Efficient High-Dimensional Importance Sampling

Jean-Francois Richard
University of Pittsburgh
and
Wei Zhang
National University of Singapore

March 31, 2005

Abstract

The paper describes a simple, generic and yet highly accurate Efficient Importance Sampling (EIS) Monte Carlo (MC) procedure for the evaluation of high-dimensional numerical integrals. EIS is based upon a sequence of auxiliary weighted regressions which actually are linear under appropriate conditions. It can be used to evaluate likelihood functions and byproducts thereof, such as ML estimators, for models which depend upon unobservable variables. A dynamic stochastic volatility model and a logit panel data model with unobserved heterogeneity (random effects) in both dimensions are used to provide illustrations of EIS high numerical accuracy, even under small number of MC draws. MC simulations are used to characterize the finite sample numerical and statistical properties of EIS-based ML estimators.

Keywords: Monte Carlo, Importance Sampling, Marginalized Likelihood, Stochastic Volatility, Random Effects.

1 Introduction

Monte Carlo (hereafter MC) simulation techniques provide a powerful tool to numerically evaluate expectations of functions of random variables for which no analytical expressions are available. See e.g., Fishman (1996) for an in-depth analysis of MC concepts and algorithms. One particular area where MC methods play a critical role is that of models which incorporate large numbers of unobserved random variables. Examples to be discussed further below are stochastic volatility models in finance and large panels with unobserved heterogeneity, where dimensions of integration can be in the thousands. It has long been recognized that the feasibility of evaluating such high-dimensional integrals by MC simulation critically depends on the selection of efficient MC samplers. Importance Sampling which consists of replacing an inefficient initial sampler by a more efficient auxiliary sampler is conceptually well suited for that purpose. Its weakness lies in the current lack of generic algorithms to construct Efficient Importance Samplers (hereafter EIS) at a sufficiently broad level of generality.

The object of the present paper is to propose a new algorithm to construct EIS, which is generic and particularly well adapted to (very) high-dimensional MC integration. In particular, under appropriate simplifying conditions it amounts to a simple recursive sequence of auxiliary least squares optimization problems which, as we shall illustrate below, can produce enormous efficiency gains at low cost of computation.

The paper is organized as follows: Section 2 introduces Importance Sampling. The generic principle of EIS is introduced in Section 3. Its high-dimensional implementation is presented in Section 4. Numerical and statistical properties of EIS based estimates are analyzed in Section 5. Pilot applications of EIS to stochastic volatility and panel models are presented in Section 6; Section 7 concludes.

2 Importance Sampling

2.1 Principle

Importance Sampling (hereafter IS) has been around for quite some time. Early descriptions of the method can be found e.g. in Kahn and Marshall (1953), Trotter and Tukey (1956) or Hammersley and Handscomb (1964,

Section 5.4). See also Fishman (1996, Section 4.1) for a more recent in-depth presentation of IS. In this section we briefly introduce the concept of IS, establishing notation for our subsequent analysis.

Assume one has to evaluate a functional integral of the form

$$G(\delta) = \int_X g(x, \delta) \cdot p(x|\delta) \cdot dx \tag{1}$$

where g denotes a function which is integrable in x with respect to (w.r.t.) a density $p(x|\delta)$ with support X. Let θ, y and λ denote parameters, observations and latent variables, respectively. Applications requiring the evaluation of integrals of the form given in (1) are: (i) The Bayesian evaluation of posterior moments for which $\delta = y$ and $x = (\theta, \lambda)$; (ii) The (classical) evaluation of marginalized likelihood functions for which $\delta = (\theta, y)$ and $x = \lambda$; (iii) The computation of Generalized Method of Moment estimators for which $\delta = \theta$ and $x = (y, \lambda)$. Examples of marginalized likelihood evaluations are presented in Section 6 below.

The factorization under the integral sign in equation (1) is typically associated with the initial formulation of one's statistical model and we shall accordingly refer to p as to an initial sampler. An initial MC estimate of $G(\delta)$ is given by

$$\overline{G}_S(\delta) = \frac{1}{S} \sum_{i=1}^{S} g(\tilde{x}_i, \delta) \tag{2}$$

where the $\tilde{x}_i$'s are independently identically distributed (i.i.d.) draws from p. If, however, the sampling variance of g w.r.t. x is (very) large, accurate estimation of G may require prohibitively large numbers of draws. Dramatic illustrations of such inefficiency will be provided in Section 6 below.

Importance Sampling consists of replacing the initial sampler $p(x|\delta)$ by an auxiliary parametric importance sampler $m(x|a)$, rewriting equation (1) as

$$G(\delta) = \int_X g(x, \delta) \cdot w(x; \delta, a) \cdot m(x|a) dx \tag{3}$$

with

$$w(x; \delta, a) = \frac{p(x|\delta)}{m(x|a)} \tag{4}$$

The corresponding IS-MC estimate of $G(\delta)$ is given by

$$\overline{G}_{S;m}(\delta, a) = \frac{1}{S} \sum_{i=1}^{S} w(\tilde{x}_i; \delta, a) \cdot g(\tilde{x}_i; \delta) \tag{5}$$

where the $\tilde{x}_i$'s now denote i.i.d. draws from $m(x|a)$. The MC sampling variance of $\overline{G}_{s;m}$ is given by

$$V[\overline{G}_{S;m}(\delta, a)] = \frac{1}{S} G(\delta) \cdot V(a; \delta) \tag{6}$$

with

$$V(a; \delta) = \frac{1}{G(\delta)} \int_X \left[\frac{\varphi(x; \delta)}{m(x|a)} - G(\delta) \right]^2 m(x|a) dx \tag{7}$$

$$\varphi(x; \delta) = g(x, \delta) \cdot p(x|\delta) \tag{8}$$

Clearly, one's objective in selecting $m(x|a)$ - or $a \in A$ once a class of samplers $M = \{m(x|a); a \in A\}$ has been preselected - should be that of minimizing $V(a; \delta)$ w.r.t. a for any given δ. Conditions for the finiteness of $V(a; \delta)$ are discussed e.g. in Geweke (1996) or Stern (1997). Nevertheless, the critical convergence issue of whether a Central Limit theorem applies to $\sqrt{s}(\overline{G}_{s;m} - G)$ remains difficult to verify in complicated and/or high-dimensional IS applications and will be discussed further below.

Note that the term Importance Sampling emphasizes the fact that $m(x|a)$ is designed to sample mostly from the important part of $S(\delta)$ i.e.,. from those parts of $S(\delta)$ which contribute most to the value of the integral. Such terminology can be somewhat misleading since it is now well recognized that large or even infinite values of $V(a; \delta)$ typically originate from the far tails of $m(x|a)$ which is precisely why IS pathologies can be empirically hard to detect.

Clearly, the two critical issues to be addressed in IS applications are (1) the selection of an appropriate class M of auxiliary samplers; and (2) the selection of an efficient sampler within M i.e. one for which $V(a; \delta)$ is as small as possible. The selection of M is bound to remain problem-specific, though our subsequent discussion will provide important guidelines for such selection. The EIS principle proposed in this paper specifically addresses the issue of selecting (near) optimal as in A.

2.2 Short literature review

Since the construction of importance samplers clearly constitutes the Achilles heel of IS, it has received much attention over the years. Let us review here some innovative proposals in this respect. While Tierney and Kadane (1986) do not specifically discuss IS, the concept of Laplace approximation they

rely upon to evaluate posterior moments can also be used to construct importance samplers. It essentially consists of locally approximating $\phi(x, \delta)$ around its modal value. Geweke (1989) explicitly discusses minimization of $V(a; \delta)$ within specific classes of fat-tail densities, typically multivariate student -t densities and skewed generalizations thereof, labeled split-t densities. Evans (1991) relies upon adaptative methods whereby earlier draws of x are used to identify large values of the weight function ω in equation (4) and to revise accordingly the sampler. Owens and Zhou (2000) discuss various improvements of the IS technique which are well suited for low-dimensional applications.

Durbin and Koopmans (1997) apply IS to evaluate the likelihood function of non-Gaussian state space models. Essentially, by constructing a Gaussian approximation to their model, they are able to express the ratio between the two likelihoods as an integral which is functionally similar to equation (3). The interest of their method is twofold. First, it shows that the selection of an importance sampler can be approached via the construction of an operational approximation to a complex model and, in this respect, offers conceptual similarities with the EIS principle proposed below; Second, it is applicable in significantly higher dimensions than the alternative methods discussed above.

Another sophisticated implementation of IS is found in Madras and Piccione (1989) where the authors use as IS the (implicit) equilibrium distribution associated with a Monte Carlo Markov Chain (MCMC) simulator. The main advantage of their method lies in the flexibility of MCMC simulations but the convergence properties of their procedure are typically difficult to assess.

An important message which emerges from this brief literature overview is that importance samplers have to be carefully tailored to the problem under consideration. This has proved to be a significant obstacle to routine applications of IS. Another problem lies in the fact that, except for the specific problem addressed by Durbin and Koopmars (1997), none of the existing IS methods appear to be applicable to (very) high-dimensional applications of the form of those considered in Section 6 below.

3 Efficient Importance Sampling

3.1 Principle

As indicated by Equation (7), if there existed $a_0(\delta) \in A$ such that

$$m(x|a_0(\delta)) = \frac{\varphi(x;\delta)}{G(\delta)} \tag{9}$$

then $V(a_0(\delta);\delta) = 0$. This reflects the fact that $\varphi(x,\delta)$ can be interpreted as a kernel of the actual posterior density of x given δ, which obviously would be the ideal sampler for x given δ. In general, however, $\varphi(x;\delta)$ is not amenable to MC simulations which is precisely why it needs to be approximated by an importance sampler. Clearly, an efficient importance sampler is one which is closest to being proportional to $\varphi(x;\delta)$ under the metric associated with equation (7). The focus being on (approximate) proportionality and the integrating constant of $\varphi(x;\delta)$ being unknown, we can usefully rephrase the IS problem as one of approximating the kernel $\varphi(x;\delta)$ by an IS-kernel $k(x;a)$ subject to the restriction that the latter be analytically integrable w.r.t. x (the importance of this requirement will become fully apparent once we discuss sequential EIS implementations). The relationship between k and m is given by

$$m(x|a) = \frac{k(x;a)}{\chi(a)}, \text{ with } \chi(a) = \int k(x;a)dx \tag{10}$$

Equation (7) can be rewritten as

$$V(a;\delta) = \int_X h[d^2(x;a,\delta)] \cdot g(x,\delta) \cdot p(x|\delta)dx \tag{11}$$

with

$$d(x;a,\delta) = ln\varphi(x;\delta) - \gamma - ln\ k(x;a) \tag{12}$$

$$\gamma = ln\ G(\delta) - ln\chi(a) \tag{13}$$

$$h(c) = e^{\sqrt{c}} + e^{-\sqrt{c}} - 2 = 2\sum_{i=1}^{\infty} \frac{c^i}{(2i)!} \tag{14}$$

Note that h is monotone and convex on $\mathbf{R}_+$. Since $G(\delta)$ is unknown we shall treat γ as an additional auxiliary parameter to be included in a. An optimal choice for a is given by the solution of the nonlinear Generalized Least Squares (hereafter GLS) problem

$$a_*(\delta) = \text{ Arg Min}_{a\in A}\ V(a;\delta) \tag{15}$$

Since an efficient sampler is one for which $d(x;a,\delta)$ is expected to be small an average we can usefully consider replacing $h(c)$ by its leading term c, which implies solving the simpler GLS problem,

$$\hat{a}(\delta) = \text{ Arg Min}_{a\in A}\ Q(a;\delta) \tag{16}$$

with

$$Q(a;\delta) = \int_X d^2(x;a,\delta)\cdot g(x;\delta)\cdot p(x|\delta)\ dx \tag{17}$$

The following lemma provides an upper bound for the relative loss of efficiency resulting from the replacement of $h(c)$ by c.

Lemma 3.1 *If, under conditions such as those proposed by Geweke (1996), $V(a;\delta)$ is finite, then*

$$V(\hat{a}(\delta);\delta) > V(a_*(\delta);\delta) > h[Q(\hat{a}(\delta);\delta)] \tag{18}$$

Proof. The proof follows from Jensen's inequality, whereby

$$V(a;\delta) > h[Q(a;\delta)] \text{ on } \mathbf{R}_+,$$

together with equations (15) and (16). □

Equation (18) enables us to compute an upper bound for the relative loss of efficiency associated with using $\hat{a}(\delta)$ in place of $a_*(\delta)$. In all EIS applications we have run, a couple of which are discussed in Section 6 below, that upper bound has never exceeded a few percents which is why we only consider the simpler optimization problem (16) in the rest of our paper.

3.2 EIS from the exponential family of distribution

If $k(x;a)$ belongs to the exponential family of distribution then its logarithm can be expressed as a linear function of a under suitable reparametrization, so that

$$ln\ k(x;a) = a'\cdot c(x) \tag{19}$$

where $c(x)$ is a vector of sufficient statistics. This produces an additional significant simplification in that the optimization problem in (16) is now linear in a.

A related concept, which can also contribute simplifying further the EIS optimization problem as well as providing operational guidelines for the selection of the class of samplers M, is that of families of kernels which are closed under multiplication.

Definition 3.1 *A class* $K = \{k(x;a); a \in A\}$ *of density kernels is close under multiplication if and only if, for any two* a_1 *and* a_2 *in* A*, there exists* a_3 *in* A *such that*

$$k(x;a_3) \propto k(x;a_1) \cdot k(x;a_2),\ x \in X \tag{20}$$

The following notation is used to represent the implicit operator which maps (a_1, a_2) in a_3

$$a_3 = a_1 * a_2 \tag{21}$$

Classes of density kernels within the exponential family of distributions typically are closed under multiplication, in connection with the following lemma, where n denotes sample size.

Lemma 3.2 *If the family of density kernels* $\{k_n(\cdot|a); a \in A\}$ *admits a sufficient statistics* T_n *of fixed dimension, i.e., if there exists a function* v_n *such that*

$$k_n(x_1, \cdots, x_n; a) \propto v_n(T(x_1, \cdots, x_n); a),$$

and if v_n *is integrable with respect to* t*, then there exists a density* $g(a|t,n)$ *on* A *such that*

$$g(a|t,n) \propto v_n(t;a)$$

and the family consisting of such g*'s is closed under multiplication.*

Proof. See Degroot (1970, Section 9.3). □

In order to illustrate the usefulness of this concept, let us assume that: (1) $K = \{k(x;a); a \in A\}$ is a class of density kernel which is closed under multiplication, and (2) there exists a factorization of $\varphi(x;\delta)$ into

$$\varphi(x;\delta) = g_0(x,\delta) \cdot k(x; a_0(\delta)),\ a_0(\delta) \in A \tag{22}$$

In such a case, $d(x;a,\delta)$ in equations (11)-(17) can be replaced by the simpler expression

$$d_1(x;a,\delta) = lng_0(x;\delta) - \gamma_1 - ln\ k(x;a_1) \tag{23}$$

Let $\hat{a}_1(\delta)$ denote the solution of the optimization problem associated with d_1. It follows from Lemma 2 that $\hat{a}(\delta)$, as defined in equation (16) is given by

$$\hat{a}(\delta) = a_0(\delta) * \hat{a}_1(\delta) \tag{24}$$

In view of the factorization in (8), $k(x; a_0(\delta))$ could be a kernel of the initial sampler $p(x|\delta)$ but it could also include any additional terms in $g(x; \delta)$ which can be added to k by taking advantage of the closedness of k. As we shall illustrate by the two applications in Section 6, closedness not only simplifies the EIS optimization problem but it also helps to select the class k of kernels.

3.3 Monte Carlo EIS implementation

In practice, integrals such as $Q(a; \delta)$ in equation (17) cannot be evaluated analytically. However, MC estimates thereof are trivially available and are of the form

$$\hat{Q}_R(a; \delta) = \frac{1}{R} \sum_{i=1}^{R} [ln\varphi(\tilde{x}_i; \delta) - \gamma - ln\ k(\tilde{x}_i; a)]^2 \cdot g(\tilde{x}_i(\delta)) \tag{25}$$

where $\{\tilde{x}_i; i : 1 \to R\}$ are i.i.d. draws from the initial sampler p. Equation (25) provides the operational basis for our (GLS) EIS algorithm and has to be minimized in (a, γ) for any given value of δ. A few additional implementation details can usefully be mentioned here.

1. Draws from p typically generate very high variance in the weight function $g(\tilde{x}_i; \delta)$ - which is precisely why EIS is required. Therefore, it is preferable to delete g from equation (25) and to solve instead the unweighted LS problem.

2. Similarly, p will generate a high percentage of drawn in the tails of $ln\varphi$. While this helps in terms of securing a global EIS approximation for $ln\varphi$, it can also produce somewhat inaccurate estimates of $\hat{a}(\delta)$ under small number of draws. Rather than increasing the number of draws R in equation (25), which turns out to be computationally inefficient, we found that it was preferable by far to iterate a small number of times (typically from 2 to 4) on the EIS-LS algorithm itself, using the EIS sampler produced at step j as initial sampler to compute the next step $\hat{a}_{j+1}(\delta)$. Actually, as the EIS sampler gets more accurate, the corresponding weight function g can usefully be reintroduced for the final iteration(s). Furthermore, by using Common Random Numbers which, as discussed further below, are required for functional evaluation, we can effectively secure the convergence of the EIS optimization algorithm toward a fixed point solution $\hat{a}_R(\delta)$.

3.4 Functional Evaluation: Common Random Numbers

There was an important reason for carrying along δ as an argument in all preceeding derivations. Specifically, statistical inference procedures typically require the evaluation of a function $G(\delta)$. As we replace $G(\delta)$ by its EIS (functional) estimate $\overline{G}_{S;m}(\delta)$ an issue of smoothness immediately arises which can be critical if, for example, $\overline{G}_{S;m}(\delta)$ has to be numerically minimized in δ. Even highly accurate EIS estimates of $G(\delta)$ will exhibit critical discontinuities if based upon draws which are independent of one another across neighboring values of δ. Fortunately, there exists a conceptually simple technique for securing the necessary smoothness of $\overline{G}_{S;m}(\delta)$, which is known as that of Common Random Numbers (hereafter CRN's).

The CRN technique requires that the random draws $\{\overline{x}_i(\delta); i : 1 \to S\}$ from a sampler $m(x|a(\delta))$ be obtained from a *common* sequence of draws $\{\tilde{u}_i; i : 1 \to S\}$, whose distribution does not depend on δ, by means of a transformation of the form

$$\tilde{x}_i = \xi(\tilde{u}_i, \delta) \tag{26}$$

which is continuous (and/or differentiable) in δ. The classical example is that of a Normal sampler whose draws are obtained by a linear transformation of standardized Normal draws. More generally, equation (26) follows by application of the inversion technique for pseudo random number generation, - see e.g. Devroye (1986). Specifically, if $F(x|a(\delta))$ denotes a univariate distribution function and F^{-1} its inverse, then i.i.d. draws from F can be obtained by the transformation

$$\tilde{x}_i = F^{-1}(\tilde{u}_i|a(\delta)) \tag{27}$$

where the $\tilde{u}_i$'s are i.i.d. draws from a uniform $(0, 1)$ distribution. However, in many cases, the inversion technique is numerically highly inefficient relative to the more performant random number generation techniques (such as acceptance/rejection) which can result in unacceptably high computing time for high-dimensional applications. In such cases, we have found that interpolation (in u and a) from a common set of $\{\tilde{u}_i$'s$\}$ provides an operational method for implementing CRN's. All functional estimation results reported below were obtained from CRN's. Note finally that, as discussed further in Section 4 below high-dimensional EIS-CRN implementations will be based upon factorization into univariate components to which formula (27) is individually applied.

3.5 A Convergence Test

It has long been recognized that cases where the (E)IS sampler $m(x|a)$ has thinner tails than the integrand $\varphi(x;\delta)$ can imply the non-existence of $V(a;\delta)$ in which case the consistency of $\overline{G}_{S;m}(\delta)$ as an estimate of $G(\delta)$ is no longer guaranteed.

A trivial example often cited in the early Bayesian literature is that where φ has fat student-t tails and one selects a Gaussian IS sampler. However, cases of non-existence of $V(a;\delta)$ can be hard to detect in more complex applications such as those presented in Section 6 below. The traditional solution which consists of evaluating $\overline{G}_{S;m}(\delta)$ under increasing number of draws and empirically verifying its convergence leaves much to desire since detection lies upon highly unlikely draws in the far tails of $m(x|a)$ and, therefore, typically requires prohibitively large numbers of draws with no guarantees of successful detection. Actually, the thinner the tails of $m(x|a)$ are the less likely brute force detection becomes.

Fortunately, we can offer here an empirical test of convergence which is an immediate byproduct of the EIS computations, requires no additional draws, and appears to be extremely sensitive to non-existence of $V(a;\delta)$. It consists of comparing two different MC estimates of $V(a(\delta);\delta)$ one computed from i.i.d. draws from the EIS sampler $m(x|\hat{a}(\delta))$ and the other one from a sampler with broader coverage (variance), whether the initial sampler $p(x|\delta)$ or another sampler from the class m with inflated variance relative to the EIS sampler. Following equation (11), the generic formula for these estimates is given by

$$\tilde{V}_R(\delta) = \frac{1}{R}\sum_{i=1}^{R} h(d^2(\tilde{x}_i, \hat{a}_R(\delta), \delta) \cdot \frac{\varphi(\tilde{x}_i;\delta)}{q(\tilde{x}_i|\delta)} \tag{28}$$

where $\{\tilde{x}_i; i: 1 \rightarrow R\}$ denote (CRN) i.i.d. draws from the selected sampler(s) $q(x|\delta)$. The key advantage of such a comparison is that, as illustrated in the next section, it is very effective in flagging out potential lack of convergence even under small numbers of MC draws. See also Koopman and Shephard (2004) for a test based upon extreme values theory.

3.6 Two Pilot Applications

First, let φ denote the following transformation of a density gamma

$$\varphi(x;\delta) = \frac{1}{\Gamma(\delta+1)} \cdot \exp(-x^{1/\delta}),\ x>0,\ \delta>0 \tag{29}$$

in which case $G(\delta)=1$ for all $\delta>0$. Let also

$$k(x;a) = \exp(-ax), \quad x>0,\ a>0 \tag{30}$$

denote an exponential density kernel. While k is not a particularly good choice of IS sampler for this problem, it enables us to illustrate a variety of situations of interest. Note, in particular, that k has thinner tails than φ for $\delta>1$ and that $V(a;\delta)$ is infinite for $\delta>2$.

In this case, the EIS auxiliary regressions, as defined in equation (25), amount to a simple least squares regression of $x^{1/\delta}$ on x. It is easy to verify that if the x's are drawn from $k(x;a_0)$, then the (population) coefficient of that regression is given by

$$\hat{a}_1(\delta) = a_0^{1-\frac{1}{\delta}} \cdot \frac{1}{\delta} \cdot \Gamma(1+\frac{1}{\delta}) \tag{31}$$

with a fixed point solution (under iterated EIS)

$$\hat{a}(\delta) = \left[\frac{1}{\delta} \cdot \Gamma(1+\frac{1}{\delta})\right]^{\delta} \tag{32}$$

Formula (22) does not apply here so that we don't have a natural initial sampler. But it is obvious from formulae (29) and (30) that $\hat{a}(\delta)$ has to be a decreasing function of δ with $\hat{a}(1)=1$. Whence we selected $a_0(\delta)=1/\delta$ for the initial sampler (other starting values such as $a_0(\delta)=1$ work equally well but may require additional EIS iterations). Results are presented in Table 1 for values of δ ranging from 0.6 to 2.6. The two samplers used for the convergence test discussed above are the EIS sampler with $\hat{a}_R(\delta)$ and one with $\hat{a}_1(\delta) = 5.\hat{a}_R(\delta)$ (equivalent to multiplying the variance of the EIS sampler by five). Obviously, for $\delta=1$ the EIS sampler satisfies equation (9) with $\hat{a}_R(\delta)=1$ and is perfect. While an exponential sampler is not particularly efficient for values of δ very different from 1, this pilot application illustrates some important points.

- The final (iterated) EIS regression coefficients $\hat{a}_R(\delta)$ are accurate LS estimates of $\hat{a}(\delta)$;

- The two variance bounds in equation (18) are always very close to one another, negating the need for solving the nonlinear LS problem in equation (15). The MC sampling variances for $\overline{G}_{S;m}(\delta)$ - obtained by rerunning the entire EIS algorithm under different seeds - slightly exceeds these bounds. This reflects additional variance due to the fact that $a(\delta)$ is being estimated by $\hat{a}_R(\delta)$ - see the discussion in Section 5 below.

- The two variance estimates for the convergence test are very close to one another for $\delta < 1$, start diverging for $1 < \delta < 2$ ($V(a;\delta)$ remains finite but the tails of the EIS sampler are thinner than those of φ) and rapidly diverge for $\delta > 2$. Note also the explosion in the MC standard deviation of the inflated variance estimate. In sharp contrast, the MC standard deviations of $\hat{G}_R(\delta)$ provide no indications of the problem.

Our second pilot application considers the classical (pathological) problem of approximating a Student$-t$ density by a Normal density. Let

$$\varphi(x;\delta) = \frac{\Gamma(\frac{\delta+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\delta}{2})} \cdot \left(\frac{\delta-2}{2}\right)^{-\frac{1}{2}} \cdot \left(1 + \frac{2x^2}{\delta-2}\right)^{-\frac{1}{2}(\delta+1)} \tag{33}$$

This particular parametrization ensures that $G(\delta) = 1$ and also that $\mathrm{Var}_\varphi(x|\delta) = 0.5$ for all $\delta > 2$. The family k of Gaussian kernels is parametrized as

$$k(x;a) = \exp(-ax^2) \tag{34}$$

so that $\mathrm{Var}_m(x|a) = (2a)^{-1}$. Note that $\varphi(x;\delta)$ converges to $m(x|1)$ for $\delta \to \infty$. Results are presented in Table 2 and here again, highlight the high sensitivity of our convergence test to the thin tail problem. In contrast MC standard deviations for $\overline{G}_{S;m}(\delta)$ fail to detect any problem.

4 High-Dimensional EIS implementation

The EIS-LS algorithm introduced in Section 3 only applies to very low-dimensional x's. In higher dimensional set-ups, feasibility requires that it be

decomposed into a sequence of low-dimensional optimization problems. We now present an operational sequential implementation of EIS which, as strikingly illustrated by the applications discussed in Section 6 below, is applicable in very high dimensional integration problems. It exploits the fact that high-dimensional models are typically specified not as a single joint distribution but as a sequence of conditional distributions whereby latent and observable variables are generated individually (sequentially in the time dimension or in parallel in cross sections). Therefore, we now assume that there exists a natural sampling preordering partition of x into low-dimensional (typically univariate) components, say $x = (x_1, \cdots, x_L)$. The functionals φ, g and p in equation (8) are partitioned conformably with x into

$$\varphi(x;\delta) = \Pi_{\ell=1}^{L}(X_\ell;\delta), \quad g(x;\delta) = \Pi_{\ell=1}^{L} g_\ell(X_\ell;\delta) \tag{35}$$

$$p(x|\delta) = \Pi_{\ell=1}^{L} p_\ell(x_\ell|X_{\ell-1},\delta) \tag{36}$$

where $X_\ell = (x_1, \cdots x_\ell)$ and X_0 denotes known initial conditions. Unknown initial conditions would be included in the x's to be integrated in which case X_0 could be empty. Our notation highlights the fact that p constitutes a sequential (initial) sampler for x while φ and g do not. It also covers time series as well as cross section applications which as illustrated in Section 6 below, translate into appropriate conditional independence assumptions (and/or exclusion restrictions) in the expressions of $(\varphi_\ell, p_\ell, g_\ell)$. The EIS sampler $m(x|a)$ is then partitioned conformably as

$$m(x|a) = \Pi_{\ell=1}^{L} m_\ell(x_\ell|X_{\ell-1}, a_\ell) \tag{37}$$

with $a = (a_1, \cdots, a_L) \in A = \Pi_{\ell=1}^{L} A_\ell$. As discussed in Section 3.1 above, EIS approximations involve kernels instead of densities since, in particular, the integral of φ_ℓ w.r.t. x_ℓ is unknown. A kernel k_ℓ for m_ℓ is a function which is proportional to m_ℓ for any given $(X_{\ell-1}, a_\ell)$. The following notation applies to a kernel k_ℓ, its integral χ_ℓ and the corresponding density m_ℓ:

$$\chi_\ell(X_{\ell-1}; a_\ell) = \int k_\ell(X_\ell; a_\ell) d\ x_\ell \tag{38}$$

$$m_\ell(x_\ell|\chi_{\ell-1}, a_\ell) = \frac{k_\ell(X_\ell; a_\ell)}{\chi_\ell(X_{\ell\cdot 1}; a_\ell)} \tag{39}$$

The very fact that the integrating constant χ_ℓ may - and actually will - depend on $X_{\ell\cdot 1}$ has critical implications for our sequential EIS implementation. Specifically if in line with equation (12), we aimed at approximating $ln\ \varphi_\ell(X_\ell;\delta)$ by a log kernel $ln\ k_\ell(X_\ell; a_\ell)$, then the integrating constant

$\chi_\ell(X_{\ell\cdot 1}; a_\ell)$ would not be accounted for. This would amount to ignoring the fundamental sequential structure of the initial model and would inevitably produce highly inefficient overall samplers. Since, however, χ_ℓ does not depend on x_ℓ and has a known analytical expression, it can be transferred back and explicitly accounted for in the $(\ell-1)$th suboptimization problem. Specifically, this amounts to reshuffling the partitioning of m as follows:

$$\begin{aligned}\frac{\varphi(x;\delta)}{m(x|a)} &= \Pi_{\ell=1}^{L}\ \frac{\varphi_\ell(X_\ell;\delta)}{m_\ell(x_\ell|X_{\ell-1},a_\ell)} \\ &= \chi_1(X_0;a_1)\cdot\Pi_{\ell=1}^{L}\ \frac{\varphi_\ell(X_\ell;\delta)\cdot\chi_{\ell+1}(X_\ell;a_{\ell+1})}{k_\ell(X_\ell;a_\ell)}\end{aligned} \tag{40}$$

with $\chi_{L+1}(\cdot)\equiv 1$. The sequential implementation of EIS immediately follows from equation (40). It consists of a back-recursive sequence of optimization subproblems, whose step ℓ consists of approximating $ln(\varphi_\ell\cdot\chi_{\ell+1})$ by lnk_ℓ. More specifically, the backward recursion on $\{\hat{a}_\ell(\delta);\ell: L\rightarrow 1\}$ is given by:

$$\begin{aligned}\hat{a}_\ell(\delta) = \text{ArgMin}_{a_\ell,\gamma_\ell}\ \frac{1}{R}\ \sum_{i=1}^{R}\ \Big[ln\Big[\varphi_\ell(\tilde{X}_\ell^{(i)};\delta)\cdot\chi_{\ell+1}(\tilde{X}_\ell^{(i)};\hat{a}_{\ell+1}(\delta))\Big] \\ -\gamma_\ell - ln\ k_\ell(\tilde{X}_\ell^{(i)};a_\ell)]^2\cdot g_\ell(\tilde{X}_\ell^{(i)};\delta)\end{aligned} \tag{41}$$

where $\tilde{X}_\ell^{(i)} = (\tilde{x}_1^{(i)},\cdots,\tilde{x}_\ell^{(i)})$ and the $\{\tilde{x}_\ell^{(i)}\}$ denote i.i.d. trajectories drawn from the initial sampler - that is to say $\tilde{x}_\ell^{(i)}$ is drawn from $p_\ell(x_\ell|\tilde{X}_{\ell-1}^{(i)},\delta)$. As above, we recommend deleting the weight factor g_ℓ, at least in the initial EIS iteration(s). Note that it is critical that lnk_ℓ approximates well $ln(\varphi_\ell,\chi_{\ell+1})$ in X_ℓ, not simply in x_ℓ. Once an EIS sequence of kernels $\{k_\ell\}$ has been obtained, the EIS estimate of $G(\delta)$ is computed by formula (5) under (CRN) trajectories $\{\hat{x}_1^{(i)},\hat{x}_2^{(i)},\cdots,\hat{x}_L^{(i)}\}$ drawn from the corresponding EIS samplers $\{m_\ell\}$. We note that the convergence test introduced in Section 3.5 can be applied individually to each EIS optimization subproblem, with the advantage that it would single out individual kernels which could be responsible for a poor global EIS approximation.

We conclude this presentation of our sequential EIS implementation with two important remarks.

First, the sequence of low dimensional optimization problem is not equivalent to the (unfeasible) joint optimization. In order to illustrate the implicit trade-off, let us consider the case where $L=2$ and $x=(x_1,x_2)$. In line with

equation (40) we can rewrite $G(\delta)$ as:

$$G(\delta) = \int \varphi_1(x_1;\delta)[\int \varphi_2(x;\delta)dx_2]dx_1, \tag{42}$$

$$= \chi_1(a_1) \cdot \int \left[\frac{\varphi_1(x_1,\delta)\cdot\chi_2(x_1;a_2)}{k_1(x_1,a_1)}\right] \cdot h_2(x_1;a_2,\delta)\cdot m_1(x_1|a_1)dx_1,$$

$$\text{with } h_2(x_1;a_2,\delta) = \int \left[\frac{\varphi_2(x;\delta)}{k_2(x;a_2)}\right] \cdot m_2(x_2|x_1,a_2)dx_2 \tag{43}$$

The terms between brackets in the right-hand sides of equations (42) and (43) correspond to the two EIS subproblems defined in equation (21). It follows that the function h_2 is not accounted for in our sequential EIS implementation. Since, however, the key to a successful EIS solution is that the ratios $\varphi_\ell\chi_{\ell+1}/k_\ell$ be near constant, h_2 itself should be near constant and its omission from the EIS optimization ought to be largely inconsequential. This heuristic justification will be fully supported by the exceptional efficiency of the EIS sequential algorithm in very high-dimensional applications such as those presented in Section 6 below.

Second, the fact that the sequential EIS algorithm runs backward could be perceived to be a significant drawback since, in particular, it requires recomputing the EIS sampler each time new dimensions of integration are added. However, such reruns are expected to be very fast since one can use as initial sampler for such a rerun the EIS sampler previously computed for the lower dimensional problem augmented by the natural sampler for the added dimensions. Under standard mixing assumptions one would expect that the impact of the added dimensions would fade out as the algorithm runs backward and that only a relatively small number of EIS subsamplers would have to be significantly adjusted.

5 Numerical and Statistical Properties of EIS Estimates

Following equation (5), an EIS functional estimate of $G(\delta)$ has the form

$$\overline{G}_{S;m}(\delta,\hat{a}_R(\delta)) = \frac{1}{S}\sum_{i=1}^{S} w(\tilde{x}_i(\delta);\delta,\hat{a}_R(\delta))\cdot g(\tilde{x}_i(\delta),\delta) \tag{44}$$

where $\hat{a}_R(\delta)$ minimizes $\hat{Q}_R(a;\delta)$ as given in equation (25) and $\{\tilde{x}_i(\delta); i:1 \rightarrow S\}$ are i.i.d. draws from $m(x|\hat{a}_R(\delta))$ obtained from a fixed matrix U of CRN by a transformation of the form given in equations (26) or (27). Therefore, $\hat{a}_R, \{\tilde{x}_i\}$ and $\overline{G}_{S;m}$ are all implicit functions of δ and U. The following shorthand notation will prove useful for our subsequent analysis

$$\overline{G}_S(\delta, U) = \overline{G}_{S;m}(\delta, \hat{a}_R(\delta)) \tag{45}$$

Note that it is a trivial matter to rerun our entire EIS algorithm under i.i.d. draws from U since it is just a matter of changing the initial seed of the random number generator. This enables us to produce i.i.d. draws from $\overline{G}_S(\delta, U)$ from which the following improved estimate of $G(\delta)$ obtains together with its MC variance

$$\overline{G}_{SL}(\delta) = \frac{1}{L}\sum_{\ell=1}^{L}\overline{G}_s(\delta, \tilde{u}_\ell) \tag{46}$$

$$\text{Var}_U\ (\overline{G}_{SL}(\delta)) = \frac{1}{L}\Big[\frac{1}{L}\sum_{\ell=1}^{L}(\delta, \tilde{u}_\ell) - \overline{G}^2_{SL}(\delta)\Big] \tag{47}$$

That same MC simulation procedure can be used to produce estimates of the numerical accuracy of any implicit transformation of $G(\delta)$. Assume for example, that $\delta = (\theta, y)$ where y is an observed sample and θ a vector of unknown parameter. Let $G(\delta)$ in equation (1) denote the corresponding likelihood function, marginalized w.r.t. a vector x of latent variables. Let $\hat{\theta}(y)$ denote the (unfeasible) Maximum Likelihood (hereafter ML) of θ, and $\hat{\theta}_s(y, U)$ its ML-EIS (numerical) estimate, respectively defined as

$$\hat{\theta}(y) = \text{ Arg Max}_\theta\ ln\ G(\theta, y) \tag{48}$$

$$\hat{\theta}_S(y, U) = \text{ Arg Max}_\theta\ ln\ \overline{G}_S(y, \theta, U) \tag{49}$$

As above, we can rerun under different seeds the entire EIS-ML algorithm now consisting of equations (44), (45) and (49), in order to obtain i.i.d. draws of $\hat{\theta}_s(y, U)$ from which to compute a final EIS-ML estimate of $\hat{\theta}(y)$ together with its numerical covariance matrix, respectively given by

$$\hat{\theta}_{SL}(y) = \frac{1}{L}\sum_{\ell=1}^{L}\hat{\theta}_s(y, \tilde{u}_\ell) \tag{50}$$

$$\text{Var}_U\ (\hat{\theta}_{SL}(y)) = \frac{1}{L}\ \Big[\frac{1}{L}\sum_{\ell=1}^{L}\ \hat{\theta}^2_S(y, \tilde{u}_\ell) - \hat{\theta}^2_{SL}(y)\Big] \tag{51}$$

Finally, we can also produce by MC simulation an EIS estimate of the statistical covariance matrix of $\hat{\theta}(Y)$ as an estimator of θ. Note that, as highlighted in Section 3.4, Y and U are independent of one another by construction. Therefore, we can fix U and use our (estimated) statistical model to produce draws of $Y|\theta$ - in practice, we would draw joint trajectories $\{\tilde{x}_i, \tilde{y}_i\}$ and only retain the $\tilde{y}_i, s$. The corresponding EIS estimate of the statistical covariance matrix of $\hat{\theta}(y)$ is given by

$$\hat{\text{Var}}_{Y|\theta}(\hat{\theta}(Y)) = \frac{1}{L}\Big[\sum_{\ell=1}^{L} \hat{\theta}_s^2(\tilde{y}_\ell, U) - \overline{\theta}_{SL}^2\Big], \text{ with} \tag{52}$$

$$\overline{\theta}_{SL} = \frac{1}{L}\sum_{\ell=1}^{L} \hat{\theta}_S(\tilde{y}_\ell, U) \tag{53}$$

In summary, once an ML-EIS program has been produced it can trivially be embedded in two independent MC simulation loops in order to compute covariance matrices for $\hat{\theta}_S(y, U)$ as a numerical estimate of $\hat{\theta}(y)$ and for $\hat{\theta}(y)$ as a statistical estimate of θ. Similar numerical procedures apply to other classes of estimators.

We conclude this section by emphasizing the fact that our recommended practice of computing numerical and statistical covariance matrices for EIS-based estimators contrasts with the classical treatment of simulation estimators in the recent econometric literature. Useful references are McFadden (1989), Pakes and Pollard (1989) or, more recently, Gourieroux and Monfort (1993, 1996). In that literature, the sampling properties of simulation estimators are derived from the joint distribution of the observable as well as latent variables. Under appropriate conditions the (fixed S) asymptotic covariance matrix of $\hat{\theta}_S(y, U)$ equals $(1 + S^{-1})$ that of $\hat{\theta}(Y)$. While such results provide a convenient characterization of numerical accuracy, they hide the fact that joint simulation of (Y, U) often constitutes a highly inefficient numerical treatment of latent variables. As dramatically illustrated by the results provided in our next section, EIS treatments of latent variables which aim at drawing latent variables (near) conditionally on the observables - result in far greater relative numerical accuracy than conventional joint simulation.

6 Two High-Dimensional Pilot Applications

We now present two high-dimensional applications illustrating the high numerical accuracy of ML-EIS estimation for two important classes of latent

variable models: stochastic volatility and panels with unobserved heterogeneity along both dimensions.

6.1 Stochastic Volatility

Stochastic volatility models play a central role in finance. Pioneering contributions are Taylor (1986), Melino and Turnbull (1990) or Duffie and Singleton (1988). It has long been recognized that natural IS, as defined in equation (2), is hopelessly inefficient - see Danielsson and Richard (1993) for a dramatic illustration of such inefficiency. In order to illustrate the full flexibility of our EIS algorithm in this context we consider here four versions of the same baseline model. Let y_t denote the daily return of a financial asset and λ_t its unobserved variance (Stochastic Volatility, hereafter SV). SV models typically consist of a stochastic equation for y_t given λ_t and another for λ_t given λ_{t-1} (which can trivially be generalized into a higher order autoregressive process). Two versions of the density $g(y_t|\lambda_t, \cdot)$ will be considered: a fat tail Student-t density (S) and a thin tail Normal density (N). For the density $p(\lambda_t|\lambda_{t-1}, \cdot)$, we shall consider in turn an Inverted-Gamma (G) density and a Lognormal density (L). These four densities are parametrized as follows:

$$g_t^1(y_t|\lambda_t, \theta) \propto \lambda_t^{\frac{1}{2}} \cdot \left[1 + \frac{(y_t - \mu)^2)}{\lambda_t(w-2)}\right]^{-\frac{1}{2}(w+1)}, \tag{54}$$

$$g_t^2(y_t|\lambda_t, \theta) \propto \lambda_t^{-\frac{1}{2}} \cdot \exp\left[-\frac{1}{2} \cdot \frac{(y_t - \mu)^2}{\lambda_t}\right], \tag{55}$$

$$p_t^1(\lambda_t|\lambda_{t-1}, \theta) \propto r_{t-1}^{\frac{1}{2}\nu} \cdot \lambda_t^{-\frac{1}{2}(\nu+2)} \cdot \exp\left(-\frac{r_{t-1}}{\lambda_t}\right), \tag{56}$$

$$\text{with } r_{t-1} = \frac{1}{2}(\gamma + \partial\lambda_{t-1}) \text{ and } E(\lambda_t|\lambda_{t-1}) = \frac{2}{v-2} r_{t-1} = q + r\lambda_{t-1}$$

$$p_t^2(v_t|v_{t-1}, \theta) \propto \exp\left[-\frac{1}{2\sigma^2}(v_t - q - rv_{t-1})^2\right] \tag{57}$$

where $v_t = ln\lambda_t$ and θ includes the corresponding subset of the parameters $(q, r, \sigma, \mu, \omega, \nu)$. We shall consider here the four pairwise combinations of q_t and p_t, respectively denoted SG, SL, NG and NL, but shall only detail the sequential EIS implementation for the SG version since the other versions only require fairly obvious modifications of the auxiliary regressions.

Since p_t^1 belongs to the exponential family of distributions and is closed under multiplication, we can proceed as outlined in Section 3.2 and define

accordingly the class K_t as consulting of products of two I-G kernels, one consisting of all I-G factors in $\varphi_t = g_t^1 p_t^1$ and the other designed to best approximate the remainder. Since the latter only depends on λ_t, we shall select the following forms for the EIS kernel k_t and its integrating constant χ_t:

$$k_t(\lambda_{(t)}, a_t) \propto \left[r_{t-1}^{\frac{1}{2}\nu} \cdot \lambda_t^{-\frac{1}{2}(\nu+3)} \exp(-\frac{r_{t-1}}{\lambda_t}) \right] \cdot \left[\lambda_t^{-b_t} \cdot \exp(-\frac{c_t}{\lambda_t}) \right] \tag{58}$$

$$\chi_t(\lambda_{t-1}, a_t) \propto r_{t-1}^{\frac{1}{2}\nu} \cdot (r_{t-1} + c_t)^{-\frac{1}{2}(\nu+1)-b_t} \tag{59}$$

with $a_t' = (b_t \ c_t)$. Note the inclusion of r_{t-1}, which only depends on λ_{t-1}, in the expression for k_t. It provides a convenient way of making sure that r_{t-1} will be accounted for in the period $t-1$ EIS auxiliary regression. By construction, all terms included between the first two brackets in equation (58) belong to both φ_t and k_t. Therefore, they cancel out in the auxiliary EIS regressions which simplify into the following OLS regressions:

$$\begin{aligned} &\text{Dependent variables: } ln\ \chi_{t+1}(\lambda_t, \hat{a}_{t+1}) - \frac{1}{2}(w+1) \cdot\ ln\ \left[1 + \frac{(y_t - \mu)^2}{\lambda_t (w-2)} \right] \\ &\text{Regressors: } ln\ \lambda_t \text{ and } \lambda_t^{-1} \text{ (plus one intercept)} \end{aligned} \tag{60}$$

Equations (58) to (60) are all we need to apply the sequential EIS algorithm described in Section 4. As discussed above, we rerun EIS under a fixed set of CRN's for each evaluation of the likelihood function. The computing time required for a very large number of inversion of the IG distribution function is considerably reduced by the initial construction of a high-accuracy bivariate interpolation table (U and degrees of freedom) for the IG inverse distribution function. For the purpose of illustration we computed ML-EIS estimates for a sample consisting of 1,447 observations of IBM daily stock price changes for the period 1/9/82-3/31/87. As indicated by the results reported in the SG row of Table 3, ML-EIS estimates are numerically extremely accurate even with as little as $R = S = 10$ MC draws (note, in particular, that the ratios between numerical and statistical standard deviations is much smaller than the classical $S^{-1/2}$ ratio). Table 3 also reports results for the SL, NG and NL cases.

Total computing time for a full ML optimization (using a simplex algorithm which is extremely robust for this type of applications) is of the order of 45 seconds on a 750 MHZ UNIX server (and about 15 seconds for the

faster NL version). The high persistence with values of r in the 0.95-0.98 range is typical of SV models. We note that the Lognormal density for λ_t is qualitatively better identified than its IG counterpart. Actually, it can be shown that the high-value of $\hat{r}$ also requires a high value for $\hat{\nu}$ in order to match the sample (stationary) variation coefficient of λ_t which is of the order of 0.4.

The high numerical accuracy of EIS in the context of SV models results from the facts that: (i) The observation $\{y_t\}$ are very informative on the underlying latent process $\{\lambda_t\}$ and EIS is designed precisely to take full advantage of such situations; (ii) A total of $2T = 2,894$ auxiliary parameters were used to construct the EIS sampler. Such high accuracy has been fully confirmed by recent applications of EIS to a wide range of SV models (univariate, bivariate, two factors, semi-parametric versions) in dimensions up to $8{,}000^{+}$. See Liesenfeld and Richard (2003a, b). Comparisons with alternative numerical evaluations of SV models are found in Liesenfeld and Richard (2003c) and Bauwens and Hautsch (2003). The combination of high numerical accuracy and ease of implementation of EIS appears to be unmatched in that class of models.

6.2 Logit with Unobserved Heterogeneity

It is commonly held in the econometric literature that ML estimation of panel data models with unobserved random heterogeneity along both dimensions is unfeasible - see e.g. the comments in McFadden (1989). This has led to the development of alternative though statistically less efficient simulation based estimation techniques. See Lerman and Manski (1981), McFadden (1989), Pakes and Pollard (1989) or Börsch-Supan and Hajivassiliou (1990). See also Gourieroux and Monfort (1993) for a survey or Gourieroux and Monfort (1996) for an in-depth analysis of simulation based inference techniques.

In this section, we demonstrate that contrary to common beliefs, highly accurate numerical evaluation of the likelihood function of such panel data models is fully operational under EIS. Consider a model consisting of a logit for the observable $y_{it} \in \{0, 1\}$ conditionally on the random effects α_i and λ_t, independent normal distributions for the α's and a multivariate normal distribution for the λ's. Let $\alpha' = (\alpha_1 \cdots \alpha_N), \lambda' = (\lambda_1 \cdots \lambda_T), y' = (y_{it}; i : 1 \to N, t : 1 \to T)$ and $\delta = (y, \theta)$ where θ regroups all unknown parameters in the model. Let x_{it} denote a vector of exogenous variables. The likelihood

function is of the form given by equation (1) together with

$$g(\alpha, \lambda, \delta) = g_0(\lambda, \delta) \cdot \Pi_{i=1}^{N} g_i(\alpha_i, \lambda, \delta) \tag{61}$$

with $g_0(\lambda, \delta) \equiv 1$ and

$$g_i(\alpha_i, \lambda_i, \delta) = \Pi_{t=1}^{T} \frac{[\exp(v_{it}]^{y_{it}}}{1 + \exp(v_{it})} \tag{62}$$

$$v_{it} = \beta' x_{it} + \alpha_i + \lambda_t \tag{63}$$

$$p(\alpha, \lambda|\delta) = p_0(\lambda|\theta) \cdot \Pi_{i=1}^{N} p_i(\alpha_i|\theta) \tag{64}$$

$$p_0(\lambda|\theta) \propto |H_\theta|^{\frac{1}{2}} \exp -\frac{1}{2}\lambda' H_\theta \lambda \tag{65}$$

$$p_i(\alpha_i|\theta) \propto \sigma_\alpha^{-1} \exp\Big[-\frac{1}{2}\Big(\frac{\alpha_i}{\sigma_\alpha}\Big)^2\Big] \tag{66}$$

It is implicitly assumed here that $N >> T$ as commonly the case. For $T >> N$ we would permute the α's and λ's in all factorizations. Extensions such as exchangeable α's are trivially handled by conditioning the densities p_i in equation (66) on a common factor α_0 and adding an additional density for α_0. Notationally this would amount to incorporating α_0 into λ.

It is obvious from equations (61) to (66) that we should partition EIS samplers conformably with p in equation (64), with the critical extension that the α_i's are now to be independent *conditionally* on λ in order to fully account for the (posterior) dependence between α and λ as induced by g. Specifically, we select for the α_i's conditionally independent kernels of the form

$$ln\ k_i(\alpha_i, \lambda, a_i) = -\frac{1}{2}\Big[2b_i' v_i + v_i' C_i v_i + \Big(\frac{\alpha_i}{\sigma_\alpha}\Big)^2\Big] \tag{67}$$

with

$$v_i = X_i \beta + \lambda + \alpha_i e,$$

$X_i' = (x_i, \cdots, x_{iT}), \lambda' = (\lambda_1 \cdots \lambda_T),\ e' = (1 \cdots 1),\ b_i \in \mathbf{R}^T,\ C_i = \text{Diag}\,(c_i), c_i = C_i e \in \mathbf{R}_+^T$ and $a_i = (b_i c_i)$. It turns out that the constraints $c_i > 0$ never bind and can safely be ignored. Note that the integrating constant χ_i associated with k_i only depends upon (λ, a_i), not upon the other α_j's. In order to facilitate subsequent EIS integration w.r.t. $\lambda, ln\ k_i$ is rewritten as a quadratic form in λ and is given by

$$ln\ \chi_i(\lambda, a_i) \propto -\frac{1}{2}\Big[\ell_i' C_i \ell_i + 2\ell_i' b_i - \Big(\frac{\overline{\alpha_i}}{\sigma_i}\Big)^2\Big] \tag{68}$$

with

$$\ell_i = \lambda + X_i\beta, \quad \sigma_i^{-2} = \sigma_\alpha^{-2} + c_i'e \text{ and } \overline{\alpha}_i = -\sigma_i^2(\ell_i'c_i + b_i'e) \tag{69}$$

In summary, the EIS auxiliary GLS regressions for the α_i's consist of regressing $ln\ g_i$ on the $2T$ regressions $\{(v_{it}, v_{it}^2); t : 1 \to T\}$ with an intercept and weights g_{it} (OLS for the first EIS iteration). Under EIS sampling and conditionally on λ, the α_i's are independently normally distributed with means $\overline{\alpha}_i$ and variances σ_i^2.

As for λ, we note that $\varphi_0 = g_0 p_0$ and $\{\chi_i; i : 1 \to N\}$ all are in the forms of Gaussian kernels. Therefore, a perfect EIS sampler for λ obtains immediately by combining together the $N+1$ quadratic forms in equations (65) and (68) - providing a remarkable example of "perfect fit" in the sense of equation (9). Rearranging terms in the usual way, we find that

$$m_0(\lambda|a_0) \propto |A|^{\frac{1}{2}} \exp\Big[-\frac{1}{2}(\lambda-\mu)'A(\lambda-\mu)\Big] \tag{70}$$

with $a_0' = (\mu',\ \text{vec}'A)$ and

$$A = H_0 + \sum_{i=1}^{N}(C_i - \sigma_i^2 c_i c_i') \tag{71}$$

$$\mu = -A^{-1} \sum_{i=1}^{N} \Big[C_i X_i\beta + b_i - \sigma_i^2(c_i'X_i\beta + b_i'e)c_i\Big] \tag{72}$$

This completes the description of the EIS algorithm for this application. Note the very large number of auxiliary parameters (of the order of $2TN$) used to produce the EIS approximation to the actual posterior density of (α, λ). A caveat applies before we present numerical results. As in Section 6.1 we shall aim at running the auxiliary EIS regressions under relatively small numbers of MC draws (of the order of three times the number of regressions). As the v_{it}'s all depend on α_i, occasional bad draws can generate very high multicollinearity in the auxiliary regressions and even crash the EIS algorithm (Production of the results reported in Table 4 required several millions of EIS regressions!). We were able to completely eliminate the problem by introducing a small amount of shrinkage in the auxiliary EIS regressions. Specifically, we used a second order Taylor Series Expansion (hereafter TSE) of g_i around $\alpha_i = \lambda_t = 0$ in order to produced exact restrictions for a_i. Let these restrictions be written as

$$a_i = R_i\delta_i + q_i \tag{73}$$

where $\delta_i \in \mathbf{R}^\ell$ with $\ell < 2T$ denotes free regression coefficients and $(R_i q_i)$ is a $2T \cdot (\ell + 1)$ matrix of known constants (which actually depend upon $\{x_{it}, y_{it}); t : 1 \to T\}$ and β which are all included in δ). An unconstrained EIS OLS estimator of the form $\hat{a}_i = G_i^{-1} h_i$ is then replaced by the shrinkage estimator

$$\tilde{a}_i = (G_i + \kappa M_i)^{-1}(R_i + \kappa M_i q_i) \tag{74}$$

with $M_i = I_{2T} - R_i(R_i' R_i)^{-1} R_i'$. This shrinkage option completely eliminates unwanted interruptions of the EIS algorithm at virtually no loss of EIS efficiency even with very low values of κ (0.01 or less in the application which follows).

In order to illustrate the impressive numerical performance of EIS within this class of models, we generated a fictitious sample of size $T = 15$ and $N = 1,000$, with no exogeneous variables and a stationary $AR(1)$ process for λ with autocorrelation coefficient ρ and stationary variance σ_λ^2. The parameters true values were set equal to $\sigma_\alpha = \sigma_\lambda = 0.3$ and $\rho = 0.5$, implying a relatively moderate amount of heterogeneity (results derived under different values produce similar qualitative results). The number of regressors in each EIS auxiliary regression equals 30 (plus one intercept). We used $R = S = 100$ MC draws and 3 EIS iterations.

In Table 4.1 we report MC estimates of the likelihood function at the parameter true values under the natural sampler (NAT), as defined by equations (64) to (66), the EIS sampler (EIS), the sampler obtained by TSE of the g_i, s around $\alpha_i = \lambda_i = 0$ (TSE_0) and an (unfeasible) sampler obtained by TSE of the g_i, s around the true values of α_i and λ_t, which we had initially stored (TSE_1). TSE_1 is closest in spirit to the Laplace approximations proposed by Tierney and Kadane (1986). Obviously, in practice the α's and λ's would have to be estimated first which would result in increased numerical inefficiency. The results in Table 4.1 illustrate the clear superiority of our EIS global approximations relative to TSE local ones. The results also indicates that the problem associated with the natural sampler in this context is one of enormous downward bias more than variance. Actually, this result is not surprising. Assume an individual α_i draw from p_i has a probability 0.75 of hitting the region of importance. The probability that $N = 1,000$ independent draws jointly hit the region of importance is then of the order of 10^{-124}.

Finally, in order to illustrate the performance of EIS within an inferential context, we computed ML-EIS estimates of θ. Full optimization using a

simplex algorithm requires of the order of 50 EIS likelihood evaluations for a computing time of the order of 1 minute on our 750 MHZ UNIX workstation. As in Section 6.1, numerical as well as statistical standard deviations were produced using 20 replications of the EIS-ML optimization (under the estimated values). The results are reported in Table 4.2. Note here again the impressive numerical accuracy of the results. The particular sample we used appears to have produced a borderline value for $\hat{\rho}$ (Note, however, that we only have $T = 15$ periods, that is to say only 15 latent λ_t's to identify ρ). In order to verify that the low value of $\hat{\rho}$ was due to that particular sample and not to an inherent EIS-ML problem, we ran 50 MC replications of our algorithm under the parameter true values. The corresponding statistical means and standard deviations equal (0.2922, 0.2965, 0.4726) and (0.0627, 0.0287, 0.1850), respectively. The (EIS) ML estimators are clearly statistically well-behaved.

This EIS-ML algorithm has recently been successfully applied by Liesenfeld and Richard (2004) to a dynamic logit model for the union participation decision of young men. The data ($N = 545, T = 8$) were taken from Vella and Verbeek (1998) who estimated the model under random individual effects and fixed time effects. Randomizing both heterogeneity components enable Liesenfeld and Richard to qualify their relative impact on agents' decisions. They find that the dynamic of the union participation decision is dominated by individual heterogeneity.

7 Conclusion

We proposed an operational recursive Least Square algorithm to construct (very) high dimensional Importance Samplers. In contrast with current procedures, which are mostly based upon local approximations of the integrand, our algorithm explicitly minimizes the variance of the MC-IS estimate in order to produce a global approximation to the posterior density of the variables to be integrated out. Our algorithm's performance in high-dimensional latent variables models is unparalleled, as illustrated in the context of two important classes of models in the modern econometric literature. Its success appears to result from a combination of three factors: (1) The availability of full sequential factorizations which reduce the optimization problem to an operational sequence of low-dimensional Least Squares problems, (2) The use of very large number of auxiliary parameters, typically a multiple of the

sample size in order to produce very good global fit between the integrand and the importance sampler; and (3), last but not least, the fact that the posterior densities of the latent variables appear to be very well-conditioned in the applications we have considered.

EIS is not meant to substitute for other methods under all circumstances. In particular, Monte Carlo Markov Chain (MCMC) algorithms appear to be well adapted to Bayesian applications when posterior densities of the parameters are ill-behaved and/or cannot be conveniently sequentially factorized as required for EIS. But neither do we believe that MCMC can be indiscriminately applied across the board. Actually, we believe the two methods can be complementary and we are currently investigating the possibility of combining both to conduct operational full Bayesian analysis of (dynamic) latent variable models, such as the ones we analyzed here within a classical ML framework.

Acknowledgements: We are particularly indebted to Roman Liesenfeld who, together with the first author, has developed several highly successful applications of EIS. This experience has led to major improvements of the present paper. Our work has been supported by the National Science Foundation (Grant SES-9223365).

8 References

Bauwens, L., and Hautsch, N. (2003), "Stochastic Conditional Intensity Processes", working paper available at http://www.econ.au.dk/ec2bologna/papers /lb sci.pdf.

Börsch-Supan, A. and Hajivassiliou, V. (1990), "Smooth Unbiased Multivariate Probability Simulation for Maximum Likelihood Estimation of Limited Dependent Variable Models," Cowles Foundation discussion paper no. 960, Yale University.

Danielsson, J. and Richard, J.F. (1993), "Accelerated Gaussian Importance Sampler with Applications to Dynamic Latent Variable Models," *Journal of Applied Econometrics* 8, 153-173.

DeGroot, M.H. (1970), "Optimal Statistical Decisions," New York: McGraw-Hill.

Devroye, L. (1986), "Non-Uniform Random Variate Generation," New York: Springer–Verlag.

Duffie, D., and Singleton, K.J. (1988), "Simulated Moment Estimation of Markov Models of Asset Prices," Stanford University mimeo.

Durbin, J., and Koopmans, S.J. (1997), "Monte Carlo Maximum Likelihood Estimation for non-Gaussian State Space Models," *Biometrika*, 84, 669-684.

Evans, M. (1991),"Adaptative Importance Sampling and Chaining," *Contemporary Mathematics (Statistical Multiple Integration)*, 115, 137-142.

Fishman, G.S. (1996), "Monte Carlo Concepts, Algorithms, and Applications," New York: Springer-Verlag.

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317-1339. - -(1996), "Monte Carlo

Simulation and Numerical Integration," in *The Handbook of Computational Economics*, Vol. 1., eds. H. Amman, D. Kendrick, J. and Rust, Amsterdam: North Holland.

Gourieroux, C., and Monfort, A. (1993), "Simulation Based Inference: A Survey with Special Reference to Panel Data Models," *The Journal of Econometrics*, 59, 5-33.

- -(1996), "Simulation Based Econometric Methods," *CORE Lecture Series*, Oxford: Oxford University Press.

Hammersley, J.M., and Handscomb, D. (1964), "Monte Carlo Methods," London: Methuen.

Kahn, H., and Marshall, A. (1953), "Methods of Reducing Sample Size in Monte Carlo Computations," *Journal of the Operations Research Society of America*, 1, 263-278.

Koopman, S. J., and Shephard, N. (2004), "Estimating the Likelihood of the Stochastic Volatility Model: Testing the Assumptions Behind Importance Sampling," Free University Amersterdam mimeo.

Lerman, S., and Manski, C. (1981), "On the Use of Simulated Frequencies to Approximate Choice Probability," in*Structural Analysis of Discrete Data with Econometric Applications*, Chap. 7, eds. C. Manski and D. McFadden, Cambridge, MIT Press,

Liesenfeld, R., and Richard, J.F. (2003a), "Monte Carlo Methods and Bayesian Computation: Importance Sampling," in *The International Encyclopedia of the Social and Behavioral Sciences*, 10000-10004, eds. N. J. Smelsev and P. B. Baltes, Oxford: Elsevier-Science.

- -(2003b), "Univariate and Multivariate Volatility Models: Estimation and Diagnostics," *The Journal of Empirical Finance*, 10, 505-531.

- -(2003c), "Estimation of Dynamic Bivariate Mixture Models: Comments on Watanabe," *The Journal of Business and Economic Statistics*, 21, 570-576.

- -(2004), "Simulation Techniques for Panels: Efficient Importance Sampling," forthcoming in the Econometrics of Panel Data, A Handbook of the Theory eds. L. Matyas and P. Seveitse, 3rd edition, Boston: Kluwer.

Madras, N., and Piccioni, M. (1994), "Importance Sampling for Families of Distributions," *The Annals of Applied Probability*, 9, 1202-1225.

McFadden, D. (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57, 995-1026.

Melino, A., and Turnbull, S. (1990), "Pricing Foreign Currency Options with Stochastic Volatility," *Journal of Econometrics*, 45, 239-265.

Owen, A., and Zhou, Y. (2000), "Safe and Effective Importance Sampling," *Journal of the American Statistical Association*, 95, 135-143.

Pakes, A., and Pollard, D. (1989), "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1057.

Stern, S. (1997), "Simulation-Based Estimators," *Journal of Economic Literature*, 35, 2006-2039.

Taylor, S. (1986), "Modeling Financial Time Series," Chichaster: John Wiley.

Tierney, L., and Kadane, J.B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82-86.

Trotter, H.F., and Tukey, J.W. (1956), "Conditional Monte Carlo for Normal Samples," in *Symposium on Monte Carlo Methods*, ed. H.A. Mayer, New York: John Wiley.

Vella, F., and Verbeek, M. (1998), "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men," *Journal of Applied Econometrics*, 13, 163-283.

Table 1: EIS Results for φ in Equation (29)

				Variance Estimates		
δ	$\hat{a}(\delta)$	$\hat{a}_R(\delta)$	$\overline{G}_{S;m}(\delta)$	low	high	inflated
0.6	1.736	1.683	1.001	0.0232	0.0235	0.0249
		(0.119)	(0.025)	(0.0039)	(0.0041)	(0.0032)
0.8	1.321	1.303	1.001	0.0103	0.0104	0.0114
		(0.043)	(0.013)	(0.0018)	(0.0019)	(0.0010)
1.2	0.747	0.756	0.997	0.0090	0.0090	0.0126
		(0.023)	(0.010)	(0.0021)	(0.0021)	(0.0018)
1.6	0.396	0.411	0.987	0.0242	0.0246	0.5371
		(0.036)	(0.028)	(0.0066)	(0.0070)	(0.6051)
2.0	0.196	0.209	0.974	0.0365	0.0376	13.78
		(0.030)	(0.044)	(0.0109)	(0.0120)	(30.43)
2.4	0.092	0.100	0.959	0.0469	0.0490	89.27
		(0.020)	(0.057)	(0.0147)	(0.0167)	(243.0)
2.6	0.061	0.068	0.952	0.0517	0.0542	170.8
		(0.015)	(0.062)	(0.0164)	(0.0188)	(491.6)

Relevant formulae:

$\hat{a}(\delta)$: Population EIS Coefficient; Formula (32)
$\hat{a}_R(\delta)$: Mean EIS-OLS Coefficient; Formula (25) - Unweighted
$\overline{G}_{S;m}(\delta)$: Mean EIS-Estimate of $G(\delta) = 1$; Formula (5)
Low: Lower Bound $h[Q(\hat{a}_R(\delta) : \delta)]$; Formula (18)
High: Upper Bound $V(\hat{a}_R(\delta); \delta)$; Formula (18)
Inflated: $V(0.2\hat{a}_R(\delta); \delta)$; Formula (28)

Additional Details:

Number of MC Draws: $R = S = 100$; Number of EIS Iterations: 15 (Needed for $\delta = 2.6$); Numbers in Parentheses are MC standard Deviations based upon 50 Replications of EIS

Table 2: EIS Results for Student t Density - Equation (33)

			Variance Estimates		
δ	$\hat{a}_R(\delta)$	$\overline{G}_{S;m}(\delta)$	low	high	inflated
3	2.003	0.9704	0.0202	0.0210	34.78
	(0.354)	(0.0306)	(0.0155)	(0.0184)	(231.1)
5	1.280	0.9857	0.0135	0.0138	6.478
	(0.148)	(0.189)	(0.0093)	(0.0104)	(41.56)
7	1.148	0.9910	0.0101	0.0102	1.663
	(0.100)	(0.0135)	(0.0065)	(0.0070)	(10.13)
9	1.096	0.9936	0.0081	0.0082	0.545
	(0.077)	(0.0105)	(0.0050)	(0.0052)	(3.084)
45	0.010	0.9991	0.0018	0.0018	0.0035
	(0.016)	(0.0021)	(0.0009)	(0.0010)	(0.0005)
145	1.000	0.9997	0.0006	0.0006	0.0010
	(0.005)	(0.0006)	(0.0003)	(0.0003)	(0.0001)

Relevant Formulae and Details: Same as for Table 1 except that we use only 5 EIS Iterations

Table 3: Stochastic Volatility; EIS-ML Estimates (IBM Data)

	$\hat{q}$	$\hat{r}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\omega}$	$\hat{\nu}$	$ln\ L(\hat{\theta}, y) \times 10^{-4}$
	0.0526	0.9698	-	0.0518	21.63	226.55	-0.10869
SG	(0.0003)	(0.0002)		(0.0005)	(0.18)	(0.62)	(0.00001)
	[0.0275]	[0.0184]		[0.0216]	[13.91]	[61.89]	
	0.0742	0.9568	-	0.0503	-	133.83	-0.10887
NG	(0.0008)	(0.0004)		(0.0004)		(1.85)	(0.00001)
	[0.0280]	[0.0177]		[0.0205]		[62.13]	
	0.0116	0.9763	0.0788	0.0535	19.82	-	-0.10864
SL	(0.0001)	(0.0002)	(0.0005)	(0.0002)	(0.09)		(0.00000^{+})
	[0.0110]	[0.0177]	[0.0261]	[0.0239]	[11.13]		
	0.0197	0.9574	0.1204	0.0514	-	-	-0.10883
NL	(0.0002)	(0.0004)	(0.0006)	(0.0005)			(0.00000^{+})
	[0.0136]	[0.0315]	[0.0311]	[0.0234]			

Notes: Number of MC Draws: $R = S = 10$; Three EIS Iterations; Numerical () and Statistical [] Standard Deviations based upon 20 MC Replications

Table 4.1: Logit: EIS Estimate of $L(\theta_0; y)$

Method	Estimate	MC St. dev.
NAT	0.8035D-74	0.4458D-74
EIS	0.9947D+54	0.0044D+54
TSE_0	0.9890D+54	0.0832D+54
TSE_1	0.9659D+54	0.0558D+54

Note: Proportionality Constants Were Ignored

Table 4.2: Logit; ML-EIS Estimates

$\hat{\sigma}_\alpha$	$\hat{\sigma}_\lambda$	$\hat{\rho}$	$L(\hat{\theta}; y)$
0.2704	0.2972	0.2683	0.1610D+55
(0.0004)	(0.0003)	(0.0013)	(0.0004D+55)
[0.0482]	[0.0226]	[0.1828]	

Notes: Sample Size: $N = 1,000$ and $T = 15$; Three EIS Iterations; Number of MC Draws: $R = S = 100$