Discussion Papers No. 540, April 2008
Statistics Norway, Research Department

Christian N. Brinch

Simulated Maximum Likelihood using Tilted Importance Sampling

Abstract:
This paper develops the important distinction between tilted and simple importance sampling as methods for simulating likelihood functions for use in simulated maximum likelihood. It is shown that tilted importance sampling removes a lower bound to simulation error for given importance sample size that is inherent in simulated maximum likelihood using simple importance sampling, the main method for simulating likelihood functions in the statistics literature. In addition, a new importance sampling technique, generalized Laplace importance sampling, easily combined with tilted importance sampling, is introduced. A number of applications and Monte Carlo experiments demonstrate the power and applicability of the methods. As an example, simulated maximum likelihood estimates from the infamous salamander mating model from McCullagh and Nelder (1989) can be found to easily satisfactory precision with an importance sample size of 100.

Keywords: Simulation based estimation, importance sampling.

JEL classification: C13, C15

Acknowledgement: Thanks to Taryn Galloway, Hans J. Skaug and Terje Skjerpen for helpful comments.

Address: Statistics Norway, Research Department and Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo, e-mail: cnb@ssb.no

Discussion Papers	comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no
http://ideas.repec.org/s/ssb/dispap.html

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
NO-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1 Introduction

Generalized linear mixed models and other nonlinear and non-Gaussian models incorporating random effects are widely used in applied sciences, see e.g. McCulloch and Searle (2001). Random effects are routinely modelled to take into account dependency structures between observations, thus relaxing the traditional requirement of independent observations in data analysis. With a slightly different perspective, random effects models are used for the study of variance components with e.g. binary or count data. In addition, state space models and many models in spatial statistics are also technically similar to random effects models.

There are still unsolved computational problems associated with parameter estimation in nonlinear or non-Gaussian models with random effects, see e.g. Robert and Casella (2004). Statistical inference may be difficult, as evaluation of the likelihood function requires the evaluation of potentially high dimensional integrals. Some estimation methods, such as simulation based Bayesian inference or the Monte Carlo EM algorithm, handle this problem by bypassing evaluation of the likelihood function. However, one of the core methods for inference in such models, simulated maximum likelihood (SML), tackles the problem head-on by applying importance sampling to simulate the likelihood function and maximizing the simulated likelihood function using numerical techniques.

This paper contributes to the literature on SML by distinguishing between tilted and simple importance sampling as techniques for simulation of (components of) likelihood functions. Simple importance sampling is based on the same importance sample for evaluation of the likelihood function for all parameter values. Tilted importance, on the other hand, applies parameter dependent importance sampling by deriving the importance samples as parameter dependent functions of a set of common random numbers. Section 2 defines simple and tilted importance sampling and demonstrates how and why tilted importance sampling allows us to escape from a lower bound on simulation error (for given importance sample size) that is inherent in SML using simple importance sampling.

SML using simple importance sampling is the main SML algorithm in the statistics literature. An authorative reference work like Robert and Casella (2004) even defines SML in terms of using simple importance sampling. However, SML using tilted importance sampling is not a new method. SML using tilted importance sampling as defined here is applied the state space literature starting with Durbin and Koopman (1997, 2000), using software presented in

Koopman, Shephard, and Doornik (1999). In addition, Skaug (2002) used SML based on tilted importance sampling for generalized linear mixed models in a paper focussing on automatic differentiation. There are also other SML implementations in the econometrics literature that use techniques that involve tilted importance sampling, such as the rich literature on estimation of multinomial probit models, see e.g. Stern (1997) or the methods presented in Richard and Zhang (2007). Although the method of SML using tilted importance sampling is, thus, not novel, the existing literature has failed to communicate the essential difference between SML based on tilted and simple importance sampling. In particular, the superior performance of SML using tilted importance sampling compared to simple importance sampling and an explanation of this performance difference has not been presented. Moreover, the impact of SML using tilted importance sampling seems to be mainly confined to the econometrics literature, except for Skaug (2002).

Section 3 discusses Laplace importance samplers, a type of importance samplers that can easily be adapted to tilted importance sampling. Laplace importance samplers essentially simulate the difference in the likelihood from the likelihood of a linear Gaussian model, by using the Gaussian distribution based on the quadratic approximation of the log penalized conditional likelihood about the maximum as an importance sampling distribution. A new type of importance sampler, the generalized Laplace importance sampler, that allows the researcher to try out different and more robust importance samplers while still only simulating the difference from the likelihood of a linear Gaussian model is developed here.

Section 4 briefly describes a prototype implementation of the algorithms, that is, evaluation of the log likelihood function and its gradient, based on tilted importance sampling using the generalized Laplace importance sampling scheme. This implementation is used for the applications that follow in Section 5. Here cases from the literature and Monte Carlo experiments are reported to demonstrate the performance of SML using tilted importance sampling. Section 6 concludes.

2 Importance sampling and simulated maximum likelihood

Importance sampling is a well known technique for estimation of an integral

$$y = \int_{\mathbb{R}^q} f(x)dx, \tag{1}$$

by considering a probability density function π with q-dimensional domain and exploiting the identity $f(x) = (f(x)/\pi(x))\pi(x)$, to estimate the integral by

$$\hat{y} = \sum_{i=1}^{n} \frac{f(x_i)}{\pi(x_i)}, \tag{2}$$

where $x_1, \ldots, x_n$ is a (pseudo-)random sample based the distribution characterized by π, the importance sampling distribution. If the sample is independent, $\hat{y}$ is an unbiased estimator of y with

$$Var(\hat{y}) = n^{-1/2} Var\left(\frac{f(X)}{\pi(X)}\right), \tag{3}$$

where X is a random variable characterized by π, provided the variance exists. The usual aim of importance sampling is that the variance of the ratio $f(X)/\pi(X)$ exists and is as small as possible, while it is must be relatively easy to obtain a random sample from π.

For use with SML, the aim is somewhat different. The main goal is to find the arg max of a likelihood function $L(\theta)$, with $\theta \in \Theta$, with Θ denoting some set. The likelihood function consists of one or more integrals that depend on θ, for simplicity of exposition, assume that the likelihood function is one such integral. The strategy labeled SML is to provide a smooth Monte Carlo estimate of the likelihood function and to use the arg max of this function as a Monte Carlo estimate of the arg max of the exact function. The starting point is thus to provide a smooth Monte Carlo estimate of the function

$$L(\theta) = \int_{\mathbb{R}^q} f(x, \theta) dx. \tag{4}$$

Simple importance sampling uses a distribution with density $\pi(x, \theta_0)$ that is a good approximation of the integrand $f(x, \theta)$ in the neighborhood of some initial guess θ_0 as an importance sampling distribution, giving

$$\hat{L}(\theta; \theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i, \theta)}{\pi(x_i; \theta_0)}, \tag{5}$$

where $x_1, \ldots, x_n$ is a random sample from the distribution characterized by $\pi(x; \theta_0)$. Clearly, with simple importance sampling, the importance sample is independent of the parameter θ. *Tilted importance sampling* can be derived from the more general expression

$$\hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i, \theta)}{\pi(x_i; \theta)}. \tag{6}$$

Suppose there exists a density $\pi_z(z)$ such that 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 (potentially vector valued) 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 (6) with

$$\hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} |J(\theta, z_i)| \frac{f(g(z_i,\theta);\theta)}{\pi_z(z_i)}. \tag{7}$$

The most obvious difference between simple and tilted importance sampling is the local nature of the simple importance sampler. Both techniques give smooth simulated likelihood functions based on sets of common random numbers. Simple importance samplers are the subset of tilted importance samplers that are parameter independent. Of course, equation (7) can be considered a simple importance sampler applied to a transformed version of the original problem.

In the following we will study the limitations of the simple importance sampler. Hence, assume a regular case, where the exact likelihood function has a unique maximum, $\tilde{\theta}$, and the simulated likelihood function has finite moments for all $\theta \in \Theta$. Let equation (6) represent the simulated likelihood function. The score function can now be written,

$$\frac{\partial \log \hat{L}(\theta)}{\partial \theta} = (\hat{L}(\theta))^{-1} \frac{1}{n}\sum_{i=1}^{n} W_i(\theta) = (\hat{L}(\theta))^{-1} \bar{W}(\theta), \tag{8}$$

where $W_i(\theta)$ for given θ are i.i.d. random variables defined by

$$W_i(\theta) = \frac{f(x_i,\theta)}{\pi(x_i;\theta)} \left(\frac{\partial \log f(x_i,\theta)}{\partial \theta} - \frac{\partial \log \pi(x_i;\theta)}{\partial \theta}\right). \tag{9}$$

At $\tilde{\theta}$ the vector $\bar{W}(\tilde{\theta})$ is asymptotically normal with expectation 0. Locally, $\bar{W}(\theta)$ can be approximated by a linear function in θ. Thus, in particular we can, if the SML estimate is $\hat{\theta}$ write

$$\bar{W}(\tilde{\theta}) = \hat{L}(\hat{\theta}) I_o (\tilde{\theta} - \hat{\theta}), \tag{10}$$

where the linear expansion is about $\hat{\theta}$, and I_o is the observed information at $\hat{\theta}$. The difference between the exact maximum likelihood estimate and the SML estimate is now asymptotically normal with expectation 0 and covariance matrix

$$\Omega = I_o^{-1} \Sigma I_o^{-1}, \tag{11}$$

where Σ is the covariance matrix of $(L(\hat{\theta}))^{-1}\bar{W}(\hat{\theta})$.

The optimal simple importance sampler is usually defined as $\pi(x;\theta_0) = f(x,\tilde{\theta})$, giving exact likelihood value at $\tilde{\theta}$. With this importance sampler, $W_i(\tilde{\theta})$ simplifies to

$$W_i(\tilde{\theta}) = \frac{\partial \log f(x_i,\theta)}{\partial \theta} \tag{12}$$

Thus, even with an optimal importance sampler for the likelihood function, the SML estimate is random if the derivative of the integrand depends on x_i, even though the likelihood function value at $\tilde{\theta}$ is exact. A simple importance sampler cannot be optimal for use with SML unless the score function is independent of the random effects.

It is possible to characterize the distribution of W_i further under a quadratic approximation. Assume that $\log f$ can be approximated by a quadratic expansion about its maximum in (x,θ), denoted (x^*,θ^*) Thus, with $\lambda = (x - x^*, \theta - \theta^*)$

$$\hat{f}(x,\theta) = f(x^*,\theta^*)e^{-\frac{1}{2}\lambda I \lambda'}, \tag{13}$$

where I can be partitioned as

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} \tag{14}$$

with dimensions corresponding to the dimensions of x and θ. Then, $I_o = I_{22} - I_{21}I_{11}^{-1}I_{21}$ and

$$W_i = (x_i - x^*)' I_{12} + I_{22}(\theta - \theta^*). \tag{15}$$

The variance of X is under this importance sampler equal to I_{11}^{-1}, and the variance of W_i follows as $I_m = I_{21}I_{11}^{-1}I_{12}$, the missing information. The missing information is the difference between the complete information, the negative of the Hessian of $\log \hat{f}$ with respect to θ, I_{22}, and the observed information. Thus, the larger the loss of information from not observing the random effects, the larger will the lower bound on simulation error under simple importance sampling be. This argument is inspired by Jank and Booth (2003), who compared SML using simple importance sampling with the Monte Carlo EM algorithm.

As a contrast, SML using tilted importance sampling gives no lower bound on Σ for given importance sample size, as clearly, the optimal importance sampler is $\pi(x;\theta) = f(x,\theta)$ which gives zero variance. This is not only a theoretical result, but a practical result that occurs when

e.g. tilted Laplace importance sampling as defined below is applied to models with quadratic log penalized conditional likelihood functions, such as linear mixed models. As we shall see in the applications below, the variance associated with tilted importance samplers can also be very low in applications where the importance samplers are not optimal.

The lesson of this section is that in the context of SML, an optimal importance sampler is an importance sampler that is tilted such that the derivatives of the importance sampling density w.r.t. the parameters of the model are equal to the derivatives of the log penalized conditional likelihood function. This is equivalent to transforming the original simulation problem to a simulation problem where the integrand is independent of the parameters of the model. Such exact tilting will usually not be feasible. However, this concept of optimality sets out rather different heuristics for choosing importance samplers than the standard concept of optimal importance sampling distributions outlined above for the simple importance sampler.

In addition to the core importance for simulating the exact ML estimates, the global nature of the tilted importance samplers are also useful for procedures based on likelihood values away from the maximum, such as likelihood ratio tests or associated interval estimators or for dimension reduction in Bayesian analysis, see Sections 5.1.5 and 5.1.6 for applications.

3 Laplace importance sampling

3.1 Standard Laplace Importance Sampling

Let the function to be estimated be specified as

$$L(\theta) = \int_{\mathbb{R}^q} e^{f(x,\theta)} dx, \tag{16}$$

and let f have a unique maximum in x and let it be twice differentiable in x for all θ. (Note that f now denotes the log integrand.) These assumptions are relevant for many cases and are guaranteed to be satisfied e.g. for the penalized conditional likelihood functions of generalized linear mixed models with canonical link functions and Gaussian random effects.

Let $x^*(\theta) = \arg\max_x f(x,\theta)$. Further let $H(\theta) = -\frac{\partial^2 f(x,\theta)}{\partial x \partial x'}$, evaluated at $x = x^*(\theta)$. The proposal is to use the normal distribution implied by the quadratic expansion of the log integrand

as importance sampling distribution, giving an importance sampler

$$\hat{L}_1(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{e^{f(x_i,\theta)}}{\phi(x_i; x^*(\theta), (H(\theta))^{-1})}. \tag{17}$$

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 necessary expression for tilted importance sampling is then 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

$$\phi(x_i; x^*(\theta), (H(\theta))^{-1}) = \frac{\phi(z_i; 0; I)}{|C(\theta)|}, \tag{18}$$

Thus, equation (17) is equivalent to

$$\hat{L}_1(\theta) = |C(\theta)|\frac{1}{n}\sum_{i=1}^{n} \frac{e^{f(\theta,u^*(\theta)+C(\theta)v_i;X)}}{\phi(v_i; 0, I)}, \tag{19}$$

which is ready for use with tilted importance sampling.

The relationship between Laplace importance sampling and the Laplace approximation can be highlighted by rewriting equation (19) as

$$\hat{f}_1(\theta) = (2\pi)^{q/2} e^{f(x^*(\theta),\theta)} |H(\theta)|^{-1/2} \frac{1}{n}\sum_{i=1}^{n} e^{f(x_i,\theta) - f(x^*,\theta) + \frac{1}{2}(x_i - x^*)' H(\theta)(x_i - x^*)}. \tag{20}$$

The Laplace approximation is based on the identity

$$f(\theta) = (2\pi)^{q/2} e^{f(x^*(\theta),\theta)} |H(\theta)|^{-1/2} E(e^{f(X,\theta) - f(x^*,\theta) + \frac{1}{2}(X - x^*)' H(\theta)(X - x^*)}), \tag{21}$$

where the expectation (with respect to X) is taken over a normal distribution with expectation x^* and precision matrix $H(\theta)$, and the variables x_i in equation (20) are draws from the same distribution. The standard Laplace approximation may thus be derived by approximating the expectation in equation (21) with one. With Laplace importance sampling, the expectation is estimated by straightforward Monte Carlo simulation.

The term "Laplace importance sampling" was introduced by Kuk (1999a). Kuk's simulated

likelihood is based on

$$\hat{f}_{\theta_0}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{e^{f(x_i,\theta)}}{\phi(x_i; x^*(\theta_0), (H(\theta_0))^{-1})}, \tag{22}$$

that is, simple importance sampling combined with Laplace importance sampling. A similar importance sampler was applied by Pinheiro and Bates (1995) for nonlinear mixed models. Skaug (2002) uses tilted Laplace importance sampling as it is defined here. The importance sampling tradition in the analysis of state space models following Durbin and Koopman (1997) uses tilted Laplace importance sampling, as defined here, though this may not be transparent due to differences in terminology. The "simulation smoothers" from the state space literature are algorithms that perform the transformation from common random numbers in the same fashion as the tilted importance sampler in equation (6). This method has been misrepresented in Kuk (1999b), as SML using simple importance sampling.

3.2 Generalized Laplace importance sampling

The main drawback with importance sampling is that rare draws may have large impact, giving heavy tailed distributions for the simulated likelihood, which may be passed on to heavy tailed distributions for the parameter estimates. In particular, it is prudent to worry about the problem that second and higher order moments may not exist for the SML estimates.

The most straightforward solution to the problem of nonexistent variance is to use a more dispersed importance sampler, by multiplying the standard deviation of the importance sampling distribution with a constant $r > 1$. We will in the following refer to r as "the excess dispersion factor" (relative to the quadratic approximation). A direct implementation is

$$\hat{f}_2(\theta) = (2\pi)^{q/2} e^{f(\theta,x^*(\theta))} |H(\theta)|^{-1/2} r^q \frac{1}{n}\sum_{i=1}^{n} \frac{e^{f(\theta,x^*(\theta)+rC(\theta)v_i)-f(\theta,x^*(\theta))}}{\phi(v_i; 0, I)}. \tag{23}$$

Clearly, for some sufficiently high r, the summand will be bounded above by one, attained at $v_i = 0$, which implies that the variance of the importance sampler is finite (for some r). However, this importance sampler has the drawback, compared to the standard Laplace importance sampler, that the results will not be exact for linear Gaussian models, so we are no longer only simulating the difference from the likelihood of a linear Gaussian model.

Exactness for linear gaussian models can be reintroduced using control variates (see Ripley, 1987). Control variates are based on the idea that instead of directly simulating $E(X)$, we can simulate $E(X - (Z - E(Z)))$, where the inner expectation can be solved analytically. Now as a

control variate, we now use the approximation to $e^{f(x,\theta)}$ based on the second order expansion of f about its mode. Then we can derive the following estimator, the *generalized Laplace importance sampler*,

$$\hat{f}_3(\theta) = e^{f(x^*(\theta),\theta)}|H(\theta)|^{-1/2}$$
$$(1 + n^{-1}\sum_{i=1}^{n} p_i^{-1}(e^{f(x^*(\theta)+C(\theta)w_i,\theta)-f(x^*(\theta),\theta)} - e^{-\frac{1}{2}w_i'w_i})), \tag{24}$$

where $w_i = rv_i$ and p_i is the draw density

$$p_i = (2\pi)^{-q/2}r^{-q}e^{-\frac{1}{2}v_i'v_i}. \tag{25}$$

Note that Laplace importance sampling appears as the special case with $r = 1$. Like the case without control variates, the generalized Laplace importance sampler has finite variance with sufficiently high excess dispersion factor. However, with too high excess dispersion factor, the variance will be high, even if it is finite. Note also the potential problem in that the simulated factor in equation (15) may turn out as negative. That anyway indicates that the importance sampler is working far from well.

The control variate technique is here used in a completely different manner than in Durbin and Koopman (1997). In that study, control variates are used to take into account deviations from the linear-Gaussian approximating model. Here, control variates are used to ensure that computations are exact in a linear-Gaussian model even when the importance sampler differs from the standard Laplace importance samplers as outlined in the previous section.

It should be added that the techniques used in this section can be applied with any importance sampling distribution that we can manipulate through equation (5). As long as we use the quadratic approximation of the log integrand about its mode as a control variate, we need not use Gaussian distributions as importance sampler to ensure exactness in the linear Gaussian case.

A simple example is to base importance sampling on (a version of) the multivariate t distribution. Multivariate t variables with v degrees of freedom may be generated by drawing from a multivariate normal distribution and scaling the distribution according to a scalar variable ρ, where $\rho = 1/\sqrt{Z/v}$, where Z is drawn from a chi-square distribution with v degrees of freedom, see Nadarajah and Kotz (2005). It is thus simple to draw w_i. The draw probability densities

involve a unidimensional integral,

$$p_i = \int x^{q/2} e^{-\frac{1}{2} w_i' w_i x} h(x) dx, \tag{26}$$

where $h(x)$ is a gamma density function with parameters $\alpha = v/2$ and $\beta = 2/v$. The draw probability densities are only necessary to compute once, not for every likelihood calculation. The w_i and p_i from this procedure can be plugged into equation (24) above.

3.3 Antithetic variables

Antithetic variables is a useful variance reduction technique in importance sampling, see e.g. Ripley (1987). If the variable v is drawn from a symmetric importance sampler, also $-v$ is used in the importance sample. More explicitly, equation (24) is substituted with

$$\begin{aligned} \hat{f}_4(\theta) &= e^{f(x^*(\theta),\theta)} |H(\theta)|^{-1/2} \\ &(1 + n^{-1} \textstyle\sum_{i=1}^{n} \sum_{j=0}^{1} p_i^{-1} (e^{f(x^*(\theta)+(-1)^j C(\theta) w_i, \theta) - f(x^*(\theta),\theta)} - e^{-\frac{1}{2} w_i' w_i})). \end{aligned} \tag{27}$$

Summing over j is always implicit before calculations that exploit the independence of the importance sample, such as computations of the variance estimator of the gradient or diagnostics. There is a particularly good reason for using antithetic variables in combination with Laplace importance samplers. If the expression within the expectation operator in equation (21) can be approximated by a third order polynomial, the Laplace approximation is exact, because first and second order terms are absorbed into the Laplace approximation and the third order moments of the normal distribution vanish. Third order terms do not vanish for the Laplace importance sampler unless antithetic variables are used.

Further types of antithetic variables as developed in Durbin and Koopman (1997) are not used in this paper. Such variables may be useful for variance reduction purposes. However, there is a tradeoff between using more antithetic variables and having available a larger independent sample, for estimation of simulation error and for diagnostic purposes.

4 Implementation

Implementation of the SML algorithm outlined above is not trivial, in particular because there is usually no closed form solution of the function $x^*(\theta)$. Therefore a brief description follows.

The implementation is a prototype coded in the MATLAB programming language and relies on the use of sparse matrix algorithms. The implementation is based on the following quite flexible implementational model, which encompasses generalized linear mixed models and also other random effects models (such as the stochastic volatility model discussed in the next section).

The model requires that the log penalized conditional likelihood, that is, the log integrand in the likelihood, can be expressed as

$$f(b,\theta) = \sum_{i=1}^{n} g(y_i, z_i\beta + w_i b, \gamma) + h(b,\theta), \tag{28}$$

where $y_1, \ldots, y_n$ are the dependent variables, $z_1, \ldots, z_n$ and $w_1, \ldots, w_n$ are vectors of covariates in the wide sense of also describing the random effects structure of the model. β are parameters associated with the covariates and b are the random effects (renamed from x not to be confused with covariates). The parameter γ is a dispersion parameter, which does in many applications not affect the model. The function g describes the observation specific conditional likelihood function, that is, the probability density or frequency of y_i, conditional on z_i, w_i, b and the parameters. The distribution of y_i and y_j are independent for $i \neq j$, (hence the sum), as any dependency is modelled through the random effects. The function h describes the (log) density function of the random effects, depending on a parameter θ.

The first step in the algorithm to evaluate the likelihood is to find the arg max of $f(b,\theta)$ with respect to b. This is achieved using a trust region Newton algorithm. Thus we need the gradient and Hessian of f with respect to b. In computing these, we use the following equations

$$\frac{\partial f}{\partial b} = \sum_{i=1}^{n} g^{(1)}(y_i; z_i\beta + w_i b, \gamma) Z_i + \frac{\partial h(b,\theta)}{\partial b}, \tag{29}$$

$$\frac{\partial^2 f}{\partial b \partial b'} = \sum_{i=1}^{n} g^{(2)}(y_i; z_i\beta + w_i b, \gamma)(Z_i \otimes Z_i) + \frac{\partial^2 h(b,\theta)}{\partial b \partial b'}, \tag{30}$$

where g^i denotes the i'th order partial derivative of g with respect to the linear predictor $z_i\beta + w_i b$. It is thus necessary that the code provides for (scalar) first and second derivatives for g and more general vector and matrix-valued gradients and Hessians for h, where no more structure has been imposed. It is then straightforward to Cholesky-factorize the Hessian and apply the formulae from Section 4 to find tilted generalized Laplace importance sampled likelihoods.

For maximum likelihood estimation, also the gradient of the log likelihood function with respect to parameters is extremely useful. Numerical derivatives are particularly troublesome

in this type of application, because the numerical optimization procedure inside the likelihood expression is not exact and introduces small errors in the function values that contaminates numerically evaluated derivatives. The gradient has been manually coded. There are three challenging issues here.

First, the likelihood function routine contains the arg max function $b^*(\theta)$. The derivative of this is

$$\frac{\partial b^*(\theta)}{\partial \theta} = -(\frac{\partial^2 f(b,\theta)}{\partial b \partial b'})^{-1} \frac{\partial^2 f(b,\theta)}{\partial b \partial \theta}. \tag{31}$$

Second, the likelihood contains the determinant of the Hessian, whose derivative with respect to a generic parameter α is given by

$$\frac{\partial 0.5 \log |H|}{\partial \alpha} = \sum_i \sum_j (H^{-1})_{ij} \frac{\partial H_{ij}}{\partial \alpha}. \tag{32}$$

Third, we need the derivative of the Cholesky factor of the inverse Hessian. Let C denote the (lower triangular) Cholesky factor of H^{-1}. Let B denote a matrix with ones at the subdiagonal, 0.5 at each entry of the diagonal and zeros on the superdiagonal. Now

$$\frac{\partial C}{\partial \alpha} = B \bullet \left(H^{-1} \frac{\partial H}{\partial \alpha} \right) C. \tag{33}$$

With these formulae, it is straightforward, but tedious, to code the gradients of the Laplace importance samplers.

The code is constructed such that different versions of g and h are interchangable. In the applications that follow, g has logit and probit variants and a variant adapted to a normal distribution where the mean is zero and the log variance is equal to the linear predictor (the stochastic volatility model below). h is implemented in variants that allow for many random effects with different (or the same) variance, a completely general covariance matrix and a state space variant with bidiagonal precision matrix.

If the likelihood function factors into different integrals, each integral is evaluated separately. For estimating simulation uncertainty, the matrix Σ from equation (11) is then estimated as the sum of the estimated Σ (based on the empirical distribution of W_i from equation (9)) for each component of the likelihood function. The simulation errors reported in the tables are the square roots of the diagonal entries of the SML variance matrix Ω defined in equation (11).

A simple diagnostic term is also reported. Let $Z_{ik}(\theta)$ denote $W_i(\theta)/\hat{L}(\theta)$ from equation (9),

for likelihood component number k. Clearly, the first order condition for maximum implies

$$\sum_k \sum_i (Z_{ik}(\hat{\theta}) - Z_{\cdot k}(\hat{\theta})) = 0. \tag{34}$$

The diagnostic term reported from the implementation is the vector valued statistic

$$\tau(Z) = \frac{\max(|Z_{ik} - Z_{\cdot k}|)}{\sum_i \sum_k (|Z_{ik} - Z_{\cdot k}|)}, \tag{35}$$

where $Z_{\cdot k} = n^{-1} \sum_i Z_{ik}$. The components of τ give parameter specific values ranging from $1/n$ to $1/2$. A high value of τ implies that the corresponding component of the gradient is dominated by a single draw and is likely to be unreliable. Unfortunately, there is no clear-cut interpretation in the other direction.

The simulated log likelihood function is maximized using quasi-Newton algorithms. Estimation is typically performed by first computing the maximum Laplace approximated likelihood, and then using the results as starting values for SML estimation.

5 Applications

The results to be reported here demonstrate the applicability and performance of SML using tilted importance sampling. The main idea is to evaluate the estimator in terms of whether it approximates the exact maximum likelihood estimator. Since the exact maximum likelihood is impractical to find to the required precision by other methods, most of the applications are evaluated by estimating the models many times with different importance samples. Arguably, such an evaluation amounts at best to showing that the SML estimate is nicely distributed about something, which may or may not be the exact maximum likelihood estimate.

One solution is to generate huge data sets according to a known model as in the probit random effects applications reported below. Then the exact ML is known to be close to the parameter values used in generating the data. However, the high computational loads associated with such an approach are not ideal for testing out different versions of the importance samplers. Another solution, that does not directly solve the problem, but allows further investigation, is to use the generalized Laplace importance sampling scheme outlined above to try out different importance samplers. If different importance samplers lead to distributions about the same values, this will strengthen the hypothesis that the distributions are concentrated about the

exact maximum likelihood estimate. A third solution is to exploit the fact that well functioning importance samplers without misleading estimates of simulation variance should give simulation variance estimates and empirical variation in simulation estimates that are proportional to n^{-1}. A combination of these techniques have been applied to the different applications below.

5.1 Salamander mating

5.1.1 Introduction

This application demonstrates how SML using tilted Laplace importance sampling neatly and efficiently solves the infamous "Salamander mating" problem described in detail in McCullagh and Nelder (1989) in the sense that the maximum likelihood estimate can be found to a very high precision using a rather small importance sample. The method is compared to SML using simple importance sampling, and the results are compared to various results in the literature.

The model is a random effects logit model, with two individual specific types of random effects, one for each male salamander involved and one for each female salamander. The outcome is whether the salamanders are able to mate successfully in a laboratory context. The experimental design is crossed, and the likelihood function factors into 20-dimensional integrals. Two random effects variances and four parameters affecting the mean probability of success give a six-parameter model. In some applications in the literature, the likelihood is the product of two such integrals, while some papers use four extra integrals based on two replications of the experiment. We will repeatedly refer to these as Salamander model 1 and Salamander model 2.

We will study SML estimation of these models in some depth. The prototype output from running these models is presented in Table 1. The headers in the table refer to the parameters in the model, with the first four headers refering to the parameters affecting the mean probability. The models are estimated with importance sample sizes 100, with the generalized Laplace importance sampler with excess dispersion factor 1.3. MLAL refers to maximum Laplace approximated likelihood. It is clear from the table that the SML corrects the MLAL estimates only a little bit, but that the correction is substantial when seen in comparison to the reported simulation errors.

The reported simulation errors are very small, and dwarfed by the standard errors (based on the observed information). Note from equation (11) that this is in spite of, and not because of, the relatively large standard errors in these models. The diagnostics rows report τ from equation (35). With an independent sample size of 50, these should not be a cause of alarm.

Table 1: Reports from single runs of Salamander models

	Intercept	WSF	WSM	WSFWSM	σ_f	σ_m
Salamander model 1						
MLAL point estimates	1.335	-2.940	-0.422	3.181	1.255	0.269
SML point estimates	1.362	-3.003	-0.439	3.261	1.320	0.409
SML standard errors	0.665	0.977	0.661	1.086	0.440	0.640
SML simulation errors	0.009	0.016	0.003	0.017	0.014	0.022
SML diagnostics	0.155	0.089	0.129	0.082	0.068	0.073
Salamander model 2						
MLAL point estimates	1.008	-2.904	-0.702	3.588	1.084	1.020
SML point estimates	1.020	-2.961	-0.701	3.635	1.174	1.119
SML standard errors	0.402	0.586	0.473	0.646	0.272	0.265
SML simulation errors	0.004	0.011	0.004	0.013	0.010	0.009
SML diagnostics	0.028	0.036	0.031	0.038	0.040	0.023

(The model is implemented with the log random effects variance as parameter, which affects the diagnostics, but not the other values reported here.)

The models converged in 2 and 4 seconds, respectively. Thus there is little reason for not using a larger importance sample. For practical purposes, however, the values reported in Table 1 indicate little need for focusing on reducing simulation errors. The main reason for wanting to use larger importance samplers would be to investigate whether the results in Table 1 are misleading. The following Monte Carlo experiment indicates that they are not.

5.1.2 Simulation experiment

Salamander model 1 was estimated by SML 1000 times each for a variety of generalised Laplace importance samplers. Importance sample size 100 was used, as above. Table 2 shows the mean parameter estimates and standard deviations for SML point estimates based on the standard Laplace importance sampler, the generalized Laplace importance sampler with excess dispersion factors 1.1 and 1.3, and the importance sampler based on excess dispersion factor 1.3, but without use of the control variate technique, from equation (23).

All of the estimators give approximately the same mean estimates and small standard deviations about the mean. The results testify to the effectiveness of SML using tilted importance sampling in solving such a problem. The full experiment reported here runs in about one and a half hours.

The estimators based on generalized Laplace importance sampling with excess dispersion factors above 1 improve slightly on the direct Laplace importance sampler in terms of variability.

Table 2: Mean and standard deviation in SML parameters for Salamander model 1, based on 1000 estimation runs

Method	Statistic	Intercept	WSF	WSM	WSFWSM	σ_f	σ_m
LIS	mean	1.368	-3.012	-0.441	3.261	1.317	0.427
	st. dev.	0.010	0.023	0.004	0.025	0.023	0.028
GLIS, r=1.1	mean	1.369	-3.012	-0.441	3.262	1.317	0.428
	st. dev.	0.007	0.016	0.003	0.018	0.016	0.018
GLIS, r=1.3	mean	1.368	-3.012	-0.441	3.262	1.316	0.427
	st. dev.	0.008	0.017	0.004	0.019	0.014	0.023
DLIS, r=1.3	mean	1.371	-3.017	-0.442	3.267	1.321	0.436
	st. dev.	0.011	0.023	0.005	0.024	0.019	0.031

Table 3: Fractiles of distribution of simulation error estimate

Method	Fractile	Intercept	WSF	WSM	WSFWSM	σ_f	σ_m
LIS	min	0.004	0.008	0.002	0.009	0.005	0.013
	0.05	0.005	0.011	0.002	0.012	0.008	0.016
	0.95	0.018	0.042	0.007	0.045	0.041	0.050
	max	0.163	0.311	0.036	0.336	0.338	0.246
GLIS, r=1.3	min	0.004	0.009	0.002	0.010	0.007	0.015
	0.05	0.006	0.013	0.003	0.014	0.009	0.019
	0.95	0.012	0.026	0.006	0.028	0.025	0.032
	max	0.023	0.058	0.009	0.065	0.071	0.045

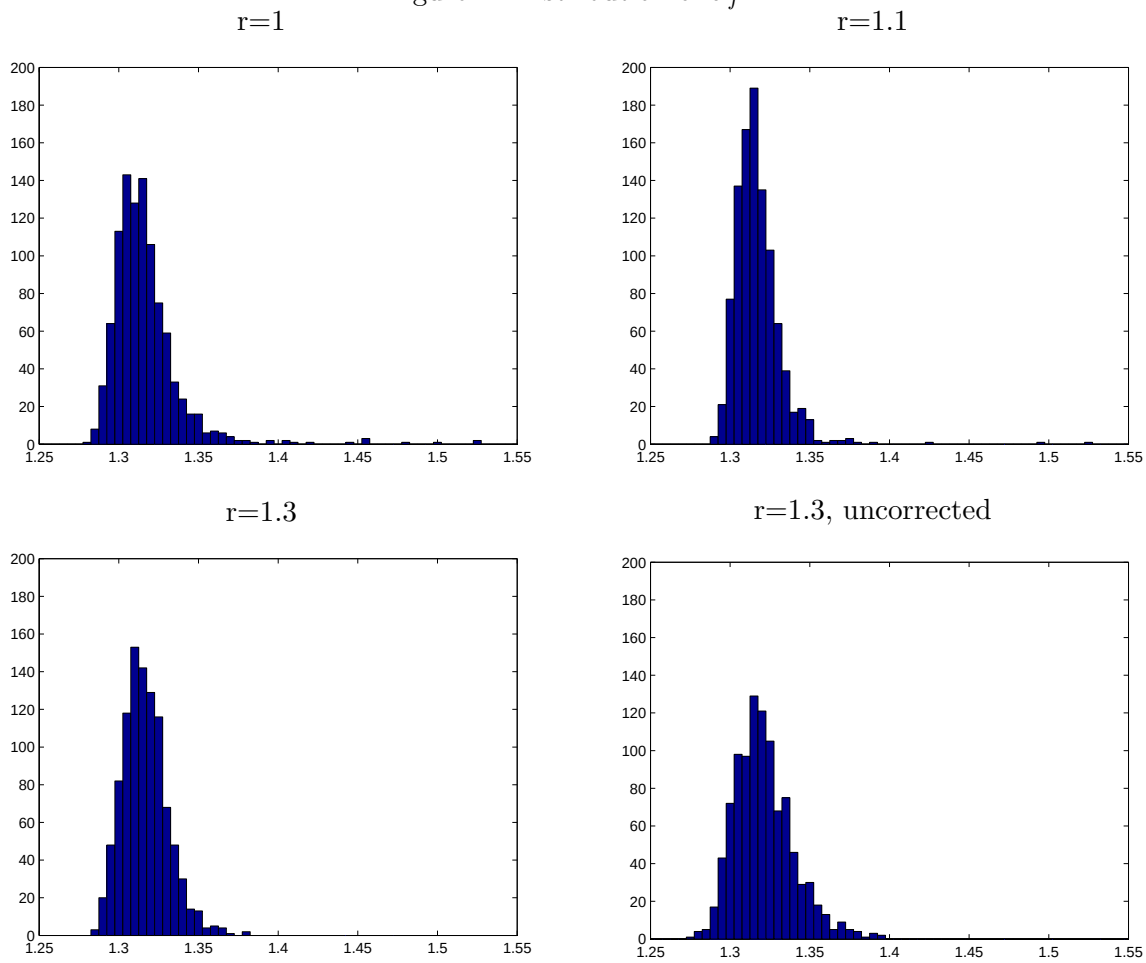
Further, histograms of the point estimates of σ_f and σ_m are provided in Figures 1 and 2. It is clear that the estimators with smaller excess dispersion factor give rise to distributions with (somewhat) heavy right hand tails. Higher excess dispersion factors effectively pull in this tail, giving SML estimators with nice bell-shaped simulation error distributions.

The second aim of this small simulation experiment is to evaluate the simulation error estimator. Table 3 reports quantiles of the estimates of simulation error (based on the same experiment) from the first and third estimators reported in Table 2. The numbers should be compared with the standard deviations in that table. The simulation error estimates are somewhat less dispersed for the generalized Laplace version of the estimator. The simulation error estimates seem quite reasonable. However, with such small importance samples, it is prudent to allow for some uncertainty also for the estimators of simulation errors.

5.1.3 Comparisons with results from the literature

Tables 4 and 5 compare parameter estimates based on tilted generalized Laplace importance sampling with other estimates from the literature. Since the precise maximum likelihood may

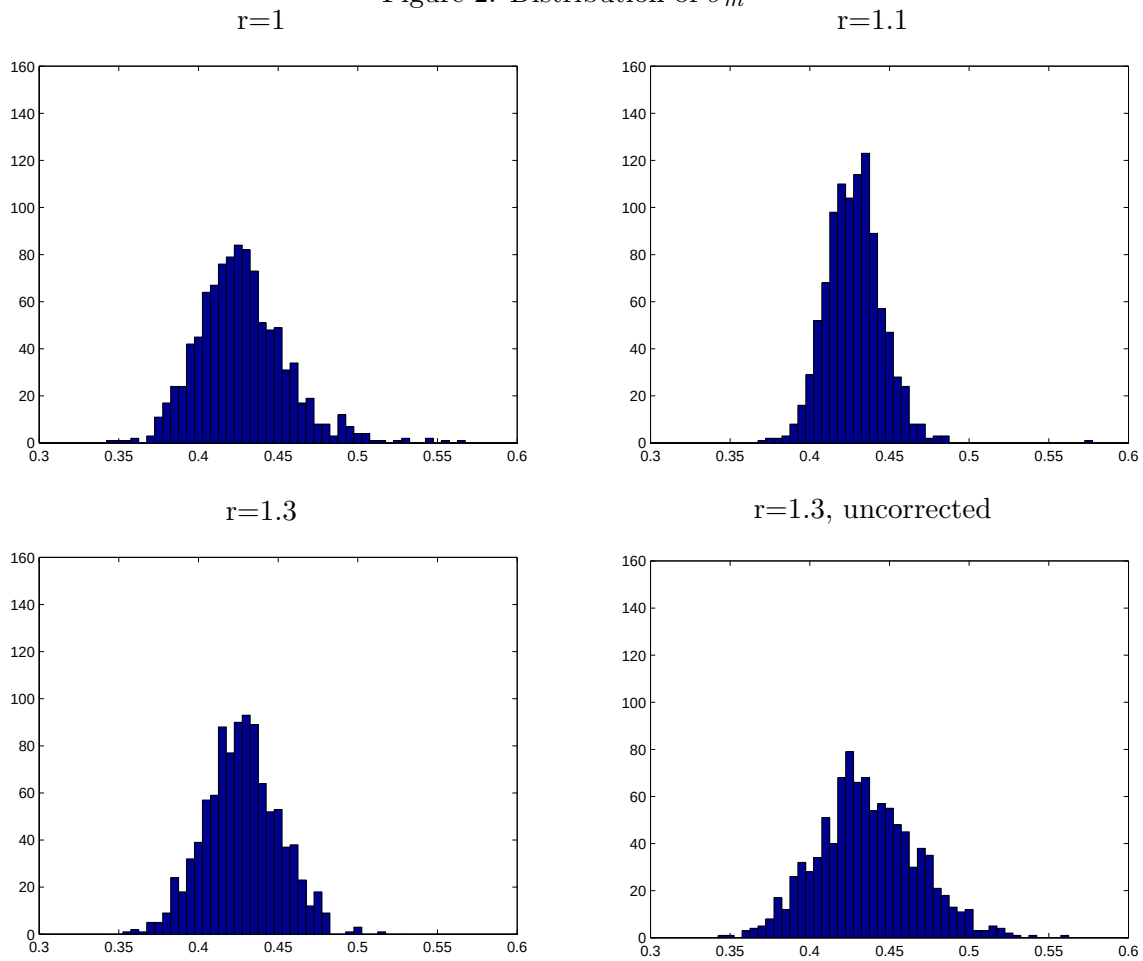
Figure 1: Distribution of σ_f



be of some interest, the importance sample size was increased from 100 to 40000. This should reduce the size of the simulation errors by a factor of 20. It does, and note that the SML estimate is practically equal to the mean of the SML estimates based on small importance samples reported in Table 2.

Table 4 reports SML estimates with standard errors based on observed information from Salamander model I and compares these with other estimates from the literature. The two first rows are from Lin and Breslow (1996) and report maximum penalized quasilikelihood and bias corrected maximum penalized quasilikelihood estimators. The third row is a posterior median estimate from Karim and Zeger (1992). The next three rows are standard and higher order maximum Laplace approximated estimators from Shun (1997). SML (Kuk) refers to the final SML using simple Laplace importance sampling from Kuk (1999a), with importance sample size 10000. SML (Millar) refers to Millar (2004), which is an MCMC variant of SML using simple importance sampling, but with importance sample size 100000.

Figure 2: Distribution of σ_m



There are two points to note here. First, the results are different from the results reported in Kuk (1999a) and to some less extent Millar (2004), indicating that ML based on simple importance sampling may be slightly misleading even when based on very large importance samples. Secondly, and as a digression, the results obtained here are, to the specified accuracy, almost equal to the estimates based on the standard higher order Laplace approximation reported in Shun (1997), but different from the estimates based on the Laplace approximation with exponentiated correction term. This is interesting because one of the main points in Shun and McCullagh (1995) and Shun (1997) is that the standard higher order Laplace approximation is invalid in the Salamander mating case (where the dimension of the integral is increasing in the sample size), and the exponentiated higher order Laplace approximation is developed to "fix" this problem. Asymptotic considerations do not always provide a good guide to numerical properties of small sample estimators.

Table 5 compares a similar exercise for Salamander model 2 to results from the literature.

Table 4: Comparisons with literature, Salamander model 1

Method	Intercept	WSF	WSM	WSFWSM	σ_f^2	σ_m^2
PQL					1.41	0.09
CPQL					1.71	0.40
Gibbs P. med.	1.48	-3.25	-0.50	3.62	1.53	0.37
MLAL	1.34	-2.94	-0.42	3.18	1.59	0.07
H.O. Laplace	1.37	-3.02	-0.44	3.27	1.72	0.18
E.H.O. Laplace	1.39	-3.06	-0.45	3.31	1.80	0.25
SML (Kuk)	1.39	-3.05	-0.45	3.29	1.72	0.25
SML (Millar)	1.375	-3.022	-0.444	3.274	1.749	0.198
SML	1.3685	-3.0121	-0.4411	3.2620	1.7333	0.1840
SML standard errors	0.68	1.01	0.69	1.08	1.14	0.54
SML simulation errors	0.0004	0.0009	0.0002	0.0010	0.0020	0.0011

Table 5: Comparisons with literature, Salamander model 2

Method	Intercept	WSF	WSM	WSFWSM	σ_f^2	σ_m^2
PQL					0.72	0.63
CPQL					0.99	0.91
Gibbs P. med.	1.03	-3.01	-0.69	3.74	1.50	1.36
MLAL	1.01	-2.91	3.59	-0.70	1.17	1.04
MCEM	1.03	-2.98	- 0.71	3.65	1.40	1.25
SML (Skaug)	1.02	-2.96	-0.70	3.63	1.38	1.23
SML	1.0182	-2.9593	-0.6970	3.6344	1.3843	1.2395
SML standard errors	0.41	0.58	0.48	0.64	0.63	0.58
SML simulation errors	0.0002	0.0006	0.0002	0.0006	0.0012	0.0011

The first four rows correspond to the rows in Table 4. Row five provide the results from Booth and Hobert (1999). They report that their MCEM algorithm used 80 minutes of computation time on some workstation. While computer power has increased during the last years, the solution using tilted importance sampling is found with comparable precision in a matter of (maybe 4) seconds. The same point has been stressed by Skaug (2002), who uses a method similar to the tilted (non-generalized) Laplace importance samplers used here and is reported in row number 6.

5.1.4 Tilted versus simple importance sampling

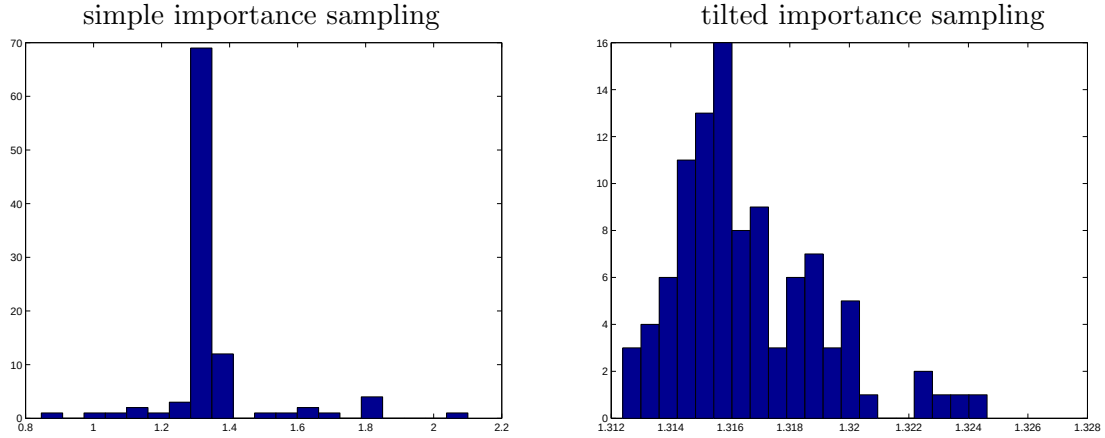
Several studies in the literature applied SML using simple importance sampling to the salamander mating problem, as discussed above. Here, the results from a small Monte Carlo experiment contrasting SML with tilted and simple importance sampling are reported.

The model was estimated 100 times with tilted and simple importance sampling, using the

same importance samples and (standard) Laplace importance sampling. First, the models were estimated using tilted importance sampling, then the SML estimates were used as parameter θ_0 in the simple importance sampling expression. Importance sample size 10000 was used.

Figure 3 illustrates the difference by comparing histograms for the simulated maximum likelihood estimates of σ_f for the simple and tilted importance sampler (note the different axes).

Figure 3: Distribution of simulated maximum likelihood estimates of σ_f



The simulated maximum likelihood based on simple importance sampling has a standard deviation that is in the order of 100 times as large as the standard deviation in the simulated maximum likelihoods based on tilted importance sampling.

Note that this result may be a bit misleading, as advocates of SML using simple importance sampling would argue for running the algorithm over again until convergence, usually in the minimalistic sense of finding a sequence of two estimates that are rather close. Thus, it is unlikely that the outliers in Figure 3 would ever be presented as the final result of such an exercise. Still, if we base the computation of standard deviation on only the central 10 percent of the empirical distributions in Figure 3, the standard deviation is still in the order of 10 times as large for the SML using simple importance sampling. Thus, SML using tilted importance sampling in this case requires sample sizes of 1 percent of the sample size of SML using simple importance sampling, to give comparable precision, even when the outlier problem of simple importance sampling is ignored. This finding is consistent with the fact that the results refered to as SML (Kuk) in Table 4, based on importance sample size 10000 seem to be no closer to the exact ML estimates than the values reported in Table 1, with importance sample size 100.

Table 6: Quantiles of posterior distribution

Method	Quantile	Intercept	WSF	WSM	WSFWSM	σ_f^2	σ_m^2
Gibbs	0.05	0.35	-4.08	-1.57	2.65	0.65	0.55
	0.5	1.03	-3.01	-0.69	3.74	1.50	1.36
	0.95	1.78	-2.10	0.09	4.91	3.04	2.89
Accept-Reject	0.05	0.35	-4.14	-1.58	2.68	0.62	0.54
	0.5	1.02	-3.05	-0.70	3.73	1.51	1.36
	0.95	1.78	-2.08	0.10	4.96	3.11	2.90

5.1.5 Bayesian analysis of salamander mating

Here, the results from a small Bayesian analysis of Salamander model 2 are reported, to illustrate how the tilted importance sampling may be used in Bayesian analysis. The idea is that, by integrating out the random effects, the dimensionality of the required sampling of the posterior distribution is reduced considerably. Flat priors are used, so the posterior distribution is a normalized likelihood. The log random effects variance is used as parameter, to be consistent with Karim and Zeger (1992).

For drawing from the normalized likelihood, the simulated log likelihood function provides us with a substantial dimension reduction, from 126 to 6 parameters in the Bayesian sense of the term. We use accept-reject sampling based on a Gaussian proposal distribution with mean at the simulated maximum likelihood estimate and variance taken from the corresponding observed information, multiplied by a factor t^2 to ensure that the maximum accept probability is at the maximum likelihood estimate and equal to one. The results here are based on $t = 1.55$ and 50000 proposals gave us a sample of 3444 independent draws, $t = 1.55$ is sufficiently high for all the accept probabilities to be below 1. The main merit of this kind of sampling is that it generates an independent sample.

Such an exercise takes about 7 hours, and it is not clear whether this is a particularly efficient way to find Bayesian estimates for this model. (In particular, we could probably have refined the sampling procedure considerably.) The sample has quantiles that are very similar to the quantiles reported in Karim and Zeger (1992), as shown in Table 6, where the first three rows are taken from that paper and the last three rows are the results of the computations outlined here.

Table 7: Different 95 percent confidence intervals for parameters in Salamander model 2

Method	limit	Intercept	WSF	WSM	WSFWSM	σ_f	σ_m
Observed Information	lower	0.24	-4.10	-1.62	2.36	0.75	0.70
	upper	1.80	-1.82	0.22	4.91	1.85	1.78
Posterior quantiles	lower	0.35	-4.14	-1.58	2.68	0.79	0.73
	upper	1.78	-2.08	0.10	4.96	1.76	1.70
Likelihood ratio	lower	0.18	-4.41	-1.81	2.28	0.65	0.60
	upper	1.99	-1.85	0.38	5.18	1.85	1.79

5.1.6 Likelihood ratio based confidence intervals

For many practical purposes, the log likelihood function is not approximately quadratic about the mode, and neither confidence intervals based on observed information nor confidence intervals based on draws from the normalized likelihood will give valid confidence intervals. (The latter are still Bayesian credible intervals based on diffuse priors.) Monotone, nonlinear transformations of the parameters, such as whether we estimate the standard error, variance or log variance of random effects, will affect estimated confidence intervals. Confidence regions based on likelihood ratio statistics however, are invariant to monotone transformations, and may provide more robust estimates.

The confidence intervals in Table 7 are produced by three different methods. The first group of intervals is based on observed information (with log random effects variance as parameter), given by the parameter estimates plus/minus 1.96 times the standard errors (recomputed to give the random effects standard error as parameter.) The second group of intervals is based on the Bayesian simulation reported above (with a flat prior in all parameters, including log random effects variance). Thus, the numbers are reproduced from Table 6. The third group of intervals is based on likelihood ratios. Likelihood profiles were computed for each parameter numerically over a grid (of 70 points with intervals equal to 0.1 standard error, centred at the parameter estimate). The likelihood was linearly approximated between the grid points, and the confidence intervals are based on where the log likelihood crosses the threshold of 2.34 less than the log likelihood at the maximum.

The differences between the computed intervals were not alarmingly large, though the likelihood ratio based intervals are generally somewhat wider than the others.

5.2 Probit variance components

The salamander mating data have been a testing ground for estimation methods for generalized linear mixed models with a crossed design. The most common application of GLMMs, however, involves nested designs, where quadrature methods are often feasible for evaluation of integrals. The purpose of the examples in this section is to show that SML using tilted importance sampling can be a competetive method also relative to quadrature methods. This is important because, unlike quadrature methods, the approach advocated here does not heavily constrain the type of problem we are able to study.

The adaptive quadrature techniques used in Rabe-Hesketh, Skrondal, and Pickles (2005), RSP in the following, are based on iterative updating of a normal approximation of the empirical Bayes posterior of the random effects, which is used for determining an appropriate grid for evaluation of the integrand. A number of simulation experiments using probit variance components models are discussed in RSP. Some of these simulation studies are replicated here with SML using tilted importance sampling.

5.2.1 Experiment 1

The first model is a simple random effects probit model with just one random effect per observation. The following model was simulated

$$y_{ij}^* = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j + \epsilon_{ij}, \tag{36}$$

where $y_{ij} = 1$ if $y_{ij}^* > 0$, $y_{ij} = 0$ otherwise. u_j is here a Gaussian random effect with expectation 0 and variance σ^2 and ϵ_{ij} is a Gaussian residual term with variance 1.

In the simulation the fixed parameters were set to $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$ with different values for σ, and x_{1ij} and x_{2j} were drawn from independent discrete distributions with equal probability for values 0 and 1. In all simulations, j runs from 1 to 1000, so the likelihood consists of a sum of 1000 one-dimensional integrals. The cluster size varies in the different simulations. 50 replications were run, as in RSP.

Table 8 shows cluster sizes, intracluster correlation, the derived value of σ and estimates based on maximum Laplace approximated likelihood, SML using tilted generalised Laplace importance sampling with excess dispersion factor 1.3, in addition to maximum likelihood based on adaptive and ordinary quadrature reproduced from RSP. SML using tilted importance sam-

Table 8: Simulation results, Rabe-Hesketh et al., Exp. 1

n_j	ρ	σ	L.A.	GLIS	AQ	OQ
10	0.3	0.655	0.6431 (0.0287)	0.6492 (0.0291)	0.659 (0.025)	0.658 (0.025)
10	0.9	3.000	3.5875 (1.3971)	2.8716 (0.1192)	2.986 (0.133)	2.812 (0.106)
100	0.3	0.655	0.6490 (0.0177)	0.6500 (0.0177)	0.653 (0.018)	0.649 (0.017)
100	0.9	3.000	2.7584 (0.0702)	2.8859 (0.0799)	2.935 (0.090)	1.768 (0.062)
500	0.3	0.655	0.6550 (0.0154)	0.6552 (0.0154)	0.654 (0.019)	0.543 (0.023)
500	0.9	3.000	2.8430 (0.0777)	2.9305 (0.0865)	2.991 (0.081)	1.224 (0.056)

pling gave excellent results, although a miniscule bias seems to apply to the cases with $\rho = 0.9$ ($\sigma = 2.9$ implies $\rho > 0.89$). It is interesting to note that maximum likelihood based on the Laplace approximation breaks down for small cluster size and large intracluster correlation. This phenomenon is a result of local maxima in the Laplace approximated likelihood function. These local maxima disappear when the Laplace approximation is corrected using importance sampling. The estimated simulation errors are very small, estimated to about 0.02 for the second row of Table 8, and less for the other rows.

While SML using tilted importance sampling is not an obvious method of choice for estimation of models that only involve one-dimensional integrals, the simulations above show that it works fine, substantially better than non-adaptive Gaussian quadrature in this context, and practically as good as adaptive quadrature. For the importance sample sizes used here, the method is less computationally demanding than the variant of adaptive quadrature advocated in RSP.

5.2.2 Experiment 2

The second simulation study is based on a hierarchical (three level) random effects probit model with random effects at two levels. The following model was simulated

$$y^*_{ijk} = \beta_0 + u^{(2)}_{jk} + u^{(3)}_{k} + \epsilon_{ijk}, \tag{37}$$

where $y_{ij} = 1$ if $y^*_{ij} > 0$, $y_{ij} = 0$ otherwise. $u^{(2)}_{jk}$ and $u^{(3)}_{k}$ are Gaussian random effects with expectation 0 and variances σ^2_2 and σ^2_3 and ϵ_{ij} is a Gaussian residual term with variance 1.

In this study, the excess dispersion factor is set to 1.3 for integrals of dimension 11 and 1.0 for integrals of dimension 101. (High excess dispersion factors are problematic in high dimensional integrals.)

Table 9: Simulation results, Rabe-Hesketh et al., Exp. 2

n_1, n_2, n_3	param	true	L.A.	GLIS	AQ
10,10,100	σ_2	1.225	1.1776 (0.0123)	1.1931 (0.0130)	1.222 (0.015)
	σ_3	1.035	1.0141 (0.0785)	1.0124 (0.0772)	1.076 (0.117)
10,100,10	σ_2	1.225	1.2115 (0.0220)	1.2245 (0.0223)	1.237 (0.040)
	σ_3	1.035	1.0171 (0.0718)	1.0188 (0.0709)	1.102 (0.092)
10,10,100	σ_2	1.225	1.1543 (0.0149)	1.1508 (0.0153)	1.226 (0.021)
	σ_3	1.936	1.9487 (0.1345)	1.9426 (0.1375)	2.021 (0.166)
10,100,10	σ_2	1.225	1.1896 (0.0333)	1.2180 (0.0356)	1.235 (0.037)
	σ_3	1.936	1.9523 (0.1354)	1.9319 (0.1304)	1.965 (0.138)
10,10,10	σ_2	1.225	1.1407 (0.0418)	1.2171 (0.0508)	1.231 (0.051)
	σ_3	1.936	1.8695 (0.1599)	1.8946 (0.1624)	1.967 (0.192)

Two phenomena are illustrated in Table 9. First, for the rows that correspond to likelihoods with 11-dimensional integrals, for models number 2, 4 and 5, tilted importance sampling seems to work as good as or better than adaptive quadrature. For the rows that correspond to 101-dimensional intergrals, tilted importance sampling does seem to correct the maximum Laplace approximated likelihood somewhat, but not to the extent that the full bias is corrected.

An extra experiment not reported in the table was performed for the 101-dimensional integrals, with importance sample sizes 1000 and not 100, and the results were somewhat better. The parameter σ_2 was now corrected from 1.1797 for the Laplace approximated likelihood to 1.2082 for the tilted importance sampled likelihood for the model corresponding to row 1, and from 1.1525 to 1.1735 for the model corresponding to row 5 in Table 9.

The experiment thus shows that SML using tilted importance sampling works very well in the models with medium sized (11-dimensional) integrals and somewhat less satisfactorily in the model with really high dimensional integrals.

However, from a more practical point of view, the SML estimators perform very well in this experiment. The reason why the SML does not fully correct the bias of the maximum Laplace approximated likelihood estimator in the high dimensional case is quite likely that the distribution is heavy tailed, as in the boat race example reported below.

While emphasizing that SML using tilted importance sampling performs extremely well here, even in the case with 101-dimensional integrals, there is no reason in general not to use methods that take into account the conditional independence structure that allows for treating the integral as 2-dimensional rather than 101-dimensional. Extensions of tilted importance sampling as a correction term of Laplace approximations for nested integrals is an important future research direction.

Table 10: Simulation results, Rabe-Hesketh et al., Exp. 3

	β_1	β_2	β_3	β_4	β_5	β_6
true	0	0.5	-1	1	-0.5	0
mean(TLA)	-0.0095	0.5044	-0.9690	0.9643	-0.5222	0.0026
std(TLA)	0.0343	0.0315	0.0329	0.0325	0.0318	0.0325
mean(TGLIS)	-0.0096	0.5044	-0.9690	0.9643	-0.5221	0.0318
std(TGLIS)	0.0343	0.0315	0.0329	0.0325	0.0318	0.0325
	σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_5^2	σ_6^2
true	0.25	0.25	0.25	0.25	0.25	0.25
mean(TLA)	0.2528	0.2619	0.2458	0.2231	0.2205	0.2600
std(TLA)	0.0494	0.0440	0.0321	0.0349	0.0312	0.0368
mean(TGLIS)	0.2543	0.2635	0.2479	0.2252	0.2221	0.2612
std(TGLIS)	0.0495	0.0441	0.0322	0.0351	0.0313	0.0368
	σ_{12}	σ_{13}	σ_{14}	σ_{15}	σ_{16}	σ_{23}
true	0.12	-0.12	0.09	-0.09	0.12	-0.16
mean(TLA)	0.1260	-0.1024	0.0816	-0.0430	0.1101	-0.1805
std(TLA)	0.0310	0.0339	0.0258	0.0328	0.0278	0.0289
mean(TGLIS)	0.1258	-0.1029	0.0812	-0.0434	0.1098	-0.1813
std(TGLIS)	0.0311	0.0340	0.0258	0.0328	0.079	0.0290

5.2.3 Experiment 3

A third application from RSP investigates a random coefficients probit model, where the latent continuous variable is defined by

$$y_{ij}^* = X_{ij}(\beta + u_j) + \epsilon_{ij}. \tag{38}$$

Here, u_j represents the variation in coefficients over clusters. β and u_j may be multidimensional of course. In this application, they are 6-dimensional and u_j is specified as multinormal with a general covariance matrix. Integrals over 6 dimensions are at the boundary of what is achievable using quadrature. It is straightforward with SML using tilted importance sampling. Therefore, we can, unlike RSP, easily do the estimation many times.

Table 10 presents results from 50 simulation and estimation runs based on this model, for most of the parameters in the model. Table 10 shows that maximum likelihood based on the Laplace approximation and SML using tilted Laplace importance sampling give almost exactly the same results. Further, the estimates are generally rather close to the true parameters for most parameters. It is likely that this is because the log penalized conditional likelihood of this model is very close to quadratic.

The simple diagnostics (τ) show that typically no single draw amounts to more than about

one percent of the correction. The simulation error estimates show simulation errors that at best may affect the last digits in Table 10.

5.3 State space applications

Durbin and Koopman (1997) and Durbin and Koopman (2000) developed an approach to SML that is (in some variants) equivalent to SML using tilted (standard) Laplace importance sampling as developed in a more general context here, in the following referred to as the DK approach.

There seems to be three main differences between the implementation outlined here and the DK approach. First, the DK approach uses an algorithm for finding the posterior mode of the random effects that is equivalent to Newton's method and not a trust region version of Newton's method as implemented here. Thus, the method is fragile with respect to local non-concavities in the function to be maximized (with respect to the random effects), and various fixes are applied to correct for this problem. Secondly, the DK approach does not provide the gradient of the log likelihood function. Thirdly, the approach outlined here can not easily handle distributions of random effects with singular covariance matrices, which often arise in state space applications.

We replicate two of the applications from Durbin and Koopman (2000), using the computational approach outlined here. The focus is on doing the estimations many times over to investigate the empirical distributions of the SML estimate and to see if there is scope for improvement using generalized Laplace importance sampling.

5.3.1 Stochastic volatility

The first application is a stochastic volatility model of exchange rate data, where the daily changes in exchange rates are modeled as Gaussian variables with zero mean and a time-dependent variance, where the log variance is modeled as a Gaussian AR(1)-process. The data originate from Harvey, Ruiz, and Shephard (1994). Note that Jungbacker and Koopman (2006) provide a similar exercise based on more advanced stochastic volatility models using the DK approach.

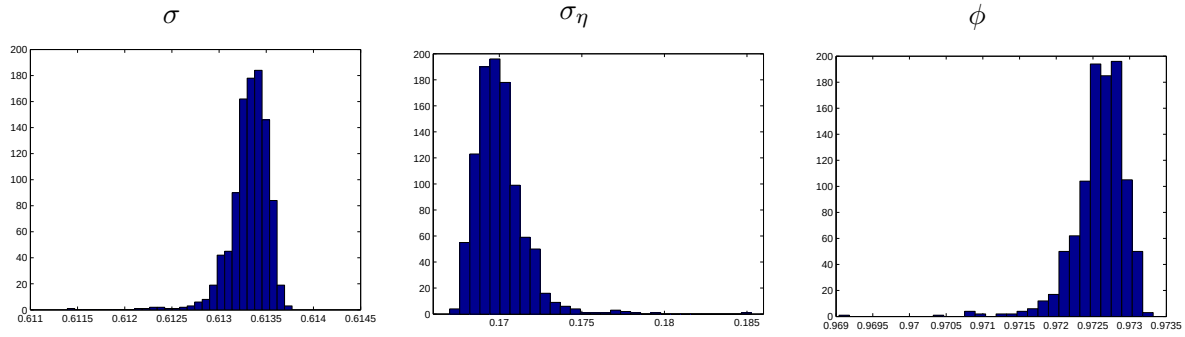
The time specific log variances are treated as the random effects in this implementation. These are Gaussian with mean zero and bidiagonal precision matrix, due to the AR(1)-structure. The log conditional likelihood, the density of the data conditional on the log variance gives a simple expression (which is quite close to quadratic in the log variance).

There are three parameters, the equilibrium level of the variance, σ, the AR-parameter ϕ

and the variance of the innovations in the AR(1)-process, σ_η^2. The likelihood consists of a 942-dimensional integral over the random volatility effects.

Figure 4 reports histograms of SML estimates based on 1000 different importance samples, based on importance sample size 1000 and excess dispersion factor 1. Each estimation run takes about one and a half minutes. There was little room for improvement based on generalized Laplace importance samplers, though the results already seem impressive enough. The estimates are somewhat heavy-tailed, but the outliers are not prominent enough to affect the mean estimates much. Further, the standard errors of the estimates (not reported here) are sufficiently large to completely dominate the simulation errors implicit in Figure 4.

Figure 4: Distribution of parameter estimates



The parameter estimates based on the Laplace approximation are 0.6136 for σ, 0.1683 for σ_η and 0.9728 for ϕ, so SML based on tilted importance sampling hardly changes the results from maximum likelihood based on the Laplace approximation.

It seems reasonable to conjecture that this application works this well because the model is extremely close to linear-Gaussian. With other parameter values, where the (posterior) variance of the random stochatic volatility terms is larger, so that the likelihood of the data can not be as well approximated by a quadratic function of the log variance, one should expect more difficulties with such high dimensional integrals.

5.3.2 Boat races

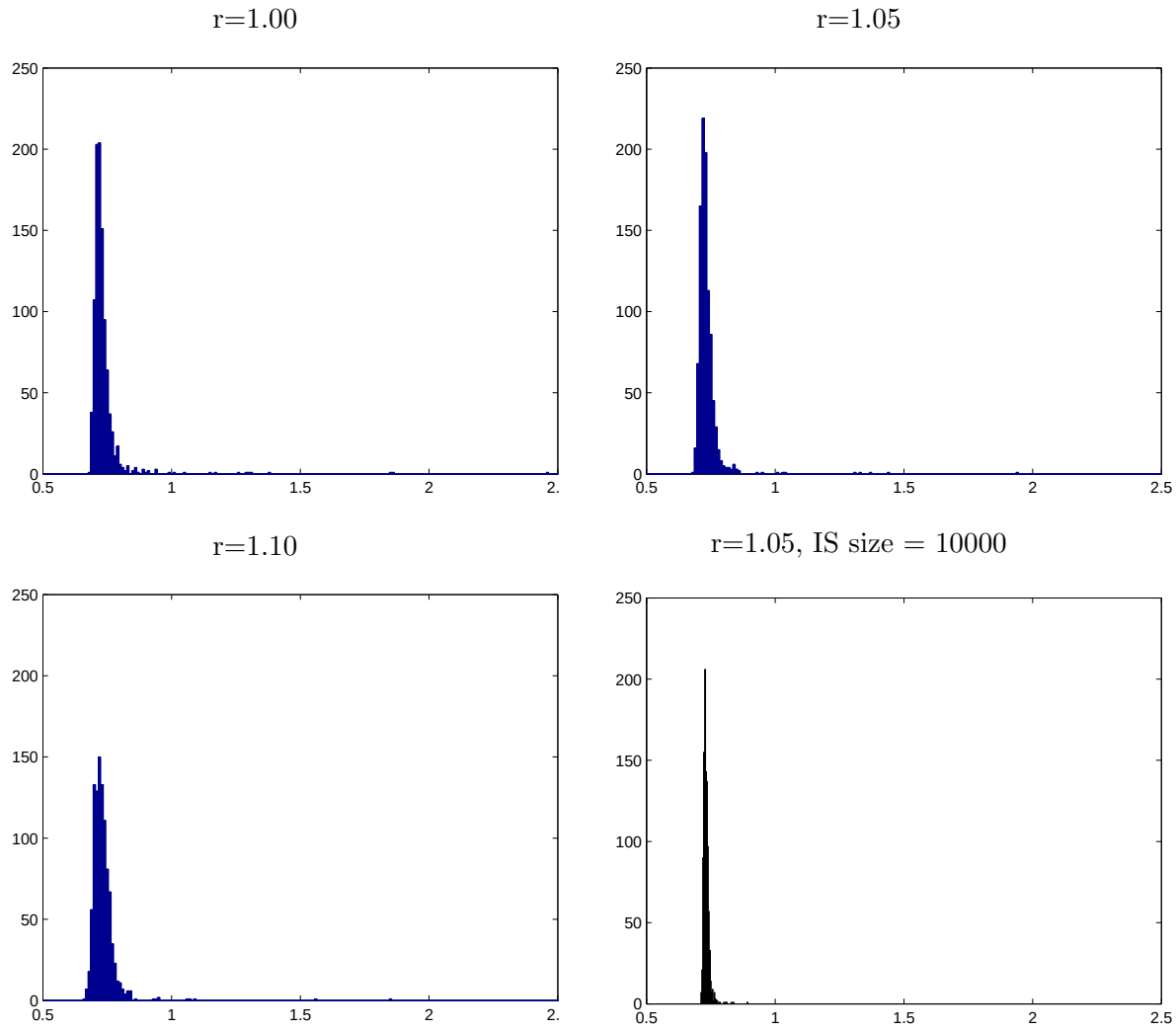
Another application in Durbin and Koopman (2000) is a model where the outcome is the winner in the famous boat race between Cambridge and Oxford Universities. The data are a dichotomous time series, which runs over 172 years, where the outcome is modeled as a logit model. The log odds of the outcome probability is modeled as a random walk with Gaussian disturbances. Only the standard deviation associated with this random walk is estimated as a parameter in

the model.

Unlike the stochastic volatility model discussed above, the difference between the maximum likelihood estimate based on the Laplace approximation and the SML estimates is substantial. Figure 5 reports histograms of the estimator under excess dispersion factors 1, 1.05 and 1.1, based on 1000 replications with importance sample size 1000.

Figure 5 shows that generalized Laplace importance sampling improves on the standard variant. Outliers are not as pronounced for r=1.05 and r=1.1, and the simulation errors (standard deviations of estimates) are about halved. The distribution becomes more dispersed again as r is set higher than 1.1.

Figure 5: Distribution of σ



A reasonable conjecture is that the SML estimates are pretty close to the exact maximum likelihood estimate, except for the outliers in Figure 5. Whether a single estimate is such an

outlier is easily found by the diagnostic measure discussed above. All the outliers are completely dominated by a single draw, giving $\tau \approx 0.5$.

An interesting check is to see whether the distribution of SML estimates is concentrated by considering larger importance samples. The lower right figure in the panel shows results corresponding to the upper right, with an importance sample ten times as large. There is little doubt that increasing the sample size reduces the variability of the estimates. The standard error is in fact reduced by a factor of five, more than the square root of ten implied by asymptotic theory, apparently because the extreme outliers become less of a problem with a larger importance sample. The mean estimate (not reported in a table) was 0.7321 for the large sample, and 0.7361 for the small sample, with 0.7377 for r=1.0 and 0.7334 for r=1.1. The point estimate reported in Durbin and Koopman (2000) is 0.7218 (implied by variance reported as 0.521).

6 Conclusion

This paper has introduced the distinction between simple and tilted importance sampling as methods for simulating likelihood functions containing integrals that depend on parameters of the likelihood function. Further, it has been demonstrated that tilted importance samplers are necessary to give near optimal importance samplers for the score function, which is what essentially needs to be simulated when using SML. Further, a new class of importance samplers has been developed, the generalized Laplace importance samplers, that is easy to combine with tilted importance sampling.

Although the exact maximum likelihood estimate in models with likelihood functions containing difficult integrals is not the easiest of targets to compare with, the methods seem to work extraordinarily well in the applications and Monte Carlo experiments considered here. The methods works extremely well for the classic salamander mating case, providing us with the tools to do practically exact simulated maximum likelihood in addition to simulated profile likelihoods and accept-reject sampling from the normalized simulated likelihood, with small computational costs.

The examples on nested random effects probit models show that the method works well in difficult low dimensional integrals, even in cases where we are not able to exploit the conditional independence structure of the integrals, forcing us to treat the integrals as higher dimensional than what is necessary with other methods. As has also previously been demonstrated, these

methods work very well for some non-Gaussian state space models. There is room for improvement in the state space applications using the generalized Laplace importance sampler developed here. However, the main contribution of this paper will hopefully be to increase the impact of SML using tilted importance sampling outside the state space domain, rather than bringing incremental improvement within the state space domain.

The results in this paper suggest that SML using tilted importance sampling should be the classical method of choice for models with likelihoods containing (relatively) high dimensional integrals, due to extremely good performance relative to computational costs, as seen in the Salamander mating model. There are certainly models for which SML using tilted importance sampling is not possible to apply efficiently. (Models with discrete random effects are an obvious class.) However, the class of models that are in a rough sense similar to the applications reported here and where current practice is to use methods such as SML using simple importance sampling, maximum penalized quasilikelihood (see McCulloch and Searle, 2001) or MCEM, that are far inferior in terms of computational efficiency and/or bias compared to SML using tilted importance sampling, is large. The small computational costs associated with SML using tilted importance sampling also open up for using numerical techniques that go beyond finding the maximum likelihood estimate and some associated confidence interval based on asymptotic theory, such as bootstrapping and bias correction of the SML estimate.

References

Booth, J. G., and J. P. Hobert (1999): "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(1), 265–285.

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

——— (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.

Harvey, A. C., E. Ruiz, and N. Shephard (1994): "Multivariate stochastic variance models," *Review of Economic Studies*, 61, 247–264.

Jank, W., and J. Booth (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.

Jungbacker, B., and S. J. Koopman (2006): "Monte Carlo likelihood estimation for three multivariate stochastic volatility models," *Econometric Reviews*, 25(2-3), 385–408.

Karim, M., and S. Zeger (1992): "Generalized Linear Models with Random Effects; Salamander Mating Revisited," *Biometrics*, 48(2), 631–644.

Koopman, S. J., N. Shephard, and J. A. Doornik (1999): "Statistical algorithms for models in state space using SsfPack 2.2," *The Econometrics Journal*, 2(1), 107–160.

Kuk, A. Y. C. (1999a): "Laplace importance sampling for generalized linear mixed models," *Journal of Statistical Computation and Simulation*, 63(2), 143–158.

Kuk, A. Y. C. (1999b): "The use of approximating models in Monte Carlo maximum likelihood estimation," *Statistics and Probability Letters*, 45(4), 325–333.

Lin, X., and N. E. Breslow (1996): "Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion," *Journal of the American Statistical Association*, 91, 1007–1016.

McCullagh, P., and J. A. Nelder (1989): *Generalized linear models, 2nd ed.* Chapman and Hall.

McCulloch, C. E., and S. R. Searle (2001): *Generalized, Linear and Mixed Models*. Wiley.

Millar, R. B. (2004): "Simulated maximum likelihood applied to non-gaussian and nonlinear mixed effects and state-space models," *Australian and New Zealand Journal of Statistics*, 46(4), 543–554.

Nadarajah, S., and S. Kotz (2005): "Sampling distributions associated with the multivariate t distribution," *Statistica Neerlandica*, 59(2), 214–234.

Pinheiro, J. C., and D. M. Bates (1995): "Approximations to the log-likelihood function in the nonlinear mixed-effects model," *Journal of Computational and Graphical Statistics*, 4(1), 12–35.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles (2005): "Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects," *Journal of Econometrics*, 128(2), 301–323.

Richard, J. F., and W. Zhang (2007): "Efficient high-dimensional importance sampling," *Journal of Econometrics*, 141(2), 1385–1411.

Ripley, B. D. (1987): *Stochastic simulation*. John Wiley & Sons, Inc. New York, NY, USA.

Robert, C., and G. Casella (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.

Shun, Z., and P. McCullagh (1995): "Laplace Approximation of High Dimensional Integrals," *Journal of the Royal Statistical Society. Series B*, 57(4), 749–760.

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

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