Brinch, C.N. (2012): " Efficient simulated maximum likelihood estimation through explicitly parameter dependent importance sampling".Computational Statistics, Volume 27, Number 1, pp.13-28

Title:	Efficient simulated maximum likelihood estimation through explicitly parameter dependent importance sampling
Author:	Christian N. Brinch
Version:	**Authors own final version / Post-print (peer reviewed)** **The original publication is available at www.springer.com**
Publisher:	Springer http://www.springerlink.com/content/jx26381513vj185v/
DOI:	http:dx.doi.org/10.1007/s00180-011-0230-z

Noname manuscript No.
(will be inserted by the editor)

Efficient simulated maximum likelihood estimation through explicitly parameter dependent importance sampling

Christian N. Brinch

the date of receipt and acceptance should be inserted later

Abstract There exists an overall negative assessment of the performance of the simulated maximum likelihood algorithm in the statistics literature, founded on both theoretical and empirical results. At the same time, there also exist a number of highly successful applications. This paper explains the negative assessment by the coupling of the algorithm with "simple importance samplers", samplers that are not explicitly parameter dependent. The successful applications in the literature are based on explicitly parameter dependent importance samplers. Simple importance samplers may efficiently simulate the likelihood function value, but fail to efficiently simulate the score function, which is the key to efficient simulated maximum likelihood. The theoretical points are illustrated by applying Laplace importance sampling in both variants to the classic salamander mating model.

Keywords importance sampling · salamander mating model · simulation based estimation

1 Introduction

In many statistical models the likelihood function consists of factors that are integrals whose solutions are not available in closed form. The main example is the random effects model in various guises, where a relatively simple model can be expressed conditional on some vector of latent random variables. The purpose of such variables is typically to introduce dependence between different outcomes. Special cases are state space models or spatial models. The most common variant in applied statistics is the straightforward sort of nonlinear

Christian N. Brinch
Statistics Norway, Research Department and University of Oslo, Centre for Ecological and Evolutionary Synthesis
E-mail: cnb@ssb.no

or generalized linear model with random effects that are associated with one or more factors in the model.

Techniques for estimation of such models is a lively research area, where the main emphasis in the recent decades have been to Bayesian analysis, in particular to Markov chain Monte Carlo techniques. Both Bayesian analyses and the Monte Carlo EM algorithm for finding maximum likelihood estimates are able to solve the problem of integrals not available in closed form through bypassing direct evaluation of the likelihood function. However, it is also possible to approach the problem more directly. The simulated maximum likelihood algorithm applies importance sampling to simulate the likelihood function and uses numerical optimization methods to maximize the simulated likelihood function.

Simulated maximum likelihood appeared in the core statistics literature with Geyer and Thompson (1992) and Gelfand and Carlin (1993). Development and assessment in the important context of generalized linear mixed models appeared in McCulloch (1997). In the nonlinear mixed model context, the method appeared in Pinheiro and Bates (1995). It is always hard to substantiate that an assessment has the character of being a general consensus assessment. However, it seems fair to state that in terms of performance, the method is seen as inferior to the Monte Carlo EM algorithm and more generally to Monte Carlo Markov chain techniques for statistical inference. By inferior I mean that the variance in estimated parameters due to the stochasticity of the simulated likelihood function is large, so that very large importance samples are required for estimates with a precision comparable of that attained by other techniques. The negative assessment is most explicit in the theoretical results in Jank and Booth (2003), and recommendations based on numerical assessments in e.g. McCulloch (1997) and Jank (2006). More important is the implicit assessment evident in the sparse treatment of the method in reference books such as Robert and Casella (2004) or McCulloch and Searle (2001) and the limited use of the method in most fields of applied statistics.

The above assessment does however not hold universally in the literature. In a series of papers, following Durbin and Koopman (1997, 2000), simulated maximum likelihood has been applied in state space models conforming to the class of generalized linear mixed models. Skaug (2002) and Skaug and Fournier (2006) apply the method with success to other generalized linear mixed models. Pinheiro and Bates (1995) do report adequate performance although they report to prefer other likelihood approximations. In addition, a literature on simulation based estimation developed early within the econometrics literature concerning models with limited dependent variables, with successful applications, see Hajivassiliou and Ruud (1994) or Stern (1997) for surveys. There is thus a discord between the negative assessment in the statistics literature and a number of successful practical applications. The problem of resolving this discord is to some extent exacerbated by the fact that simulated maximum likelihood estimator may well be heavy tailed, so that an apparently successful practical application may actually be misleading.

The contribution of this article is to point out that the perceived computational inferiority of simulated maximum likelihood is closely related to definitions or implementations of the simulated likelihood functions in terms of importance sampling distributions that are not explicitly parameter dependent, "simple importance sampling" in the terminology adopted below. The heuristics for developing importance samplers for use with simulated maximum likelihood should be different from the heuristics for developing importance samplers for simulating scalars, as typefied by the "optimal importance sampler" discussed in expositions of simulated maximum likelihood in e.g. McCulloch and Searle (2001) or Jank (2006). The reason is that the key to efficient simulated maximum likelihood estimation is efficient simulation of the score function, not the likelihood function value. Efficient simulation of the score function requires explicitly parameter dependent (EPD) importance samplers. Simple importance sampling leads to a positive lower bound for the stochastic variability of the simulated maximum likelihood estimator, originally derived by Jank and Booth (2003), where it is interpreted as a limitation of simulated maximum likelihood per se. There is no such bound for simulated maximum likelihood based on EPD importance sampling. EPD importance samplers giving smooth estimates of smooth functions may be constructed using a simple transformation formula. The successful applications of simulated maximum likelihood in the literature are characterized by use of EPD importance sampling.

The theoretical points are illustrated by estimation and simulation results based on the well known salamander mating model from McCullagh and Nelder (1989). The simulated likelihood functions are based on Laplace importance samplers, with both simple and EPD versions. It is intrinsically difficult to evaluate simulated maximum likelihood based on simple importance sampling, because the method is prescribed as an iterative method with informal convergence criteria. However, the asymptotic efficiencies (in the importance sample size) of both estimators are functions of the variances of the simulated score functions evaluated at the exact maximum likelihood, and these can easily be compared.

I first demonstrate that maximum likelihood estimates can easily be found to high precision using EPD Laplace importance sampling. The asymptotic efficiency of both estimators are then compared based on repeated evaluations of score function at the approximate maximum likelihood. Even with an importance sample of 100 times the size, 5 out of 6 parameters would vary more based on the simple importance sampler compared to the corresponding EPD importance sampler. The variability of simulated maximum likelihood estimates based on simple Laplace importance sampling is of the same order of size as the estimated theoretical lower bound for simple importance samplers, giving standard deviations of 2-5 times the lower bound.

2 Simulated maximum likelihood

The reason for using simulation based estimation methods is typically that the likelihood function is not available is closed form. Let the exact likelihood function $L(\theta)$ be defined for a vector of parameters $\theta \in \Theta$, by the multidimensional integral $\int \exp(f(x,\theta))dx$, with x vector valued. The dependence of the likelihood function on data is supressed.

The simulated maximum likelihood algorithm proceeds by simulating the likelihood function, and finding the maximum of the simulated function, typically by numerical optimization methods. A general formula for simulation of functions through importance sampling is

$$\hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(f(x_i,\theta))}{\pi(x_i;\theta)}. \tag{1}$$

where $x_1,\ldots,x_n$ is a random sample from the distribution characterized by the density function $\pi(x;\theta)$. The role of this importance sampling distribution is traditionally to provide a good estimate of the integrand to be simulated.

Because formula (1) requires different importance samples for different θ, properties such as continuity and differentiability of the exact likelihood function is not preserved in the simulated likelihood function. Since such properties are important for finding maxima and evaluating parameter uncertainty, it is essential in the simulated maximum likelihood context to use some technique that bases the simulated likelihood function on common random numbers for different θ.

The approach that I will refer to as *simple importance sampling* in the following is based on the formula

$$\hat{L}_s(\theta;\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(f(x_i,\theta))}{\pi(x_i;\theta_0)}, \tag{2}$$

where $x_1,\ldots,x_n$ is a random sample from the distribution characterized by $\pi(x;\theta_0)$, where the role of θ_0 is to ensure that π is a good importance sampler for θ in the neighborhood of θ_0. Clearly, the importance sample is not explicitly dependent on the parameter θ. In a series of expositions of simulated maximum likelihood, such as Kuk and Cheng (1999), Kuk (1999), McCulloch and Searle (2001) and Robert and Casella (2004), simulated maximum likelihood is defined in terms simple importance sampling.

Suppose random variables X with distribution $\pi(x;\theta)$ may be generated by $X = g(Z;\theta)$, where Z is a random variable with density function $\pi_z(z)$, and g is a continuously differentiable function with nonsingular Jacobian w.r.t. Z, denoted $J(\theta, Z)$. We can then, by transformation from the random variable X to the random variable Z, replace equation (1) with

$$\hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} |J(\theta,z_i)| \frac{\exp(f(g(z_i,\theta);\theta))}{\pi_z(z_i)}, \tag{3}$$

where $z_1, \ldots, z_n$ is a random sample from the distribution characterized by the density function $\pi_z(z)$. I refer to this approach as *explicitly parameter dependent (EPD) importance sampling*, although the crucial point is not explicitness per se, but the combination of explicitness with common random numbers. The technique has not been described at such a general level before, but is still widely applied without noting the conceptual difference from simple importance sampling, see Section 3 for examples and references to applications. The main novelty of this article is pointing out that EPD importance sampling is a distinct technique compared to simple importance sampling, and that this technique is crucial to computationally efficient implementation of simulated maximum likelihood.

EPD importance sampling based on common random numbers is widely applicable, see below for examples, and does not usually require the explicit use of the formula in equation (3). The technique does however impose some limitations on the importance sampler. Since g must be a continuous function of θ, accept-reject algorithms and more advanced samplers that use accept-reject sampling as building blocks, such as the Metropolis algorithm, are not compatible with EPD importance sampling.

The properties of SML estimators for a regular case with a well behaved EPD importance sampler can be summarized as follows:

Assumption 1 (Regular, interior, local maximum likelihood.) L has an interior strict local maximum at $\tilde{\theta} \in \Theta$. L is two times continuously differentiable in a neighborhood of $\tilde{\theta}$.

Assumption 2 (Sufficiently dispersed importance sampler.) The random variable derived from Z as

$$\frac{\partial(|J(\theta, Z)| \exp(f(g(Z,\theta);\theta)))/\pi_z(Z)}{\partial\theta}|_{\theta=\tilde{\theta}} \tag{4}$$

has finite variance.

Theorem 1 *Under Assumptions 1 and 2, a simulated likelihood function as defined in equation (3) has a (random) stationary point $\hat{\theta}_n$ such that $\sqrt{n}(\hat{\theta}_n - \tilde{\theta})$ converges in distribution to $N(0, \Omega)$ as $n \to \infty$.*

$$\Omega = I^{-1} \Sigma I^{-1}, \tag{5}$$

where I is the observed information at $\tilde{\theta}$ and Σ is the limit as $n \to \infty$ of the variance matrix of the simulated score function, $\partial \log \hat{L}(\theta)/\partial\theta$, evaluated at $\tilde{\theta}$, scaled by $\sqrt{n}$.

Σ can be expressed as the limit as $n \to \infty$ of the variance matrix of $L(\tilde{\theta})^{-1} n^{-1/2} \sum_{i=1}^{n} W_i$, with W_i being realizations of a random variable W derived from Z as

$$W = \frac{\exp(f(x,\theta))}{\pi(x;\theta)} \left(\frac{\partial f(x,\theta)}{\partial\theta} - \frac{\partial \log \pi(x;\theta)}{\partial\theta} + \left(\frac{\partial f(x,\theta)}{\partial x} - \frac{\partial \log \pi(x,\theta)}{\partial x}\right) \frac{\partial g(z,\theta)}{\partial\theta} \right), \tag{6}$$

evaluated at $\theta = \tilde{\theta}$, $x = g(Z, \tilde{\theta})$ and $z = Z$.

Theorem 1 is partially a concise description of widely known results, with a straightforward proof assigned to the appendix. E.g. the formula for asymptotic variance of simulated maximum likelihood estimators is used for assessments of stochastic variability in software such as SSF-pack (Koopman et al, 1999) or AD model builder (Skaug and Fournier, 2006). However, the spelling out of the components of W in equation (6) is to my knowledge new. Assumption 2 also differs somewhat from the usual specifications of sufficiently dispersed importance samplers. The performance of simulated maximum likelihood under importance samplers that do not achieve finite variance and the assessment of whether importance samplers do achieve finite variance is an important issue with simulated maximum likelihood in general, see e.g. Geweke (1989) or Koopman et al (2009), but is not the topic of the discussion here.

The main topic here is that the asymptotic variance of the simulated maximum likelihood estimate depends on the asymptotic variance of the simulated score function and not the simulated likelihood function value. The key to computationally efficient implementation of SML is hence importance samplers that give W as defined in equation (6) with low variance. A simple importance sampler can be characterized by $\partial \log \pi(x_i;\theta)/\partial\theta = 0$ and $\partial g(z,\theta)/\partial\theta = 0$. The optimal simple importance sampler is usually defined as $\pi(x;\theta_0) = C \exp(f(x,\tilde{\theta}))$, with C a normalizing constant, giving exact likelihood value at $\tilde{\theta}$. With this importance sampler, equation (6) simplifies to

$$W = C^{-1}\frac{\partial f(x,\theta)}{\partial\theta} \tag{7}$$

Thus, even with an optimal importance sampler, giving *exact* estimates for the likelihood function, the simulated maximum likelihood estimate varies as long as the derivative of the integrand with respect to θ depends on x. Hence, a simple importance sampler cannot really be "optimal" for use with simulated maximum likelihood, unless the score function is independent of the random effects.

The optimal importance sampler for EPD importance sampling is characterized by $\pi(x;\theta) = C(\theta)\exp(f(x,\theta))$, with $C(\theta)$ implicitly defined by

$$\int C(\theta)\exp(f(x,\theta))dx = 1. \tag{8}$$

Such an importance sampler gives zero variance for simulated maximum likelihood estimates. Substituting for $\pi(x;\theta)$ in equation (6) gives

$$W = C(\theta)^{-1}\frac{\partial \log C(\theta)}{\partial\theta}, \tag{9}$$

which is deterministic for all values of θ. W is equal to zero with $\theta = \tilde{\theta}$ because

$$\frac{\partial \log C(\theta)}{\partial\theta} = \int \exp(f(x,\theta))\frac{\partial f(x,\theta)}{\partial\theta}dx, \tag{10}$$

by implicit differentiation of equation (8). The right hand side of equation (10) is the derivative of the exact likelihood function, which is of course zero evaluated at the exact maximum likelihood estimate $\tilde{\theta}$.

A real example of an optimal EPD importance sampler is the EPD Laplace importance sampler as defined below applied to Gaussian linear mixed models. There are of course other, well known, ways of computing exact estimates in that case, see e.g. McCulloch and Searle (2001), but as a contrast simulated maximum likelihood based on a simple importance sampler would fail to find the exact estimates. More generally an efficient importance sampler for simulated maximum likelihood should be defined as a sampler leading to low variance in the simulated score function, translating into low variance in the simulated maximum likelihood estimates, rather than a low variance in the likelihood function value.

A lower bound for the variance associated with simulated maximum likelihood (implicitly, using simple importance sampling) appeared in Jank and Booth (2003). In short, in our terminology, assume f can be approximated by a quadratic expansion about its joint maximum in (x, θ), denoted (x^*, θ^*). Thus, with $\lambda = (x - x^*, \theta - \theta^*)$

$$\hat{f}(x, \theta) \approx f(x^*, \theta^*) - \frac{1}{2}\lambda I_f \lambda', \tag{11}$$

where I_f can be partitioned as

$$I_f = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} \tag{12}$$

with dimensions corresponding to the dimensions of x and θ. Then, the observed information can be specified as $I = I_{22} - I_{21} I_{11}^{-1} I_{21}$ and

$$W = (X - x^*)' I_{12} + I_c(\theta - \theta^*). \tag{13}$$

The variance of X is under the optimal simple importance sampler equals I_{11}^{-1}, and the variance of W follows as $I_m = I_{21} I_{11}^{-1} I_{12}$, the missing information. The missing information is the difference between what the observed information would have been if the random effects had been observed, the complete information I_{22}, and the observed information.

3 Examples of EPD importance sampling

I will now discuss three examples of EPD importance sampling. The first example, which is the one I will pursue in the numerical examples below, is based on Laplace importance sampling. The other two examples illustrate the broad applicability of EPD importance sampling.

Let $x^*(\theta) = \arg\max_x f(x, \theta)$. Further let $H(\theta) = -\partial^2 f(x, \theta)/\partial x \partial x'$, evaluated at $x = x^*(\theta)$. Laplace importance sampling proceeds by using the normal distribution implied by the quadratic expansion of the log integrand about the maximum as importance sampling distribution, giving an importance sampler

$$\hat{L}_1(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{e^{f(x_i,\theta)}}{\phi(x_i; x^*(\theta), (H(\theta))^{-1})}. \tag{14}$$

where $x_1, \ldots, x_n$, is a random sample based on the normal distribution with mean $x^*(\theta)$ and precision matrix $H(\theta)$. The expression for EPD importance sampling is found by noting that the sample $x_1, \ldots, x_n$ may be generated by $x_i = x^* + C(\theta)z_i$, where $z_1, \ldots, z_n$ are draws from a multivariate, independent, standard normal distribution, and $C(\theta)$ is the Cholesky factor of $H(\theta)^{-1}$. The density of x_i can be expressed using the density of z_i, by

$$\phi(x_i; x^*(\theta), (H(\theta))^{-1}) = \frac{\phi(z_i; 0; I)}{|C(\theta)|}. \tag{15}$$

Thus, equation (14) is equivalent to

$$\hat{L}_1(\theta) = |C(\theta)|\frac{1}{n}\sum_{i=1}^{n} \frac{e^{f(\theta, x^*(\theta)+C(\theta)z_i; X)}}{\phi(z_i; 0, I)}, \tag{16}$$

which is ready for use with EPD importance sampling.

The term Laplace importance sampling was introduced by Kuk (1999) in the context of the corresponding simple importance sampler, found by substituting θ_0 for θ in the denominator in equation (14) and using the density from this denominator as an importance sampler.) The main version of the simulated likelihood in Durbin and Koopman (1997, 2000) is equivalent to the EPD Laplace importance sampler. EPD Laplace importance samplers were also applied in Pinheiro and Bates (1995), Skaug (2002) and Skaug and Fournier (2006).

The second example is a very simple EPD importance sampler. Assume that $f(x, \theta) = f_1(x, \theta) + f_2(x, \theta)$, where f_1 is the log likelihod function, conditional on the random effects and f_2 is the log density function of random effects. Assume that draws based on this density function may be generated by $x_i = g(z_i, \theta)$, where $z_1, \ldots, z_n$ are some random numbers independent of θ and g is a smooth function. Now,

$$\hat{L}_2(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_1(x_i, \theta). \tag{17}$$

The sampling density associated with x_i is canceled out by f_2. While it might be stretching the term to describe this as importance sampling, the sampling procedure is certainly explicitly parameter dependent. Still, the performance in a simulated maximum likelihood context should be expected to be poor. Although such results are not reported, it fares far worse than simple Laplace importance sampling in the numerical examples below. There are three main points with this example. First, the definition of EPD importance samplers above is broad. Secondly, EPD importance samplers may be completely straightforward. Thirdly, using an EPD importance sampler is not sufficient for achieving good performance in simulated maximum likelihood estimation.

The third example is the GHK simulation algorithm, developed by Geweke (1991), Hajivassiliou (1990) and Keane (1993) for simulation of the likelihood of multinomial probit models. The likelihood under consideration can be expressed as a product of probabilities of the type

$$Pr(U_1 \leq 0, \ldots, U_m \leq 0), \tag{18}$$

where $U_1, \ldots, U_m$ are joint normal with expectation and variance matrix depending on the parameters of the model, with the diagonal elements of the variance set to one without loss of generality. In the following, let $m = 2$ for expositional purposes, even though m would necessarily be higher, maybe 4, for simulation to be necessary for likelihood evaluation. The simplest variant of the GHK algorithm exploits

$$Pr(U_1 \leq 0, U_2 \leq 0) = Pr(U_1 \leq 0) Pr(U_2 \leq 0 | U_1 \leq 0), \tag{19}$$

where the first factor is straightforward to evaluate and the second factor is not. However, if $z_1, \ldots, z_n$ are independent draws from a uniform distribution on $(0, 1)$,

$$x_i = \Phi^{-1}(z_i \Phi(-\mu_1)) + \mu_1, i = 1, \ldots, n, \tag{20}$$

where Φ is the distribution function of the standard normal distribution and μ_1 is the expectation of U_1, will give a sample from the distribution of U_1, conditional on $U_1 \leq 0$. Hence,

$$Pr(U_1 \leq 0, U_2 \leq 0) \approx Pr(U_1 \leq 0) n^{-1} \sum_{i=1}^{n} Pr(U_2 \leq 0 | U_1 = x_i). \tag{21}$$

The extension to $m > 2$ is straightforward. It is easy to draw values for U_2, conditional on $U_2 \leq 0$ and conditional on exact values of U_1, with a formulas similar to equation (20). The importance sampler is EPD, as the function defining x_i is parameter dependent and smooth. In addition to providing an example of the diversity of EPD importance sampling as defined here, the GHK algorithm is also an example of an EPD importance sampler that has been applied to simulated maximum likelihood estimation with documented success. The numerical adequacy of the GHK algorithm in simulated maximum likelihood has been documented in a number of studies, see e.g. Stern (1997) for an overview and more in-depth discussion of the algorithm.

4 Numerical illustration - salamander mating

This numerical illustration is confined to the infamous salamander mating model that first appeared in McCullagh and Nelder (1989). I first briefly illustrate that the exact maximum likelihood can be found to very high precision using EPD Laplace importance sampling. I then evaluate the difference

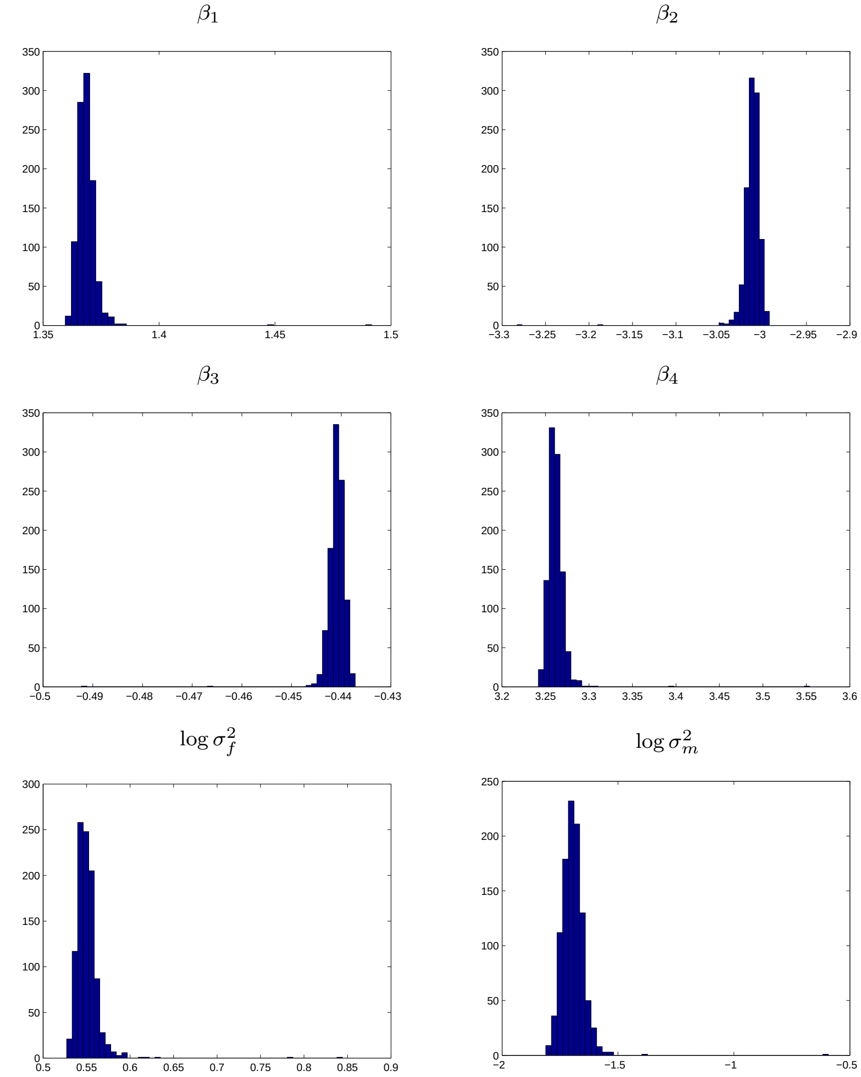

Fig. 1 Histograms of parameter estimates from 1000 replications of simulated maximum likelihood estimation using EPD Laplace importance sampling with importance sample size 1000.

between simple and EPD Laplace importance sampling by studying the distributions of score function at the simulated maximum likelihood estimate. Thirdly, I briefly study how the variability of the estimates based on the simple Laplace importance sampler relates to the lower bound discussed above.

The salamander mating model originates from an experiment where the aim was to study how the mating success of salamanders depends on whether they originate from different populations. A small sample of males and females

Table 1 Mean, median and standard deviation in 1000 replications of simulated maximum likelihood estimation using EPD Laplace importance sampling with importance sample size 1000, corresponding to histograms in Figure 1.

Statistic	β_1	β_2	β_3	β_4	$\log \sigma_f^2$	$\log \sigma_m^2$
Mean	1.3687	-3.0122	-0.4410	3.2619	0.5505	-1.6943
Median	1.3682	-3.0111	-0.4408	3.2606	0.5487	-1.6971
Standard deviation	0.0057	0.0125	0.0023	0.0128	0.0160	0.0554

Table 2 Simulated maximum likelihood estimate of salamander mating model based on EPD Laplace importance sampling with importance sample size 100 000.

Statistic	β_1	β_2	β_3	β_4	$\log \sigma_f^2$	$\log \sigma_m^2$
Parameter estimate	1.3688	-3.0120	-0.4409	3.2615	0.5497	-1.6943
Standard error	0.6808	1.0152	0.6925	1.0858	1.6763	0.3515
Simulation error	0.0004	0.0008	0.0002	0.0008	0.0010	0.0046

from two populations were repeatedly matched with potential mating partners from the same and the other population - and success in mating was recorded - measured as a binary variable. The mating success is specified as a logit model. However, since the same salamanders enter several mating attempts, observations are not independent. To take into account this interdependency between observations, the model is specified with two random effects associated with each mating attempt, one for the male and one for the female salamander. Because the sampling design is not nested, the likelihood function of the model contains high dimensional integrals.

Specifically, let $x_{i,j}$ be a four-dimensional vector describing which out of two populations the male j and the female i originate from, with crossed effects. Let $U_1, \ldots, U_{10}$ be independent normal variables, the female random effects, with expectiation 0 and variance σ_f^2, and $V_1, \ldots, V_{10}$ independent normal variables, the male random effects, with expectiation 0 and variance σ_m^2. $y_{i,j}$ are Bernoulli random variables with conditional expectation given by

$$E(y_{i,j}|U_i, V_j) = h^{-1}(x'_{i,j}\beta + U_i + V_j), \tag{22}$$

where h is the logit link function. The data used are the "summer experiment only" variant, where the likelihood function factors into two such 20-dimensional integrals. The model has been studied in a vast range of articles, e.g. Karim and Zeger (1992), Lin and Breslow (1996), Shun (1997) and Kuk (1999), with similar variants also studied in Booth and Hobert (1999) and Skaug (2002). The maximum likelihood estimate is also found to high precision in Skaug (2002).

Because simulated maximum likelihood can produce heavy-tailed distributions of estimates, a documentation of successful application of the method should arguably include histograms of replicated parameter estimates. Figure 1 shows histograms based on 1000 parameter estimates of the salamander mating model based on simulated maximum likelihood using EPD Laplace

Table 3 Mean and standard deviation in 1000 simulated score functions at approximate maximum likelihood, based on simple and EPD Laplace importance sampling with importance sample size 1000.

Method	Statistic	β_1	β_2	β_3	β_4	$\log \sigma_f^2$	$\log \sigma_m^2$
Simple	Mean	0.0051	0.0002	0.0031	0.0007	-0.0082	-0.0024
EPD	Mean	0.0009	0.0003	0.0004	0.0010	-0.0033	-0.0003
Simple	Std. dev.	0.0025	0.0016	0.0015	0.0010	0.0042	0.0043
EPD	Std. dev.	0.0002	0.0001	0.0002	0.0002	0.0010	0.0002

importance sampling with importance sample size 1000. (Here and in the following, standard antithetic variables are used in importance samplers, see e.g. Durbin and Koopman (1997)) Clearly, the parameters are estimated rather precisely, and the stochastic variability seems to be adequately good-natured, although with somewhat heavy one-sided tails. Table 1 gives the corresponding means, medians and standard deviations. It is clear that even though there are outliers in the histograms indicating heavy one-sided tails, the means are not much affected, being very close to the medians. The method works adequately for the purposes in this numerical example.

Table 2 presents corresponding parameter estimates based on EPD Laplace importance sampling, but with importance sample size increased to 100 000, together with standard errors based on the observed information. Further, the table presents predicted standard deviations of the parameters due to randomness in the importance sampling, based on the asymptotic limit in Theorem 1, *simulation errors* in the following. The parameter estimates are practically identical to the mean out of 1000 parameter estimates in Table 1. Simulation errors are very low and broadly consistent with the standard deviations from Table 1.

It is not trivial how to set up a comparison between simple and EPD importance sampling. Simulated maximum likelihood based on simple importance sampling is usually prescribed as an iterative method, with new importance samplers with θ_0 in equation (2) based on current θ. This procedure continues until convergence, at least in the informal sense of having a sequence of two similar estimates. Simulated maximum likelihood based on EPD importance sampling is not iterative in the same sense. The solution to this comparison problem is to assess the asymptotic efficiency of the methods, based on the variance of the score functions evaluated at the approximate maximum likelihood found above.

Table 3 presents the mean and variances of 1000 simulated score functions based on simple and EPD importance sampling evaluated at the simulated maximum likelihood estimate presented in Table 2. The score functions are simulated as the (analytical) derivatives of the simple and EPD Laplace importance samplers. The importance samples used for computation are identical for the simple and EPD importance samplers - hence the simulated likelihood function values are identical, but the scores differ. The means of the score functions are about 0 for both types of importance sampler, corroborating

Table 4 Asymptotic "simulation errors" derived from simulated score functions and approximated lower bounds based on "missing information".

Method	β_1	β_2	β_3	β_4	$\log\sigma_f^2$	$\log\sigma_m^2$
Sim. err. (simple)	0.0680	0.1262	0.0403	0.1375	0.0637	1.2527
Sim. err. (EPD)	0.0046	0.0098	0.0018	0.0097	0.0135	0.0525
Simple lower bound, appr. 1	0.0331	0.0534	0.0146	0.0450	0.0285	0.3082
Simple lower bound, appr. 2	0.0357	0.0593	0.0155	0.0525	0.0336	0.4214

that we have found the approximate maximum likelihood. The standard deviations of the elements of the score function are from 4 to 20 times as high for the simple importance sampler. This demonstrates that the disadvantage with simple importance sampling is not only that the importance sampler "may be inaccurate when θ is not close to θ_0", to paraphrase Skaug (2002). It is inaccurate in the relevant sense also at θ_0.

The variance of the score function translates into variance in the simulated maximum likelihood estimates asymptotically through equation (5). The predicted simulation errors, the square roots of the diagonal entries of Ω, are presented in Table 4. The variability varies quite a lot between parameters. For $\log\sigma_f^2$ the simulation error is only between 4 and 5 times as high for the simple importance sampler. For the other parameters, the simulation error is 13-25 times as high. Thus, even with an importance sample of 100 or 150 times the size, the simulated maximum likelihood estimate based on simple importance sampling will have higher simulation error than a simulated maximum likelihood based on EPD importance sampling for 5 out of 6 parameters.

A final question to assess is to what extent the inferior performance of the simulated maximum likelihood based on simple importance sampling can be explained by the lower bound for simple importance samplers discussed in Section 2. There are two easily available approximations to the missing information. Both are based on the log integrand in the likelihood function taken as a function of both parameters and random effects, evaluated at the parameter values reported in Table 2, together with the posterior mode of the random effects parameters, $x^*(\theta)$. Approximation 1 is the missing information computed solely on the basis of the Hessian of this penalized conditional likelihood function. Approximation 2 uses the observed information that is the basis for standard errors in Table 2 in place of the observed information from the penalized conditional likelihood function. The approximated lower bounds to simulation errors are presented in Table 4 and are quite similar. The variability of estimates based on the simple Laplace importance sampler is of the same order of size as the lower bounds, giving about 2-4 times as high standard deviations as the lower bound. Hence, the poor performance of the simple Laplace importance sampling algorithm in this example most likely reflects the drawbacks discussed in Section 2.

5 Discussion

Simulated maximum likelihood is a powerful method for approximating the exact maximum likelihood estimates in models where the likelihood function contains moderately difficult integrals. A necessary prerequisite for a powerful simulated maximum likelihood algorithm is that it is based on explicitly parameter dependent importance sampling - otherwise efficient simulation of the score function is not possible. Hence, assessments of simulated maximum likelihood in the statistics literature based on the simple importance sampling algorithm are correct in the assessment that this is not a powerful method. This is however only due to an important deficiency in the importance sampling algorithms applied, not the potential performance of simulated maximum likelihood per se.

The lack of a clear distinction between simple and EPD importance sampling, or even the lack of consciousness about the existence of the technique of EPD importance sampling, has hampered research on simulation based estimation. Simulated maximum likelihood based on EPD Laplace importance sampling performs extremely well in a number of cases, including the salamander mating model used here and in examples in Skaug (2002) and Durbin and Koopman (1997, 2000). There is considerable scope for application of the method and the method is applied in practice. There are important further research questions associated with simulated maximum likelihood, in particular in terms of controlling and diagnosing heavy-tailed variability in estimates, see Koopman et al (2009). It is important that research on simulated maximum likelihood recognizes the huge performance difference between simple and EPD importance sampling. As an example, a recent contribution, Richard and Zhang (2007), develops a new technique for importance sampling of likelihood functions containing high dimensional integrals. Their importance sampler is EPD in the sense defined here, with impressive results. However, it is difficult to assess whether the results are impressive because an EPD importance sampler is applied or because their new importance sampling technique performs well compared to e.g. EPD Laplace importance sampling.

References

Booth JG, Hobert JP (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. Journal of the Royal Statistical Society Series B 61(1):265–285

Durbin J, Koopman SJ (1997) Monte Carlo maximum likelihood estimation for non-Gaussian state space models. Biometrika 84(3):669–684

Durbin J, Koopman SJ (2000) Time series analysis on non-Gaussian observations based on state space models from both classical and Bayesian perspectives. Journal of the Royal Statistical Society, Series B 62:3–56

Gelfand AE, Carlin BP (1993) Maximum-likelihood estimation for constrained-or missing-data models. The Canadian Journal of Statistics 21(3):303–311

Geweke J (1989) Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57(6):1317–1339

Geweke J (1991) Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In: Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface, pp 571–78

Geyer CJ, Thompson EA (1992) Constrained Monte Carlo Maximum Likelihood for Dependent Data. Journal of the Royal Statistical Society Series B 54(3):657–699

Hajivassiliou V (1990) Smooth Simulation Estimation of Panel Data LDV Models. Department of Economics, Yale University

Hajivassiliou V, Ruud P (1994) Classical estimation methods for ldv models using simulation. In: Engle RF, McFadden D (eds) Handbook of Econometrics, vol. 4, Amsterdam: North-Holland

Jank W (2006) Efficient simulated maximum likelihood with an application to online retailing. Statistics and Computing 16(2):111–124

Jank W, Booth J (2003) Efficiency of Monte Carlo EM and simulated maximum likelihood in two-stage hierarchical models. Journal of Computational and Graphical Statistics 12(1):214–229

Karim MR, Zeger SL (1992) Generalized Linear Models with Random Effects; Salamander Mating Revisited. Biometrics 48(2):631–644

Keane MP (1993) Simulation Estimation for Panel Data Models with Limited Dependent Variables. Handbook of Statistics 11:545–571

Koopman SJ, Shephard N, Doornik JA (1999) Statistical algorithms for models in state space using ssfpack 2.2. The Econometrics Journal 2(1):107–160

Koopman SJ, Shephard N, Creal D (2009) Testing the assumptions behind importance sampling. Journal of Econometrics 149(1):2–11

Kuk AYC (1999) Laplace importance sampling for generalized linear mixed models. Journal of Statistical Computation and Simulation 63(2):143–158

Kuk AYC, Cheng YW (1999) Pointwise and functional approximations in Monte Carlo maximum likelihood estimation. Statistics and Computing 9(2):91–99

Lin X, Breslow NE (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of the American Statistical Association 91:1007–1016

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd ed. Chapman and Hall

McCulloch CE (1997) Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association 92:162–170

McCulloch CE, Searle SR (2001) Generalized, Linear and Mixed Models. Wiley

Pinheiro JC, Bates DM (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics 4(1):12–35

Richard JF, Zhang W (2007) Efficient high-dimensional importance sampling. Journal of Econometrics 141(2):1385–1411

Robert CP, Casella G (2004) Monte Carlo Statistical Methods. Springer

Shun Z (1997) Another Look at the Salamander Mating Data: A Modified Laplace Approximation Approach. Journal of the American Statistical Association 92(437):341–349

Skaug HJ (2002) Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. Journal of Computational and Graphical Statistics 11:458–470

Skaug HJ, Fournier DA (2006) Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. Computational Statistics and Data Analysis 51(2):699–709

Stern S (1997) Simulation-Based Estimation. Journal of Economic Literature 35(4):2006–2039

A Appendix A: Proof of Theorem

Proof As is easily verified through equation (3), the gradient of the simulated likelihood function at $\tilde{\theta}$ is an importance sampler of the gradient of the exact likelihood function at $\tilde{\theta}$ and thus converges a.s. to 0 as $n \to \infty$.

The gradient of the simulated likelihood function at $\tilde{\theta}$ is the mean of i.i.d. random variables, W from equation (6). (Found by taking the derivatives in equation (3) and substituting terms). By Assumption 2, W has finite variance, and the central limit theorem can be applied: The gradient of the simulated likelihood function, scaled by $\sqrt{n}$ is asymptotically normal with expectation 0.

The simulated score function is the ratio of the gradient of the simulated likelihood to the simulated likelihood. As the simulated likelihood converges a.s. to the exact likelihood, it also converges to the exact likelihood in probability and the Slutsky theorem can be applied. Thus also the simulated score function, scaled by $\sqrt{n}$, is asymptotically normal.

Denote the simulated score function at $\tilde{\theta}$ by S and its asymptotic variance matrix Σ. In a neighborhood of $\tilde{\theta}$, the simulated score function can be approximated by a linear function in θ. Thus

$$s(\theta) = S + I(\theta - \tilde{\theta}) \tag{23}$$

where I is the observed information at $\tilde{\theta}$, which is non-singular by Assumption 1. The solution to the first order condition of the maximization problem is

$$\theta_n - \tilde{\theta} = -I^{-1}S \tag{24}$$

and the simulated maximum likelihood estimate is (asymptotically) a linear function of the simulated score at the exact maximum likelihood estimate.