Gunawan, David; Dang, Khue-Dung; Quiroz, Matias; Kohn, Robert; Tran, Minh-Ngoc

**Working Paper**

# Subsampling Sequential Monte Carlo for static Bayesian models

Sveriges Riksbank Working Paper Series, No. 371

**Provided in Cooperation with:**
Central Bank of Sweden, Stockholm

*Suggested Citation:* Gunawan, David; Dang, Khue-Dung; Quiroz, Matias; Kohn, Robert; Tran, Minh-Ngoc (2019) : Subsampling Sequential Monte Carlo for static Bayesian models, Sveriges Riksbank Working Paper Series, No. 371, Sveriges Riksbank, Stockholm

This Version is available at:
https://hdl.handle.net/10419/215449

SVERIGES RIKSBANK
WORKING PAPER SERIES 371


# Subsampling Sequential Monte Carlo for Static Bayesian Models

*David Gunawan, Khue-Dung Dang, Matias Quiroz, Robert Kohn and Minh-Ngoc Tran*

April 2019

WORKING PAPERS ARE OBTAINABLE FROM

www.riksbank.se/en/research
Sveriges Riksbank • SE-103 37 Stockholm
Fax international: +46 8 21 05 31
Telephone international: +46 8 787 00 00

The Working Paper series presents reports on matters in
the sphere of activities of the Riksbank that are considered
to be of interest to a wider public.
The papers are to be regarded as reports on ongoing studies
and the authors will be pleased to receive comments.

# Subsampling Sequential Monte Carlo for Static Bayesian Models

David Gunawan[1,2], Khue-Dung Dang[1,2], Matias Quiroz[1,2,3], Robert Kohn[1,2] and Minh-Ngoc Tran[2,4]

Sveriges Riksbank Working Paper Series
No. 371
April 2019

**Abstract**

We show how to speed up Sequential Monte Carlo (SMC) for Bayesian inference in large data problems by data subsampling. SMC sequentially updates a cloud of particles through a sequence of distributions, beginning with a distribution that is easy to sample from such as the prior and ending with the posterior distribution. Each update of the particle cloud consists of three steps: reweighting, resampling, and moving. In the move step, each particle is moved using a Markov kernel and this is typically the most computationally expensive part, particularly when the dataset is large. It is crucial to have an efficient move step to ensure particle diversity. Our article makes two important contributions. First, in order to speed up the SMC computation, we use an approximately unbiased and efficient annealed likelihood estimator based on data subsampling. The subsampling approach is more memory efficient than the corresponding full data SMC, which is an advantage for parallel computation. Second, we use a Metropolis within Gibbs kernel with two conditional updates. A Hamiltonian Monte Carlo update makes distant moves for the model parameters, and a block pseudo-marginal proposal is used for the particles corresponding to the auxiliary variables for the data subsampling. We demonstrate the usefulness of the methodology for estimating three generalized linear models and a generalized additive model with large datasets.

**Keywords.** Hamiltonian Monte Carlo, Large datasets, Likelihood annealing
**JEL classification.** C11, C15, C55.

[1]:*School of Economics, UNSW Business School, University of New South Wales.* [2]:*ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).* [3]:*Research Division, Sveriges Riksbank. Email: quiroz.matias@gmail.com.* [4]:*Discipline of Business Analytics, University of Sydney.* The opinions expressed in this article are the sole responsibility of the authors and should not be interpreted as reflecting the views of Sveriges Riksbank.

# 1 Introduction

The main problem of Bayesian inference is to estimate the expectation of a function of the unknown parameters with respect to their posterior distribution. This is typically resolved by obtaining a simulation approximation of the expectation using samples from the posterior distribution. Exact approaches such as Markov Chain Monte Carlo (MCMC) (Brooks et al., 2011) methods have been the main methods used for sampling from complex posterior distributions. Despite this, MCMC methods have some notable drawbacks and limitations. One drawback often overlooked by practitioners when fitting complex models, is the failure to converge caused by poorly mixing chains. While Hamiltonian Monte Carlo (Neal, 2011, HMC) is a remedy in many cases, it can be notoriously difficult to tune. Limitations of MCMC methods include the difficulties of assessing convergence, parallelizing the computation, and estimating the marginal likelihood efficiently from MCMC output, the latter being useful for model selection (Kass and Raftery, 1995). Sequential Monte Carlo (see Doucet et al., 2001 for an introductory overview) methods provide an alternative exact simulation approach to MCMC methods and overcome some of their drawbacks. Moreover, in contrast to MCMC methods, SMC can provide online updates of the parameters as data is collected, which is particularly useful for dynamic (time-varying parameters) models. SMC is also useful for static (non time-varying parameters) models (Chopin, 2002; Del Moral et al., 2006), and can in such cases more easily explore multimodal posterior distributions than MCMC.

Despite the advantages of SMC, it is remarkably less used than MCMC for static models. One possible explanation is that, while amenable to computer parallelization, it is still very computationally expensive and particularly so for large datasets. Another obstacle caused by large datasets is that they prevent efficient computer parallelization of SMC, as the full dataset needs to be available for each worker which is infeasible as it consumes too much Random-Access Memory (RAM). We propose an efficient data subsampling approach which significantly reduces both the computational cost of the algorithm and the memory requirements when parallelizing: see Section 3.6 for a detailed explanation of the latter. Our approach utilizes the methods previously developed for Subsampling MCMC (Quiroz et al., 2018a; Dang et al., 2018) and places them within the SMC framework. See Quiroz et al. (2018c) for an introductory text in Subsampling MCMC.

In a Bayesian context, SMC is a method to traverse a cloud of particles through a sequence of distributions, with the initial distribution both easy to sample from and to evaluate, while the final distribution is the posterior distribution. The cloud of particles at step $p$ is an estimate of the $p$th distribution in the sequence. The

particles consist of the unknown parameters and any additional latent variables that are part of the model. The evolution of the particle cloud consists of three steps: reweighting, resampling and moving. Of these, the first two steps are common to all SMC schemes and are straightforward. The move step is the most expensive and is critical to ensure that the particle cloud is representative of the distribution it aims to estimate.

To the best of our knowledge, data subsampling has not been explored in SMC for static models. While Wang et al. (2019) term their algorithm Subsampling SMC, their approach is distinct as they combine data annealing and likelihood annealing, whereas we use data subsampling to estimate the likelihood. Specifically, we consider a likelihood annealing approach in which we estimate the annealed likelihood efficiently using an approximately unbiased estimator. Likelihood estimates for SMC in a non-subsampling context have been used in Duan and Fulop (2015), who propose to estimate the likelihood unbiasedly using a particle filter in a time series state space model application. However, Duan and Fulop (2015) use a random walk MCMC kernel for the move step of the model parameters, which is inefficient in high dimensions and we now turn to this issue.

The literature has focused on accelerating SMC algorithms by designing efficient MCMC kernels for the move step to achieve efficient particle diversity. The efficiency concept here is the ability of the MCMC kernel to generate distant proposals which have a high probability of being accepted, so as to move the particle efficiently using as few iterations of the kernel as possible. Various approaches exist to achieve this. For example, adaptive SMC adapts the tuning parameters of the kernel to improve its efficiency (Jasra et al., 2011; Fearnhead and Taylor, 2013; Buchholz et al., 2018). A different approach is explored in South et al. (2016), who use SMC with a flexible independent proposal based on copulas models. Finally, the use of derivatives to construct efficient proposals through the Metropolis Adjusted Langevin Algorithm (Roberts and Stramer, 2002, MALA) have been explored (Sim et al., 2012; South et al., 2017). It is now well-known that the MALA proposal is a special case of the more general proposal utilizing Hamiltonian dynamics proposed in Duane et al. (1987) (see Neal 2011; Betancourt 2017 for introductory overviews). Although South et al. (2017) mention HMC in their introduction, they only consider MALA in their paper and show how neural networks can be applied to adaptively choose its tuning parameters. Daviet (2016) considers HMC proposals for particle diversity, however, HMC is painfully slow for very large datasets and therefore this approach does not scale well in the number of data observations.

We propose data subsampling to improve the computational efficiency and a HMC type of kernel for efficient particle diversity, while leveraging on data subsampling

in order to achieve scalability in the number of observations. As a by-product, data subsampling lowers the memory requirements of the algorithm (see Section 3.6), making it amenable for computer parallelism on very large datasets. Our framework combines that of Duan and Fulop (2015) for carrying out SMC with an estimated likelihood, Quiroz et al. (2018a) for estimating the likelihood and controlling the error in the target density and Dang et al. (2018) for constructing efficient proposals for high-dimensional targets in a subsampling context.

Our article is organized as follows. Section 2 reviews sequential Monte Carlo for static models. Section 3 outlines our methodology, which scales SMC to large datasets and high-dimensional models by combining efficient data subsampling and Hamiltonian Monte Carlo to sample from an accurate approximate target density. Section 4 applies our methodology in a variety of settings for both real and simulated data and shows that it gives accurate estimates of both the posterior density and the marginal likelihood. Section 5 concludes.

# 2 Sequential Monte Carlo

## 2.1 SMC for static Bayesian models

Denote the observed data $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top)^\top$, with $\boldsymbol{y}_k \in \mathcal{Y} \subset \mathbb{R}^{d_{\boldsymbol{y}}}$. Let $\boldsymbol{\theta}$ be the vector of unknown parameters, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^{d_{\boldsymbol{\theta}}}$, with $p(\boldsymbol{\theta})$ and $p(\boldsymbol{y}|\boldsymbol{\theta})$ the prior and likelihood. In Bayesian inference, the uncertainty about the unobserved $\boldsymbol{\theta}$ is specified by the posterior density $\pi(\boldsymbol{\theta})$, which by Bayes' theorem is

$$\pi(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})}, \quad \text{where} \quad p(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{1}$$

is the marginal likelihood, also known as the evidence, and is often used for Bayesian model selection.

The main problem in Bayesian inference is to estimate the posterior expectation of a function of $\boldsymbol{\theta}$,

$$\mathbb{E}_\pi(\varphi(\boldsymbol{\theta})) = \int_{\Theta} \varphi(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{2}$$

In simulation based inference, this is typically achieved by sampling from (1) and computing (2) by Monte Carlo integration. A second problem is to compute the marginal likelihood in (1). However, it is well known that standard Monte Carlo integration is very inefficient for this task.

In a Bayesian context, SMC (Doucet et al., 2001) is a collection of methods that can approximately sample from (1) and in addition provide an efficient estimator of

the marginal likelihood. Their early use was for inference in dynamic systems (Gordon et al., 1993; Liu and Chen, 1998; Gilks and Berzuini, 2001), but more recently their potential has been realized for static (non-dynamic) models (Chopin, 2002; Del Moral et al., 2006), in which they generalize importance sampling approaches such as Annealed Importance Sampling (Neal, 2001, AIS).

SMC specifies a sequence of $P$ densities, connecting the density of the prior $p(\boldsymbol{\theta})$ to the density of the posterior $\pi(\boldsymbol{\theta})$ in (1). The sequence is usually obtained either through data annealing (Chopin, 2002), in which the data is introduced sequentially, or temperature annealing (Neal, 2001), in which the likelihood is tempered $p\left(\boldsymbol{y}|\boldsymbol{\theta}\right)^{a_p}$ with $a_0 = 0 < a_1 < \cdots < a_P = 1$. Our article considers the latter and we note at the outset that we propose to estimate the tempered likelihood $p\left(\boldsymbol{y}|\boldsymbol{\theta}\right)^{a_p}$ by data subsampling, see Section 3. The tempered posterior is

$$\pi_p(\boldsymbol{\theta}) = \frac{\eta_p(\boldsymbol{\theta})}{Z_p}, \text{ where } \eta_p(\boldsymbol{\theta}) = p\left(\boldsymbol{y}|\boldsymbol{\theta}\right)^{a_p} p(\boldsymbol{\theta}) \quad \text{and} \quad Z_p = \int_{\boldsymbol{\Theta}} p\left(\boldsymbol{y}|\boldsymbol{\theta}\right)^{a_p} p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (3)$$

SMC proceeds by sampling a set of $M$ particles from the prior $p(\boldsymbol{\theta})$ and traverses them through the sequence of densities $\pi_p(\boldsymbol{\theta}), p = 1, \ldots, P$ by, for each $p$, (i) reweighting, (ii) resampling and (iii) moving the particles. At the final $p = P$, the particles are a (weighted) sample from $\pi(\boldsymbol{\theta})$. We now discuss this in more detail.

The initial particle cloud $\left\{\boldsymbol{\theta}_{1:M}^{(0)}, W_{1:M}^{(0)}\right\}$ is obtained by generating the $\left\{\boldsymbol{\theta}_{1:M}^{(0)}\right\}$ from $p\left(\boldsymbol{\theta}\right)$, and giving them equal weight, i.e., $W_{1:M}^{(0)} = 1/M$. The weighted particles $\left\{\boldsymbol{\theta}_{1:M}^{(p-1)}, W_{1:M}^{(p-1)}\right\}$ at the $(p-1)$st stage, $p = 1, \ldots, P$, are (weighted) samples from $\pi_{p-1}\left(\boldsymbol{\theta}\right)$. At the $p$th stage, the transition from $\pi_{p-1}\left(\boldsymbol{\theta}\right)$ to $\pi_p\left(\boldsymbol{\theta}\right)$ is obtained by the *reweighting step*,

$$w_i^{(p)} = W_i^{(p-1)} \frac{\eta_p\left(\boldsymbol{\theta}_i^{(p-1)}\right)}{\eta_{p-1}\left(\boldsymbol{\theta}_i^{(p-1)}\right)} = W_i^{(p-1)} p\left(\boldsymbol{y}|\boldsymbol{\theta}_i^{(p-1)}\right)^{a_p - a_{p-1}},$$

and then normalizing $W_i^{(p)} = w_i^{(p)} / \sum_{i'=1}^{M} w_{i'}^{(p)}$. The reweighting will, when $p$ increases, assign vanishingly smaller weights to particles which are unlikely under the tempered likelihood, causing the so-called particle degeneracy problem, in which the weight mass is concentrated only on a small fraction of the particles, causing a small effective sample size (explained in Section 2.2). This is resolved by the *resampling step*, in which the particles $\boldsymbol{\theta}_{1:M}^{(p)}$ are sampled with a probability equal to their normalized weights $W_{1:M}^{(p)}$, and subsequently setting $W_{1:M}^{(p)} = 1/M$. While this ensures that the particles with small weights are eliminated, it causes so-called particle depletion because the particles with large weights may duplicate. This is resolved by the

*move step*, in which a $\pi_p$-invariant Markov kernel $K_p$ is applied to move each of the particles $R$ steps. Notice that since a particle at stage $p$ is approximately distributed as $\pi_p(\theta)$ and $K_p$ is $\pi_p$-invariant, no burn-in period is required as in MCMC methods, where often a very large number of burn-in iterations are required. Finally, we note that the algorithm is easy to parallelize with respect to the particle dimension, because the computations required for each particle do not depend on that of the other particles. Thus, provided that $p(\boldsymbol{y}|\boldsymbol{\theta})$ can be computed at each worker without storage issues, it is straightforward to implement the parallel version.

Del Moral et al. (2006) provide consistency results and central limit theorems for estimating (2) based on the SMC output.

## 2.2 Statistical efficiency of SMC

The statistical efficiency of the $p$th stage of the SMC reweighting part is measured through the Effective Sample Size (ESS) defined as (for example Liu, 2001)

$$\mathrm{ESS}_p = \left( \sum_{i=1}^{M} \left( W_i^{(p)} \right)^2 \right)^{-1}.$$

The ESS varies between 1 and $M$, where a low value of ESS indicates that the weights are concentrated only on a few particles. A common problem in SMC is the choice of tempering sequence $\{a_p, p = 1, \ldots, P\}$, which has a substantial impact on ESS and therefore needs careful choice. We follow Del Moral et al. (2012) and choose the tempering sequence adaptively to ensure a sufficient level of particle diversity by selecting the next value of $a_p$ such that ESS stays close to some target value $\mathrm{ESS}_{\mathrm{target}}$. We do so by evaluating the ESS over a grid points $a_{1:G,p}$ of potential values of $a_p$ for a given $p$ and select $a_p$ as that value of $a_{g,p}$, $g = 1, \ldots, G$, whose ESS is closest to $\mathrm{ESS}_{\mathrm{target}}$. Throughout our article $\mathrm{ESS}_{\mathrm{target}} = 0.8M$.

For this adaptive choice of tempering sequence, Beskos et al. (2016) establish consistency results and central limit theorems for estimating (2) based on the SMC output.

## 2.3 Marginal likelihood estimation with SMC

The marginal likelihood $p(\boldsymbol{y})$ is often used in the Bayesian literature to compare models by their posterior model probabilities (Kass and Raftery, 1995). An advantage of SMC is that it automatically produces an estimate of $p(\boldsymbol{y})$.

Using the notation of Section 2.1, we note that $Z_P = p(\boldsymbol{y})$ and $Z_0 = 1$,

$$p(\boldsymbol{y}) = \prod_{p=1}^{P} \frac{Z_p}{Z_{p-1}} \text{ with } \frac{Z_p}{Z_{p-1}} = \int \left( \frac{\eta_p(\boldsymbol{\theta})}{\eta_{p-1}(\boldsymbol{\theta})} \right) \pi_{p-1}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}.$$

Because the particle cloud $\left\{ \boldsymbol{\theta}_{1:M}^{(p-1)}, W_{1:M}^{(p-1)} \right\}$ at the $(p-1)$st stage is an approximate sample from $\pi_{p-1}(\boldsymbol{\theta})$, the ratio above is estimated by

$$\widehat{\frac{Z_p}{Z_{p-1}}} = \sum_{i=1}^{M} W_i^{(p-1)} \frac{\eta_p\left(\boldsymbol{\theta}_i^{(p-1)}\right)}{\eta_{p-1}\left(\boldsymbol{\theta}_i^{(p-1)}\right)}.$$

The estimate of the marginal likelihood is then

$$\widehat{p}(\boldsymbol{y}) = \prod_{p=1}^{P} \widehat{\frac{Z_{a_p}}{Z_{a_{p-1}}}}. \tag{4}$$

# 3 Methodology

## 3.1 Sequence of target densities

Suppose that $\boldsymbol{y}_k, k = 1, ..., n$ are independent given $\boldsymbol{\theta}$ so that the likelihood and log-likelihood can be written as

$$L(\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\boldsymbol{y}_k|\boldsymbol{\theta}) \text{ and } \ell(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ell_k(\boldsymbol{\theta}), \tag{5}$$

where $\ell_k(\boldsymbol{\theta}) = \log p(\boldsymbol{y}_k|\boldsymbol{\theta})$. We are concerned with the case where the log-likelihood is computationally very costly, because $n$ is so large that repeatedly computing this sum is impractical.

Quiroz et al. (2018a) propose to subsample $m$ observations and estimate $\ell(\boldsymbol{\theta})$ by $\widehat{\ell}_m(\boldsymbol{\theta})$ in (9) and subsequently estimate $L(\boldsymbol{\theta})$ by

$$\widehat{L}(\boldsymbol{\theta}) = \exp\left( \widehat{\ell}_m(\boldsymbol{\theta}) - \frac{1}{2} \widehat{\sigma}_m^2(\boldsymbol{\theta}) \right), \tag{6}$$

where $\widehat{\sigma}_m^2(\boldsymbol{\theta})$ is an estimate of $\sigma^2(\boldsymbol{\theta}) = \mathbb{V}\left( \widehat{\ell}_m(\boldsymbol{\theta}) \right)$. The motivation for (6) is that $\exp\left( \widehat{\ell}_m(\boldsymbol{\theta}) - \sigma^2(\boldsymbol{\theta})/2 \right)$ is unbiased for $L(\theta)$ when $\widehat{\ell}_m(\boldsymbol{\theta})$ is normal (Ceperley and Dewing, 1999). Otherwise, if it is not normal or if the variance $\sigma^2(\boldsymbol{\theta})$ is estimated, it is unbiased for a perturbed likelihood $L_{(m,n)}(\boldsymbol{\theta})$. Quiroz et al. (2018a) show that when using the control variate in Section 3.2 in the estimator $\widehat{\ell}_m(\boldsymbol{\theta})$, the fractional

error of the perturbed likelihood is

$$\left|\frac{L_{(m,n)}(\boldsymbol{\theta}) - L(\boldsymbol{\theta})}{L(\boldsymbol{\theta})}\right| = \mathcal{O}\left(\frac{1}{nm^2}\right).$$

Our approach is based on extending the target at the $p$th density, i.e. $\pi_p(\boldsymbol{\theta})$ in (3), to include the set of subsampling indices $\boldsymbol{u} = (u_1, \ldots, u_m)$, where $\boldsymbol{u} \in \mathcal{U} \subset \{1, \ldots, n\}^m$ when sampling data observations with replacement. Let $\widehat{L}_p(\boldsymbol{\theta})$ be an estimator of the tempered likelihood $L(\boldsymbol{\theta})^{a_p}$. Similar to Quiroz et al. (2018a), we can unbiasedly estimate $a_p\ell(\boldsymbol{\theta})$ with $a_p\widehat{\ell}(\boldsymbol{\theta})$, and since $\mathbb{V}\left(a_p\widehat{\ell}(\boldsymbol{\theta})\right) = a_p^2\sigma^2(\boldsymbol{\theta})$ and motivated by (6), we propose the annealed likelihood estimator

$$\widehat{L}_p\left(\boldsymbol{\theta}\right) = \exp\left(a_p\widehat{\ell}_m(\boldsymbol{\theta}) - \frac{1}{2}a_p^2\widehat{\sigma}_m^2(\boldsymbol{\theta})\right). \tag{7}$$

The extended target at the $p$th density is

$$\overline{\pi}_p\left(\boldsymbol{\theta}, \boldsymbol{u}\right) \propto \widehat{L}_p(\boldsymbol{\theta}) p\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{u}\right) = \exp\left(a_p\widehat{\ell}_m(\boldsymbol{\theta}) - \frac{1}{2}a_p^2\widehat{\sigma}_m^2(\boldsymbol{\theta})\right) p\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{u}\right), \tag{8}$$

where $p(\boldsymbol{u})$ is the density of $\boldsymbol{u}$ (or, more strictly, a probability mass function since $\boldsymbol{u}$ is discrete). At the final annealing step, (8) becomes $\overline{\pi}_P\left(\boldsymbol{\theta}, \boldsymbol{u}\right) \propto \widehat{L}\left(\boldsymbol{\theta}\right) p(\boldsymbol{\theta}) p(\boldsymbol{u})$, which is the target considered in Quiroz et al. (2018a). Quiroz et al. (2018a) show that the perturbed marginal density for $\boldsymbol{\theta}$, $\pi_{(m,n)}(\boldsymbol{\theta}) = \int_{\boldsymbol{\mathcal{U}}} \overline{\pi}_P\left(\boldsymbol{\theta}, \boldsymbol{u}\right) d\boldsymbol{u}$ converges in the total variation metric to $\pi(\boldsymbol{\theta})$ at the rate $\mathcal{O}\left(1/(nm^2)\right)$. Hence, our proposed approach is approximate but can be very accurate while also scaling well with respect to the subsample size. For example, if we take $m = \mathcal{O}(\sqrt{n})$, then by Quiroz et al. (2018a, Part (i) of Theorem 1)

$$\int_{\boldsymbol{\Theta}} \left|\pi_{(m,n)}(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta})\right| d\boldsymbol{\theta} = \mathcal{O}\left(\frac{1}{n^2}\right).$$

Moreover, suppose that $\varphi(\boldsymbol{\theta})$ is a scalar function with finite second moment. Then, by Quiroz et al. (2018a, Part (ii) of Theorem 1)

$$\left|\mathbb{E}_{\pi_{(m,n)}}\left(\varphi(\boldsymbol{\theta})\right) - \mathbb{E}_{\pi}\left(\varphi(\boldsymbol{\theta})\right)\right| = \mathcal{O}\left(\frac{1}{n^2}\right).$$

This gives our algorithm the theoretical guarantees of converging at a very fast rate to the truth as $n$ increases, both with respect to the posterior density (as measured by total variation) and with respect to the posterior moments. We confirm empirically that we get very accurate inference in our application in Section 4, even for a very small $m$ relative to $n$.

The next section describes the approach in Quiroz et al. (2018a) for obtaining efficient estimators of the log-likelihood. Section 3.3 describes the reweighting and resampling steps. Section 3.4 describes the Markov move step. Section 3.5 shows how to estimate the marginal likelihood. Finally, Section 3.6 outlines the memory advantage of our method for parallel computation compared to standard (non-subsampling) SMC. Algorithm 2 summarizes our approach.

## 3.2 Efficient estimator of the log-likelihood

Quiroz et al. (2018a) propose to estimate $\ell(\boldsymbol{\theta})$ in (5) by the unbiased difference estimator,

$$\widehat{\ell}_m(\boldsymbol{\theta}) = \sum_{k=1}^{n} q_k(\boldsymbol{\theta}) + \frac{n}{m}\sum_{i=1}^{m} \ell_{u_j}(\boldsymbol{\theta}) - q_{u_j}(\boldsymbol{\theta}), \quad u_j \in \{1,\ldots,n\} \text{ iid}, \tag{9}$$

where

$$\Pr(u_j = k) = \frac{1}{n} \text{ for all } k = 1,\ldots,n \text{ and } j = 1,\ldots,m,$$

and $q_k(\theta) \approx \ell_k(\theta)$ are control variates. The estimator is based on writing

$$\ell\left(\boldsymbol{\theta}\right) = \sum_{k=1}^{n} q_k\left(\boldsymbol{\theta}\right) + \sum_{k=1}^{n} d_k\left(\boldsymbol{\theta}\right) = q\left(\boldsymbol{\theta}\right) + d\left(\boldsymbol{\theta}\right),$$

with $d_k\left(\boldsymbol{\theta}\right) = \ell_k\left(\boldsymbol{\theta}\right) - q_k\left(\boldsymbol{\theta}\right)$, $q\left(\boldsymbol{\theta}\right) = \sum_k q_k\left(\boldsymbol{\theta}\right)$, and $d\left(\boldsymbol{\theta}\right) = \sum_k d_k\left(\boldsymbol{\theta}\right)$. The last term on the right hand side of (9) is an unbiased estimator of $d(\boldsymbol{\theta})$. We now discuss a choice of control variates due to Bardenet et al. (2017), which computes $q(\boldsymbol{\theta})$ in $\mathcal{O}(1)$ time. Hence, the cost of computing the estimator is $\mathcal{O}(m)$ and we can take $m = \mathcal{O}(\sqrt{n})$ in order to achieve the convergence rates $\mathcal{O}(1/n^2)$ for both the perturbed density and its moments as discussed in Section 3.1.

Let $\overline{\boldsymbol{\theta}}$ be some posterior location estimate of $\boldsymbol{\theta}$, for example the mean, obtained from a current particle cloud from $\overline{\pi}_p\left(\boldsymbol{\theta}, \boldsymbol{u}\right)$. A second order Taylor series expansion of the log-density around $\overline{\boldsymbol{\theta}}$ is

$$\ell_k\left(\boldsymbol{\theta}\right) = \ell_k\left(\overline{\boldsymbol{\theta}}\right) + \nabla_{\boldsymbol{\theta}}\ell_k\left(\overline{\boldsymbol{\theta}}\right)^{\top}\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right) + \frac{1}{2}\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right)^{\top}\left(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^{\top}}\ell_k\left(\overline{\boldsymbol{\theta}}\right)\right)\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right) + o\left(||\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}||\right),$$

and we therefore approximate $\ell_k(\boldsymbol{\theta})$ by

$$q_k\left(\boldsymbol{\theta}\right) = \ell_k\left(\overline{\boldsymbol{\theta}}\right) + \nabla_{\boldsymbol{\theta}}\ell_k\left(\overline{\boldsymbol{\theta}}\right)^{\top}\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right) + \frac{1}{2}\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right)^{\top}\left(\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^{\top}}\ell_k\left(\overline{\boldsymbol{\theta}}\right)\right)\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right),$$

and $o(\delta)$ denotes the small order of $\delta$, meaning $o(\delta)/\delta \to 0$ as $\delta \to 0$. Then,

$$q(\boldsymbol{\theta}) \;=\; A\left(\overline{\boldsymbol{\theta}}\right) + B\left(\overline{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right) + \frac{1}{2}\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right)^{\top} C\left(\overline{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right),$$

where

$$A\left(\overline{\boldsymbol{\theta}}\right) = \sum_k \ell_k\left(\overline{\boldsymbol{\theta}}\right), B\left(\overline{\boldsymbol{\theta}}\right) = \sum_k \nabla_{\boldsymbol{\theta}} \ell_k\left(\overline{\boldsymbol{\theta}}\right)^{\top} \text{ and } C\left(\overline{\boldsymbol{\theta}}\right) = \sum_k \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^{\top}} \ell_k\left(\overline{\boldsymbol{\theta}}\right).$$

Note that the sums $A\left(\overline{\boldsymbol{\theta}}\right)$, $B\left(\overline{\boldsymbol{\theta}}\right)$, and $C\left(\overline{\boldsymbol{\theta}}\right)$ are computed only once at every stage of the SMC, regardless of the number of particles. Then, for each particle, estimating $d(\boldsymbol{\theta})$ by $\widehat{d}_m(\boldsymbol{\theta}) = (n/m)\sum_j d_{u_j}(\boldsymbol{\theta})$ is computed in $\mathcal{O}(m)$ time and so is (9) because $q(\boldsymbol{\theta})$ is $\mathcal{O}(1)$. We can estimate $\sigma^2(\boldsymbol{\theta}) = \mathbb{V}\left(\widehat{\ell}_m(\boldsymbol{\theta})\right)$ by

$$\widehat{\sigma}^2_m(\boldsymbol{\theta}) = \frac{n^2}{m^2}\sum_{j=1}^{m}\left(d_{u_j}(\boldsymbol{\theta}) - \overline{d}_{\boldsymbol{u}}(\boldsymbol{\theta})\right)^2,$$

where $\overline{d}_{\boldsymbol{u}}(\boldsymbol{\theta})$ denotes the mean of the $d_{u_j}$ for the sample $\boldsymbol{u} = (u_1, ..., u_m)$. We note that $\widehat{\sigma}^2_m(\boldsymbol{\theta})$ comes at virtually no cost since it involves terms that are already evaluated when computing $\widehat{d}_m(\boldsymbol{\theta})$.

Finally, we note that the variance of $a_p^2\widehat{\sigma}^2_m(\boldsymbol{\theta})$ is much smaller than that of $\widehat{\sigma}^2_m(\boldsymbol{\theta})$ for small $a_p$ $(0 \le a_p \le 1)$. We can then consider a less accurate and faster control variate obtained using only a first order Taylor series expansion. We experiment with this in Section 4 and find that our approach is robust to a less accurate control variate.

### 3.3 The reweighting and resampling steps

The initial particle cloud is now $\left\{\boldsymbol{\theta}^{(0)}_{1:M}, \boldsymbol{u}^{(0)}_{1:M}, W^{(0)}_{1:M}\right\}$, obtained by generating the $\left\{\boldsymbol{\theta}^{(0)}_{1:M}, \boldsymbol{u}^{(0)}_{1:M}\right\}$ from $p(\boldsymbol{\theta})$ and $p(\boldsymbol{u})$, and assigning equal weights, i.e., $W^{(0)}_{1:M} = 1/M$. The weighted particles $\left\{\boldsymbol{\theta}^{(p-1)}_{1:M}, \boldsymbol{u}^{(p-1)}_{1:M}, W^{(p-1)}_{1:M}\right\}$ at the $(p-1)$st stage are a sample from $\overline{\pi}_{p-1}(\boldsymbol{\theta}, \boldsymbol{u})$ and are propagated to $\overline{\pi}_p(\boldsymbol{\theta}, \boldsymbol{u})$, by updating the weights $W^{(p)}_{1:M} = w^{(p)}_{1:M}/\sum_{i=1}^{M} w^{(p)}_i$, where

$$w^{(p)}_i = W^{(p-1)}_i \exp\left(\left(a_p - a_{p-1}\right)\widehat{\ell}_m(\boldsymbol{\theta}^{(p-1)}_i) - \frac{1}{2}\left(a_p^2 - a_{p-1}^2\right)\widehat{\sigma}^2_m(\boldsymbol{\theta}^{(p-1)}_i)\right).$$

The resampling step is described in Section 2.1.

### 3.4 The Markov move step

We now outline the Markov move step of our approach, which utilizes Hamiltonian dynamics to propose distant particle moves and data subsampling in order to efficiently compute the dynamics. Similarly to Section 2.1, the Markov move is designed to leave each of the sequence target densities $\overline{\pi}_p(\boldsymbol{\theta}, \boldsymbol{u})$, for $p = 0, ..., P$ invariant. To accommodate subsampling, it is divided into two parts and is described in Algorithm 1. We refer the reader to Dang et al. (2018) for the details.

**Algorithm 1** Single Markov move with a kernel invariant for $\overline{\pi}_p(\boldsymbol{\theta}, \boldsymbol{u})$ in (8).

For $i = 1, ..., M$,

1. Sample $\boldsymbol{u}_i|\boldsymbol{\theta}_i, \boldsymbol{y}$: Propose $\boldsymbol{u}^* \sim p(\boldsymbol{u})$, and set $\boldsymbol{u}_i = \boldsymbol{u}^*$, with probability

$$\alpha_{\boldsymbol{u}} = \min\left(1, r := \frac{\exp\left(a_p \widehat{\ell}_m(\boldsymbol{\theta}_i, \boldsymbol{u}^*) - \frac{1}{2} a_p^2 \widehat{\sigma}_m^2(\boldsymbol{\theta}_i, \boldsymbol{u}^*)\right)}{\exp\left(a_p \widehat{\ell}_m(\boldsymbol{\theta}_i, \boldsymbol{u}_i) - \frac{1}{2} a_p^2 \widehat{\sigma}_m^2(\boldsymbol{\theta}_i, \boldsymbol{u}_i)\right)}\right), \tag{10}$$

   The $\boldsymbol{u}^*$ is proposed from the prior and is independent of the current value of $\boldsymbol{u}_i$, so the difference between the log of the numerator and log of the denominator of the ratio $r$ in (10) can be highly variable. This move might get stuck when the numerator is significantly overestimated. A remedy is to induce a high correlation $\rho$ between the log of the estimated annealed likelihood at the current and proposed draws in (10). This can be achieved either through correlating the $\boldsymbol{u}$ as in Deligiannidis et al. (2018) (see Quiroz et al. 2018a for discrete $\boldsymbol{u}$) or by block updates of $\boldsymbol{u}$ as in Tran et al. (2017); Quiroz et al. (2018b). We implement the block updates with $G$ blocks, which gives a correlation $\rho \approx 1 - \frac{1}{G}$.

2. Sample $\boldsymbol{\theta}_i|\boldsymbol{u}_i, \boldsymbol{y}$: Given a subset of data $\boldsymbol{u}_i$, we move the particle $\boldsymbol{\theta}_i$ using a Hamiltonian Monte Carlo (HMC) proposal in a Metropolis-Hastings (MH) algorithm. This becomes a standard HMC move for a given subset $\boldsymbol{u}$.

Note that the above is a Gibbs update of $\boldsymbol{\theta}_i, \boldsymbol{u}_i|\boldsymbol{y}$. The MH within Gibbs performed in Step 1. is valid (Johnson et al., 2013) and so is the HMC within Gibbs (Neal, 2011) in Step 2. Therefore, this kernel has $\overline{\pi}_p(\boldsymbol{\theta}, \boldsymbol{u})$ as its invariant distribution. Dang et al. (2018) previously proposed an MCMC version of this algorithm.

The HMC proposal has a few parameters that need to be determined, such as the mass matrix $\boldsymbol{M}$, the step size $\epsilon$, and the number of leapfrog steps $L$. We follow Buchholz et al. (2018), who develop a tuning procedure for all the parameters in a HMC proposal within a SMC framework.

**Algorithm 2** Subsampling Sequential Monte Carlo

1. Initially, sample the particles $\left\{\boldsymbol{\theta}_i^{(0)}, \boldsymbol{u}_i^{(0)}\right\}$ from the prior densities $p(\boldsymbol{\theta})$ and $p(\boldsymbol{u})$ and give all particles equal weights, $W_i = 1/M$, $i = 1, ..., M$. Initialize $p = 0$.

2. While the tempering sequence $a_p \neq 1$ do

   (a) Set $p \leftarrow p + 1$

   (b) Find $a_p$ adaptively to maintain the ESS around $\text{ESS}_{\text{target}}$ (Section 2.2).

   (c) Reweighting: compute the unnormalized weights

$$\begin{aligned} w_i^{(p)} &= W_i^{(p-1)} \frac{\eta_{a_p}\left(\boldsymbol{\theta}_i^{(p-1)}, \boldsymbol{u}_i^{(p-1)}\right)}{\eta_{a_{p-1}}\left(\boldsymbol{\theta}_i^{(p-1)}, \boldsymbol{u}_i^{(p-1)}\right)} \\ &= W_i^{(p-1)} \exp\left(\left(a_p - a_{p-1}\right) \widehat{\ell}_m(\boldsymbol{\theta}_i^{(p-1)}) - \frac{1}{2}\left(a_p^2 - a_{p-1}^2\right) \widehat{\sigma}_m^2(\boldsymbol{\theta}_i^{(p-1)})\right), \end{aligned}$$

   and normalize as $W_i^{(p)} = \frac{w_i}{\sum_{i'=1}^M w_{i'}}$, $i = 1, ..., M$.

   (d) Compute $\overline{\boldsymbol{\theta}}$ as $\overline{\boldsymbol{\theta}} = \sum_{i=1}^M W_i^{(p)} \boldsymbol{\theta}_i^{(p-1)}$ and then obtain

$$\sum_{k=1}^n \ell_k\left(\overline{\boldsymbol{\theta}}\right), \sum_{k=1}^n \nabla_{\boldsymbol{\theta}} \ell_k\left(\overline{\boldsymbol{\theta}}\right), \sum_{k=1}^n \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} \ell_k\left(\overline{\boldsymbol{\theta}}\right)$$

   and the mass matrix $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}\left(\overline{\boldsymbol{\theta}}\right)$. Note that this step is based on the full dataset.

   (e) Resample the particles $\left\{\boldsymbol{\theta}_i^{(p-1)}, \boldsymbol{u}_i^{(p-1)}\right\}_{i=1}^M$ using the weights $\left\{W_i^{(p)}\right\}_{i=1}^M$ to obtain resampled particles $\left\{\boldsymbol{\theta}_i^{(p)}, \boldsymbol{u}_i^{(p)}\right\}_{i=1}^M$ and set $W_i^{(p)} = 1/M$.

   (f) Apply $R$ Markov moves to each particle $\boldsymbol{\theta}_i^{(p)}, \boldsymbol{u}_i^{(p)}$ using Algorithm 1.

## 3.5 Estimating the Marginal Likelihood

Our approach can naturally be extended from Section 2.3 by considering the augmented target density $\overline{\pi}_p(\boldsymbol{\theta}, \boldsymbol{u})$ in (8). First, write the ratio of marginal likelihoods as

$$\frac{Z_p}{Z_{p-1}} = \int \gamma_p(\boldsymbol{\theta}) \pi_{p-1}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \text{ with } \gamma_p(\boldsymbol{\theta}) = \frac{\eta_p(\boldsymbol{\theta})}{\eta_{p-1}(\boldsymbol{\theta})},$$

and we wish to estimate $\frac{Z_p}{Z_{p-1}}$, i.e. we need $\gamma_p(\boldsymbol{\theta}, \boldsymbol{u})$ such that

$$\frac{Z_p}{Z_{p-1}} = \int_{\mathcal{U}} \int_{\boldsymbol{\Theta}} \gamma_p(\boldsymbol{\theta}, \boldsymbol{u}) \pi_{p-1}(\boldsymbol{\theta}, \boldsymbol{u}) d\boldsymbol{\theta} d\boldsymbol{u}.$$

If we take

$$\gamma_p(\boldsymbol{\theta}, \boldsymbol{u}) = \frac{\eta_p(\boldsymbol{\theta}, \boldsymbol{u})}{\eta_{p-1}(\boldsymbol{\theta}, \boldsymbol{u})},$$

then

$$\begin{aligned} \int_{\mathcal{U}} \int_{\boldsymbol{\Theta}} \gamma_p(\boldsymbol{\theta}, \boldsymbol{u}) \pi_{p-1}(\boldsymbol{\theta}, \boldsymbol{u}) d\boldsymbol{\theta} d\boldsymbol{u} &= \int_{\mathcal{U}} \int_{\boldsymbol{\Theta}} \frac{\eta_p(\boldsymbol{\theta}, \boldsymbol{u})}{\eta_{p-1}(\boldsymbol{\theta}, \boldsymbol{u})} \frac{\eta_{p-1}(\boldsymbol{\theta}, \boldsymbol{u})}{Z_{p-1}} p(\boldsymbol{\theta}) p(\boldsymbol{u}) d\boldsymbol{\theta} d\boldsymbol{u} \\ &= \frac{Z_p}{Z_{p-1}}. \end{aligned}$$

Thus, if $\left\{ \boldsymbol{\theta}_{1:M}^{(p-1)}, \boldsymbol{u}_{1:M}^{(p-1)}, W_{1:M}^{(p-1)} \right\}$ at the $(p-1)$st sequence is an approximate sample from $\overline{\pi}_{a_{p-1}}(\boldsymbol{\theta}, \boldsymbol{u})$, we estimate the ratio $Z_p/Z_{p-1}$ by

$$\widehat{\frac{Z_p}{Z_{p-1}}} = \sum_{i=1}^{M} W_i^{(p-1)} \frac{\eta_p\left(\boldsymbol{\theta}_i^{(p-1)}, \boldsymbol{u}_i^{(p-1)}\right)}{\eta_{p-1}\left(\boldsymbol{\theta}_i^{(p-1)}, \boldsymbol{u}_i^{(p-1)}\right)},$$

and the marginal likelihood estimate is obtained using this expression in (4).

### 3.6 Efficient memory management by data subsampling

We now explain in detail how data subsampling helps parallel computing from a memory efficiency point of view. Suppose first that we perform standard SMC (using all the data) and that we apply computer parallelism using $N$ workers, so that each worker deals, on average, with $M/N$ particles. In this case, for each stage $p$, the computations performed for each particle require repeated likelihood evaluations (using all $n$ data) when applying $R$ Markov move steps. Hence, each worker needs to have access to the full dataset.

Suppose now that we use our data subsampling approach in the same setting using $M/N$ particles for each of the $N$ workers. Then, at the beginning of each stage $p$ of the algorithm, we still require a full data evaluation for computing $A(\overline{\boldsymbol{\theta}}), B(\overline{\boldsymbol{\theta}})$ and $C(\overline{\boldsymbol{\theta}})$ in Section 3.2. However, at each $p$, we can now subsample the data according to $\boldsymbol{u}_i^{(p-1)}$ for each particle and subsequently perform the $R$ Markov move steps, which now require repeated estimated likelihood evaluations (using $m \ll n$ observations) and in addition $A(\overline{\boldsymbol{\theta}}), B(\overline{\boldsymbol{\theta}})$ and $C(\overline{\boldsymbol{\theta}})$. Now each worker needs to have access only to the subsampled dataset, as well as $A(\overline{\boldsymbol{\theta}}), B(\overline{\boldsymbol{\theta}})$ and $C(\overline{\boldsymbol{\theta}})$. However, these are only

summaries of the full dataset and are therefore very memory efficient.

# 4 Evaluations

## 4.1 Experiments

We now evaluate our methodology through the following experiments.

- *Experiment 1: The usefulness of the Hamiltonian Monte Carlo kernel.*
  We show how effective a HMC kernel for the Markov move step is compared to a random walk proposal and a MALA proposal.

- *Experiment 2: Evaluating the speed and the accuracies of the marginal likelihood estimate and the approximate posterior density.*
  We show that our subsampling approach is accurate by comparing the estimates of the marginal likelihood and posterior density to those obtained by the full data SMC, which represent the gold standard estimates.

- *Experiment 3: Evaluating the effect of the control variate.*
  We show that our subsampling approach can further improve the speed by using a first order control variate instead of the second order alternative (see Section 3.2 for details). We conclude that the results in terms of accuracies of marginal posterior densities and marginal likelihood estimates are robust to this choice of control variates.

- *Application: Non-linear bankruptcy modeling.*
  We perform inference in a non-linear bankruptcy model for a large dataset of Swedish firms. The results are compared against the MCMC competitor Subsampling MCMC (Quiroz et al., 2018a) implemented with a Hamiltonian data subsampling proposal (Dang et al., 2018). We use the estimate of the marginal likelihood to perform model selection between a non-linear model and a linear model.

We tune all SMC algorithms following Buchholz et al. (2018), whom provide a tuning procedure for all parameters in the Markov kernel. We use $M = 280$ particles, a choice motivated by our cluster with 28 cores with each core dealing (on average) with 10 particles. We repeat this 10 times (on different machines) to computed the standard error of the log marginal likelihood estimator.

## 4.2 Models and data

We use the following models and datasets to evaluate our methodology.

**Logistic regression.** The model for the response $y_i \in \{0, 1\}$ given a set of covariates $\boldsymbol{x}_i \in \mathbb{R}^{d_{\boldsymbol{x}}}$ and parameters $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}$, with $d_{\boldsymbol{x}} = d_{\boldsymbol{\theta}}$, is

$$p\left(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}\right) = \left(\frac{1}{1 + \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\theta}\right)}\right)^{y_i} \left(\frac{1}{1 + \exp\left(-\boldsymbol{x}_i^\top \boldsymbol{\theta}\right)}\right)^{1-y_i}.$$

For Experiment 1 and 3 in, respectively, Section 4.3 and 4.5, we use the HIGGS dataset (Baldi et al., 2014) that contains $n = 11,000,000$ observations and 28 covariates. The response is "detected particle" and 21 of the covariates are kinematic properties measured by particle detectors, while 7 are high-level features to capture non-linearities. Together with the intercept this forms $d_{\boldsymbol{\theta}} = 29$ and we assign the prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\boldsymbol{\theta}}})$ where $\mathbf{I}_d$ is the $d \times d$ identity matrix. For the application in Section 4.6, we use a Swedish firm bankruptcy dataset that contains $n = 4,748,089$ observations with firm default as the response variable and eight firm-specific and macroeconomic covariates, which gives 9 covariates after adding an intercept. We consider a generalized additive model by basis expansions of the covariates, see Section 4.6 for details. This example uses the prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, 10^2\mathbf{I}_{d_{\boldsymbol{\theta}}})$. We show how to perform model selection using our methodology.

**Student-t regression.** We consider a univariate Student-t regression

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta} + e_i,\ e_i \sim t(\nu = 5),$$

where $t$ is the Student-t distribution with $\nu$ degrees of freedom. For Experiment 2, we simulate a dataset with $n = 500,000$ and with $d_{\boldsymbol{\theta}} = 50$, where the covariates are simulated such that the marginal variances are 1 and their pairwise correlation is 0.9. The parameters are simulated independently from Uniform$(-5, 5)$. We assign the prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, 10\mathbf{I}_{d_{\boldsymbol{\theta}}})$.

**Poisson regression.** Our final model is a Poisson regression where the univariate $y$ follows a Poisson distribution with an expectation that is log-linear, i.e.

$$y_i|\boldsymbol{x}_i \sim \text{Poisson}(\exp(\boldsymbol{x}_i^\top \boldsymbol{\theta})).$$

We generate $n = 200,000$ observations with $d_{\boldsymbol{\theta}} = 30$ covariates simulated from $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{29})$ (the intercept is $\mathbf{1}$). The parameters are simulated independently from Uniform$(-0.2, 0.2)$ and are assigned the prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, 0.1\mathbf{I}_{d_{\boldsymbol{\theta}}})$.

Table 1: Comparing the performances of three kernels for the Markov move, Hamiltonian Monte Carlo (HMC), Metropolis Adjusted Langevin Algorithm (MALA) and Random Walk (RW). The table shows the estimate of the log of the marginal likelihood with standard error in parenthesis, the CPU time, the number of annealing steps $P$ (tuned to maintain $ESS \approx 0.8M$) and the number of Markov moves $R$ (tuned as in Buchholz et al., 2018). The results are for the logistic regression model estimated using the HIGGS data and $M = 280$ particles. All methods use the second order control variate in Section 3.2. The results are averaged over 10 runs, which are used to compute the standard error of the estimator.

| | log marginal likelihood | CPU time (hrs) | $P$ | $R$ |
|---|---|---|---|---|
| HMC | -7,013,460.90 (0.32) | 2.31 | 106 | 5 |
| MALA | -7,013,462.49 (0.26) | 4.77 | 106 | 20 |
| RW | -7,013,461.43 (0.32) | 33.43 | 106 | 200 |

## 4.3 Experiment 1: Evaluating the Markov move kernel

We now evaluate how effectively the Hamiltonian Monte Carlo Markov move step addresses the particle depletion problem compared to a random walk kernel and a MALA kernel. To this end, we use the logistic regression model estimated using the HIGGS data. We recall that the tuning parameters are set following Buchholz et al. (2018). The mass matrix $M$ in both HMC and MALA is taken as $\widehat{\Sigma}^{-1}$, which is the estimated inverse covariance matrix of the tempered posterior. We note that each step in the sequence has a corresponding estimate of this inverse covariance matrix, obtained using the corresponding particles from that step. For the random walk, the optimal scaling $(2.38^2/d_{\boldsymbol{\theta}})\widehat{\Sigma}^{-1}$ (Roberts et al., 1997) resulted in numerical errors, which is why we further scaled with 0.1.

Table 1 shows the results obtained using the second order Taylor series expansion control variate in Section 3.2. The log-likelihood estimator has $m = 5,000$ subsamples and the block-pseudo marginal is carried out using $G = 100$. It is clear that the Hamiltonian approach is computationally faster, and this is because it needs to take a smaller number of Markov steps $R$. The table also shows that the estimate of the log marginal likelihood is very similar for all methods. For the rest of our article we use the HMC kernel.

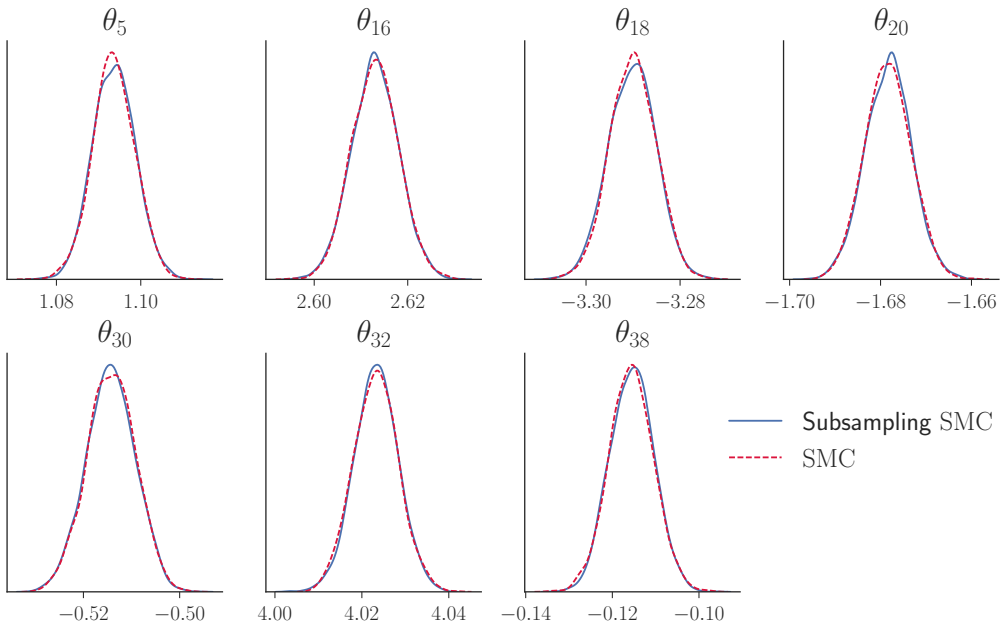


Figure 1: Kernel density estimates of a subset of the marginal posterior densities of $\boldsymbol{\theta}$ for the Student-t regression model with simulated data explained in Section 4.2. The density estimates are obtained by Subsampling MCMC and Subsampling SMC.

## 4.4 Experiment 2: Evaluating the speed and accuracy

We note at the outset that obtaining a gold standard estimate of the marginal likelihood to evaluate against is not feasible for the two largest datasets HIGGS and bankruptcy. This is because the full dataset needs to be available at each worker (we use 28) in order to compute the likelihood together with its gradient and Hessian, which quickly consumes the RAM of the computer. We therefore consider the Student-t and Poisson models and datasets in Section 4.2 for this experiment. For both examples we use $G = 100$ blocks and the second order Taylor series control variates and set $m$ to correspond to a sample fraction of about 0.0025. The results are shown in Table 2, which shows that our method is about 6.5 to 10.5 times faster and, moreover, confirms the accuracy of the marginal likelihood estimate of our method. The table also shows the results from the Laplace approximation to the marginal likelihood which, while arguably easier to compute, might not provide an accurate approximation. Finally, Figures 1 and 2 show that the marginal posterior densities are very well approximated for both the Student-t regression and the Poisson regression (we have confirmed this accuracy for all parameters but omitted due to space restrictions).

Table 2: Comparing the performances of Subsampling SMC and full data SMC. The table shows the estimate of the log of the marginal likelihood with standard error in parenthesis, the CPU time, the number of annealing steps $P$ (tuned to maintain $ESS \approx 0.8M$) and the number of Markov moves $R$ (tuned as in Buchholz et al., 2018). The results are for the Student-t regression and Poisson regression models estimated using the simulated datasets explained in Section 4.2. We use $M = 280$ particles. All methods use the second order control variate in Section 3.2. The results are averaged over 10 runs, which are used to compute the standard error of the estimator.

| | log marginal likelihood | CPU time (hrs) | $P$ | $R$ |
|---|---|---|---|---|
| **Student-t regression** | | | | |
| $(n = 500,000, m = 1,200)$ | | | | |
| Full data SMC | -815,775.82 (0.39) | 5.92 | 126 | 4 |
| Subsampling SMC | -815,773.49 (0.59) | 0.57 | 127 | 4 |
| Laplace approximation | -815,683.52 | | | |
| **Poisson regression** | | | | |
| $(n = 200,000, m = 500)$ | | | | |
| SMC | -260,888.69 (1.40) | 0.94 | 80 | 4 |
| Subsampling SMC | -260,887.87 (0.27) | 0.14 | 80 | 5 |
| Laplace approximation | -260,895.78 | | | |

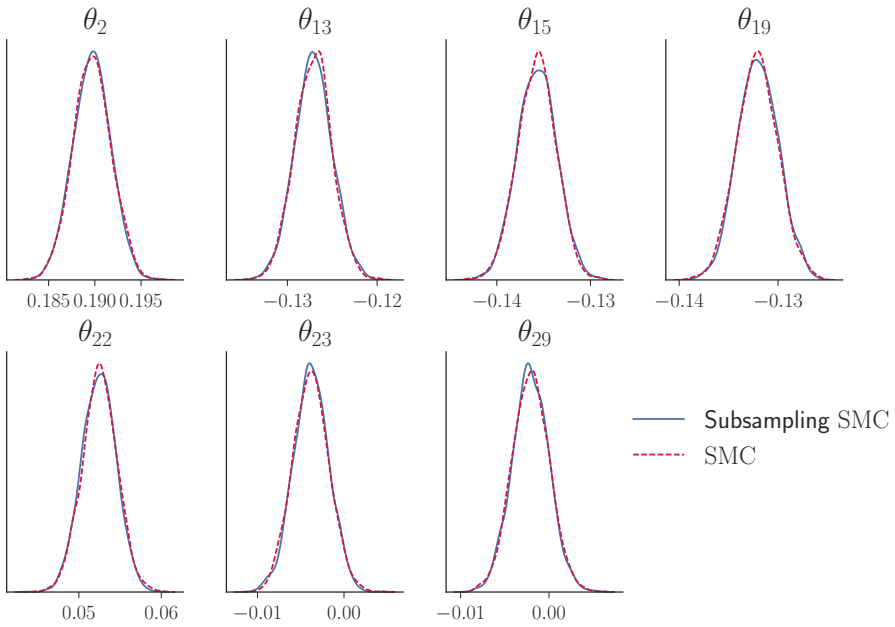


Figure 2: Kernel density estimates of a subset of the marginal posterior densities of $\boldsymbol{\theta}$ for the Poisson regression model with simulated data explained in Section 4.2. The density estimates are obtained by Subsampling MCMC and Subsampling SMC.

## 4.5 Experiment 3: Evaluating the effect of the control variate

We have previously shown that our approach provides accurate estimates of the marginal likelihood and marginal posterior densities using a second order Taylor series expansion. A natural question is: how robust are these results to the accuracy of the control variate? Table 3 shows the result for the HIGGS dataset, which confirms that the marginal likelihood estimator remains accurate when using a first order control variate, and further improves the speed by a factor of about 5. Figure 3 shows that the marginal posterior densities remain accurate (we have confirmed similar accuracy for all parameters). We make a final remark that it is possible to switch between the control variates. For example, a sensible strategy is to start with a faster but less accurate control variate when $a_p$ is small as the variance of $a_p\widehat{\ell}_m(\boldsymbol{\theta})$ might then be small even if the variance of $\widehat{\ell}_m(\boldsymbol{\theta})$ is large. For larger $a_p$, when the variance reduction from the multiplication is less pronounced, one might switch to the more accurate second order control variate.

## 4.6 Application: Non-linear modeling of firm bankruptcy

We now illustrate how to use our method for model selection using the large firm bankruptcy dataset explained in Section 4.1. We compare our marginal posterior

Table 3: Comparing the performances of the less accurate control variate (1st order) and more accurate control variate (2nd order). The table shows the estimate of the log of the marginal likelihood with standard error in parenthesis, the CPU time, the number of annealing steps $P$ (tuned to maintain $ESS \approx 0.8M$) and the number of Markov moves $R$ (tuned as in Buchholz et al. (2018)). The results are for the logistic regression model, estimated with the HIGGS dataset explained in Section 4.2. We use $M = 280$ particles. The results are averaged over 10 runs, which are used to compute the standard error of the estimator.

| | log marginal likelihood | CPU time (hrs) | $P$ | $R$ |
|---|---|---|---|---|
| 1st order | -7,013,461.07 (0.46) | 0.47 | 106 | 5 |
| 2nd order | -7,013,460.90 (0.32) | 2.31 | 106 | 5 |

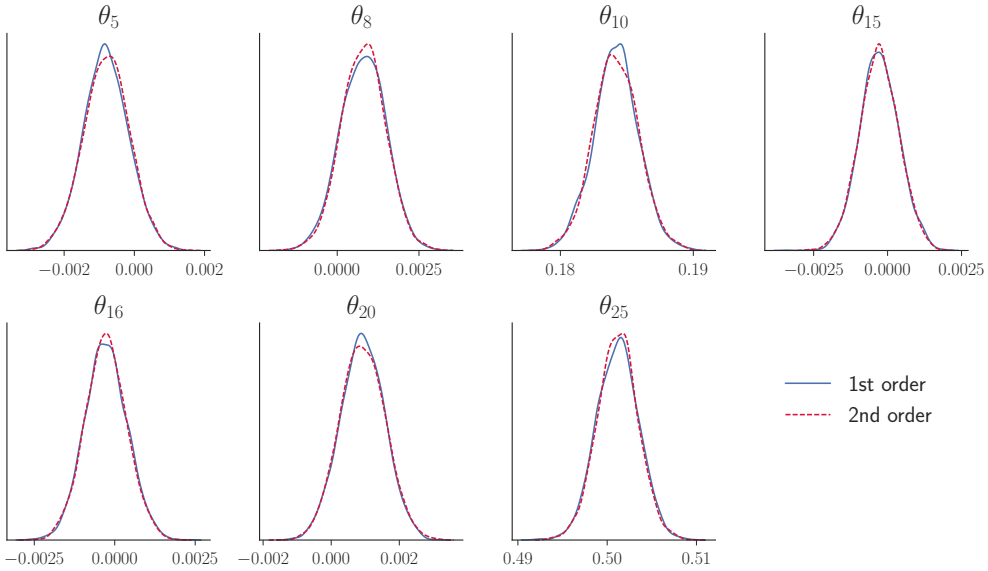


Figure 3: Kernel density estimates of a subset of the marginal posterior densities of $\boldsymbol{\theta}$. The density estimates are both obtained by Subsampling SMC, using different control variates based on a 1st and 2nd order Taylor series expansion as explained in Section 3.2.

density estimates against those of Subsampling MCMC (Quiroz et al., 2018a) as implemented by Dang et al. (2018) and find them nearly indistinguishable. We also compare both methods to the ground truth full data MCMC as in Dang et al. (2018). However, note that Subsampling MCMC cannot be used for model selection. Common methods such as Chib and Jeliazkov (2001) are not useful for Subsampling MCMC since the (perturbed) likelihood cannot be evaluated. This is a major advantage of Subsampling SMC compared to Subsampling MCMC.

We perform model selection between two models. The first model $\mathcal{M}_1$ is linear in data (in logit scale) with $d_{\boldsymbol{\theta}} = 9$. The second model $\mathcal{M}_2$ is non-linear in data (in logit scale) using $B$-splines as in Dang et al. (2018), with a total of $d_{\boldsymbol{\theta}} = 81$ coefficients. Non-linear bankruptcy models for this dataset have been analyzed in Quiroz and Villani (2013) and Giordani et al. (2014). Let $\Pr(\mathcal{M}_a)$ denote the prior probability of model $a$, $a = 1, 2$. Then the posterior probability of model $\mathcal{M}_a$ is

$$\Pr(\mathcal{M}_a|\boldsymbol{y}) \propto p(\boldsymbol{y}|\mathcal{M}_a)\Pr(\mathcal{M}_a),$$

where $p(\boldsymbol{y}|\mathcal{M}_a)$ is the marginal likelihood of model $\mathcal{M}_a$. We estimate $p(\boldsymbol{y}|\mathcal{M}_a)$ with the method outlined in Section 3.5. Given the marginal likelihood of each model, we can compute the Bayes Factor (BF) for the non-linear model $\mathcal{M}_2$ vs the linear model $\mathcal{M}_1$ as

$$\text{BF}_{21} = \frac{\Pr(\mathcal{M}_2|\boldsymbol{y})}{\Pr(\mathcal{M}_1|\boldsymbol{y})}. \tag{11}$$

The non-linear model is favored if $\text{BF}_{21} > 1$. We use the strength of evidence in Jeffreys (1961, p. 438), in which $10^{3/2} < \text{BF}_{21} < 10^2$ is considered very strong evidence and $\text{BF}_{21} > 10^2$ is decisive evidence. We use the uniform prior $\Pr(\mathcal{M}_1) = \Pr(\mathcal{M}_2) = 1/2$.

We let the number of blocks $G = 100$ and set the subsample size $m = 3,000$. For Subsampling MCMC we set these tuning parameters as in Dang et al. (2018). The estimates from the full data MCMC is considered as the "gold standard" when we assess the accuracy of the algorithms. This is achieved through an MCMC chain of $2,000$ post burn-in MCMC samples, with burn-in set to $1,000$ iterations. We have confirmed that the MCMC mixes well and the iterates are therefore an adequate representation of the true posterior.

Table 4 shows the estimate of the log marginal likelihood for both models and the corresponding Bayes factors obtained by Subsampling SMC. The table shows decisively that the non-linear model is superior. We again stress that producing marginal likelihood estimates is very convenient by SMC, whereas not possible with Subsampling MCMC.

Table 4: Estimates of the log marginal likelihoods and Bayes factors $\mathrm{BF}_{21}$ in (11) for selecting between $\mathcal{M}_1$ and $\mathcal{M}_2$. The estimates of the Standard Error (SE) for the marginal likelihood estimates are in parenthesis. The SE is computed using the 10 independent parallel runs. The prior probabilities are $\Pr(\mathcal{M}_1) = \Pr(\mathcal{M}_2) = 1/2$.

| | $\log \widehat{p}(\boldsymbol{y}\|\mathcal{M}_1)$ | $\log \widehat{p}(\boldsymbol{y}\|\mathcal{M}_2)$ | $\mathrm{BF}_{21}$ |
|---|---|---|---|
| Bankruptcy | $-208,517.79$ (0.21) | $-200,215.13$ (6.57) | $\exp(8,302.66)$ |

Figures 4 shows the kernel density estimates of the marginal posterior of selected parameters of the non-linear model for the bankruptcy dataset. It is evident that both Subsampling SMC and Subsampling MCMC are very accurate and we have confirmed the accuracy of the kernel density estimates for all the parameters, which we do not show here to save space. Instead, Figure 5 shows the estimated marginal posterior expectations and posterior variances by the two algorithms for all the parameters in the non-linear models. This confirms the accuracy of the estimates of each single parameter. We have confirmed that the kernel density estimates and the estimated marginal posterior expectations and posterior variances are accurate also for the linear model (not shown here).

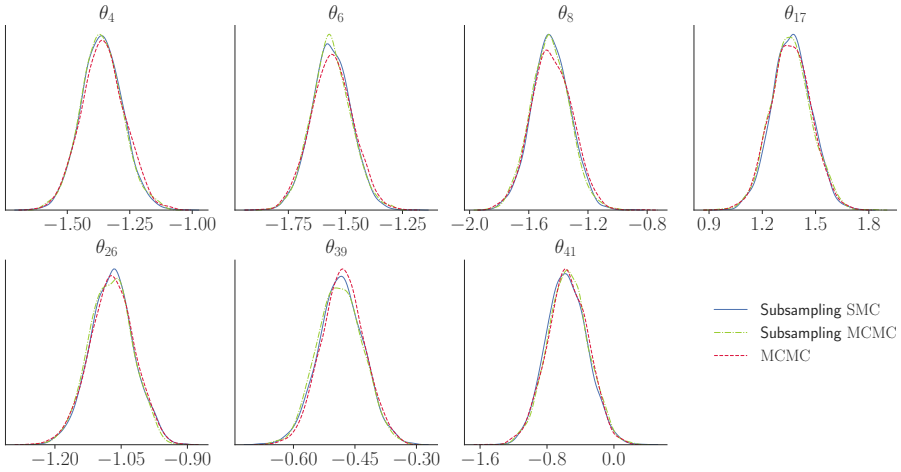


Figure 4: Kernel density estimates of a subset of the marginal posterior densities of $\boldsymbol{\theta}$ for the logistic model $\mathcal{M}_2$ for the bankruptcy dataset. The density estimates are obtained by MCMC, Subsampling MCMC and Subsampling SMC. MCMC represents the ground truth.

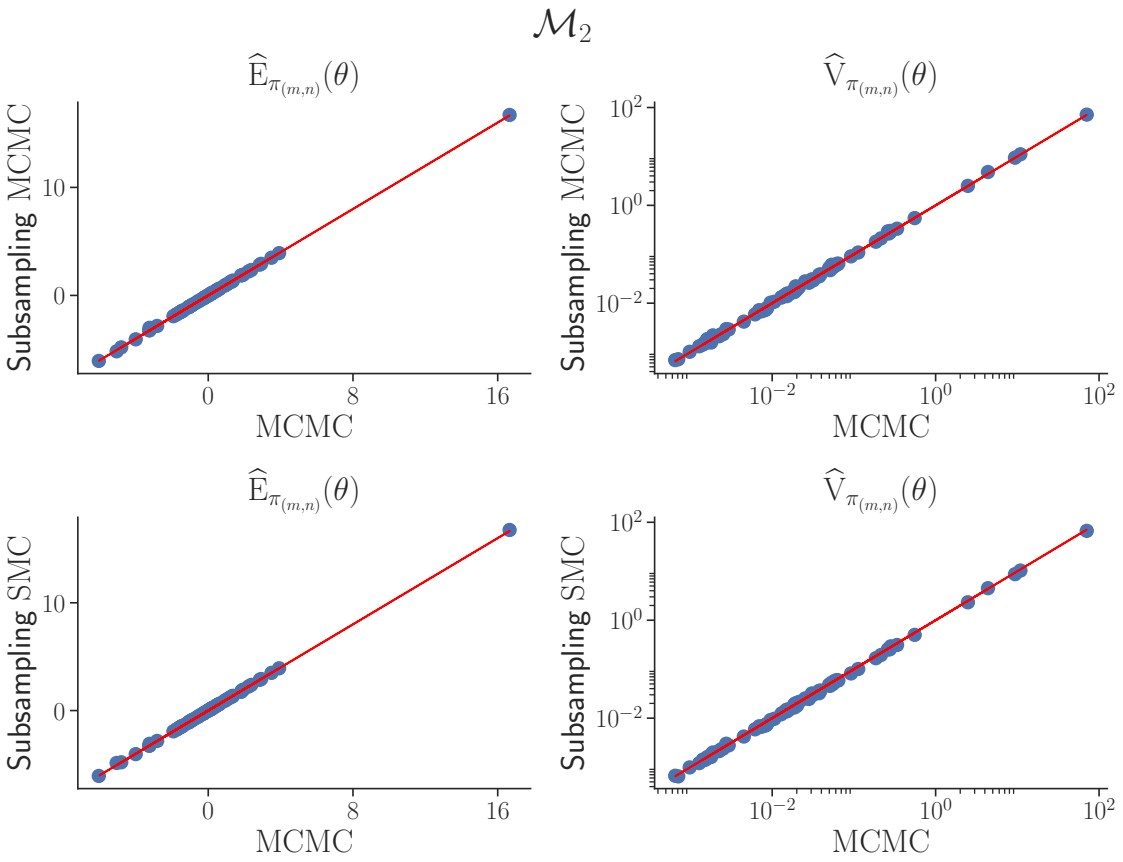


Figure 5: Estimates of marginal posterior means (left panel) and posterior variances (right panel) of $\boldsymbol{\theta}$ for the logistic model $\mathcal{M}_2$ for the bankruptcy dataset. The estimates are obtained by Subsampling MCMC and Subsampling SMC and plotted as dots, together with a 45 degree line. This line corresponds to estimates that are in perfect agreement.

Finally, the superiority of the non-linear model is well understood for the bankruptcy data from Figure 6, which shows that the relationship between the bankruptcy probability and the covariate Size is not a logistic function of the covariate as the linear model suggests. The figure shows again that the results obtained from Subsampling SMC are indistinguishable from those of Subsampling MCMC.

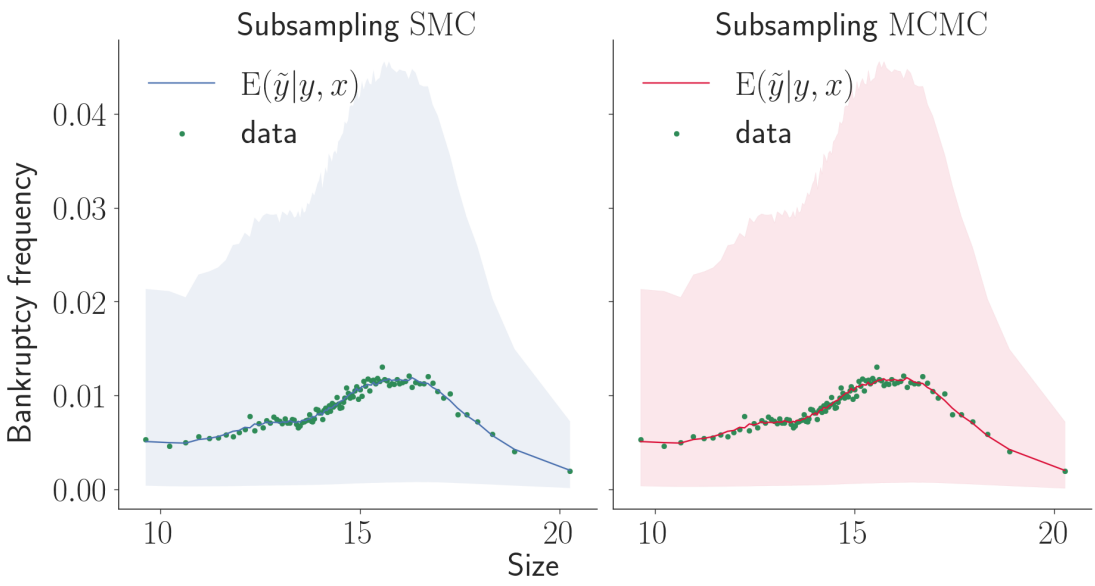


Figure 6: Realized and estimated bankruptcy probabilities. The figure shows the results with respect to the size variable (logarithm of deflated sales) for Subsampling SMC (left panel) and Subsampling MCMC (right panel). The data are divided into 100 equally sized groups based on the size variable. For each group, the empirical estimate of the bankruptcy probability is the fraction of bankrupt firms. These empirical estimates are represented as dots, where the corresponding $x$-value (size) has been set to the mean within the group. The model estimates for each of the 100 groups are obtained by, for each posterior sample $\boldsymbol{\theta}$, averaging the posterior predictive $\Pr(\widetilde{y}_k = 1|\boldsymbol{y}, x_k)$ for all observations $k$ in a group, and subsequently computing the posterior predictive mean $\mathbb{E}(\widetilde{y}_k = 1|\boldsymbol{y}, x_k)$ (solid line) and 90% prediction interval (quantiles 5-95, shaded region).

# 5 Conclusions

We proposed a simple and effective approach to speed up sequential Monte Carlo for static Bayesian models using data subsampling. The key ingredients of our approach are an efficient annealed likelihood estimator and an effective Markov kernel move step based on Hamiltonian Monte Carlo which boosts particle diversity. This kernel is computationally expensive for large datasets and data subsampling is crucial to obtain a feasible approach. We argued that the subsampling approach is also very convenient for managing computer memory when implementing SMC using parallel computing, because it avoids the need for each worker to store the full dataset. We demonstrated that the method performs efficient and accurate inference for three generalized linear models and a generalized additive model. Moreover, we showed that it allows Bayesian model selection through accurate estimates of the marginal likelihood, which is a major advantage compared to Subsampling MCMC.

## Acknowledgements

David Gunawan, Matias Quiroz and Robert Kohn were partially supported by Australian Research Council Center of Excellence grant CE140100049.

## References

Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for exotic particle in high energy physics with deep learning. *Nature Communications*, 5.

Bardenet, R., Doucet, A., and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557.

Beskos, A., Jasra, A., Kantas, N., and Thiery, A. (2016). On the convergence of adaptive sequential Monte Carlo methods. *The Annals of Applied Probability*, 26(2):1111–1146.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC press.

Buchholz, A., Chopin, N., and Jacob, P. E. (2018). Adaptive tuning of Hamiltonian Monte Carlo within sequential Monte Carlo. *arXiv preprint arXiv:1808.07730*.

Ceperley, D. and Dewing, M. (1999). The penalty method for random walks with uncertain energies. *The Journal of Chemical Physics*, 110(20):9812–9820.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of American Statistical Association*, 96(453):270–281.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539–552.

Dang, K. D., Quiroz, M., Kohn, R., Tran, M. N., and Villani, M. (2018). Hamiltonian Monte Carlo with energy conserving subsampling. *arXiv preprint arXiv: 1708.00955v2*.

Daviet, R. (2016). Inference with Hamiltonian sequential Monte Carlo simulators. `http://www.remidaviet.com/files/HSMC-paper.pdf`. [Online; accessed 1-May-2018].

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society, Series B*, 68:411–436.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive Sequential Monte Carlo for approximate Bayesian computation. *Statistics and Computing*, pages 1009–1020.

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018). The correlated pseudomarginal method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):839–870.

Doucet, A., De Freitas, N., and Gordon, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer.

Duan, J. C. and Fulop, A. (2015). Density-tempered marginalised sequential Monte Carlo samplers. *Journal of Business and Economics Statistics*, 33(2):192–202.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Fearnhead, P. and Taylor, B. M. (2013). An adaptive sequential Monte Carlo sampler. *Bayesian Analysis*, 8(2):411–438.

Gilks, W. R. and Berzuini, C. (2001). Following a moving target–Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):127–146.

Giordani, P., Jacobson, T., Von Schedvin, E., and Villani, M. (2014). Taking the twists into account: Predicting firm bankruptcy risk with splines of financial ratios. *Journal of Financial and Quantitative Analysis*, 49(4):1071–1099.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET.

Jasra, A., Stephens, D. A., Doucet, A., and Tsagaris, T. (2011). Inference for Lévy-driven stochastic volatility models via adaptive Sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38(1):1–22.

Jeffreys, H. (1961). *The Theory of Probability*. OUP Oxford.

Johnson, A. A., Jones, G. L., and Neath, R. C. (2013). Component-wise Markov chain Monte Carlo: Uniform and geometric ergodicity under mixing and composition. *Statistical Science*, 28(3):360–375.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of American Statistical Association*, 90(430):773–795.

Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. New York: Springer.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044.

Neal, R. (2001). Annealed importance sampling. *Statistics and Computing*, 11:125–139.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*.

Quiroz, M., Kohn, R., Villani, M., and Tran, M. N. (2018a). Speeding up MCMC by efficient data subsampling. *Journal of American Statistical Association*, To appear.

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., and Dang, K.-D. (2018b). The block-Poisson estimator for optimally tuned exact subsampling MCMC. *arXiv preprint arXiv:1603.08232v5*.

Quiroz, M. and Villani, M. (2013). Dynamic mixture-of-experts models for longitudinal and discrete-time survival data. `https://github.com/mattiasvillani/Papers/raw/master/DynamicMixture.pdf`. [Online; accessed 2-May-2018].

Quiroz, M., Villani, M., Kohn, R., Tran, M.-N., and Dang, K.-D. (2018c). Subsampling MCMC: An introduction for the survey statistician. *Sankhya A*, To appear.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis-Hastings. *Annals of Applied Probability*, 7:110–120.

Roberts, G. O. and Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357.

Sim, A., Filippi, S., and Stumpf, M. P. (2012). Information geometry and sequential Monte Carlo. *arXiv preprint arXiv:1212.0764*.

South, L. F., Pettitt, A. N., and Drovandi, C. C. (2016). Sequential Monte Carlo for static Bayesian models with independent MCMC proposals.

South, L. F., Pettitt, A. N., Friel, N., and Drovandi, C. C. (2017). Efficient use of derivative information within SMC methods for static Bayesian Models.

Tran, M. N., Kohn, R., Quiroz, M., and Villani, M. (2017). The block-pseudo marginal sampler. *preprint arXiv:1603.02485v5*.

Wang, L., Wang, S., and Bouchard-Côté, A. (2019). An annealed sequential Monte Carlo method for Bayesian phylogenetics. *arXiv preprint arXiv:1806.08813v3*.

# Earlier Working Papers:

For a complete list of Working Papers published by Sveriges Riksbank, see www.riksbank.se

| | |
|---|---|
| Estimation of an Adaptive Stock Market Model with Heterogeneous Agents<br>*by Henrik Amilon* | 2005:177 |
| Some Further Evidence on Interest-Rate Smoothing: The Role of Measurement Errors in the Output Gap<br>*by Mikael Apel and Per Jansson* | 2005:178 |
| Bayesian Estimation of an Open Economy DSGE Model with Incomplete Pass-Through<br>by Malin Adolfson, Stefan Laséen, Jesper Lindé and Mattias Villani | 2005:179 |
| Are Constant Interest Rate Forecasts Modest Interventions? Evidence from an Estimated Open Economy DSGE Model of the Euro Area<br>*by Malin Adolfson, Stefan Laséen, Jesper Lindé and Mattias Villani* | 2005:180 |
| Inference in Vector Autoregressive Models with an Informative Prior on the Steady State<br>*by Mattias Villani* | 2005:181 |
| Bank Mergers, Competition and Liquidity<br>*by Elena Carletti, Philipp Hartmann and Giancarlo Spagnolo* | 2005:182 |
| Testing Near-Rationality using Detailed Survey Data<br>*by Michael F. Bryan and Stefan Palmqvist* | 2005:183 |
| Exploring Interactions between Real Activity and the Financial Stance<br>*by Tor Jacobson, Jesper Lindé and Kasper Roszbach* | 2005:184 |
| Two-Sided Network Effects, Bank Interchange Fees, and the Allocation of Fixed Costs<br>*by Mats A. Bergman* | 2005:185 |
| Trade Deficits in the Baltic States: How Long Will the Party Last?<br>*by Rudolfs Bems and Kristian Jönsson* | 2005:186 |
| Real Exchange Rate and Consumption Fluctuations follwing Trade Liberalization<br>*by Kristian Jönsson* | 2005:187 |
| Modern Forecasting Models in Action: Improving Macroeconomic Analyses at Central Banks<br>*by Malin Adolfson, Michael K. Andersson, Jesper Lindé, Mattias Villani and Anders Vredin* | 2005:188 |
| Bayesian Inference of General Linear Restrictions on the Cointegration Space<br>*by Mattias Villani* | 2005:189 |
| Forecasting Performance of an Open Economy Dynamic Stochastic General Equilibrium Model<br>*by Malin Adolfson, Stefan Laséen, Jesper Lindé and Mattias Villani* | 2005:190 |
| Forecast Combination and Model Averaging using Predictive Measures<br>*by Jana Eklund and Sune Karlsson* | 2005:191 |
| Swedish Intervention and the Krona Float, 1993-2002<br>*by Owen F. Humpage and Javiera Ragnartz* | 2006:192 |
| A Simultaneous Model of the Swedish Krona, the US Dollar and the Euro<br>*by Hans Lindblad and Peter Sellin* | 2006:193 |
| Testing Theories of Job Creation: Does Supply Create Its Own Demand?<br>*by Mikael Carlsson, Stefan Eriksson and Nils Gottfries* | 2006:194 |
| Down or Out: Assessing The Welfare Costs of Household Investment Mistakes<br>*by Laurent E. Calvet, John Y. Campbell and Paolo Sodini* | 2006:195 |
| Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models<br>*by Paolo Giordani and Robert Kohn* | 2006:196 |
| Derivation and Estimation of a New Keynesian Phillips Curve in a Small Open Economy<br>*by Karolina Holmberg* | 2006:197 |
| Technology Shocks and the Labour-Input Response: Evidence from Firm-Level Data<br>*by Mikael Carlsson and Jon Smedsaas* | 2006:198 |
| Monetary Policy and Staggered Wage Bargaining when Prices are Sticky<br>*by Mikael Carlsson and Andreas Westermark* | 2006:199 |
| The Swedish External Position and the Krona<br>*by Philip R. Lane* | 2006:200 |

| | |
|---|---|
| Price Setting Transactions and the Role of Denominating Currency in FX Markets<br>*by Richard Friberg and Fredrik Wilander* | 2007:201 |
| The geography of asset holdings: Evidence from Sweden<br>*by Nicolas Coeurdacier and Philippe Martin* | 2007:202 |
| Evaluating An Estimated New Keynesian Small Open Economy Model<br>*by Malin Adolfson, Stefan Laséen, Jesper Lindé and Mattias Villani* | 2007:203 |
| The Use of Cash and the Size of the Shadow Economy in Sweden<br>*by Gabriela Guibourg and Björn Segendorf* | 2007:204 |
| Bank supervision Russian style: Evidence of conflicts between micro- and macro-prudential concerns<br>*by Sophie Claeys and Koen Schoors* | 2007:205 |
| Optimal Monetary Policy under Downward Nominal Wage Rigidity<br>*by Mikael Carlsson and Andreas Westermark* | 2007:206 |
| Financial Structure, Managerial Compensation and Monitoring<br>*by Vittoria Cerasi and Sonja Daltung* | 2007:207 |
| Financial Frictions, Investment and Tobin's q<br>*by Guido Lorenzoni and Karl Walentin* | 2007:208 |
| Sticky Information vs Sticky Prices: A Horse Race in a DSGE Framework<br>*by Mathias Trabandt* | 2007:209 |
| Acquisition versus greenfield: The impact of the mode of foreign bank entry on information and bank lending rates<br>*by Sophie Claeys and Christa Hainz* | 2007:210 |
| Nonparametric Regression Density Estimation Using Smoothly Varying Normal Mixtures<br>*by Mattias Villani, Robert Kohn and Paolo Giordani* | 2007:211 |
| The Costs of Paying – Private and Social Costs of Cash and Card<br>*by Mats Bergman, Gabriella Guibourg and Björn Segendorf* | 2007:212 |
| Using a New Open Economy Macroeconomics model to make real nominal exchange rate forecasts<br>*by Peter Sellin* | 2007:213 |
| Introducing Financial Frictions and Unemployment into a Small Open Economy Model<br>*by Lawrence J. Christiano, Mathias Trabandt and Karl Walentin* | 2007:214 |
| Earnings Inequality and the Equity Premium<br>*by Karl Walentin* | 2007:215 |
| Bayesian forecast combination for VAR models<br>*by Michael K. Andersson and Sune Karlsson* | 2007:216 |
| Do Central Banks React to House Prices?<br>*by Daria Finocchiaro and Virginia Queijo von Heideken* | 2007:217 |
| The Riksbank's Forecasting Performance<br>*by Michael K. Andersson, Gustav Karlsson and Josef Svensson* | 2007:218 |
| Macroeconomic Impact on Expected Default Freqency<br>*by Per Åsberg and Hovick Shahnazarian* | 2008:219 |
| Monetary Policy Regimes and the Volatility of Long-Term Interest Rates<br>*by Virginia Queijo von Heideken* | 2008:220 |
| Governing the Governors: A Clinical Study of Central Banks<br>*by Lars Frisell, Kasper Roszbach and Giancarlo Spagnolo* | 2008:221 |
| The Monetary Policy Decision-Making Process and the Term Structure of Interest Rates<br>*by Hans Dillén* | 2008:222 |
| How Important are Financial Frictions in the U S and the Euro Area<br>*by Virginia Queijo von Heideken* | 2008:223 |
| Block Kalman filtering for large-scale DSGE models<br>*by Ingvar Strid and Karl Walentin* | 2008:224 |
| Optimal Monetary Policy in an Operational Medium-Sized DSGE Model<br>*by Malin Adolfson, Stefan Laséen, Jesper Lindé and Lars E. O. Svensson* | 2008:225 |
| Firm Default and Aggregate Fluctuations<br>*by Tor Jacobson, Rikard Kindell, Jesper Lindé and Kasper Roszbach* | 2008:226 |
| Re-Evaluating Swedish Membership in EMU: Evidence from an Estimated Model<br>*by Ulf Söderström* | 2008:227 |

| | |
|---|---|
| The Effect of Cash Flow on Investment: An Empirical Test of the Balance Sheet Channel<br>*by Ola Melander* | 2009:228 |
| Expectation Driven Business Cycles with Limited Enforcement<br>*by Karl Walentin* | 2009:229 |
| Effects of Organizational Change on Firm Productivity<br>*by Christina Håkanson* | 2009:230 |
| Evaluating Microfoundations for Aggregate Price Rigidities: Evidence from Matched Firm-Level Data on Product Prices and Unit Labor Cost<br>*by Mikael Carlsson and Oskar Nordström Skans* | 2009:231 |
| Monetary Policy Trade-Offs in an Estimated Open-Economy DSGE Model<br>*by Malin Adolfson, Stefan Laséen, Jesper Lindé and Lars E. O. Svensson* | 2009:232 |
| Flexible Modeling of Conditional Distributions Using Smooth Mixtures of Asymmetric Student T Densities<br>*by Feng Li, Mattias Villani and Robert Kohn* | 2009:233 |
| Forecasting Macroeconomic Time Series with Locally Adaptive Signal Extraction<br>*by Paolo Giordani and Mattias Villani* | 2009:234 |
| Evaluating Monetary Policy<br>*by Lars E. O. Svensson* | 2009:235 |
| Risk Premiums and Macroeconomic Dynamics in a Heterogeneous Agent Model<br>*by Ferre De Graeve, Maarten Dossche, Marina Emiris, Henri Sneessens and Raf Wouters* | 2010:236 |
| Picking the Brains of MPC Members<br>*by Mikael Apel, Carl Andreas Claussen and Petra Lennartsdotter* | 2010:237 |
| Involuntary Unemployment and the Business Cycle<br>*by Lawrence J. Christiano, Mathias Trabandt and Karl Walentin* | 2010:238 |
| Housing collateral and the monetary transmission mechanism<br>*by Karl Walentin and Peter Sellin* | 2010:239 |
| The Discursive Dilemma in Monetary Policy<br>*by Carl Andreas Claussen and Øistein Røisland* | 2010:240 |
| Monetary Regime Change and Business Cycles<br>*by Vasco Cúrdia and Daria Finocchiaro* | 2010:241 |
| Bayesian Inference in Structural Second-Price common Value Auctions<br>*by Bertil Wegmann and Mattias Villani* | 2010:242 |
| Equilibrium asset prices and the wealth distribution with inattentive consumers<br>*by Daria Finocchiaro* | 2010:243 |
| Identifying VARs through Heterogeneity: An Application to Bank Runs<br>*by Ferre De Graeve and Alexei Karas* | 2010:244 |
| Modeling Conditional Densities Using Finite Smooth Mixtures<br>*by Feng Li, Mattias Villani and Robert Kohn* | 2010:245 |
| The Output Gap, the Labor Wedge, and the Dynamic Behavior of Hours<br>*by Luca Sala, Ulf Söderström and Antonella Trigari* | 2010:246 |
| Density-Conditional Forecasts in Dynamic Multivariate Models<br>*by Michael K. Andersson, Stefan Palmqvist and Daniel F. Waggoner* | 2010:247 |
| Anticipated Alternative Policy-Rate Paths in Policy Simulations<br>*by Stefan Laséen and Lars E. O. Svensson* | 2010:248 |
| MOSES: Model of Swedish Economic Studies<br>*by Gunnar Bårdsen, Ard den Reijer, Patrik Jonasson and Ragnar Nymoen* | 2011:249 |
| The Effects of Endogenuos Firm Exit on Business Cycle Dynamics and Optimal Fiscal Policy<br>*by Lauri Vilmi* | 2011:250 |
| Parameter Identification in a Estimated New Keynesian Open Economy Model<br>*by Malin Adolfson and Jesper Lindé* | 2011:251 |
| Up for count? Central bank words and financial stress<br>*by Marianna Blix Grimaldi* | 2011:252 |
| Wage Adjustment and Productivity Shocks<br>*by Mikael Carlsson, Julián Messina and Oskar Nordström Skans* | 2011:253 |

| | |
|---|---|
| The timing of uncertainty shocks in a small open economy<br>*by Hanna Armelius, Isaiah Hull and Hanna Stenbacka Köhler* | 2016:334 |
| Quantitative easing and the price-liquidity trade-off<br>*by Marien Ferdinandusse, Maximilian Freier and Annukka Ristiniemi* | 2017:335 |
| What Broker Charges Reveal about Mortgage Credit Risk<br>*by Antje Berndt, Burton Hollifield and Patrik Sandåsi* | 2017:336 |
| Asymmetric Macro-Financial Spillovers<br>*by Kristina Bluwstein* | 2017:337 |
| Latency Arbitrage When Markets Become Faster<br>*by Burton Hollifield, Patrik Sandås and Andrew Todd* | 2017:338 |
| How big is the toolbox of a central banker? Managing expectations with policy-rate forecasts: Evidence from Sweden<br>*by Magnus Åhl* | 2017:339 |
| International business cycles: quantifying the effects of a world market for oil<br>*by Johan Gars and Conny Olovsson l* | 2017:340 |
| Systemic Risk: A New Trade-Off for Monetary Policy?<br>*by Stefan Laséen, Andrea Pescatori and Jarkko Turunen* | 2017:341 |
| Household Debt and Monetary Policy: Revealing the Cash-Flow Channel<br>*by Martin Flodén, Matilda Kilström, Jósef Sigurdsson and Roine Vestman* | 2017:342 |
| House Prices, Home Equity, and Personal Debt Composition<br>*by Jieying Li and Xin Zhang* | 2017:343 |
| Identification and Estimation issues in Exponential Smooth Transition Autoregressive Models<br>*by Daniel Buncic* | 2017:344 |
| Domestic and External Sovereign Debt<br>*by Paola Di Casola and Spyridon Sichlimiris* | 2017:345 |
| The Role of Trust in Online Lending<br>*by Christoph Bertsch, Isaiah Hull, Yingjie Qi and Xin Zhang* | 2017:346 |
| On the effectiveness of loan-to-value regulation in a multiconstraint framework<br>*by Anna Grodecka* | 2017:347 |
| Shock Propagation and Banking Structure<br>*by Mariassunta Giannetti and Farzad Saidi* | 2017:348 |
| The Granular Origins of House Price Volatility<br>*by Isaiah Hull, Conny Olovsson, Karl Walentin and Andreas Westermark* | 2017:349 |
| Should We Use Linearized Models To Calculate Fiscal Multipliers?<br>*by Jesper Lindé and Mathias Trabandt* | 2017:350 |
| The impact of monetary policy on household borrowing – a high-frequency IV identification<br>*by Maria Sandström* | 2018:351 |
| Conditional exchange rate pass-through: evidence from Sweden<br>*by Vesna Corbo and Paola Di Casola* | 2018:352 |
| Learning on the Job and the Cost of Business Cycles<br>*by Karl Walentin and Andreas Westermark* | 2018:353 |
| Trade Credit and Pricing: An Empirical Evaluation<br>*by Niklas Amberg, Tor Jacobson and Erik von Schedvin* | 2018:354 |
| A shadow rate without a lower bound constraint<br>*by Rafael B. De Rezende and Annukka Ristiniemi* | 2018:355 |
| Reduced "Border Effects", FTAs and International Trade<br>*by Sebastian Franco and Erik Frohm* | 2018:356 |
| Spread the Word: International Spillovers from Central Bank Communication<br>*by Hanna Armelius, Christoph Bertsch, Isaiah Hull and Xin Zhang* | 2018:357 |
| Predictors of Bank Distress: The 1907 Crisis in Sweden<br>*by Anna Grodecka, Seán Kenny and Anders Ögren* | 2018:358 |

| | |
|---|---|
| Diversication Advantages During the Global Financial Crisis<br>*by Mats Levander* | 2018:359 |
| Towards Technology-News-Driven Business Cycles<br>*by Paola Di Casola and Spyridon Sichlimiris* | 2018:360 |
| The Housing Wealth Effect: Quasi-Experimental Evidence<br>*by Dany Kessel, Björn Tyrefors and Roine* | 2018:361 |
| Identification Versus Misspecification in New Keynesian Monetary Policy Models<br>*by Malin Adolfson, Stefan Laseén, Jesper Lindé and Marco Ratto* | 2018:362 |
| The Macroeconomic Effects of Trade Tariffs: Revisiting the Lerner Symmetry Result<br>*by Jesper Lindé and Andrea Pescatori* | 2019:363 |
| Biased Forecasts to Affect Voting Decisions? The Brexit Case<br>*by Davide Cipullo and André Reslow* | 2019:364 |
| The Interaction Between Fiscal and Monetary Policies: Evidence from Sweden<br>*by Sebastian Ankargren and Hovick Shahnazarian* | 2019:365 |
| Designing a Simple Loss Function for Central Banks: Does a Dual Mandate Make Sense?<br>*by Davide Debortoli, Jinill Kim and Jesper Lindé* | 2019:366 |
| Gains from Wage Flexibility and the Zero Lower Bound<br>*by Roberto M. Billi and Jordi Galí* | 2019:367 |
| Fixed Wage Contracts and Monetary Non-Neutrality<br>*by Maria Björklund, Mikael Carlsson and Oskar Nordström Skans* | 2019:368 |
| The Consequences of Uncertainty: Climate Sensitivity and Economic Sensitivity to the Climate<br>*by John Hassler, Per Krusell and Conny Olovsson* | 2019:369 |
| Does Inflation Targeting Reduce the Dispersion of Price Setters' Inflation Expectations?<br>*by Charlotte Paulie* | 2019:370 |


Sveriges Riksbank
Visiting address: Brunkebergs torg 11
Mail address: se-103 37 Stockholm

Website: www.riksbank.se
Telephone: +46 8 787 00 00, Fax: +46 8 21 05 31
E-mail: registratorn@riksbank.se