapid for the achieved value of $\kappa = 2.26$. The empirical distributions of Z under m and $\bar{\pi}$ are plotted (Fig. 8(c)) and are close to the theoretical distributions $\mathcal{N}(-\sigma^2/2, \sigma^2)$ and $\mathcal{N}(\sigma^2/2, \sigma^2)$ respectively, where $\sigma = 7.5$. Figs 8(d)–8(f) show the draws of R, its empirical distribution and the associated correlogram arising from the correlated pseudomarginal scheme. It is clear that R is approximately distributed according to $\mathcal{N}(-\kappa^2/2, \kappa^2)$, which is overlaid, but the correlogram decays slower than for random-effect models and one-dimensional state space models. The gain over the pseudomarginal method is around σ^2, meaning that we need around 50 times as many particles in the pseudomarginal method to achieve similar results to the correlated pseudomarginal scheme. When $T = 400$, we obtained $\kappa = 1.92$ and $\sigma = 4.30$, resulting in gains over the pseudomarginal of approximately 18 fold. When $k = 2$, the gains are more impressive and are around 25 fold for $T = 400$ and 80 fold when $T = 6400$.

6. Discussion

The correlated pseudomarginal method is an extension of the pseudomarginal method using an estimator of the likelihood ratio appearing in its acceptance probability obtained by correlating estimators of its numerator and denominator. We have detailed implementations of this general idea for random-effects and state space models. For random-effects models, we have provided theory to apply this methodology efficiently and have also verified empirically its efficacy for state space models. In our examples, the computational gains over the pseudomarginal method increase with T and can be over two orders of magnitude for large data sets. The correlated pseudomarginal method is particularly useful for partially observed diffusions where sophisticated Markov chain Monte Carlo alternatives, such as particle Gibbs techniques, are inefficient.

From a theoretical point of view, in the random-effects scenario, we have obtained a result

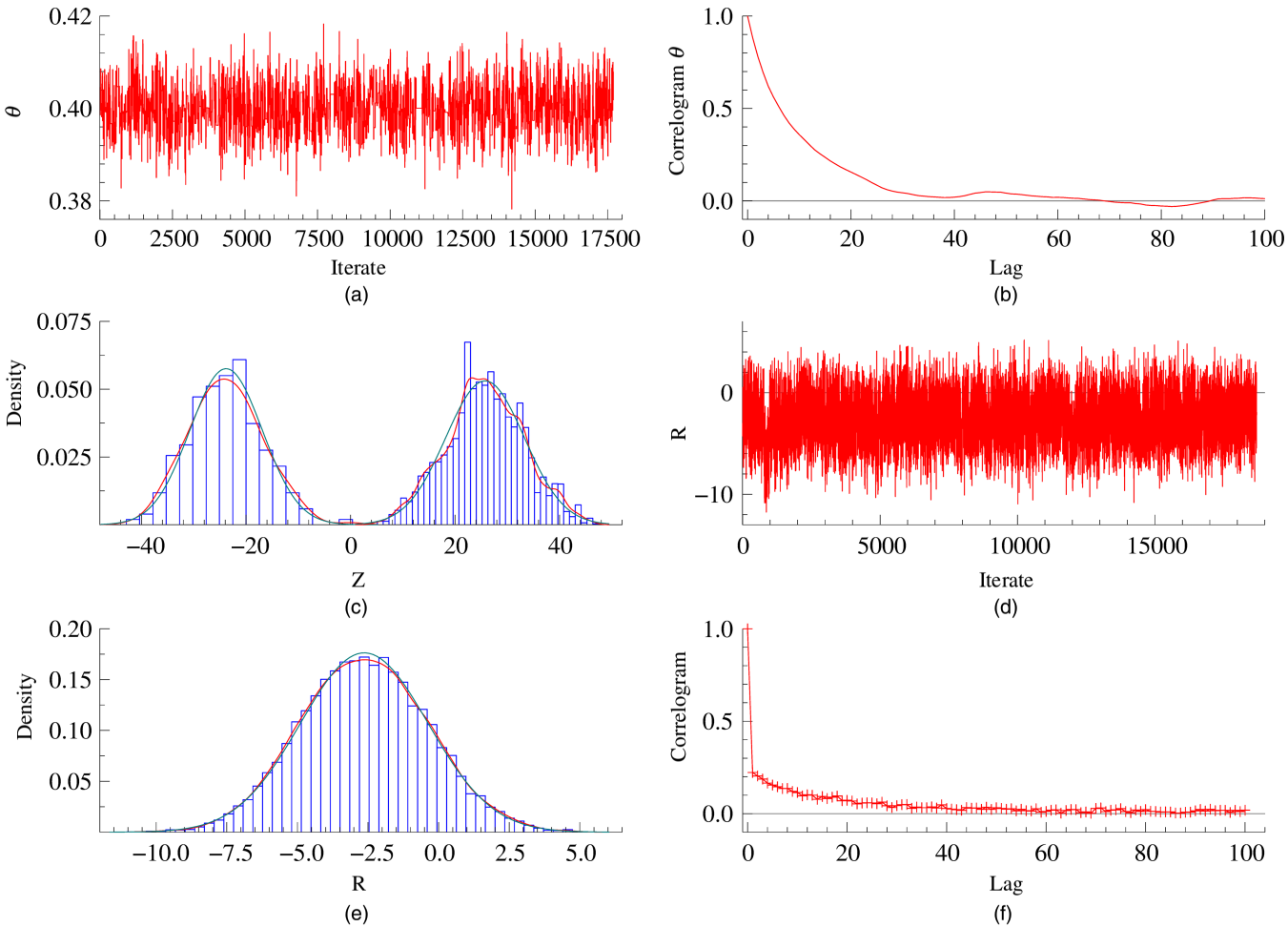

Fig. 8. The correlated pseudomarginal results for the three-dimensional state space model with $T = 6400$: (a) parameter samples θ and (b) corresponding correlogram; (c) histograms of Z arising from m and $\bar{\pi}$ (———, $s = 6.93$; ———, $s = 7.52$); (d) draws of R; (e) histogram of R (———, $s = 2.26$); (f) correlogram

suggesting that a necessary condition to ensure finiteness of the integrated auto-correlation time of the correlated pseudomarginal chain, as T increases, is to have N_T growing at least at rate $\sqrt{T}$. Our experimental results suggest that this condition is also sufficient and thus that the computational cost per iteration of the correlated pseudomarginal method is $O(T^{3/2})$ *versus* $O(T^2)$ for the pseudomarginal method. For state space models, our empirical results indicate that this scaling degrades with the state dimension k and that we need N_T to grow at rate $T^{k/(k+1)}$, leading to a computational cost per iteration of order $O(T^{(2k+1)/(k+1)})$, up to a logarithmic factor (the particle filter with Hilbert sort has computational complexity $N_T \log(N_T)$ per observation), for the correlated pseudomarginal method *versus* $O(T^2)$ for the pseudomarginal method. It would be of interest but technically very involved to establish these results rigorously.

From a methodological point of view, it is possible in the state space context to use alternatives to the Hilbert resampling sort to implement the correlated pseudomarginal algorithm (Lee, 2008; Malik and Pitt, 2011; L'Ecuyer *et al.*, 2018) and several such methods have been proposed following the first version of this work (Doucet *et al.*, 2015); see, for example Jacob *et al.* (2016) and Sen *et al.* (2018). Empirical results in Jacob *et al.* (2016) and L'Ecuyer *et al.* (2018) and our own experiments indicate that all these procedures provide roughly similar improvements over the pseudomarginal method. One direction of interest is to use the sequential randomized quasi-Monte-Carlo algorithm, proposed and analysed by Gerber and Chopin (2015), within the correlated pseudomarginal scheme by correlating the single uniform distribution that was used to randomize the quasi-Monte-Carlo grid. This is one motivation behind choosing the Hilbert sort procedure over alternative schemes, since this algorithm comes with theoretical guarantees. In a random-effects context, the use of quasi-Monte-Carlo methods has already been demonstrated to provide significant improvements (Tran *et al.*, 2016a). Finally, a sequential extension of the particle marginal MH algorithm (Andrieu *et al.*, 2010), a pseudomarginal method, has been proposed in Chopin *et al.* (2013) and it would be interesting to develop an efficient sequential version of the correlated pseudomarginal scheme.

Acknowledgements

The authors are grateful to the Joint Editor, Associate Editor, reviewers and Sebastian Schmon for their useful comments which have helped to improve the manuscript. Arnaud Doucet's research is partially supported by the Engineering and Physical Sciences Research Council, grant EP/K000276/1.

References

Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc.* B, **72**, 269–342.

Andrieu, C., Doucet, A. and Lee, A. (2012) Discussion on 'Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation', by P. Fearnhead and D. Prangle. *J. R. Statist. Soc.* B, **74**, 451–452.

Andrieu, C. and Roberts G. O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, **37**, 697–725.

Andrieu, C. and Vihola, M. (2015) Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.*, **25**, 1030–1077.

Beaumont, M. (2003) Estimation of population growth or decline in genetically monitored populations. *Genetics*, **164**, 1139–1160.

Bérard, J., Del Moral, P. and Doucet, A. (2014) A lognormal central limit theorem for particle approximations of normalizing constants. *Electron. J. Probab.*, **19**, 1–28.

Berti, P., Pratelli, L. and Rigo, P. (2006) Almost sure weak convergence of random probability measures. *Stochastics*, **78**, 91–97.

Carpenter, J., Clifford, P. and Fearnhead, P. (1999) Improved particle filter for nonlinear problems. *IEE Proc.* F, **146**, 2–7.
Ceperley, D. M. and Dewing, M. (1999) The penalty method for random walks with uncertain energies. *J. Chem. Phys.*, **110**, 9812–9820.
Chopin, N. and Gerber, M. (2017) Sequential quasi-Monte Carlo: introduction for non-experts, dimension reduction, application to partly observed diffusion processes. *Preprint arXiV:1706.05305* Ecole Nationale de la Statistique et de l'Administration, Malakoff.
Chopin, N., Jacob, P. E. and Papaspiliopoulos, O. (2013) SMC^2: an efficient algorithm for sequential analysis of state space models. *J. R. Statist. Soc.* B, **75**, 397–426.
Cotter, S. L., Roberts, G. O., Stuart, A. M. and White, D. (2013) MCMC methods for functions: modifying old algorithms to make them faster. *Statist. Sci.*, **28**, 424–446.
Crauel, H. (2003) *Random Probability Measures on Polish Spaces*. Boca Raton: CRC Press.
Dahlin, J., Lindsten, F., Kronander, J. and Schön, T. B. (2015) Accelerating pseudo-marginal Metropolis-Hastings by correlating auxiliary variables. *Preprint arXiv:1511.05483*. Department of Electrical Engineering, Linköping University, Linköping.
Del Moral, P. (2004) *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. New York: Springer.
Doornik, J. A. (2007) *Object-oriented Matrix Programming using Ox*, 3rd edn London: Timberlake Consultants.
Doucet, A., Deligiannidis, G. and Pitt, M. K. (2015) The correlated-pseudo-marginal method. *Preprint arXiv:1511.04992v2*. Department of Statistics, University of Oxford, Oxford.
Doucet, A., Jacob, P. E. and Rubenthaler, S. (2013) Derivative-free estimation of the score vector and observed information matrix with applications to state-space models. *Preprint arXiv:1304:5768*. Department of Statistics, University of Oxford, Oxford.
Doucet, A., Pitt, M. K., Deligiannidis, G. and Kohn, R. (2015) Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, **102**, 295–313.
Flury, T. and Shephard, N. (2011) Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models. *Econmetr. Theory*, **27**, 933–956.
Gentil, I. and Rémillard, B. (2008) Using systematic sampling selection for Monte Carlo solutions of Feynman-Kac equations. *Adv. Appl. Probab.*, **40**, 454–472.
Gerber, M. and Chopin, N. (2015) Sequential quasi Monte Carlo (with discussion). *J. R. Statist. Soc.* B, **77**, 509–579.
Geyer, C. J. (1992) Practical Markov chain Monte Carlo. *Statist. Sci.*, **7**, 473–483.
Guarniero, P., Johansen, A. M. and Lee, A. (2017) The iterated auxiliary particle filter. *J. Am. Statist. Ass.*, **112**, 1636–1647.
Heston, S. L. (1993) A closed-form solution for options with stochastic volatility with applications to bound and currency options. *Rev. Finan. Stud.*, **6**, 327–343.
Ionides, E. L., Breto, C. and King, A. A. (2006) Inference for nonlinear dynamical systems. *Proc. Natn. Acad. Sci. USA*, **103**, 18438–18443.
Jacob, P. E., Lindsten, F. and Schön, T. B. (2016) Coupling of particle filters. *Preprint arXiv:1606.01156*. Department of Statistics, Harvard University, Cambridge.
Johndrow, J. E., Smith, A., Pillai, N. S. and Dunson, D. B. (2016) Inefficiency of data augmentation for large sample imbalanced data. *Preprint arXiv:1605.05798*. Department of Statistical Science, Duke University, Durham.
Koop, J. C. (1972) On the derivation of expected value and variance of ratios without the use of infinite series expansions. *Metrika*, **19**, 156–170.
L'Ecuyer, P., Munger, D., Lécot, C. and Tuffin, B. (2018) Sorting methods and convergence rates for array-RQMC: some empirical comparisons. *Math. Comput. Simuln*, **143**, 191–201.
Lee, A. (2008) Towards smooth particle filters for likelihood estimation with multivariate latent variables. *MSc Thesis*. Department of Computer Science, University of British Columbia, Vancouver.
Lee, A. and Holmes, C. (2010) Discussion on 'Particle Markov chain Monte Carlo methods', by C. Andrieu, A. Doucet and R. Holenstein. *J. R. Statist. Soc.* B, **72**, 327.
Lin, L., Liu, K. F. and Sloan, J. (2000) A noisy Monte Carlo algorithm. *Phys. Rev.* D, **61**, article 074505.
Lindsten, F. and Doucet, A. (2016) Pseudo-marginal Hamiltonian Monte Carlo. *Preprint arXiv:1607.02516*. Department of Information Technology, Uppsala University, Uppsala.
Lindsten, F., Jordan, M. I. and Schön, T. B. (2014) Particle Gibbs with ancestor sampling. *J. Mach. Learn. Res.*, **15**, 2145–2184.
Malik, S. and Pitt, M. K. (2011) Particle filters for continuous likelihood evaluation and maximisation. *J. Econmetr.*, **165**, 190–209.
Murray, I. and Graham, M. M. (2016) Pseudo-marginal slice sampling. In *Proc. 19th Conf. Artificial Intelligence and Statistics, Cadiz, May 9th–11th* (eds A. Gretton and C. P. Robert), pp. 911–919.
Nicholls, G. K., Fox, C. and Watt, A. M. (2012) Coupled MCMC with a randomized acceptance probability. Department of Statistics, University of Oxford, Oxford.
Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007) A general framework for the parametrization of hierarchical models. *Statist. Sci.*, **22**, 59–73.

Pitt, M. K., Silva, R., Giordani, P. and Kohn, R. (2012) On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Econmetr.*, **171**, 134–151.
Roberts, G. O., Gelman, A. and Gilks, W. R. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, **7**, 110–120.
Sen, D., Thiery, A. H. and Jasra, A. (2018) On coupling particle filter trajectories. *Statist. Comput.*, **28**, 461–475.
Sherlock, C., Thiery, A., Roberts, G. O. and Rosenthal, J. S. (2015) On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.*, **43**, 238–275.
Titsias, M. K. and Papaspiliopoulos, O. (2018) Auxiliary gradient-based sampling algorithms. *J. R. Statist. Soc.* B, **80**, 749–767.
Tran, M.-N., Kohn, R., Quiroz, M. and Villani, M. (2016a) Block-wise pseudo-marginal Metropolis–Hastings. *Preprint arXiv:1603.02485*. Discipline of Business Analytics, University of Sydney, Sydney.
Tran, M.-N., Pitt, M. K. and Kohn, R. (2016b) Adaptive Metropolis-Hastings sampling using reversible dependent mixture proposals. *Statist. Comput.*, **26**, 361–381.
van der Vaart, A. W. (2000) *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material to "The correlated pseudo-marginal method"'.