Importance Sampling Squared for Bayesian Inference in Latent Variable Models

M.-N. Tran* M. Scharth* M. K. Pitt† R. Kohn*

July 16, 2014

Abstract

This article describes an approach to Bayesian inference that uses importance sampling to generate the parameters for models where the likelihood is analytically intractable but can be unbiasedly estimated. We refer to the proposed procedure as importance sampling squared (IS^2), as we can often estimate the likelihood itself by importance sampling or sequential importance sampling. We provide a formal justification for IS^2 and study its convergence properties. We analyze the effect of estimating the likelihood on the resulting inference and provide guidelines on how to set up the precision of the likelihood estimate in order to obtain an optimal tradeoff between computational cost and accuracy for posterior inference on the model parameters. We also show how to use IS^2 to accurately and optimally estimate the marginal likelihood, irrespective of whether the model is estimated by MCMC or IS^2. The IS^2 procedure is illustrated empirically for a generalized multinomial logit model and a stochastic volatility model. The results show that the IS^2 method can lead to fast and accurate posterior inference under the optimal implementation. The proposed approach tackles the same set of problems as the pseudo marginal Metropolis Hastings approach of Andrieu and Roberts (2009) but has several advantages over it.

KEYWORDS: Efficient importance sampling, Marginal likelihood, Multinomial logit, Pseudo marginal Metropolis-Hastings, Optimal number of particles, Stochastic volatility.

1 Introduction

In many statistical models the likelihood is computationally intractable, but the density of the observations can be computed conditionally on the parameters and a vector of latent variables. Then importance sampling (IS) and, more generally, sequential IS can be used to estimate the likelihood unbiasedly. For example, in a nonlinear and non-Gaussian state space model, the likelihood is intractable but the density of the observations conditional on the parameters and the states is available and the likelihood can be estimated unbiasedly using the particle filter (Andrieu et al., 2010). Generalized linear mixed models (Fitzmaurice et al., 2011) provide a

*Australian School of Business, University of New South Wales
†Department of Economics, University of Warwick

second example, with the likelihood estimated unbiasedly by IS. Our article refers to N as the number of particles used to estimate the likelihood, although in some application N can be more accurately called the number of samples.

We consider IS for Bayesian inference when working with an estimated likelihood. Our first contribution shows that IS is still valid for estimating expectations with respect to the exact posterior when the likelihood is estimated unbiasedly, and prove a law of large numbers and a central limit theorem for the IS estimators. We refer to this procedure as importance sampling squared (IS^2). An important question is how much asymptotic efficiency is lost when working with an estimated likelihood compared to the efficiency we would obtain when working with the likelihood. The second contribution of this article addresses this question by comparing the asymptotic variance obtained when the likelihood is estimated with that obtained when the likelihood is available. We show that the ratio of the former to the latter is an inflation factor that is greater than or equal to 1, and is equal to 1 if and only if the estimate of the likelihood is exact. The inflation factor increases exponentially with the variance of the log-likelihood estimate. This result allows us to understand how much relative accuracy is lost when working with an estimated likelihood.

A critical issue in practice is the choice of the number of particles N. A large N gives a more accurate estimate of the likelihood at greater computational cost, while a small N can lead to a very large estimator variance. The third contribution of the article is to provide a theory and practical guidelines on selecting N to obtain an optimal tradeoff between accuracy and computational cost. Our results show that the efficiency of IS^2 is weakly sensitive to the number of particles around its optimal value. Moreover, the loss of efficiency decreases at worse linearly when we choose N higher than the optimal value, whereas the efficiency can deteriorate exponentially when N is below the optimal. We therefore advocate a conservative choice of N in practice. We propose two approaches for selecting the number of particles. The first approach is static because selects the same number of particles for all parameter values. The second approach is dynamic because it selects the optimal number of particles depending on the value of the parameter. We show both theoretically and empirically that the dynamic approach is more efficient than the static approach.

A fourth contribution is to show how IS^2 can estimate the marginal likelihood accurately and efficiently. It is important to point out that it is still computationally attractive to estimate the marginal likelihood using IS^2 even if the model itself is estimated by MCMC methods.

In general, selecting a reliable proposal density in IS^2 may sometimes be difficult. We describe two approaches for designing efficient proposal densities when only an estimated likelihood is available. This article therefore describes a complete solution for Bayesian inference using importance sampling for models with intractable likelihood.

Beaumont (2003) develops a pseudo marginal Metropolis Hastings (PMMH) scheme to carry out Bayesian inference with an estimated likelihood. Andrieu and Roberts (2009) formally study Beaumont's method and give conditions under which the chain converges. Andrieu et al. (2010) use MCMC for inference in state space models where the likelihood is estimated by the particle filter, and Pitt et al. (2012) and Doucet et al. (2014) discuss the issue of the optimal number of particles to be used in likelihood estimation. Chopin et al. (2013)

consider a sequential Monte Carlo procedure for sequential inference in space state models in which the intractable likelihood increments are estimated by sequential Monte Carlo, calling their method SMC^2.

Given a statistical model with an unbiased estimate of the likelihood, one can choose to carry out Bayesian inference using either MCMC as in Andrieu et al. (2010) and Pitt et al. (2012) or IS^2 as proposed in our article. There are several advantages in following the IS^2 approach. First, our empirical results suggest that IS^2 is statistically more efficient than PMMH for a given computational cost. Second, MCMC requires computationally expensive burn-in draws and the assessment of the convergence of the Markov chain. Third, IS^2 makes use of all the draws from the proposal for the parameters, whereas MCMC loses the information from rejected values. Fourth, it is simple to implement variance reduction methods such as antithetic sampling and stratified mixture sampling for IS^2, as well as using Quasi-Monte Carlo techniques to improve numerical efficiency. Fifth, the IS^2 method can be used to automatically estimate the marginal likelihood of a model. Estimating the marginal likelihood accurately using MCMC can be much more difficult, when Metropolis-Hastings moves are involved. Finally, it is much simpler to parallelize the IS^2 procedure than PMMH unless only independent Metropolis Hastings moves are used in PMMH.

It is simple to implement the IS^2 method. We illustrate the method in empirical applications for the generalized multinomial logit (GMNL) model of Fiebig et al. (2010) and a two factor stochastic volatility (SV) model with leverage effects. For the GMNL model, we show that the method for optimally selecting the number of particles leads to improved performance, with inadequate choices of N leading to inaccurate estimates under IS^2, and that IS^2 leads to accurate estimation of the posterior distribution when optimally implemented. We also show how to use Quasi-Monte Carlo techniques to improve the numerical efficiency of IS^2. The SV application is based on daily returns of the S&P 500 index between 1990 and 2012. We show that the variance of theh log-likelihood estimates based the particle efficient importance sampling method of Scharth and Kohn (2013) is small for this problem, despite the long time series. Hence, IS^2 leads to highly accurate estimates of the posterior statistics for this example in a short amount of computing time. As suggested by the theory, the efficiency of IS^2 is insensitive to the choice of N in this example, as long as the number of particles is not substantially higher than the optimal value of 8. We find that as few as two particles (including an antithetic draw) lead to a highly efficient procedure for this model.

The rest of the paper is organized as follows. Section 2 sets out the class of models with latent variables that we wish to estimate. Section 3 presents the main results. Section 4 shows how to estimate the marginal likelihood optimally. Section 5 obtains the large sample properties of the log of the estimator of the likelihood that justifies the crucial assumption that it is normally distributed as well as the large sample properties of the static and dynamic estimators of the number of particles. Applications are presented in Section 6. Section 7 concludes. An appendix contains the proofs.

2 Models with latent variables

This section sets out the main class of models with latent variables that we wish to estimate. However, the application of the method presented in this article is not limited to this class of models and in Section 6.2, we apply the IS^2 method to a time series stochastic volatility model. The method can also be applied in the context of panel data estimation for 'Big Data' by using the subsampling approach of Quiroz et al. (2014).

We observe n individuals and for individual i we have the T_i observations $y_i = \{y_{i1}, ..., y_{iT_i}\}$. Corresponding to individual i, y_i there exists a latent vector α_i. Let $y = \{y_1, \ldots, y_n\}$ and $\alpha = \{\alpha_1, \ldots, \alpha_n\}$. For each observation y_{it}, there is a covariate vector x_{it}. The vector of unknown parameters in the model is denoted by θ.

Assuming that the individuals are independent, the joint density of the latent variables and the observations is decomposed as

$$p(y, \alpha|\theta) = \prod_{i=1}^{n} p_i(y_i, \alpha_i|\theta) = \prod_{i=1}^{n} p_i(y_i|\alpha_i, \theta)p(\alpha_i|\theta).$$

By assigning a prior $p(\theta)$ for θ, standard MCMC approaches can be used to sample from the joint posterior

$$p(\alpha, \theta|y) \propto p(y, \alpha|\theta)p(\theta)$$

by cycling between $p(\alpha|\theta,y)$ and $p(\theta|\alpha,y)$. These methods require computationally expensive burn-in draws and the assessment of the convergence of the Markov chain. They also suffer from the problem of slow mixing and are inefficient when the latent vector α is high dimensional.

In this article, we are interested in inference when the likelihood is analytically intractable and given by

$$p(y|\theta) = \prod_{i=1}^{n} p_i(y_i|\theta), \quad p_i(y_i|\theta) = \int p_i(y_i|\alpha_i, \theta)p(\alpha_i|\theta)d\alpha_i.$$

More specific structures are considered in Section 6. We will use an unbiased IS estimator $\widehat{p}_N(y|\theta)$ of $p(y|\theta)$ based upon a simulation sample size of N, which we shall call the number of particles. Let $h_i(\alpha_i|y,\theta)$ be an importance density for α_i. The density $p_i(y_i|\theta)$ is estimated unbiasedly by

$$\widehat{p}_{N,i}(y_i|\theta) = \frac{1}{N}\sum_{j=1}^{N} w_i(\alpha_i^{(j)}, \theta), \quad w_i(\alpha_i^{(j)}, \theta) = \frac{p(y_i|\alpha_i^{(j)}, \theta)p(\alpha_i^{(j)}|\theta)}{h_i(\alpha_i^{(j)}|y, \theta)}, \quad \alpha_i^{(j)} \stackrel{iid}{\sim} h_i(\cdot|y, \theta). \tag{1}$$

Hence,

$$\widehat{p}_N(y|\theta) = \prod_{i=1}^{n} \widehat{p}_{N,i}(y_i|\theta)$$

is an unbiased estimator of the likelihood $p(y|\theta)$. We note that it is possible to use a different number of particles N_i for each individual i and we shall do so in the empirical examples in Section 6

3 Importance sampling squared

Let $p(\theta)$ be the prior for θ, $p(y|\theta)$ the likelihood and $\pi(\theta)\propto p(\theta)p(y|\theta)$ the posterior distribution defined on the space $\Theta\subset\mathbb{R}^d$. Bayesian inference typically requires computing an integral of the form

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

for some function φ on Θ such that the integral (2) exists. When the likelihood can be evaluated, IS is a popular method for estimating this integral (see, e.g., Geweke, 1989). Let $g_{\text{IS}}(\theta)$ be such an importance density. Then, $\mathbb{E}_\pi(\varphi)$ can be estimated by the IS estimator

$$\widehat{\varphi}_{\text{IS}} = \frac{\frac{1}{M}\sum_{i=1}^M \varphi(\theta_i)w(\theta_i)}{\frac{1}{M}\sum_{i=1}^M w(\theta_i)}, \text{ with weights } w(\theta_i) = \frac{p(\theta_i)p(y|\theta_i)}{g_{\text{IS}}(\theta_i)}, \quad \theta_i \overset{iid}{\sim} g_{\text{IS}}(\theta). \tag{3}$$

It can be shown that $\widehat{\varphi}_{\text{IS}} \xrightarrow{a.s.} \mathbb{E}_\pi(\varphi)$ as $M\to\infty$, and a central limit theorem also holds for $\widehat{\varphi}_{\text{IS}}$; see Geweke (1989) and Remark 1 below.

When the likelihood cannot be evaluated, standard IS is no longer used because the weights $w(\theta_i)$ in (3) are unavailable. We now present the IS^2 scheme for estimating the integral (2) when the likelihood $p(y|\theta)$ is estimated unbiasedly by $\widehat{p}_N(y|\theta)$.

Assumption 1. *$\mathbb{E}[\widehat{p}_N(y|\theta)]=p(y|\theta)$ for every $\theta\in\Theta$, where the expectation is with respect to the random variables occurring in the process of estimating the likelihood.*

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

1. *Generate $\theta_i \overset{iid}{\sim} g_{\text{IS}}(\theta)$ and compute the likelihood estimate $\widehat{p}_N(y|\theta_i)$.*
2. *Compute the weight $\widetilde{w}(\theta_i)=p(\theta_i)\widehat{p}_N(y|\theta_i)/g_{\text{IS}}(\theta_i)$.*

Then, the IS^2 estimator of $\mathbb{E}_\pi(\varphi)$ is

$$\widehat{\varphi}_{\text{IS}^2} = \frac{\frac{1}{M}\sum_{i=1}^M \varphi(\theta_i)\widetilde{w}(\theta_i)}{\frac{1}{M}\sum_{i=1}^M \widetilde{w}(\theta_i)}. \tag{4}$$

Algorithm 1 is identical to the standard IS scheme, except that the likelihood is replaced by its estimate. To see that $\widehat{\varphi}_{\text{IS}^2}$ is a valid estimator of $\mathbb{E}_\pi(\varphi)$, we follow Pitt et al. (2012) and write $\widehat{p}_N(y|\theta)$ as $p(y|\theta)e^z$, where $z=\log\widehat{p}_N(y|\theta)-\log p(y|\theta)$ is a random variable whose distribution is governed by the randomness occurring when estimating the likelihood $p(y|\theta)$. Let $g_N(z|\theta)$ be the density of z. Assumption 1 implies that $\mathbb{E}(e^z|\theta)=\int_{\mathbb{R}} e^z g_N(z|\theta)dz=1$. We define

$$\widetilde{\pi}_N(\theta,z) := p(\theta)g_N(z|\theta)p(y|\theta)e^z/p(y), \tag{5}$$

as the joint posterior density of θ and z on the extended space $\widetilde{\Theta}=\Theta\otimes\mathbb{R}$. Then, $\widetilde{\pi}_N(\theta)=\pi(\theta)$ because

$$\int_{\mathbb{R}} \widetilde{\pi}_N(\theta,z)dz = p(\theta)p(y|\theta)/p(y) = \pi(\theta).$$

The integral (2) can be written as

$$\mathbb{E}_\pi(\varphi) = \int_{\widetilde{\Theta}} \varphi(\theta)\pi_N(\theta,z)d\theta dz = \frac{1}{p(y)}\int_{\widetilde{\Theta}} \varphi(\theta)\frac{p(\theta)p(y|\theta)e^z g_N(z|\theta)}{\widetilde{g}_{\mathrm{IS}}(\theta,z)}\widetilde{g}_{\mathrm{IS}}(\theta,z)d\theta dz, \tag{6}$$

with $\widetilde{g}_{\mathrm{IS}}(\theta,z) = g_{\mathrm{IS}}(\theta)g_N(z|\theta)$ an importance density on $\widetilde{\Theta}$. Let $(\theta_i,z_i) \overset{iid}{\sim} \widetilde{g}_{\mathrm{IS}}(\theta,z)$, i.e. generate $\theta_i \sim g_{\mathrm{IS}}(\theta)$ and then $z_i \sim g_N(z|\theta_i)$. It is straightforward to see that the estimator $\widehat{\varphi}_{\mathrm{IS}^2}$ defined in (4) is exactly an IS estimator of the integral defined in (6), with importance density $\widetilde{g}_{\mathrm{IS}}(\theta,z)$ and weights

$$\widetilde{w}(\theta_i,z_i) = \frac{p(\theta_i)p(y|\theta_i)e^{z_i}g_N(z_i|\theta_i)}{\widetilde{g}_{\mathrm{IS}}(\theta_i,z_i)} = \frac{p(\theta_i)\widehat{p}_N(y|\theta_i)}{g_{\mathrm{IS}}(\theta_i)} = \widetilde{w}(\theta_i). \tag{7}$$

This formally justifies Algorithm 1. Theorem 1 gives some asymptotic properties (in M) of IS^2 estimators. Its proof is in the Appendix.

Theorem 1. *Suppose that Assumption 1 holds, $\mathbb{E}_\pi(\varphi)$ exists and is finite, and* $\mathrm{Supp}(\pi) \subseteq \mathrm{Supp}(g_{IS})$, *where* $\mathrm{Supp}(\pi)$ *denotes the support of the distribution π. Then,*

(i) $\widehat{\varphi}_{\mathrm{IS}^2} \xrightarrow{a.s.} \mathbb{E}_\pi(\varphi)$ *as $M \to \infty$, for any $N \geq 1$.*

(ii) In addition, if

$$\int h(\theta)^2\left(\frac{\pi(\theta)}{g_{\mathrm{IS}}(\theta)}\right)^2\left(\int \exp(2z)g_N(z|\theta)dz\right)g_{\mathrm{IS}}(\theta)d\theta \tag{8}$$

is finite for $h(\theta) = \varphi(\theta)$ and $h(\theta) = 1$ for all N, then

$$\sqrt{M}\Big(\widehat{\varphi}_{\mathrm{IS}^2} - \mathbb{E}_\pi(\varphi)\Big) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2_{\mathrm{IS}^2}(\varphi)\big), \quad M \to \infty,$$

where the asymptotic variance in M for fixed N is given by

$$\sigma^2_{\mathrm{IS}^2}(\varphi) = \mathbb{E}_\pi\left\{\big(\varphi(\theta) - \mathbb{E}_\pi(\varphi)\big)^2\frac{\pi(\theta)}{g_{\mathrm{IS}}(\theta)}\mathbb{E}_{g_N}[\exp(2z)]\right\}. \tag{9}$$

(iii) Define

$$\widehat{\sigma^2_{\mathrm{IS}^2}(\varphi)} = \frac{M\sum_{i=1}^M\big(\varphi(\theta_i) - \widehat{\varphi}_{\mathrm{IS}^2}\big)^2\widetilde{w}(\theta_i)^2}{\Big(\sum_{i=1}^M \widetilde{w}(\theta_i)\Big)^2}. \tag{10}$$

Then, under the conditions in (ii), $\widehat{\sigma^2_{\mathrm{IS}^2}(\varphi)} \xrightarrow{a.s.} \sigma^2_{\mathrm{IS}^2}(\varphi)$ *as $M \to \infty$, for given N.*

We note that both $\sigma^2_{\mathrm{IS}^2}(\varphi)$ and $\widehat{\sigma^2_{\mathrm{IS}^2}(\varphi)}$ depend on N, but, for simplicity, we suppress the explicit dependence in the notation. Here, all the probabilistic statements, such as the almost sure convergence, must be understood on the extended probability space which takes into account the extra randomness occurring when estimating the likelihood. The result (iii) is practically useful, because (10) allows us to estimate the asymptotic variance of the estimator.

Remark 1. *The standard IS scheme using the likelihood is a special case of Algorithm 1. In this case,* $\widehat{p}_N(y|\theta)=p(y|\theta)$*, the weights* $\widetilde{w}(\theta_i)=w(\theta_i)$ *and therefore* $\widehat{\varphi}_{\mathrm{IS}^2}=\widehat{\varphi}_{\mathrm{IS}}$*. Note that* $g_N(z|\theta)=\delta_0(z)$*, the delta Dirac distribution concentrated at zero. From Theorem 1,* $\widehat{\varphi}_{\mathrm{IS}} \xrightarrow{a.s.} \mathbb{E}_\pi(\varphi)$*, and*

$$\sqrt{M}\Big(\widehat{\varphi}_{\mathrm{IS}}-\mathbb{E}_\pi(\varphi)\Big)\xrightarrow{d}\mathcal{N}(0,\sigma^2_{\mathrm{IS}}(\varphi)), \tag{11}$$

with

$$\sigma^2_{\mathrm{IS}}(\varphi)=\mathbb{E}_\pi\left\{\big(\varphi(\theta)-\mathbb{E}_\pi(\varphi)\big)^2\frac{\pi(\theta)}{g_{\mathrm{IS}}(\theta)}\right\}, \tag{12}$$

which can be estimated by

$$\widehat{\sigma^2_{\mathrm{IS}}(\varphi)}=\frac{M\sum_{i=1}^M\big(\varphi(\theta_i)-\widehat{\varphi}_{\mathrm{IS}}\big)^2w(\theta_i)^2}{\Big(\sum_{i=1}^M w(\theta_i)\Big)^2}. \tag{13}$$

These convergence results are well known in the literature; see, e.g. Geweke (1989).

3.1 The effect of estimating the likelihood in IS^2

The results in the previous section show that it is straightforward to use importance sampling even when the likelihood is intractable but unbiasedly estimated. This section addresses the question of how much asymptotic efficiency is lost when working with an estimated likelihood. We follow Pitt et al. (2012) and Doucet et al. (2014) and make the following idealized assumption to make it possible to develop some theory. This assumption is later justified by Proposition 3 of Section 5.

Assumption 2.
1. *There exists a function* $\gamma^2(\theta)$ *such that the density* $g_N(z|\theta)$ *of* z *is* $\mathcal{N}(-\frac{\gamma^2(\theta)}{2N},\frac{\gamma^2(\theta)}{N})$*, where* $\mathcal{N}(a,b^2)$ *is a univariate normal density with mean* a *and variance* b^2*.*
2. *For a given* $\sigma^2>0$*, let* $N_{\sigma^2}(\theta)=\gamma^2(\theta)/\sigma^2$*. Then,* $\mathbb{V}(z|\theta,N=N_{\sigma^2}(\theta))\equiv\sigma^2$ *for all* $\theta\in\Theta$*.*

Note that if $g_N(z|\theta)$ is Gaussian, then its mean must be $-\frac{1}{2}$ times its variance in order that $\mathbb{E}_{g_N}(\exp(z))=1$, i.e., for Assumption 1 to hold. Assumption 2 (ii) keeps the variance $\mathbb{V}(z|\theta,N)$ constant across different values of θ, thus making it easy to associate the IS^2 asymptotic variances with σ. Under Assumption 2, the density $g_N(z|\theta)$ depends only on σ and is written as $g(z|\sigma)$.

Lemma 1. *If Assumption 2 holds for a fixed* σ^2*, then* (8) *becomes*

$$\int h(\theta)^2\left(\frac{\pi(\theta)}{g_{\mathrm{IS}}(\theta)}\right)^2 g_{\mathrm{IS}}(\theta)d\theta<\infty. \tag{14}$$

for both $h=\varphi$ *and* $h=1$*.*

These are the standard conditions for IS (Geweke, 1989). The proof of this lemma is straightforward and omitted.

Recall that $\sigma^2_{\rm IS}(\varphi)/M$ and $\sigma^2_{\rm IS^2}(\varphi)/M$ are respectively the asymptotic variances of the IS estimators we would obtain when the likelihood is available and when the likelihood is estimated. We refer to the the ratio $\sigma^2_{\rm IS^2}(\varphi)/\sigma^2_{\rm IS}(\varphi)$ as the inflation factor. Theorem 2 obtains an expression for the inflation factor, shows that it is independent of φ, greater than 1 and increases exponentially with σ^2. Its proof is in the Appendix.

Theorem 2. *Under Assumption 2 and the conditions in Theorem 1,*

$$\frac{\sigma^2_{\rm IS^2}(\varphi)}{\sigma^2_{\rm IS}(\varphi)} = \exp(\sigma^2). \tag{15}$$

3.2 An optimal choice of the number of particles N

In practice, it is necessary to tradeoff the cost of estimating the likelihood and the accuracy of the IS^2 estimator $\widehat{\varphi}_{\rm IS^2}$ in (4). A large number of particles N results in a precise likelihood estimate, and therefore an accurate estimate of $\mathbb{E}_\pi(\varphi)$, but at a greater computational cost. A small N leads to a large variance of the likelihood estimator, so that it is necessary to have a larger number of importance samples M in order to obtain a desired accuracy of the IS^2 estimator. In either case, the computation is expensive. It is important to select an optimal value of N that minimizes the computational cost.

The time to compute the likelihood estimate $\widehat{p}_N(y|\theta)$ is as $\tau_0+N\tau_1$, where $\tau_0\geq 0$ and $\tau_1>0$. For example, if $\widehat{p}_N(y|\theta)$ is given in (1), then τ_0 is the overhead cost for designing n importance densities $h_i(\cdot|y,\theta), i=1,...,n$; $\tau_0=0$ if these densities are fixed. Given these densities, $N\tau_1$ is the computing time used to compute $\widehat{p}_N(y|\theta)$ based on N particles. It is reasonable to assume that τ_0 and τ_1 are constants independent of θ. Under Assumption 2, N depends on θ as $N=N_{\sigma^2}(\theta)=\gamma^2(\theta)/\sigma^2$.

From Theorems 1 and 2, the variance of the estimator $\widehat{\varphi}_{\rm IS^2}$ based on M importance samples from $g_{\rm IS}(\theta)$ is approximated by

$$\mathbb{V}(\widehat{\varphi}_{\rm IS^2}) \approx \frac{\sigma^2_{\rm IS^2}(\varphi)}{M} = \frac{\sigma^2_{\rm IS}(\varphi)\exp(\sigma^2)}{M}. \tag{16}$$

Let P^* be a prespecified precision. Then, we need $M=\sigma^2_{\rm IS}(\varphi)\exp(\sigma^2)/P^*$ in order for the estimator to have that precision. From Algorithm 1 and the law of large numbers, the required computing time is

$$\sum_{i=1}^M \left(\tau_0+\frac{\gamma^2(\theta_i)}{\sigma^2}\tau_1\right) \approx M\left(\tau_0+\frac{\tau_1}{\sigma^2}\mathbb{E}_{g_{\rm IS}}(\gamma^2(\theta))\right) = \frac{\sigma^2_{\rm IS}(\varphi)}{P^*}\left(\tau_0+\frac{\tau_1}{\sigma^2}\mathbb{E}_{g_{\rm IS}}(\gamma^2(\theta))\right)\exp(\sigma^2).$$

Therefore, the product

$$\mathrm{CT}^*(\sigma^2) = \exp(\sigma^2)\left(\tau_0+\frac{\tau_1}{\sigma^2}\bar{\gamma}^2\right) \tag{17}$$

can be used to characterize the computing time as a function of σ^2, where $\bar{\gamma}^2 = \mathbb{E}_{g_{\text{IS}}}(\gamma^2(\theta))$. It is straightforward to check that $\text{CT}^*(\sigma^2)$ is convex and minimized at

$$\sigma^2_{\text{opt}} = \frac{-\tau_1 + \sqrt{\tau_1^2 + 4\tau_0\tau_1/\bar{\gamma}^2}}{2\tau_0/\bar{\gamma}^2} \text{ if } \tau_0 \neq 0 \text{ and } \sigma^2_{\text{opt}} = 1 \text{ if } \tau_0 = 0. \tag{18}$$

The optimal number of particles N is such that $\mathbb{V}(z|\theta, N = N_{\sigma^2}(\theta)) = \mathbb{V}(\log \widehat{p}_N(y|\theta, N = N_{\sigma^2}(\theta)) = \sigma^2_{\text{opt}}$, for $\sigma^2 = \sigma^2_{\text{opt}}$.

Practical guidelines for selecting an optimal N. Let $\mathbb{V}_{N,\theta}(z) = \mathbb{V}(z|N,\theta)$ and let $\widehat{\mathbb{V}}_{N,\theta}(z)$ be an estimate of $\mathbb{V}_{N,\theta}(z)$. We suggest the following practical guidelines for tuning the optimal number of particles N. Section 5 gives another strategy. Note that N generally depends on θ, but this dependence is suppressed for notational simplicity.

If $\tau_0 = 0$, from (18), it is necessary to tune N such that $\widehat{\mathbb{V}}_{N,\theta}(z) = 1$. A simple strategy is to start with some small N and increase it if $\widehat{\mathbb{V}}_{N,\theta}(z) > 1$.

If $\tau_0 > 0$, we first need to estimate $\bar{\gamma}^2 = \mathbb{E}_{g_{\text{IS}}}(\gamma^2(\theta))$. Let $\{\theta^{(1)}, ..., \theta^{(J)}\}$ be J draws from the importance density $g_{\text{IS}}(\theta)$. Then, starting with some large N_0 , $\overline{\gamma^2}$ can be estimated by

$$\widehat{\bar{\gamma}}^2 = \frac{1}{J}\sum_{j=1}^{J} \widehat{\gamma}^2(\theta^{(j)}) = \frac{N_0}{J}\sum_{j=1}^{J} \widehat{\mathbb{V}}_{N_0,\theta^{(j)}}(z), \tag{19}$$

as $\widehat{\mathbb{V}}_{N_0,\theta^{(j)}}(z) = \widehat{\gamma}^2(\theta^{(j)})/N_0$. By substituting $\widehat{\bar{\gamma}}^2$ into (18), we obtain an estimate $\widehat{\sigma}^2_{\text{opt}}$ of σ^2_{opt}. Now, for each draw of θ_i in Algorithm 1, we start with some small N and increase N if $\widehat{\mathbb{V}}_{N,\theta_i}(z) > \widehat{\sigma}^2_{\text{opt}}$. Note that the new draw θ_i with the estimate $\widehat{\mathbb{V}}_{N,\theta_i}(z)$ can be added to (19) to improve the estimate $\widehat{\bar{\gamma}}^2$.

In the applications in Section 6, we use the *time normalized variance* (TNV) of an IS^2 estimator $\widehat{\varphi}_{\text{IS}^2}$ as a measure of inefficiency. We define TNV of an IS^2 estimator $\widehat{\varphi}_{\text{IS}^2}$ as

$$\text{TNV}(M, N) = \mathbb{V}(\widehat{\varphi}_{\text{IS}^2}) \times \Upsilon(M, N). \tag{20}$$

In (20), $N = N_1 + \cdots + N_M$, $\Upsilon(M,N) = M\tau_0 + N\tau_1$ is the total computing time of an IS^2 run with M importance samples $\{\theta_i, i=1,...,M\}$, generated by g_{IS}, for θ, and using N_i particles to estimate the likelihood at the value θ_i. To motivate TNV as a measure of inefficiency, we note that it increases with both the variance of the estimator and the total time taken, and that it is proportional to $CT^*(\sigma^2)$ when M is large and the N_i are tuned with θ, i.e. $N_i = N^2_\sigma(\theta)$. In this case, we denote the time normalized variance by $\text{TNV}(M,\sigma^2)$. We now show that $\text{TNV}(M,\sigma^2)$ is proprotional to $CT^*(\sigma^2)$.

$$\text{TNV}(M,\sigma^2) = \mathbb{V}(\widehat{\varphi}_{\text{IS}^2}) \times \Big(M\tau_0 + (\tau_1/\sigma^2)\sum_{i=1}^{M}\gamma^2(\theta_i)\Big) \approx \sigma^2_{IS}(\widehat{\varphi})\exp(\sigma^2)\Big(\tau_0 + (\tau_1/\sigma^2)E_{g_{IS}}(\gamma^2)\Big)$$

3.3 Effective sample size

The efficiency of the proposal density $g_{\text{IS}}(\theta)$ in the standard IS procedure (3) is often measured by the effective sample size defined as (Liu, 2001, p.35)

$$\text{ESS}_{\text{IS}} = \frac{M}{1 + \mathbb{V}_{g_{\text{IS}}}(w(\theta)/\mathbb{E}_{g_{\text{IS}}}[w(\theta)])} = \frac{M\big(\mathbb{E}_{g_{\text{IS}}}[w(\theta)]\big)^2}{\mathbb{E}_{g_{\text{IS}}}[w(\theta)^2]}. \tag{21}$$

A natural question is how to measure the efficiency of $g_{\mathrm{IS}}(\theta)$ in the IS^2 procedure. In the IS^2 context, the proposal density is $\widetilde{g}_{\mathrm{IS}}(\theta,z)=g_{\mathrm{IS}}(\theta)g_N(z|\theta)$ with the weight $\widetilde{w}(\theta,z)=w(\theta)e^z$ given in (7). Hence, the ESS for IS^2 is

$$\mathrm{ESS}_{\mathrm{IS}^2}=\frac{M\left(\mathbb{E}_{\widetilde{g}_{\mathrm{IS}}}[\widetilde{w}(\theta,z)]\right)^2}{\mathbb{E}_{\widetilde{g}_{\mathrm{IS}}}[\widetilde{w}(\theta,z)^2]}, \tag{22}$$

which can be estimated by

$$\widehat{\mathrm{ESS}}_{\mathrm{IS}^2}=\left(\sum_{i=1}^M\widetilde{w}(\theta_i)\right)^2\Big/\sum_{i=1}^M\widetilde{w}(\theta_i)^2.$$

Under Assumption 2, $\mathbb{E}_{\widetilde{g}_{\mathrm{IS}}}[\widetilde{w}(\theta,z)]=\mathbb{E}_{g_{\mathrm{IS}}}[w(\theta)]$ and $\mathbb{E}_{\widetilde{g}_{\mathrm{IS}}}[\widetilde{w}(\theta,z)^2]=e^{\sigma^2}\mathbb{E}_{g_{\mathrm{IS}}}[w(\theta)^2]$. Hence,

$$\mathrm{ESS}_{\mathrm{IS}}=\exp(\sigma^2)\times\mathrm{ESS}_{\mathrm{IS}^2}. \tag{23}$$

If the number of particles $N(\theta)$ is tuned to target a constant variance σ^2 of z, then (23) enables us to estimate $\mathrm{ESS}_{\mathrm{IS}}$ as if the likelihood was given. This estimate is a useful measure of the efficiency of the proposal density $g_{\mathrm{IS}}(\theta)$ in the IS^2 context.

3.4 Construction of the importance density $g_{\mathrm{IS}}(\theta)$

As is typical of IS algorithms, a potential drawback with IS^2 is that its performance depends on the proposal density $g_{\mathrm{IS}}(\theta)$ for θ, which may be difficult to obtain in complex models. This section outlines two approaches for designing efficient and reliable proposal densities for IS^2.

The first is based on the approach called Mixture of t by Importance Sampling Weighted Expectation Maximization (MitISEM) of Hoogerheide et al. (2012), which constructs a mixture of t densities for approximating a target distribution. The MitISEM works by minimizing the Kullback-Leibler divergence between the target and the t mixture. This approach can handle target distributions that have non-standard shapes such as multimodality and skewness.

The second is based on annealed importance sampling (AIS) of Neal (2001), which can explore the parameter space of θ efficiently and is useful in cases where the target distribution is multimodal. The AIS produces a set of weighted samples θ's which are first drawn from an easily-generated distribution and then moved towards the target distribution through Markov kernels. A very good proposal density $g_{\mathrm{IS}}(\theta)$ can then be obtained by fitting a mixture of t densities to the sample θ's generated by the AIS approach.

When the likelihood is available, the MitISEM and AIS can be used to design proposal densities that approximate the posterior accurately. It is natural to use an estimated likelihood when the exact likelihood is not available. Let us write $\widehat{p}_N(y|\theta)=\widehat{p}_N(y|\theta,u)$ with u a fixed random number stream for all θ. The target distribution in the MitISEM and AIS will be $p(\theta)\widehat{p}_N(y|\theta,u)/p(y)$ which can be considered as the posterior $p(\theta|y,u)$ conditional on the common random numbers u and y. Our procedure is analogous to using common random numbers u to obtain simulated maximum likelihood estimates of θ (see, e.g., Gourieroux and Monfort, 1995), except that we obtain a histogram estimate of the 'posterior' $p(\theta|u,y)\propto p(y|\theta,u)p(u)$. This 'posterior' is biased, but sufficiently good to obtain a good proposal density.

4 IS^2 for estimating the marginal likelihood

The marginal likelihood $p(y)=\int_\Theta p(\theta)p(y|\theta)d\theta$ is useful for model choice (see, e.g., Kass and Raftery, 1995). This section discusses how to estimate $p(y)$ optimally, reliably and unbiasedly by IS^2 when the likelihood is intractable. The IS^2 estimator of $p(y)$ is

$$\widehat{p}_{\mathrm{IS}^2}(y) = \frac{1}{M}\sum_{i=1}^{M}\widetilde{w}(\theta_i), \tag{24}$$

with the samples θ_i and the weights obtained from Algorithm 1. The following proposition shows the unbiasedness and robustness of this IS^2 estimator. The proof is in the Appendix.

Proposition 1. *(i) Under Assumption 1, $\mathbb{E}[\widehat{p}_{\mathrm{IS}^2}(y)]=p(y)$. (ii) If the condition in* (8) *holds, then there exists a finite constant K such that $\mathbb{V}(\widehat{p}_{\mathrm{IS}^2}(y))\leq K/M$.*

The standard IS estimator of $p(y)$ when the likelihood is given is $\widehat{p}_{\mathrm{IS}}(y)=\sum_{i=1}^{M}w(\theta_i)/\mathrm{M}$, with the weights $w(\theta_i)$ from (3). Let $W_i=w(\theta_i)/p(y)$ and $\widetilde{W}_i=\widetilde{w}(\theta_i)/p(y)=e^{z_i}W_i$. Under Assumption 2, $\mathbb{E}_{g_{\mathrm{IS}}}[W_i]=\mathbb{E}_{\widetilde{g}_{\mathrm{IS}}}[\widetilde{W}_i]=1$ and

$$\mathbb{V}_{\widetilde{g}_{\mathrm{IS}}}[\widetilde{W}_i] = \mathbb{E}_{\widetilde{g}_{\mathrm{IS}}}\left[e^{2z_i}W_i\right] - 1 = e^{\sigma^2}\mathbb{E}_{g_{\mathrm{IS}}}[W_i^2] - 1 = e^{\sigma^2}(\mathbb{V}_{g_{\mathrm{IS}}}[W_i]+1) - 1.$$

Then the relative inefficiency of the IS^2 and IS estimators is

$$\frac{\mathbb{V}_{\widetilde{g}_{\mathrm{IS}}}(\widehat{p}_{\mathrm{IS}^2}(y))}{\mathbb{V}_{g_{\mathrm{IS}}}(\widehat{p}_{\mathrm{IS}}(y))} = \frac{\mathbb{V}_{\widetilde{g}_{\mathrm{IS}}}[\widetilde{W}_i]}{\mathbb{V}_{g_{\mathrm{IS}}}[W_i]} = \frac{e^{\sigma^2}(\mathbb{V}_{g_{\mathrm{IS}}}[W_i]+1)-1}{\mathbb{V}_{g_{\mathrm{IS}}}[W_i]}. \tag{25}$$

Similarly to (15), the right side of (25) can be referred to as an inflation factor that shows the relative inefficiency of the IS^2 estimator of the marginal likelihood in comparison to the standard IS estimator.

Table 1: Sensitivity of the computing time to $\mathbb{V}_{g_{\mathrm{IS}}}[W_i]$ in estimating the marginal likelihood.

$v=\mathbb{V}_{g_{\mathrm{IS}}}[W_i]$	$\sigma^2_{\mathrm{min}}(v)$	$\mathrm{CT}^*_{\widehat{p}_{\mathrm{IS}^2}(y)}(\sigma^2_{\mathrm{opt}})/\mathrm{CT}^*_{\widehat{p}_{\mathrm{IS}^2}(y)}(\sigma^2_{\mathrm{min}})$
1	0.12	1.0199
5	0.16	1.0012
10	0.16	1.0003
100	0.17	1.0000
∞	0.17	1

Optimal N for estimating the marginal likelihood. Let P^* be a precision. As $\mathbb{V}_{\widetilde{g}_{\mathrm{IS}}}(\widehat{p}_{\mathrm{IS}^2}(y))=\mathbb{V}_{\widetilde{g}_{\mathrm{IS}}}(\widetilde{w})/M$, we need $M=\mathbb{V}_{\widetilde{g}_{\mathrm{IS}}}(\widetilde{w})/P^*$ in order for the variance equal to P^*. Using the notation in Section 3.2, the computing time is

$$\sum_{i=1}^{M}\left(\tau_0+\frac{\gamma^2(\theta_i)}{\sigma^2}\tau_1\right) \approx \frac{1}{P^*}\left(\tau_0+\frac{\tau_1}{\sigma^2}\bar{\gamma}^2\right)\mathbb{V}_{\widetilde{g}_{\mathrm{IS}}}(\widetilde{w}) = \frac{p(y)^2}{P^*}\left(\tau_0+\frac{\tau_1}{\sigma^2}\bar{\gamma}^2\right)(e^{\sigma^2}(\mathbb{V}_{g_{\mathrm{IS}}}[W_i]+1)-1).$$

Hence, +

$$\mathrm{CT}^*_{\widehat{p}_{\mathrm{IS}^2}(y)}(\sigma^2) = \left(\tau_0 + \frac{\tau_1}{\sigma^2}\bar{\gamma}^2\right)\left(e^{\sigma^2}(\mathbb{V}_{g_{\mathrm{IS}}}[W_i]+1)-1\right) \tag{26}$$

can be used to characterize the computing time as a function of σ^2. Denote $v = \mathbb{V}_{g_{\mathrm{IS}}}[W_i]$ and let $\sigma^2_{\min}(v)$ be the minimizer of $\mathrm{CT}^*_{\widehat{p}_{\mathrm{IS}^2}(y)}(\sigma^2)$. The following proposition summarizes some properties of $\sigma^2_{\min}(v)$. Its proof is in Appendix.

Proposition 2. *(i) For any value of v, $\mathrm{CT}^*_{\widehat{p}_{\mathrm{IS}^2}(y)}(\sigma^2)$ is a convex function of σ^2, therefore $\sigma^2_{\min}(v)$ is unique.*

(ii) $\sigma^2_{\min}(v)$ increases as v increases and $\sigma^2_{\min}(v) \longrightarrow \sigma^2_{opt}$ in (18) as $v \longrightarrow \infty$.

(iii) $\sigma^2_{\min}(v)$ is insensitive to v in the sense that it is a flat function of v for large v.

To illustrate, Table 1 shows $\sigma^2_{\min}(v)$ and the ratio $\mathrm{CT}^*_{\widehat{p}_{\mathrm{IS}^2}(y)}(\sigma^2_{\mathrm{opt}})/\mathrm{CT}^*_{\widehat{p}_{\mathrm{IS}^2}(y)}(\sigma^2_{\min})$ for some values of v, obtained for the values of τ_0, τ_1 and $\bar{\gamma}^2$ from the example in Section 6. The table shows that the values of $\sigma^2_{\min}$ and the ratios are insensitive to v, and $\sigma^2_{\min} \to \sigma^2_{\mathrm{opt}} = 0.17$ as v increases. From Proposition 2 and Table 1, we advocate using σ^2_{opt} in (18) as the optimal value of the variance of the log likelihood estimates in estimating the marginal likelihood.

5 Large sample results for the likelihood estimator

This section studies the large sample (in n) properties of the likelihood estimator for panel data that is described in Section 2, and obtains two results assuming that the number of particles $N = O(n)$. First, we justify Assumption 2 by showing the asymptotic normality of the error z in the log-likelihood estimator. Second, we obtain convergence rates of two estimators of the optimal variance, σ^2_{opt} of the error z in the log likelihood estimator. We show that the dynamic estimator of σ^2_{opt} that chooses the number of particles or samples N depending on θ is much more efficient for large n than the static estimator of σ^2_{opt} that chooses a single number of particles N for all θ. Using the notation in Section 2, the error in the log-likelihood estimator is

$$z = \sum_{i=1}^{n}\{\log \widehat{p}_{N,i}(y_i|\theta) - \log p_i(y_i|\theta)\}, \text{ where } \widehat{p}_{N,i}(y_i|\theta) = N^{-1}\sum_{j=1}^{N}\omega_i(\alpha_i^{(j)},\theta), \tag{27}$$

$\omega_i(\alpha_i,\theta) = p(y_i|\alpha_i,\theta)p(\alpha_i|\theta)/h_i(\alpha_i|y,\theta)$, and $\alpha_i^{(j)} \overset{iid}{\sim} h_i(\alpha_i|y,\theta)$.

Assumption 3 is used to prove Proposition 3.

Assumption 3. *We assume that for all θ and $i = 1,...,n$,*

(i) The proposal density $h_i(\alpha_i|y,\theta)$ is sufficiently heavy tailed so that $\mathbb{E}_{h_i}\{\omega_i(\alpha_i,\theta)^4\} < K_1$, where $K_1 > 1$.

(ii) $K_2^{-1} < p(y_i|\theta) \le K_2$, where $K_2 > 1$, so that $p(y_i|\theta)$ is bounded and also bounded away from zero.

(iii) $\mathbb{V}_{h_i}\{\omega_i(\alpha_i,\theta)^2\} > K_3 > 0$*, so that the importance weights are not all equal, i.e., the proposal is not perfect.*

Define,

$$\sigma_i^2(\theta) := \mathbb{V}_{h_i}\left(\frac{\omega_i(\alpha_i,\theta)}{p_i(y_i|\theta)}\Big|\theta\right) = \int \frac{\omega_i(\alpha_i,\theta)^2}{p_i(y_i|\theta)^2} h_i(\alpha_i|y,\theta)d\alpha_i - 1. \tag{28}$$

$$\psi(n,\theta) := \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2(\theta), \quad \widehat{\psi}_N(n,\theta) := \frac{1}{n}\sum_{i=1}^{n}\widehat{\sigma}_i^2(\theta), \text{ where } \widehat{\sigma}_i^2(\theta) := \frac{1}{N}\sum_{j=1}^{N}\Big(\frac{\omega_i(\alpha_i^{(j)},\theta)^2}{\widehat{p}_{N,i}(y_i|\theta)^2} - 1\Big), \tag{29}$$

$$\widehat{\mathbb{E}}(z) := -\frac{n\widehat{\psi}_N(\theta,n)}{2N} \text{ and } \widehat{\mathbb{V}}(z) := \frac{n\widehat{\psi}_N(\theta,n)}{N}. \tag{30}$$

Proposition 3 gives the following asymptotic results for the error z. Its proof is in the appendix.

Proposition 3. *Suppose Assumptions 1 and 3 hold. Then,*

(i)

$$\mathbb{E}(z|n,N,\theta) = -n\psi(\theta,n)/(2N) + O(n/N^2) \quad \textit{and} \quad \mathbb{V}(z|n,N,\theta) = n\psi(\theta,n)/N + O(n/N^2). \tag{31}$$

(ii)

$$\widehat{\psi}_N(\theta,n) = \psi(\theta,n) + O\big(N^{-1}\big) + O_p(n^{-\frac{1}{2}}N^{-\frac{1}{2}}). \tag{32}$$

(iii)

$$\widehat{\mathbb{E}}(z) = \mathbb{E}(z) + O(nN^{-2}) + O_p(n^{1/2}N^{-3/2}) \textit{ and } \widehat{\mathbb{V}} = \mathbb{V}(z) + O(nN^{-2}) + O_p(n^{1/2}N^{-3/2}). \tag{33}$$

(iv)

$$\mathbb{V}(z)^{-1/2}\{z - \mathbb{E}(z)\} \xrightarrow{d} \mathcal{N}(0,1) \textit{ and } \widehat{\mathbb{V}}(z)^{-1/2}\{z - \widehat{\mathbb{E}}(z)\} \xrightarrow{d} \mathcal{N}(0,1). \tag{34}$$

Remark 2. *(1) Proposition 3 justifies Assumption 2, with* $\gamma^2(\theta) = n\psi(n,\theta)$*.*

(2) The convergence of z *to normality will be fast, because* $N = O(n)$ *and from Part (vii) of Lemma 2 and* (43)*,* z *is a sum of* n *terms, each of which converges to normality.*

Proposition 4 provides two consistent estimators of the number of particles necessary to obtain an error z whose variance is approximately equal to a prespecified target variance κ^2. We would typically choose the constant $\kappa^2 = \sigma^2_{\text{opt}}$. We call the first a static estimator because the number of particles is fixed before the IS^2 scheme begins. We call the second a dynamic estimator because the number of particles depends on the proposed value of θ. The static estimator converges, to the target variance, at the rate $O_p(n^{-1/2})$ while the dynamic estimator converges at the faster rate $O_p(n^{-1})$.

Proposition 4. *We assume that:*

(i) For any in $\theta\in\Theta$,

$$\mathbb{V}(z|\theta,n,N)=n\psi(\theta,n)/N+O(n/N^2), \tag{35}$$

where $\psi(\theta,n)$ *is* $O(1)$ *in* n *uniformly in* $\theta\in\Theta$.

(ii) A consistent estimator $\widehat{\psi}_N(\theta,n)$ *(in* N*) of* $\psi(\theta,n)$ *is available such that, uniformly in* θ,

$$\widehat{\psi}_N(\theta,n)=\psi(\theta,n)+O\left(N^{-1}\right)+O_p(n^{-\frac{1}{2}}N^{-\frac{1}{2}}).$$

(iii) The posterior $\pi(\theta)$ *satisfies* $\lim_{n\to\infty}\mathbb{E}_\pi(\theta)=\overline{\theta}$ *and* $\lim_{n\to\infty}n\times\mathbb{V}_\pi(\theta)=\Sigma$.

(iv) We have a consistent estimator (in n*)* $\widehat{\theta}_n$ *of* $\overline{\theta}$.

(v) $\psi(\theta,n)$ *is a continuously differentiable function of* $\theta\in\Theta$.

Then,

A. Under Assumptions (i) to (v), suppose that an initial starting value $N_S=O(n)$ *is used and the number of particles* $N_{\widehat{\theta}_n}$ *is chosen statically as* $N_{\widehat{\theta}_n}:=n\widehat{\psi}_{N_S}(\widehat{\theta}_n,n)/\kappa^2$, *where* κ^2 *is a prespecified target variance. Then we have* $\mathbb{V}(z|\theta,n,N_{\widehat{\theta}_n})=\kappa^2+O_p(n^{-1/2})$.

B. Under Assumptions (i) and (ii), if N_θ *is chosen dynamically as* $N_\theta:=n\widehat{\psi}_{N_{\widehat{\theta}_n}}(\theta,n)/\kappa^2$ *where* $N_{\widehat{\theta}_n}=O(n)+O_p(1)$, *e.g. as in result A, then* $\mathbb{V}(z|\theta,n,N_\theta)=\kappa^2+O_p(n^{-1})$.

We note that assumptions (i) and (ii) are justified for the IS estimator by Parts (i) and (ii) of Proposition 3. The dynamic estimator is preferred because it converges faster and does not require any assumption on the behavior of the posterior $\pi(\theta)$. The efficiency of the dynamic estimator is shown empirically in Section 6.1.3. For part (iv) we note that any consistent estimator, such as a method of moments estimator, may be employed. The dynamic estimator is more expensive than the static estimator as it requires an extra computational cost to compute $\widehat{\psi}_{N_{\widehat{\theta}_n}}(\theta,n)$.

6 Applications

6.1 Mixed logit and generalized multinomial logit

6.1.1 Model and data

The generalized multinomial logit (GMNL) model of Fiebig et al. (2010) specifies the probability of individual i choosing alternative j at occasion t as

$$p(i\text{ chooses }j\text{ at }t|X_{it},\beta_i)=\frac{\exp(\beta_{0ij}+\sum_{k=1}^{K}\beta_{ki}x_{kijt})}{\sum_{h=1}^{J}\exp(\beta_{0ih}+\sum_{k=1}^{K}\beta_{ki}x_{kiht})}, \tag{36}$$

where $\beta_i = (\beta_{0i1},...,\beta_{0iJ},\beta_{1i},...,\beta_{Ki})'$ and $X_{it} = (x_{1i1t},...,x_{Ki1t},...,x_{1iJt},...,x_{KiJt})'$ are the vectors of utility weights and choice attributes respectively. The GMNL specifies the alternative specific constants as $\beta_{0ij} = \beta_{0j} + \eta_{0ij}$ with $\eta_{0ij} \sim \mathrm{N}(0,\sigma_{0j}^2)$ and the attribute weights as

$$\beta_{ki} = \lambda_i \beta_k + \gamma \eta_{ki} + (1-\gamma)\lambda_i \eta_{ki}, \qquad \lambda_i = \exp(-\delta^2/2 + \delta\zeta_i), \qquad k = 1,\ldots,K,$$

with $\eta_{ki} \sim \mathrm{N}(0,\sigma_k^2)$, $\zeta_i \sim \mathrm{N}(0,1)$. The expectation value of the scaling coefficients λ_i is one, which implies that $\mathrm{E}(\beta_{ki}) = \beta_k$.

When $\delta = 0$ (so that $\lambda_i = 1$ for all individuals) the GMNL model reduces to the mixed logit (MIXL) model, which we also consider in our analysis. The MIXL model captures heterogeneity in consumer preferences by allowing individuals to weight the choice attributes differently. By introducing taste heterogeneity, the MIXL specification avoids the restrictive independence of irrelevant alternatives property of the standard multinomial logit model (Fiebig et al., 2010). The GMNL model additionally allows for scale heterogeneity through the random variable λ_i, which changes all attribute weights simultaneously. In this model, choice behavior can therefore be more random for some consumers compared to others. The γ parameter weights the specification between two alternative ways of introducing scale heterogeneity into the model.

The parameter vector is $\theta = (\beta_{01},...\beta_{0J},\sigma_0^2,\beta_1,...,\beta_K,\sigma_1^2,...,\sigma_K^2,\delta^2,\gamma)'$, while the vector of latent variables for each individual is $x_i = (\eta_{0i},...,\eta_{Ki},\lambda_i)$. The likelihood is therefore

$$p(y|\theta) = \prod_{i=1}^{I} p(y_i|\theta) = \prod_{i=1}^{I} \left[\int \left(\prod_{t=1}^{T} p(y_{it}|x_i) \right) p(x_i) \mathrm{d}x_i \right], \tag{37}$$

where y_{it} is the observed choice, $y = (y_{11},...,y_{1T},...,y_{I1},...,y_{IT})'$ and $p(y_{it}|x_i)$ is given by the choice probability (36).

We consider an empirical application to the Pap smear data set used for simulated maximum likelihood estimation in Fiebig et al. (2010). In this data set, $I = 79$ women choose whether or not to have a Pap smear test ($J = 2$) on $T = 32$ choice scenarios. We let the observed choice for individual i at occasion t be $y_{it} = 1$ if the woman chooses to take the test and $y_{it} = 0$ otherwise. Table 2 lists the choice attributes and the associated coefficients. We impose the restriction that $\sigma_5^2 = 0$ in our illustrations since we have found no evidence of heterogeneity for this attribute beyond the scaling effect. The model (36) is identified by setting the coefficients as zero when not taking the test.

We specify the priors as $\beta_{01} \sim \mathrm{N}(0,100)$, $\sigma_{01} \propto (1+\sigma_{01}^2)^{-1}$, $\beta_k \sim \mathrm{N}(0,100)$, $\sigma_k \propto (1+\sigma_k^2)^{-1}$, for $k = 1,...,K$, $\delta \propto (1+\delta/0.2)^{-1}$, and $\gamma \sim \mathrm{U}(0,1)$. The standard deviation parameters have half-Cauchy priors, see Gelman (2006).

6.1.2 Implementation details

We calculate the likelihood (37) by integrating out the vector of latent variables for each individual separately and follow different approaches for the MIXL and GMNL models. For the MIXL model, we combine the efficient importance sampling (EIS) method of Richard and Zhang (2007) with the defensive sampling approach of Hesterberg (1995). The importance density is the two component defensive mixture

$$h(x_i|y_{i1},...,y_{iT}) = \pi h^{\mathrm{EIS}}(x_i|y_{i1},...,y_{iT}) + (1-\pi)p(x_i),$$

Table 2: Choice attributes for the pap smear data set

Choice attributes	Values	Associated parameters
Alternative specific constant for test	1	β_{01},σ_{01}
Whether patient knows doctor	0 (no), 1 (yes)	β_1,σ_1
Whether doctor is male	0 (no), 1 (yes)	β_2,σ_2
Whether test is due	0 (no), 1 (yes)	β_3,σ_3
Whether doctor recommends test	0 (no), 1 (yes)	β_4,σ_4
Test cost	{0, 10, 20, 30}/10	β_5

where $h^{\text{EIS}}(x_i|y_{i1},...,y_{iT})$ is a multivariate Gaussian importance density obtained using the EIS. We use the multivariate Gaussian EIS implementation of Koopman et al. (2013), which substantially reduces the computing time for obtaining the importance density (τ_0). Following Hesterberg (1995), the inclusion of the natural sampler $p(x_i)$ in the mixture ensures that the importance weights are bounded. We set the mixture weight as $\pi = 0.5$. For the GMNL model, we follow Fiebig et al. (2010) and use the model density $p(x_i)$ as an importance sampler. We implement this simpler approach for the GMNL model because the occurrence of large values of λ_i causes the defensive mixture estimates of the log-likelihood to be pronouncedly right skewed in this case.

The panel structure of the problem implies that the log-likelihood estimates are sums of independent estimates $\log \widehat{p}(y_i|\theta)$ for each individual. In order to target a certain variance σ^2 for the log-likelihood estimate, we choose the number of particles for each individual (N_i) and parameter combination θ such that $\mathbb{V}(\log \widehat{p}(y_i|x_i)) \approx \sigma^2/I$. We implement this scheme by using a certain number of initial importance samples and the jackknife method to estimate $\gamma_i^2(\theta)$, the asymptotic variance of $\log \widehat{p}(y_i|x_i)$, and selecting $N_i(\theta) = \widehat{\gamma}_i^2(\theta)I/\sigma^2$. The preliminary number of particles is $N = 20$ for the MIXL model and $N = 2{,}500$ for the GMNL model.

To obtain the parameter proposals $g_{\text{IS}}(\theta)$, we use the MitISEM approach with common random numbers as described in Section 3.4. The approach approximated the posterior of the two models as two component mixtures of multivariate Student's t distributions.

We implement two standard variance reduction methods at each IS stage: stratified mixture sampling and antithetic sampling. The first consists of sampling from each component at the exact proportion of the mixture weights. For example, when estimating the likelihood for the MIXL model we generate exactly πN draws from the efficient importance density $h^{\text{EIS}}(x_i|y_{i1},...,y_{iT})$ and $(1-\pi)N$ draws from $p(x_i)$. The antithetic sampling method consists of generating pairs of perfectly negatively correlated draws from each mixture component, see, e.g. Ripley (1987).

6.1.3 Optimal number of particles

We start by investigating the performance of IS for estimating the likelihood of each model using different schemes for selecting the number of particles. We generate 1,000 draws for the parameter vector θ from $g_{\text{IS}}(\theta)$. For each parameter combination, we obtain 100 independent

Table 3: The table shows the average variance, skewness and kurtosis of log-likelihood estimates over 1,000 draws of θ. The JB rejections row report the proportion of replications in which the Jarque-Bera tests rejects the null hypothesis of normality of the log-likelihood estimates at the 5% level.

	MIXL (EIS)			GMNL (natural sampler)		
	$N=24$	$N=48$	$\sigma^2\approx 0.5$	$N=8{,}000$	$N=16{,}000$	$\sigma^2\approx 1$
Variance	1.068	0.545	0.482	1.986	1.029	1.023
SD of variance	0.159	0.079	0.055	0.500	0.295	0.162
Skewness	0.035	0.051	0.024	0.002	0.002	-0.043
Kurtosis	3.017	3.035	2.991	2.929	2.928	2.953
JB rejections (5%)	0.070	0.090	0.060	0.046	0.037	0.051
Time (sec)	0.069	0.071	0.072	0.916	1.860	2.084

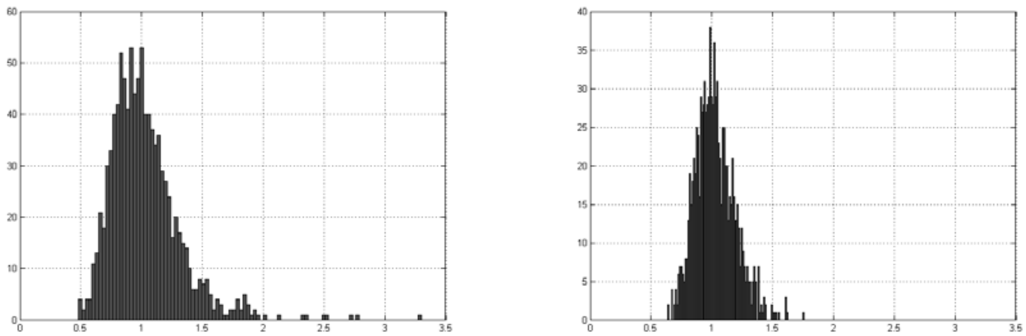
likelihood estimates using the same fixed number of particles for all individuals in the panel and by targeting a log-likelihood variance of 0.5 for the MIXL model and 1 for the GMNL model as described in the previous subsection. The number of particles in the fixed cases is $N=24$ and $N=48$ for the MIXL model and $N=8{,}000$ and $N=16{,}000$ for the GMNL model.

Table 3 shows the average sample variance, the standard deviation of the sample variance, the skewness and kurtosis of the log-likelihood estimates across the different draws for θ, as well as the proportion of parameter combinations for which the Jarque-Bera test rejects the normality of the log-likelihood estimates at the 5% level. The last row on the table reports computing times based on a computer equipped with an Intel Xeon 3.40 GHz processor with eight cores. Figure 1 displays the histograms of the sample variances for the GMNL model when $N=16{,}000$ and $\sigma^2\approx 1$ across the 1,000 draws for θ.

The results show that the algorithm for targeting constant σ^2 is effective in practice, approximately leading to the correct variance for the log-likelihood estimates on average without substantial deviations from the target precision across different parameter combinations. In comparison, a fixed number of particles leads to greater variability in precision across different draws of θ, especially for the GMNL model. The results also show that the log-likelihood estimates are consistent with the normality assumption as required by the theory. Figure 2 shows the average number of particles required for obtaining the target variance $\mathbb{V}(\log\widehat{p}(y_i|\theta))\approx\sigma^2/I$ for each individual. The figure reveals substantial variation in the average optimal number of particles across individuals, in particular for the GMNL model.

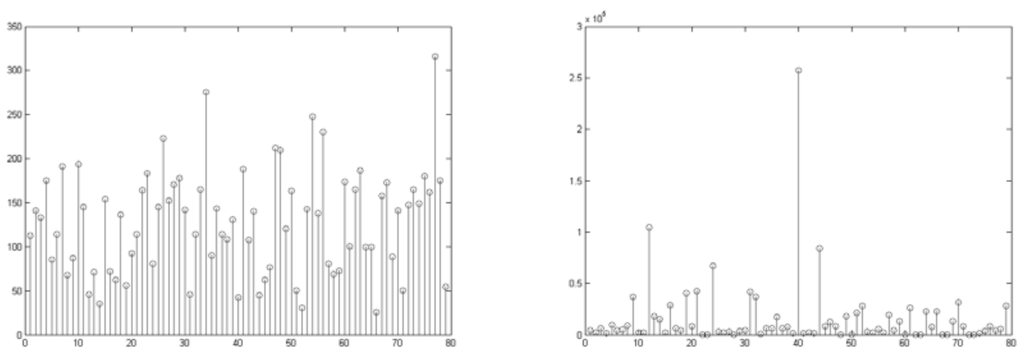
Table 3 also provides estimates of the required quantities for calculating the optimal σ^2 in (18). For the MIXL model, since the average variance when $N=24$ is 1.068, we estimate $\bar{\gamma}^2=1.068\times 24=25.63$. As the likelihood evaluation time using N particles is determined as $\tau_0+\tau_1 N$, from the computing times at $N=24$ and $N=48$, we have that $\tau_0=0.067$ and $\tau_1=8.97\times 10^{-5}$ seconds. Using (18), we conclude that σ^2_{opt} for the MIXL model is approximately 0.17. For the GMNL case, $\tau_0=0$ as there is no overhead cost in obtaining the state proposal, $\sigma^2_{\text{opt}}=1$.

To compare different implementations of IS^2, we generate $M=50{,}000$ draws from the



(a) $N = 16{,}000$ (b) Targeting $\sigma^2 = 1$ for each draw of θ

Figure 1: Histogram of the variance of the log-likelihood estimates for the GMNL model across different draws of the parameters and different schemes to select N.



(a) MIXL (b) GMNL

Figure 2: Average number of particles for each individual when targeting σ^2_{opt}.

Table 4: The MIXL model: The table shows the relative variances for posterior mean estimation using IS^2 based on $M = 50{,}000$ draws for θ and different σ^2.

	$\sigma^2 \approx 0.05$	$\sigma^2 \approx 0.17$	$\sigma^2 \approx 0.5$	$\sigma^2 \approx 0.75$
β_{01}	0.642	1.000	2.142	2.533
β_1	0.665	1.000	2.098	2.301
β_2	0.695	1.000	1.985	2.408
β_3	0.579	1.000	1.705	1.982
β_4	0.649	1.000	1.730	2.174
β_5	0.630	1.000	1.605	3.060
σ_{01}	0.590	1.000	2.232	2.668
σ_1	0.597	1.000	1.774	2.103
σ_2	0.635	1.000	1.898	2.440
σ_3	0.619	1.000	1.802	3.006
σ_4	0.926	1.000	1.420	2.757
Average MC	0.657	1.000	1.854	2.494
Theoretical MC	0.861	1.000	1.350	1.733
Time (minutes)	15.85	10.32	9.27	8.93
Estimated TNV	1.01	1.00	1.67	2.16
Theoretical TNV	1.25	1.00	1.23	1.54

parameter proposal $g_{\text{IS}}(\theta)$ and run the IS^2 by targeting different variances of the log-likelihood estimates. The target variances are $\sigma^2 = 0.05,\ 0.2,\ 0.75,\ 1$ for the MIXL model and $\sigma^2 = 0.5,\ 1,\ 1.5,\ 2$ for the GMNL model. We estimate the Monte Carlo (MC) variances of the posterior means under each implementation by the bootstrap method. Tables 4 and 5 show the relative estimated MC variances, the total actual computing time in minutes, and the relative estimated time normalized variance (20) for the MIXL and GMNL respectively. We also report for reference the theoretical relative MC variances and TNV

$$\frac{\mathbb{V}(\widehat{\varphi}_{\text{IS}^2})|_{\sigma^2=\sigma_1^2}}{\mathbb{V}(\widehat{\varphi}_{\text{IS}^2})|_{\sigma^2=\sigma_2^2}} = \frac{\exp(\sigma_1^2)}{\exp(\sigma_2^2)} \quad \text{and} \quad \frac{\text{TNV}(M,\sigma_1^2)}{\text{TNV}(M,\sigma_2^2)} = \frac{\exp(\sigma_1^2)(\tau_0 + \tau_1\bar{\gamma}^2/\sigma_1^2)}{\exp(\sigma_2^2)(\tau_0 + \tau_1\bar{\gamma}^2/\sigma_2^2)}.$$

See (16) and (20).

The results show that the relative estimated MC variances are on average close to their theoretical values, even though there is a large degree of variability in the estimates across the parameters (especially for higher values of σ^2). The estimated relative TNV are approximately consistent with their theoretical values for the optimal precision of the log-likelihood estimates.

6.1.4 Posterior analysis

Table 6 presents selected posterior statistics estimated by the IS^2 method under the optimal schemes (targeting $\sigma^2_{\text{opt}} = 0.17$ for the MIXL model and $\sigma^2_{\text{opt}} = 1$ for the GMNL model). We estimated the posterior distribution using $M = 50{,}000$ importance samples for the parameters.

Table 5: The GMNL model: The table shows the relative variances for posterior mean estimation using IS^2 based on $M = 50{,}000$ draws for θ and different σ^2.

	$\sigma^2 \approx 0.5$	$\sigma^2 \approx 1$	$\sigma^2 \approx 1.5$	$\sigma^2 \approx 2$
β_0	0.618	1.000	1.200	2.045
β_1	0.744	1.000	1.295	2.220
β_2	0.766	1.000	1.440	1.736
β_3	0.873	1.000	1.553	2.398
β_4	0.714	1.000	1.368	2.019
β_5	0.658	1.000	1.198	2.060
σ_0	0.638	1.000	1.342	1.620
σ_1	0.614	1.000	1.490	2.061
σ_2	0.729	1.000	1.412	1.942
σ_3	0.884	1.000	1.507	2.435
σ_4	0.824	1.000	1.425	2.229
δ	0.700	1.000	1.383	2.005
γ	0.686	1.000	1.324	1.776
Average MC	0.727	1.000	1.380	2.042
Theoretical MC	0.607	1.000	1.649	2.718
Time (minutes)	535	262	178	138
Estimated TNV	1.486	1.000	0.939	1.081
Theoretical TNV	1.213	1.000	1.099	1.359

Table 6: The table presents estimates of selected posterior statistics based on $M = 50{,}000$ importance samples for the parameters. The MC standard errors are in brackets.

	MIXL				GMNL			
	Mean	SD	90% CI		Mean	SD	90% CI	
β_{01}	−1.032 [0.002]	0.430 [0.002]	−1.731 [0.006]	−0.324 [0.006]	−1.183 [0.006]	0.458 [0.004]	−1.927 [0.010]	−0.412 [0.015]
β_1	0.831 [0.001]	0.234 [0.001]	0.451 [0.003]	1.223 [0.003]	1.092 [0.005]	0.348 [0.007]	0.584 [0.006]	1.702 [0.014]
β_2	−1.405 [0.001]	0.352 [0.002]	−1.986 [0.005]	−0.827 [0.005]	−1.763 [0.006]	0.480 [0.005]	−2.596 [0.013]	−1.039 [0.007]
β_3	3.249 [0.002]	0.400 [0.002]	2.618 [0.005]	3.931 [0.007]	4.091 [0.014]	0.812 [0.021]	3.023 [0.013]	5.562 [0.043]
β_4	1.323 [0.001]	0.224 [0.001]	0.963 [0.003]	1.696 [0.003]	1.658 [0.006]	0.380 [0.008]	1.106 [0.012]	2.358 [0.023]
β_5	−0.219 [0.000]	0.069 [0.000]	−0.334 [0.001]	−0.106 [0.001]	−0.271 [0.001]	0.093 [0.001]	−0.430 [0.003]	−0.128 [0.002]
σ_{01}	3.015 [0.002]	0.419 [0.002]	2.386 [0.004]	3.742 [0.007]	3.121 [0.006]	0.464 [0.005]	2.418 [0.009]	3.927 [0.015]
σ_1	1.453 [0.001]	0.228 [0.001]	1.103 [0.002]	1.845 [0.003]	1.693 [0.005]	0.360 [0.006]	1.195 [0.005]	2.320 [0.013]
σ_2	2.564 [0.002]	0.377 [0.002]	1.996 [0.003]	3.231 [0.006]	2.856 [0.009]	0.587 [0.014]	2.075 [0.008]	3.938 [0.034]
σ_3	2.656 [0.001]	0.352 [0.002]	2.124 [0.003]	3.277 [0.007]	3.005 [0.009]	0.592 [0.012]	2.195 [0.007]	4.057 [0.024]
σ_4	1.136 [0.001]	0.233 [0.001]	0.768 [0.004]	1.531 [0.003]	1.247 [0.005]	0.337 [0.005]	0.747 [0.013]	1.836 [0.010]
δ					0.621 [0.004]	0.183 [0.004]	0.318 [0.020]	0.928 [0.007]
γ					0.170 [0.004]	0.144 [0.008]	0.014 [0.001]	0.452 [0.018]
$\log p(y)$	−981.24 [0.003]				−978.27 [0.012]			
$\widehat{\text{ESS}}_{\text{IS}}$	0.774				0.3045			

We also estimated the Monte Carlo standard errors by bootstrapping the importance samples. We highlight two results. The MC standard errors are low in all cases, illustrating the high efficiency of the IS^2 method for carrying out Bayesian inference for these two models. The estimates of the log of the marginal likelihood allow us to calculate a Bayes factor of approximately 20 for the GMNL model over the MIXL model, so that our analysis provides strong evidence in favor of scale heterogeneity for this dataset.

6.1.5 Comparing IS^2 and pseudo marginal Metropolis-Hastings

The IS^2 and PMMH methods can tackle the same set of problems. This section compares the performance of IS^2 and PMMH holding both the independent proposal for the parameter vector θ and likelihood estimation fixed, which leads to a direct comparison. To accurately measure the Monte Carlo efficiency of the two methods for the MIXL model, we run 500 independent replications of the IS^2 algorithm using $M = 5{,}000$ importance samples and 500 independent PMMH Markov chains with 5,000 iterations (after a burn-in of 1,000 iterations) to estimate the posterior mean. We obtain the initial draw for the PMMH Markov chain by using the discrete IS^2 approximation of the posterior. The number of replications for the GMNL model was 200. We additionally consider two straightforward improvements to the baseline IS^2 method: Quasi-Monte Carlo (QMC) sampling of the parameters and trimmed means. We implement QMC by replacing the standard normal draws used to generate samples from $q(\theta|y)$ by scrambled Sobol sequences converted to quasi-random $N(0,1)$ draws through the inverse CDF method. The trimmed mean estimate consists of deleting the particle associated with the largest importance weight and its corresponding antithetic sample.

Table 7: IS^2 and PMMH comparison for the MIXL model.

	Monte Carlo mean-square error				Variance of MC SE estimate		
	IS^2	IS^2 (trim.)	IS^2/QMC (trim.)	PMMH	IS^2	IS^2 (trim.)	PMMH
β_{01}	0.374	0.260	0.171	1.000	0.141	0.008	1.000
β_1	0.300	0.217	0.196	1.000	0.149	0.011	1.000
β_2	0.345	0.244	0.189	1.000	0.178	0.008	1.000
β_3	0.369	0.301	0.215	1.000	0.071	0.014	1.000
β_4	0.304	0.267	0.224	1.000	0.022	0.008	1.000
β_5	0.351	0.291	0.243	1.000	0.104	0.009	1.000
σ_0	0.343	0.279	0.207	1.000	0.108	0.012	1.000
σ_1	0.397	0.293	0.199	1.000	0.169	0.010	1.000
σ_2	0.356	0.242	0.267	1.000	0.225	0.009	1.000
σ_3	0.348	0.276	0.249	1.000	0.071	0.008	1.000
σ_4	0.277	0.097	0.078	1.000	0.190	0.001	1.000

Tables 7 and 8 present the ratio between the MC mean squared errors and the ratios between the mean squared errors of the MC standard error estimates. We approximate the posterior mean using all particles generated in the replications for the standard IS^2 method. We approximate the actual MC standard error by computing the standard deviation of the posterior mean estimates across replications. The results support the argument that the IS^2 method leads to improved efficiency due to the use of independent samples and the simple implementation of variance reduction methods. The table shows that the standard IS^2 method leads to approximately 65% and 75% average reductions in the MC variance compared to the PMMH method for MIXL and GMNL models respectively. When considering trimmed means, the improvements are 75% and 94% respectively. The use of QMC leads to a modest incremental improvement of approximately 20% for both models (at virtually no extra computational effort). Moreover, we note that that we ignored the additional computational cost of the burn-in period for the PMMH method in this comparison. The computational costs are otherwise equivalent.

The results also suggest that the estimates of the MC variance within each of the replica-

Table 8: IS^2 and PMMH comparison for the GMNL model.

	Monte Carlo mean-square error				Variance of MC SE estimate		
	IS^2	IS^2 (trim.)	IS^2/QMC (trim.)	PMMH	IS^2	IS^2 (trim.)	PMMH
β_{01}	0.254	0.065	0.060	1.000	0.199	0.003	1.000
β_1	0.348	0.059	0.058	1.000	0.425	0.008	1.000
β_2	0.165	0.044	0.043	1.000	0.067	0.002	1.000
β_3	0.331	0.067	0.047	1.000	0.246	0.004	1.000
β_4	0.199	0.069	0.051	1.000	0.097	0.003	1.000
β_5	0.322	0.155	0.131	1.000	0.124	0.007	1.000
σ_0	0.215	0.077	0.070	1.000	0.094	0.004	1.000
σ_1	0.110	0.043	0.034	1.000	0.059	0.005	1.000
σ_2	0.207	0.060	0.056	1.000	0.120	0.003	1.000
σ_3	0.196	0.083	0.070	1.000	0.106	0.005	1.000
σ_4	0.188	0.042	0.026	1.000	0.129	0.002	1.000
δ	0.110	0.014	0.011	1.000	0.078	0.001	1.000
γ	0.239	0.053	0.024	1.000	0.244	0.005	1.000

tions are substantially more accurate under IS^2, highlighting the robustness of the algorithm. We emphasize that it is simple to obtain MC standard errors for any posterior statistic using IS^2, which is not always the case with MCMC.

6.2 Stochastic volatility

6.2.1 Model, data and importance sampling

We consider the univariate two-factor stochastic volatility (SV) model with leverage effects

$$y_t = \exp([c + x_{1,t} + x_{2,t}]/2)\varepsilon_t, \qquad t = 1,...,n$$
$$x_{i,t+1} = \phi_1 x_{i,t} + \rho_1 \sigma_{i,\eta}\varepsilon_t + \sqrt{1-\rho_i^2}\sigma_{i,\eta}\eta_{i,t}, \qquad i = 1,2$$

where the return innovations are i.i.d. and have the standardized Student's t distribution with ν degrees of freedom and $1 > \phi_1 > \phi_2 > -1$ for stationarity and identification. This SV specification incorporates the most important empirical features of a volatility series, allowing for fat-tailed return innovations, a negative correlation between lagged returns and volatility (leverage effects) and quasi-long memory dynamics through the superposition of AR(1) processes (see, e.g., Barndorff-Nielsen and Shephard, 2002). The two volatility factors have empirical interpretations as long term and short term volatility components.

We consider the following priors $\nu \sim \mathrm{U}(2,100)$, $c \sim \mathrm{N}(0,1)$, $\phi_1 \sim \mathrm{U}(0,1)$, $\sigma_1^2 \sim \mathrm{IG}(2.5,0.0075)$, $\rho_1 \sim \mathrm{U}(-1,0)$, $\phi_2 \sim \mathrm{U}(0,\phi_1)$, $\sigma_2^2 \sim \mathrm{IG}(2.5,0.045)$, and $\rho_2 \sim \mathrm{U}(-1,0)$. We estimate the two factor SV model using daily returns from the S&P 500 index obtained from the CRSP database. The sample period runs from January 1990 to December 2012, for a total of 5,797 observations. As in the previous example, we obtain the parameter proposal $g_{\mathrm{IS}}(\theta)$ using the MitISEM method

Table 9: Stochastic volatility - log-likelihood evaluation.

	N=10	N=20
Variance	0.009	0.005
Skewness	0.485	0.394
Kurtosis	3.796	3.640
Time (s)	1.069	1.087

of Hoogerheide et al. (2012), where we allow for two mixture components. To estimate the likelihood, we apply the particle efficient importance sampling (PEIS) method of Scharth and Kohn (2013). The PEIS method is a sequential Monte Carlo procedure for likelihood estimation that uses the EIS algorithm of Richard and Zhang (2007) to obtain a high-dimensional importance density for the states. We implement the PEIS method using antithetic variables as described in Scharth and Kohn (2013).

6.2.2 Optimal number of particles

To investigate the accuracy of the likelihood estimates for the stochastic volatility model, we replicate the Monte Carlo exercise of Section 6.1.3 for this example. Table 9 displays the average variance, skewness and kurtosis of the log-likelihood estimates based on $N=10$ and $N=20$ particles, as well as the computing times. We base the analysis on 1,000 draws from the proposal for θ and 200 independent likelihood estimates for each parameter value.

The results show that the PEIS method is highly accurate for estimating the likelihood of the SV model, despite the large number of observations. When $N = 10$, the inflation factor of (15) is approximately 1.01, indicating that performing IS with the PEIS likelihood estimate has essentially the same efficiency as if the likelihood was known. On the other hand, the log-likelihood estimates display positive skewness and excess kurtosis and are clearly non-Gaussian, suggesting that the theory of Section 3.2 only holds approximately for this example.

We approximate the optimal number of particles by assuming that the variance of log-likelihood estimates is constant across different values of θ, as it is not possible to use the jackknife method for estimating the variance of log-likelihood estimators based on particle methods. Based on the average variance of the log-likelihood estimate with $N = 20$, we estimate the asymptotic variance of the log-likelihood estimate to be $\widehat{\gamma}^2(\theta)=0.1$ on average. The last row of the table allows us to calculate that the overhead of estimating the likelihood is $\tau_0=1.051$ seconds, while the computational cost of each particle is $\tau_1=0.018\times10^{-1}$ seconds. Assuming normality of the log-likelihood estimate and that $\gamma^2(\theta) = \bar{\gamma}^2 = 0.1$ for all θ, the optimal number of particles is $N_{\text{opt}}=8$.

Figure 3 plots the predicted relative time normalized variance $\text{TNV}(M,N)/\text{TNV}(M,N_{\text{opt}})$ as a function of N. The figure suggests that the efficiency of IS^2 for this example is fairly insensitive to the number of particles at the displayed range $N\leq40$, with even the minimal number of particles $N=2$ (including an antithetic draw) leading to a highly efficient procedure.

Table 10 displays estimates of the relative variances for the posterior means based on $M=50{,}000$ importance samples for the parameters. We estimate the Monte Carlo variance of

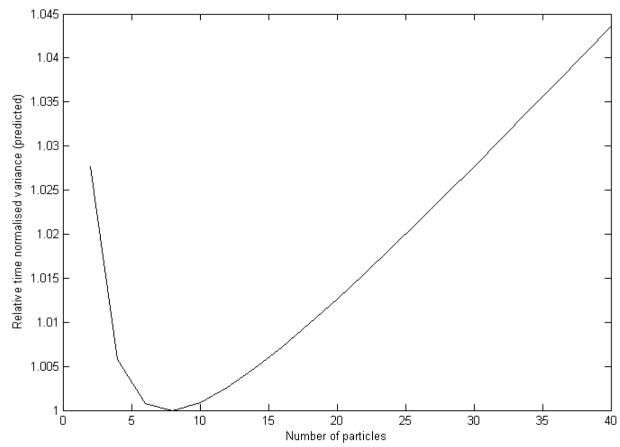


Figure 3: Predicted relative time normalized variance against the number of particles for the stochastic volatility model.

the posterior statistics by bootstrapping the importance samples. In line with Figure 3, the results indicate that the efficiency of the IS^2 method is insensitive to the number of particles in this example, with no value of N emerging as a clear optimal in the table. In light of this result, we recommend a conservative number of particles in practice as there is no important efficiency cost in choosing N moderately above the optimal theoretical value.

6.2.3 Posterior analysis

Table 11 presents estimates of selected posterior distribution statistics estimated by the IS^2 method. We estimated the posterior distribution using $M = 50{,}000$ importance samples for the parameters and $N = 20$ particles to estimate the likelihood. As before, we estimate the Monte Carlo standard errors by bootstrapping the importance samples. The results show that IS^2 leads to highly accurate estimates of the posterior statistics. The results show that the short term volatility component is almost entirely driven by leverage effects.

7 Conclusions

This article proposes the importance sampling squared method for Bayesian inference when the likelihood is computationally intractable but unbiasedly estimated, and studies the convergence properties of the IS^2 estimators. We examine the effect of estimating the likelihood on Bayesian inference and provide practical guidelines on how to optimally carry out likelihood estimation in order to minimize the computational cost. The applications illustrate that the IS^2 method can lead to fast and accurate posterior inference when optimally implemented. We believe that the theory and methodology presented in this paper are useful for practitioners who are working on models with an intractable likelihood.

Table 10: Stochastic volatility: relative variances for posterior inference. The table shows the relative variances for IS^2 for different numbers of importance samples used for estimating the likelihood.

	N=2	N=4	N=8	N=12	N=16	N=20
ν	0.963	0.873	1.000	0.937	0.893	0.885
c	0.901	0.978	1.000	0.996	0.945	0.966
ϕ_1	1.007	1.019	1.000	0.991	0.980	0.996
σ_1^2	0.995	1.000	1.000	1.022	0.974	1.032
ρ_1	1.086	1.004	1.000	1.034	0.994	1.015
ϕ_2	0.952	0.947	1.000	0.999	0.991	1.007
σ_2^2	1.016	1.030	1.000	1.005	0.969	0.972
ρ_2	1.039	1.058	1.000	1.025	1.028	1.011
Average	0.995	0.989	1.000	1.001	0.972	0.986
Theoretical	1.038	1.013	1	0.996	0.994	0.993
Time (minutes)	258	260	262	263	264	265
Estimated TNV	0.977	0.978	1.000	1.005	0.978	0.994
Theoretical	1.019	1.002	1.000	0.999	1.001	1.001

Acknowledgment

The research of Minh-Ngoc Tran, Marcel Scharth and Robert Kohn was partially supported by Australian Research Council grant DP0667069. We would like to thank Professor Denzil Fiebig for the use of the pap smear data used in Section 6

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010), "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society, Series B*, 72, 1–33.

Andrieu, C. and Roberts, G. (2009), "The pseudo-marginal approach for efficient Monte Carlo computations," *The Annals of Statistics*, 37, 697–725.

Barndorff-Nielsen, O. and Shephard, N. (2002), "Econometric analysis of realized volatility and its use in estimating stochastic volatility models," *Journal of the Royal Statistical Society B*, 64, 253–280.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2006), *Nonlinear Programming*, New Jersey: Wiley, 3rd ed.

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

Table 11: Stochastic Volatility: posterior statistics. The table presents estimates of selected posterior distribution statistics based on $M = 50{,}000$ importance samples for the parameters. The Monte Carlo standard errors are in brackets.

	Mean	Std. Dev.	Skew.	Kurt.	90% Credible Interval	
ν	13.666 [0.016]	2.619 [0.024]	1.097 [0.036]	5.026 [0.173]	10.203 [0.019]	18.543 [0.067]
c	0.044 [0.001]	0.178 [0.001]	0.311 [0.034]	3.652 [0.103]	−0.233 [0.002]	0.346 [0.004]
ϕ_1	0.994 [0.000]	0.002 [0.000]	−0.387 [0.015]	3.152 [0.041]	0.992 [0.000]	0.997 [0.000]
σ_1^2	0.007 [0.000]	0.002 [0.000]	0.654 [0.016]	3.653 [0.061]	0.004 [0.000]	0.010 [0.000]
ρ_1	−0.490 [0.001]	0.106 [0.001]	0.422 [0.014]	2.852 [0.031]	−0.646 [0.001]	−0.295 [0.003]
ϕ_2	0.871 [0.000]	0.037 [0.000]	−0.902 [0.026]	4.205 [0.124]	0.803 [0.001]	0.921 [0.000]
σ_2^2	0.028 [0.000]	0.006 [0.000]	0.516 [0.019]	3.548 [0.066]	0.019 [0.000]	0.038 [0.000]
ρ_2	−0.950 [0.000]	0.040 [0.000]	1.254 [0.020]	4.806 [0.127]	−0.995 [0.000]	−0.871 [0.001]
$\log p(y)$	-7.72×10^{-3}					
$\widehat{\text{ESS}}_{\text{IS}}$	0.592					

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013), "SMCˆ2: an efficient algorithm for sequential analysis of state-space models," *Journal of the Royal Statistical Society B*, 75, 397–426.

Doucet, A., Pitt, M. P., Deligiannidis, G., and Kohn, R. (2014), "Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator," http://arxiv.org/abs/1210.1871.

Fiebig, D. G., Keane, M. P., Louviere, J., and Wasi, N. (2010), "The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity," *Marketing Science*, 29, 393–421.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011), *Applied Longitudinal Analysis*, New Jersey: John Wiley & Sons, Ltd, 2nd ed.

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, 1, 515–534.

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317–1339.

Gourieroux, C. and Monfort, A. (1995), *Statistics and Econometric Models*, vol. 2, Melbourne: Cambridge University Press.

Hesterberg, T. (1995), "Weighted Average Importance Sampling and Defensive Mixture Distributions," *Technometrics*, 37, 185–194.

Hoogerheide, L., Opschoor, A., and van Dijk, H. K. (2012), "A class of adaptive importance sampling weighted EM algorithms for efficient and robust posterior and predictive simulation," *Journal of Econometrics*, 171, 101–120.

Kass, R. E. and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Koopman, S. J., Lucas, A., and Scharth, M. (2013), "Static and dynamic multivariate Gaussian efficient importance sampling," Working paper.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer-Verlag.

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

Pitt, M. K., Silva, R. S., Giordani, P., and Kohn, R. (2012), "On some properties of Markov chain Monte Carlo simulation methods based on the particle filter," *Journal of Econometrics*, 171, 134–151.

Quiroz, M., Villani, M., and Kohn, R. (2014), "Speeding Up MCMC by Efficient Data Subsampling," url http://arxiv.org/abs/1404.4178.

Richard, J.-F. and Zhang, W. (2007), "Efficient High-Dimensional Importance Sampling," *Journal of Econometrics*, 141, 1385–1411.

Ripley, B. (1987), *Stochastic Simulation*, New York: Wiley.

Scharth, M. and Kohn, R. (2013), "Particle Efficient Importance Sampling," Working paper.

Stirzaker, G. G. D. and Grimmett, D. (2001), *Probability and random processes*, vol. 19, Oxford, UK: Oxford University Press.

Appendix

Proof of Theorem 1. Condition $\text{Supp}(\pi) \subseteq \text{Supp}(g_{\text{IS}})$ implies $\text{Supp}(\pi_N) \subseteq \text{Supp}(\widetilde{g}_{\text{IS}})$. This, together with the existence and finiteness of $\mathbb{E}_\pi(\varphi)$ ensures that $\mathbb{E}_{\widetilde{g}_{\text{IS}}}[\varphi(\theta)\widetilde{w}(\theta,z)] = p(y)\mathbb{E}_\pi(\varphi)$ and $\mathbb{E}_{\widetilde{g}_{\text{IS}}}[\widetilde{w}(\theta,z)] = p(y)$ exist and are finite. Then, result (i) follows immediately from the strong law of large numbers.

To prove (ii), write

$$\widehat{\varphi}_{\text{IS}^2} - \mathbb{E}_\pi(\varphi) = \frac{\frac{1}{M}\sum_{i=1}^M \big(\varphi(\theta_i) - \mathbb{E}_\pi(\varphi)\big)\widetilde{w}(\theta_i, z_i)}{\frac{1}{M}\sum_{i=1}^M \widetilde{w}(\theta_i, z_i)}.$$

Let $X_i = \big(\varphi(\theta_i) - \mathbb{E}_\pi(\varphi)\big)\widetilde{w}(\theta_i,z_i)$, $i=1,...,M$, $S_M = \frac{1}{M}\sum_{i=1}^M X_i$ and $Y_M = \frac{1}{M}\sum_{i=1}^M \widetilde{w}(\theta_i,z_i)$. The X_i are independently and identically distributed with $\mathbb{E}_{\widetilde{g}_{\text{IS}}}(X_i) = 0$ and

$$\begin{aligned} \nu_N^2 = \mathbb{V}_{\widetilde{g}_{\text{IS}}}(X_i) = \mathbb{E}_{\widetilde{g}_{\text{IS}}}(X_i^2) &= \int_{\widetilde{\Theta}} \big(\varphi(\theta) - \mathbb{E}_\pi(\varphi)\big)^2 \widetilde{w}(\theta, z) p(y) \pi_N(\theta, z) d\theta dz \\ &= p(y)\mathbb{E}_{\pi_N}\left\{\big(\varphi(\theta) - \mathbb{E}_\pi(\varphi)\big)^2 \widetilde{w}(\theta, z)\right\} < \infty. \end{aligned}$$

By the CLT for a sum of independently and identically distributed random variables with a finite second moment, $\sqrt{M}S_M \xrightarrow{d} \mathcal{N}(0,\nu_N^2)$. By the strong law of large numbers, $Y_M \xrightarrow{P} \mathbb{E}_{\widetilde{g}_{\text{IS}}}(\widetilde{w}(\theta,z)) = p(y)$. Then, by Slutsky's theorem, we have that

$$\sqrt{M}\Big(\widehat{\varphi}_{\text{IS}^2} - \mathbb{E}_\pi(\varphi)\Big) = \frac{\sqrt{M}S_M}{Y_M} \xrightarrow{d} \mathcal{N}(0, \nu_N^2/p(y)^2).$$

The asymptotic variance is given by

$$\begin{aligned} \sigma^2_{\text{IS}^2}(\varphi) = \frac{\nu_N^2}{p(y)^2} &= \frac{1}{p(y)}\mathbb{E}_{\pi_N}\left\{\big(\varphi(\theta) - \mathbb{E}_\pi(\varphi)\big)^2 \widetilde{w}(\theta, z)\right\} \\ &= \mathbb{E}_\pi\left\{\big(\varphi(\theta) - \mathbb{E}_\pi(\varphi)\big)^2 \frac{\pi(\theta)}{g_{\text{IS}}(\theta)} \mathbb{E}_{g_N(z|\theta)}[\exp(2z)]\right\}. \end{aligned}$$

To prove (iii), write

$$
\begin{aligned}
\widehat{\sigma^2_{\mathrm{IS}^2}(\varphi)} &= \frac{\frac{1}{M}\sum_{i=1}^M \big(\varphi(\theta_i)-\widehat{\varphi}_{\mathrm{IS}^2}\big)^2 \widetilde{w}(\theta_i,z_i)^2}{\Big(\frac{1}{M}\sum_{i=1}^M \widetilde{w}(\theta_i,z_i)\Big)^2} \\
&\stackrel{a.s.}{\to} \frac{\mathbb{E}_{\widetilde{g}_{\mathrm{IS}}}\Big\{\big(\varphi(\theta)-\mathbb{E}_\pi(\varphi)\big)^2\widetilde{w}(\theta,z)^2\Big\}}{\Big(\mathbb{E}_{\widetilde{g}_{\mathrm{IS}}}(\widetilde{w}(\theta,z))\Big)^2} \\
&= \sigma^2_{\mathrm{IS}^2}(\varphi).
\end{aligned}
$$

□

Proof of Theorem 2. Under Assumption 2, $g_N(z|\theta)=\mathcal{N}(-\sigma^2/2,\sigma^2)$ and $\mathbb{E}_{g_N(z|\theta)}[\exp(2z)] = \exp(\sigma^2)$. From (9) and (12),

$$
\sigma^2_{\mathrm{IS}^2}(\varphi) = \mathbb{E}_\pi\left\{\big(\varphi(\theta)-\mathbb{E}_\pi(\varphi)\big)^2\frac{\pi(\theta)}{g_{\mathrm{IS}}(\theta)}\exp(\sigma^2)\right\} = \exp(\sigma^2)\sigma^2_{\mathrm{IS}}(\varphi).
$$

□

Proof of Proposition 1. Proof of (i) is straightforward. To prove (ii), we have

$$
\begin{aligned}
M\mathbb{V}_{\widetilde{g}_{\mathrm{IS}}}(\widehat{p}_{\mathrm{IS}^2}(y)) = \mathbb{V}_{\widetilde{g}_{\mathrm{IS}}}(\widetilde{w}(\theta)) &\le \mathbb{E}_{\widetilde{g}_{\mathrm{IS}}}[\widetilde{w}(\theta)^2] \\
&= p(y)^2\int\left(\int e^{2z}g_N(z|\theta)dz\right)\left(\frac{\pi(\theta)}{g_{\mathrm{IS}}(\theta)}\right)^2 g_{\mathrm{IS}}(\theta)d\theta \\
&\le p(y)^2 C\int\left(\frac{\pi(\theta)}{g_{\mathrm{IS}}(\theta)}\right)^2 g_{\mathrm{IS}}(\theta)d\theta < \infty.
\end{aligned}
$$

□

Proof of Proposition 2. Denote $a=\tau_0$, $b=\tau_1\bar{\gamma}^2$, and $x=\sigma^2$. Then

$$
f(x) = \mathrm{CT}^*_{\widehat{p}_{\mathrm{IS}^2}(y)}(\sigma^2) = \big(a+\frac{b}{x}\big)\big((v+1)e^x-1\big).
$$

We write $f(x)=f_1(x)+bf_2(x)$ with $f_1(x)=a((v+1)e^x-1)$ and $f_2(x)=((v+1)e^x-1)/x$. As $f_1(x)$ is convex, it is sufficient to show that $f_2(x)$ is convex. To do so, we will prove that $f_2''(x)>0$ for $x>0$, because a differentiable function is convex if and only if its second derivative is positive (see, e.g., Bazaraa et al., 2006, Chapter 3).

After some algebra

$$
f_2''(x) = v\frac{e^x}{x}\left[\left(1-\frac{1}{x}\right)^2+\frac{1}{x^2}\right]+\frac{e^x}{x}\left[\left(1-\frac{1}{x}\right)^2+\frac{1}{x^2}\left(1-2e^{-x}\right)\right].
$$

The first term is clearly positive. We consider the second term in the square brackets

$$
\left(1-\frac{1}{x}\right)^2+\frac{1}{x^2}\left(1-2e^{-x}\right)=1-\frac{2}{x}+\frac{2(1-e^{-x})}{x^2}.
$$

Note that $1-e^{-x}>x-\frac{x^2}{2}$ for $x>0$ and so

$$1-\frac{2}{x}+\frac{2}{x^2}(1-e^{-x})>1-\frac{2}{x}+\frac{2}{x^2}\left(x-\frac{x^2}{2}\right)=0.$$

This establishes that $f_2''(x)>0$ for all $x>0$.

To prove (ii), for any fixed v, let $x_{\min}(v)$ be the minimizer of $f(x)$. By writing

$$f(x)=(v+1)\mathrm{CT}^*(x)\times\left(1-\frac{a+b/x}{(v+1)\mathrm{CT}^*(x)}\right),$$

with $\mathrm{CT}^*(x)$ from (20), we can see that $f(x)$ is driven by the factor $(v+1)\mathrm{CT}^*(x)$ as $v\to\infty$. Hence $x_{\min}(v)$ tends to the σ^2_{opt} in (18) that minimizes $\mathrm{CT}^*(x)$, as $v\to\infty$.

Because $f'(x_{\min}(v))=0$ for any $v>0$,

$$e^{x_{\min}(v)}(ax^2_{\min}(v)+bx_{\min}(v)-b)=-\frac{b}{v+1},\ \ \text{for } v>0. \tag{38}$$

By taking the first derivative of both sides of (38), we have

$$x'_{\min}(v)=\frac{b}{(v+1)^2}\frac{e^{-x_{\min}(v)}}{ax^2_{\min}(v)+(b+2a)x_{\min}(v)}>0,\ \ \text{for any } v>0.$$

This follows that $x_{\min}(v)$ is an increasing function of v. Furthermore, $x'_{\min}(v)\to 0$ as $v\to\infty$, which establishes (iii). $\square$

To obtain Proposition 3 we first obtain some preliminary results. Define,

$$\zeta_{ij}(\theta):=\omega_i(\alpha_i^{(j)},\theta)/p(y_i|\theta)-1 \tag{39}$$

$$\varepsilon_{N,i}:=\frac{\sqrt{N}}{\sigma_i(\theta)}\left(\frac{\widehat{p}_{N,i}(y_i|\theta)}{p_i(y_i|\theta)}-1\right)=\frac{1}{\sqrt{N}\sigma_i(\theta)}\sum_{j=1}^{N}\zeta_{ij}(\theta) \tag{40}$$

so that $\varepsilon_{N,i}$ is a sum of the i.i.d. random variables.

Lemma 2. *Suppose Assumptions 1 and 3 hold. Then, for all $i=1,...,n$, and θ,*

(i) $\mathbb{E}(\zeta_{ij}(\theta))=0$ *and* $\mathbb{V}(\zeta_{ij}(\theta))=\sigma_i^2(\theta)$.

(ii) $\mathbb{E}_{h_i}(|\zeta_{ij}(\theta)|^k)\le K_4$ *for all* $j=1,...,N$, *and* $k=1,...,4$, *with* $K_4>0$.

(iii)

$$N\times\mathbb{V}_{h_i}\{\widehat{p}_{N,i}(y_i|\theta)/p(y_i|\theta)\}=\sigma_i^2(\theta) \tag{41}$$

(iv) $1/K_5<\sigma_i(\theta)^2<K_5$, *where* $K_5>1$ *is a constant.*

(v) $\mathbb{E}(\varepsilon_{N,i}|\theta)=0$, $\mathbb{V}(\varepsilon_{N,i}|\theta)=1$, $\mathbb{E}(\varepsilon^3_{N,i}|\theta)=O(N^{-1/2})$, *and* $\mathbb{E}(|\varepsilon_{N,i}|^k|\theta)<K_5$, *for* $k=1,...,4$, *where* $K_5>0$.

(vi)

$$\frac{\mathbb{E}(\varepsilon_{N,i}^3|\theta)}{\left(\mathbb{E}(\varepsilon_{N,i}^2|\theta)\right)^{3/2}} \to 0 \ \text{as } N\to\infty. \tag{42}$$

(vii) $\varepsilon_{N,i} \xrightarrow{d} \mathcal{N}(0,1)$ *as* $N\to\infty$.

Proof. Parts (i) and (ii) follow from Assumption 3, Assumption 1 and (28). Part (iii) follows from the definition (40). Part (iv) follows from Assumption 3. To obtain part (v), we note that $\mathbb{E}(\varepsilon_{N,i}|\theta)=0$, $\mathbb{V}(\varepsilon_{N,i}|\theta)=1$, follow from the definition of $\varepsilon_{N,i}$ and Part (i). It is straightforward to show that $\mathbb{E}(\varepsilon_{N,i}^3|\theta)=N\mathbb{E}(\zeta_{ij}(\theta)^3)/(N^{3/2}\sigma_i(\theta)^3)=O(N^{-1/2})$ by Part (i). Part (v) follows from the definition of $\varepsilon_{N,i}$ Part (i), and Assumption 3. The denominator on the left side of (42) is $O(1)$. The numerator is $O(N^{-1/2})$ by part (v). Part (vi) follows. Part (vii) holds because $\varepsilon_{N,i}$ is a sum of i.i.d. random variables, Part (v) and Theorem 5, p. 194, of Stirzaker and Grimmett (2001). □

We write the error z in (27) in terms of the $\varepsilon_{N,i}$ and expand as a third order Taylor series approximation with a remainder term,

$$\begin{aligned} z &= \sum_{i=1}^{n} \log\left(1+\sigma_i(\theta)\varepsilon_{N,i}/\sqrt{N}\right) \\ &= \sum_{i=1}^{n}\left(\frac{\sigma_i(\theta)\varepsilon_{N,i}}{\sqrt{N}} - \frac{\sigma_i^2(\theta)\varepsilon_{N,i}^2}{2N} + \frac{\sigma_i^3(\theta)\varepsilon_{N,i}^3}{3N\sqrt{N}} + R_{N,i}(\theta)\right). \end{aligned} \tag{43}$$

The remainder term is $R_{N,i}(\theta) \le \sigma_i^4(\theta)\varepsilon_{N,i}^4/4N^2$.

Lemma 3. *We can express* $\widehat{\sigma}_i^2(\theta)=\sigma_i^2(\theta)+c_i(\theta)/N+\xi_i(\theta)/\sqrt{N}$, *where* $\mid c_i(\theta)\mid < K_5$ *and* $\mathbb{V}_{h_i}(\xi_i(\theta)) < K_5$, *with* $K_5>0$, *for all* $i=1,\ldots,n$ *and* θ.

Proof.

$$\left(\frac{\widehat{p}_{N,i}(y_i|\theta)}{p_i(y_i|\theta)}\right)^2 = \left(\frac{1}{N}\sum_{j=1}^{N}\left(1+\zeta_{ij}(\theta)\right)\right)^2 = 1+\frac{1}{N^2}\left(\sum_{j=1}^{N}\zeta_{ij}(\theta)\right)^2 + \frac{2}{N}\sum_{j=1}^{N}\zeta_{ij}(\theta)$$

$$\text{so that} \quad \mathbb{E}_{h_i}\left(\frac{\widehat{p}_{N,i}(y_i|\theta)}{p_i(y_i|\theta)}\right)^2 = 1+\frac{\sigma_i^2(\theta)}{N}$$

$$\text{and therefore we can write} \quad \left(\frac{\widehat{p}_{N,i}(y_i|\theta)}{p_i(y_i|\theta)}\right)^2 = 1+\frac{\sigma_i^2(\theta)}{N}+\frac{\xi_i^*(\theta)}{\sqrt{N}}$$

where $\mathbb{V}_{h_i}\left(\xi^*(\theta)\right) < K_6$, with $K_6>0$ for all θ and $i=1,\ldots,N$. Define,

$$\widetilde{\sigma}_i^2(\theta) := \frac{1}{N}\sum_{j=1}^{N}\left(\frac{w_i(\alpha_i^{(j),\theta}}{p_i(y_i(\theta)}\right)^2 - 1 = \frac{1}{N}\sum_{j=1}^{N}\zeta_{ij}(\theta)^2 + \frac{2}{N}\sum_{j=1}^{N}\zeta_{ij}(\theta).$$

We can then write $\widetilde{\sigma}_i^2(\theta) = \sigma_i^2(\theta) + N^{-1/2}\widetilde{\xi}(\theta)$, where $\mathbb{V}_{h_i}\big(\widetilde{\xi}(\theta)\big) < K_7$, with $K_7 > 0$ for all θ and $i = 1,...,n$.

With some algebra, we can write

$$\widehat{\sigma}_i^2(\theta) = \Big(\frac{\widehat{p}_{N,i}(y_i|\theta)}{p_i(y_i|\theta)}\Big)^{-2}\Bigg(\widetilde{\sigma}_i^2(\theta) + \Bigg(\Big(1 - \frac{\widehat{p}_{N,i}(y_i|\theta)}{p_i(y_i|\theta)}\Big)^2\Bigg)$$

and the result follows. □

Proof of Proposition 3 . Part (i): The expression for $\mathbb{E}(z|n,N,\theta)$ follows from (43). To obtain the expression for $\mathbb{V}(z|n,N,\theta)$, we write z as a first order Taylor series expansion plus remainder, similarly to (43),

$$z = z_1 - \sum_{i=1}^{n} S_{N,i}(\theta), \quad \text{where} \quad z_1 = N^{-1/2}\sum_{i=1}^{n}\sigma_i(\theta)\varepsilon_{N,i} \quad \text{and} \quad S_{N,i}(\theta) \leq \sigma_i^2(\theta)\varepsilon_{N,i}^2/(2N). \tag{44}$$

Part (ii) follows from Lemma 3. Part (iii) follows from Part (ii). To obtain Part (iv), it is sufficient to prove the central limit theorem for z_1, ,which is a sum of independent random variables. Now, by Parts (iv) and (v) of Lemma 2,

$$\frac{\sum_{i=1}^{n}\sigma_i^3(\theta)\mathbb{E}(\varepsilon_{N,i}^3)/N^{3/2}}{\Big(\sum_{i=1}^{n}\sigma_i^2(\theta)\mathbb{E}(\varepsilon_{N,i}^3)/N^{1/2}\Big)^{3/2}} = O(N^{-1/2})\frac{\sum_{i=1}^{n}\sigma_i^3(\theta)}{\Big(\sum_{i=1}^{n}\sigma_i^2(\theta)\Big)^{3/2}} \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty.$$

The central limit theorem now follows from Theorem 5, p. 194, of Stirzaker and Grimmett (2001) □

We note that as N is $O(n)$, the central limit operates very quickly in n upon the expression for z_1 in (44) as the summation is over an increasing number of terms, each of which satisfies a central limit.

Proof of Proposition 4. Assumption (ii) yields

$$N_{\widehat{\theta}_n} = n\widehat{\psi}_{N_S}(\widehat{\theta}_n,n)/\kappa^2 = n\psi(\widehat{\theta}_n,n)/\kappa^2 + O(n/N_S) + O_p(n^{\frac{1}{2}}N_S^{-\frac{1}{2}}).$$

As N_S is chosen proportional to n then consequently, $N_{\widehat{\theta}_n}$ is of order n. Using assumption (i), the achieved variance at any ordinate θ with $N_{\widehat{\theta}_n}$ particles is

$$\mathbb{V}(z|\theta,n,N_{\widehat{\theta}_n}) = n\psi(\theta,n)/N_{\widehat{\theta}_n} + O(n/N_{\widehat{\theta}_n}^2) = \kappa^2\psi(\theta,n)/\psi(\widehat{\theta}_n,n) + O_p(n^{-1}). \tag{45}$$

Using Assumptions (iii) to (v), we obtain from a Taylor expansion of $\psi(\theta,n)$ around $\widehat{\theta}_n$ in (45) that

$$\mathbb{V}(z|\theta,n,N_{\widehat{\theta}_n}) = \kappa^2 + O_p(n^{-\frac{1}{2}}).$$

This establishes part A. For part B, we may write

$$N_\theta = n\widehat{\psi}_{N_{\widehat{\theta}_n}}(\theta,n)/\kappa^2 = n\psi(\theta,n)/\kappa^2 + O_p(1),$$

since $N_{\widehat{\theta}_n}$ is of order n. Hence, the achieved variance at a given θ and n is given by

$$\mathbb{V}(z|\theta,n,N_\theta) = n\psi(\theta,n)/N_\theta + O(n/N_\theta^2) = \kappa^2 + O_p(n^{-1}).$$

□