Supporting Text

Equation numbers continue those in the article; citation numbers correspond to references given in the article.

S1 Comments on Procedure 1

Remark 1. For a stationary time series, if $\sigma > 0$ is fixed and T grows, one expects (under suitable mixing conditions such as those of Ref. 44) that $V_t(\sigma) \to V_\infty(\sigma)$. If $V_t \approx V_\infty$ for $t = 1, 2, \ldots$ then Procedure 1 gives $\hat{\theta}^{(n)} \approx \hat{\theta}_T^{(n-1)}$. On the other hand, fixing T, letting $\sigma \to 0$ and using Eq. 14, gives a rather different result of $V_t = (c^2 + t)\sigma^2\Sigma + o(\sigma^2)$. In this case,

$$\hat{\theta}^{(n)} \approx \sum_{t=1}^{T-1} \hat{\theta}_t^{(n-1)} \frac{c^2+1}{(c^2+t)(c^2+t+1)} + \hat{\theta}_T^{(n-1)} \frac{c^2+1}{c^2+T}. \tag{15}$$

A consequence of Eq. 15 is that, for sufficiently small σ, all the weights in the weighted average representation of Procedure 1 are positive. Eq. 15 also helps to explain why small values of c may lead to slow convergence, since small values of c lead to low weights for large t.

Remark 2. If the assumption in Eq. 1 is relaxed to

$$E[\theta_t|\theta_{t-1}] = \theta_{t-1} + O(\sigma^2) \tag{16}$$

then Theorem 1 holds with $\hat{\theta}_{t-1}$ in Eq. 5 replaced by $E[\theta_t|y_{1:t-1}]$. The weaker assumption in Eq. 16 may be appropriate if θ lies in a bounded set, and θ_t is constrained to stay in this set. In this case, the weighted average interpretation of Procedure 1 is lost. Our solution to boundary issues for θ_t is to reparameterize to remove the difficulty, or just to ignore the difficulty if it disappears by itself for sufficiently small σ.

S2 Initial values

The property that Procedure 1 updates as a weighted average of local parameter estimates is less appropriate when the information about a parameter is not spread out across time. A good example of such a parameter is an initial value parameter (IVP). Other situations where information about a parameter is concentrated in time, such as modeling a structural break, can be treated in a similar way. We describe θ as an IVP if $f(x_0) = f(x_0|\theta)$, but $f(x_t|x_{t-1})$ and $f(y_t|x_t)$ do not depend on θ for $t > 0$. As a particular case, if x_0 is supposed to be fixed and unknown then one can take $\theta = x_0$. There may not be any IVP in a model; for example, if x_0 is drawn from the stationary distribution of a time homogeneous Markov transition density $f(x_t|x_{t-1}, \theta)$.

For IVPs, we develop Procedure 2 based on Lemma 1. To maximize the likelihood, we introduce a prior distribution $f(\theta)$ with prior variance $\mathrm{Var}(\theta) = \sigma^2\Sigma$.

Lemma 1. *Let $\hat{\theta}_0$ be the prior mode, i.e., $\hat{\theta}_0 = \mathrm{argmax} f(\theta)$. Let $\hat{\theta}_T$ be the posterior mode, i.e., $\hat{\theta}_T = \mathrm{argmax} f(\theta|y_{1:T})$. Then*

$$f(y_{1:T}|\hat{\theta}_T) \geq f(y_{1:T}|\hat{\theta}_0).$$

Proof of Lemma 1.

$$\frac{f(y_{1:T}|\theta{=}\hat{\theta}_T)}{f(y_{1:T}|\theta{=}\hat{\theta}_0)} = \frac{f(\theta{=}\hat{\theta}_T|y_{1:T})}{f(\theta{=}\hat{\theta}_0|y_{1:T})} \times \frac{f(\theta{=}\hat{\theta}_0)}{f(\theta{=}\hat{\theta}_T)} \geq 1$$

The inequality holds by the definition of $\hat{\theta}_0$ and $\hat{\theta}_T$, since both terms in the product are at least one. □

Procedure 2. (MIF for initial values)

1. Select starting values $\hat{\theta}^{(1)}$ and σ_1, a discount factor $0 < \alpha < 1$, a fixed lag T_0 and the number of iterations N.
2. For n in $1, \ldots, N$
   (i) Evaluate $\hat{\theta}^{(n)}_{T_0}$ using $\hat{\theta}_0 = \hat{\theta}^{(n)}$ and $\sigma = \sigma_1 \alpha^{n-1}$.
   (ii) Set $\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)}_{T_0}$.
3. Take $\hat{\theta}^{(N+1)}$ to be an estimate of θ.

Approximating $f(\theta|y_{1:T})$ by $f(\theta|y_{1:T_0})$ in step 2(i) of Procedure 2 is a standard method to facilitate nonlinear filtering, termed fixed lag smoothing (1). It is certainly necessary for a particle filter implementation. The fixed lag smoothing approximation to $f(\theta|y_{1:T})$ is only reliable when the information in the data about θ is concentrated at small t values. Applying Procedure 2 to non-IVP parameters with $T_0 = T$ is a direct way to attempt inference for time-constant parameters. The difficulty of doing this in practice was exactly the motivation for developing Procedure 1. Procedure 2 is essentially an exhaustive search over a sequence of increasingly refined IVP values. An advantage of this procedure is that it fits in computationally with Procedure 1, allowing IVPs to be estimated simultaneously with other parameters.

S3 MIF via sequential Monte Carlo

S3.1 A basic SMC algorithm

Sequential Monte Carlo (SMC), also known as the "particle filter", is a numerical method for filtering and prediction. SMC has aroused considerable practical and theoretical interest since its development in the 1990s (9–13). Here we present a basic version, which is sufficient for the purposes of this article. A Monte Carlo filter draws a sample from $f(x_t|y_{1:t}, \theta)$, and similarly one-step prediction involves drawing from $f(x_{t+1}|y_{1:t}, \theta)$. SMC is based on the identities

$$\begin{aligned} f(x_t|y_{1:t}, \theta) &= \frac{f(x_t|y_{1:t-1}, \theta) f(y_t|x_t, \theta)}{\int f(x_t|y_{1:t-1}, \theta) f(y_t|x_t, \theta) dx_t} \\ f(x_{t+1}|y_{1:t}, \theta) &= \int f(x_{t+1}|x_t, \theta) f(x_t|y_{1:t}, \theta) dx_t \end{aligned}$$

which give rise to the following algorithm:

1. Suppose recursively that $X^F_{t,1}, \ldots, X^F_{t,J}$ have (approximately) a marginal density of $f(x_t|y_{1:t}, \theta)$.

2. Make $X^P_{t+1,j}$ a draw from $f(x_{t+1}|x_t{=}X^F_{t,j}, \theta)$. Then $X^P_{t+1,j}$ has (approximately) a marginal density of $f(x_{t+1}|y_{1:t}, \theta)$.

3. Now draw $X^F_{t+1,j}$ from $\{X^P_{t+1,k}\}$ with probabilities proportional to the resampling weights $w_k = f(y_t|x_t{=}X^P_{t,k}, \theta)$. $X^F_{t+1,j}$ has (approximately) a marginal density of $f(x_{t+1}|y_{1:t+1}, \theta)$. Independent draws can be used, but we use a more efficient systematic scheme (Ref. 13, Algorithm 2).

4. The conditional log likelihood at time t, defined as $\ell_t(\theta) = \log f(y_t|y_{1:t-1}, \theta)$, is estimated by $\log\left(J^{-1} \sum_{j=1}^{J} w_j\right)$.

The log likelihood is calculated via the identity $\ell(\theta) = \log f(y_{1:T}|\theta) = \sum_{t=1}^{T} \ell_t(\theta)$.

When applying Procedure 1, the time varying parameter θ_t is included in the state space, so x_t is replaced by (x_t, θ_t). $\hat{\theta}_t$ and V_t are calculated as the sample mean over the filter particles $X^F_{t,j}$ and the sample variance over the prediction particles $X^P_{t,j}$ respectively.

We used $J = 10^4$ for MIF in Table 1 and $J = 3 \times 10^4$ for MIF in Fig. 4.

S3.2 Numerical stability

If the number J of particles is not sufficiently large, the conditional distribution $f(x_t|y_{1:t})$ may not be well sampled by $\{X^F_{t,j}, j = 1, \ldots, J\}$. Put another way, there may be few (or zero) particles $X^P_{t,j}$ consistent with the observation y_t. The few consistent particles get relatively large resampling weights and dominate the evolution of the state process — an effect known as particle depletion (13). In the context of MIF, the particle filter estimates of $\hat{\theta}_t$ and V_t (say, $\hat{\theta}^e_t$ and V^e_t) then become poor. Procedure 1 is more stable if $[V_t]_{ij}$ is approximated by 0 for $i \neq j$ and by $[V^e_t]_{ii}$ for $i = j$. This forces V_t away from singularity. Supposing Σ is diagonal, Eq. 14 reassures us that V_t/σ^2 is asymptotically diagonal as $\sigma \to 0$, so the approximation is justified by theory for small σ and by practical stability concerns for large σ. For successful maximum likelihood estimation, depletion should become a negligible issue as θ approaches $\hat{\theta}$, and that matches what we found for the example of Sec. 3. When tackling problems that stretch available computational capacity, particle depletion can still be common in the early iterations of MIF, where θ may still be far from the MLE.

Even more algorithmic stability can be achieved by using the updating rule

$$\hat{\theta}^{(n+1)} = \frac{1}{T} \sum_{t=1}^{T} \hat{\theta}^{(n)}_t. \tag{17}$$

Although Eq. 17 is attractively simple and robust to particle depletion, it does not have the theoretical property of producing a sequence of estimators converging to the MLE. We found empirically that employing Eq. 17 on the first 5 iterations of MIF added stability without adversely affecting the final estimator.

S3.3 Prediction residuals via particle filters

The prediction residuals, $u_t(\hat{\theta}) = [\mathrm{Var}(y_t|y_{1:t-1},\hat{\theta})]^{-1/2}(y_t - E[y_t|y_{1:t-1},\hat{\theta}])$, can be calculated via

$$\begin{aligned}
E[y_t|y_{1:t-1}] &\approx \frac{1}{J}\sum_{j=1}^{J} E[y_t|x_t = X^P_{t,j}] \\
\mathrm{Var}(y_t|y_{1:t-1}) &= E[\mathrm{Var}(y_t|x_t)|y_{1:t-1}] + \mathrm{Var}(E[y_t|x_t] \mid y_{1:t-1}) \\
&\approx \frac{1}{J}\sum_{j=1}^{J} \mathrm{Var}[y_t|x_t = X^P_{t,j}] + \frac{1}{J-1}\sum_{j=1}^{J} (\hat{y}_{t,j} - \hat{y}_{t,\bullet})\,(\hat{y}_{t,j} - \hat{y}_{t,\bullet})'
\end{aligned}$$

where $\hat{y}_{t,j} = E[y_t|x_t = X^P_{t,j}]$ and $\hat{y}_{t,\bullet} = (1/J)\sum_{j=1}^{J} \hat{y}_{t,j}$.

S4 Some recommendations for stochastic likelihood maximization

This section describes our approach to carrying out inference based on Procedure 1. When investigating a likelihood surface, there is a trade-off between effort spent on global searching and local searching. An effective way to investigate large-scale properties of the likelihood, and simultaneously to check that the maximization procedure is successful, it to initialize the maximization at a range of parameter values. This approach is formalized in Procedure 3, below:

Procedure 3. (Investigating the likelihood surface)

1. Pick K starting values (for example, by sampling each component of θ uniformly within an assigned plausible range) and apply Procedure 1 to get K pairs $\{(\hat{\theta}_k, \ell_k)\}$ of estimates and associated log likelihoods.

2. If there is a clear global maximum – i.e., there are many pairs (θ_k, ℓ_k) with $(\max_j \ell_j - \ell_k)$ small and $|\hat{\theta}_{\mathrm{argmax}_j \ell_j} - \hat{\theta}_k|$ small – then take the MLE to be the average of these global maximum estimates.

3. If there is not a clear global maximum – many pairs (θ_k, ℓ_k) have $(\max_j \ell_j - \ell_k)$ small but $|\hat{\theta}_{\mathrm{argmax}_j \ell_j} - \hat{\theta}_k|$ not small – then some combination of the parameters is poorly identifiable. Investigate this by plotting the components of $\{\hat{\theta}_k\}$ and calculating correlations. Perhaps make extra assumptions to improve identifiability and return to step 1.

Procedure 3 requires manual oversight. This is appropriate for diagnostic checking of the maximization procedure and investigation of the global structure of the likelihood. Manual intervention is not necessary for each maximization of a profile likelihood or parametric bootstrap computation, since these require only local optimization in the neighborhood of the MLE (which is also the true parameter vector for bootstrap simulations). The only situation where local searches would be inappropriate for profile likelihood or bootstrap computations arise when the global likelihood has two (or more) separated modes of almost equal likelihood. These modes should be identified by Procedure 3 and require local maximization about each mode. Procedure 1 can be adapted for local maximization by decreasing α, c and Σ. This also demands a smaller value of N, the number of iterations, which is helpful for implementing these computationally intensive finite sample procedures.

One subtlety in Procedure 3 is the use of the average in step 2. In our applications, the Monte Carlo error in evaluating the likelihood is typically large compared to the actual difference in the likelihood between MIF estimates that have converged to the same mode. This occurs because MIF seeks the maximum by averaging Monte Carlo error over many iterations. Thus, we chose to average MIF estimates rather than to take the one with the highest evaluated likelihood.

To implement step 2 of Procedure 3 one must determine what is meant by "small". As this procedure is intended to be used on a broad variety of models, we think automation is premature. A general observation is that "small" differences in the likelihood are of the order of one unit of log likelihood.

Some simple methods are available to check that the likelihood is being maximized effectively on simulated data, with a known parameter vector θ^*. Setting $\hat{\theta} = \arg\max \ell(\theta)$, an asymptotic result for regular parametric models is that $2(\ell(\hat{\theta}) - \ell(\theta^*))$ has approximately the distribution of $\chi^2(d_\theta)$, a chi-squared random variable on d_θ degrees of freedom (34). Thus, beyond the basic property that $\ell(\hat{\theta}) \geq \ell(\theta^*)$, one can expect $\ell(\hat{\theta}) - \ell(\theta^*) \approx d_\theta/2$. If estimates of the maximized log likelihood compared with the likelihood at θ^* are not unusual for $(1/2)\chi^2(d_\theta)$, we view this as some evidence for successful maximization. The sliced likelihood plots described in Sec. 7 give the formal demonstration of successful maximization, but require extra computation.

S5 Sufficient conditions for convergence of iterated filtering

Theorem 2 provides a complementary result to Theorem 1, giving sufficient conditions on the sequence $\sigma_n \to 0$ for Procedure 1 to convergence successfully. Although stated as a global result, Theorem 2 implies corresponding local behavior that is more relevant in practice.

Theorem 2. *Suppose that $\ell(\theta)$ is twice continuously differentiable, with a uniform convexity property that there exist $0 > a > b$ such that*

$$a > u'\nabla^2\ell(\theta)u > b \quad \textit{for all } \theta \textit{ and all unit vectors, } |u| = 1. \tag{18}$$

Define the sequence $\{\hat{\theta}^{(n)}\}$ by a stochastic difference equation,

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} + \sigma_n^2 M(\nabla\ell(\hat{\theta}^{(n)}) + \eta_n). \tag{19}$$

Take $M = (c^2+1)\Sigma$, so that M is a positive definite symmetric matrix and $\sigma_n^2 M = V_{1,n}$ in the notation of Theorem 1. Suppose that $\lim_n \sigma_n^2 n^{1-\beta} > 0$ for some $\beta \in (0,1)$. Suppose also that the sequence $\{\eta_n\}$ has $E[\eta_n] = o(1)$, $\mathrm{Var}(\sigma_n^2\eta_n) = o(1)$ and $\mathrm{Cov}(\eta_m, \eta_n) = 0$ for $m \neq n$. If there is a $\hat{\theta}$ with $\nabla\ell(\hat{\theta}) = 0$ then $\hat{\theta}^{(n)}$ converges in probability to $\hat{\theta}$.

To see how Theorem 2 applies to MIF, implemented using a Monte Carlo filter, we need some assumptions. We suppose that the Monte Carlo filter is unbiased: this is not quite true for sequential Monte Carlo with a finite sample size, but it become exactly true if we accept the goal of maximizing the expected Monte Carlo log likelihood rather than the true log likelihood. Theorem 1 then gives $E[\eta_n] = o(1)$ as long as $\sigma_n \to 0$; we have to assume that this convergence is uniform over θ. A reasonable model for the variance of a derivative based on Monte Carlo likelihood evaluations in a neighborhood of size σ_n is $\mathrm{Var}(\eta_n) = O(\sigma_n^{-2})$, implying the condition $\mathrm{Var}(\sigma_n^2 \eta_n) = o(1)$. Formally, to apply Theorem 2, one must assume that this rate is also uniform over θ. If the Monte Carlo filter uses independent sequences of random numbers for each iteration, $\mathrm{Cov}(\eta_m, \eta_n) = 0$ for $m \neq n$.

Proof of Theorem 2. The fundamental theorem of calculus gives

$$\nabla \ell(\theta) = \int_0^1 \nabla^2 \ell(s\theta + (1-s)\hat{\theta})(\theta - \hat{\theta})\, ds.$$

This can also be written as $\nabla \ell(\theta) = H(\theta)(\theta - \hat{\theta})$ where $H(\theta) = \int_0^1 \nabla^2 \ell(s\theta + (1-s)\hat{\theta})\, ds$. We re-write Eq. 19 as

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} + \sigma_n^2 M(H_n(\hat{\theta}^{(n)} - \hat{\theta}) + \eta_n) \tag{20}$$

where $H_n = H(\hat{\theta}^{(n)})$. Eq. 20 can be written as

$$\begin{aligned} \hat{\theta}^{(n+1)} - \hat{\theta} &= \prod_{k=1}^{n} (I + \sigma_k^2 M H_k)(\hat{\theta}^{(1)} - \hat{\theta}) \\ &\quad + \sum_{m=1}^{n-1} \Big\{ \prod_{k=m+1}^{n} (I + \sigma_k^2 M H_k) \Big\} \sigma_m^2 M \eta_m + \sigma_n^2 M \eta_n. \end{aligned} \tag{21}$$

$H(\theta)$ satisfies the same inequality in Eq. 18 as $\nabla^2 \ell(\theta)$, which guarantees a uniform bound on the eigenvalues of $\sigma_k^2 M H_k n^{1-\beta}$. Lemma 2, with A taken to be $\sigma_k^2 M H_k$, then secures the existence of a constant $c > 0$ such that, for sufficiently large k,

$$\log |I + \sigma_k^2 M H_k| < -c k^{\beta - 1}.$$

A comparison of $\sum_{k=m}^{n} k^{\beta-1}$ with $\int_m^n x^{\beta-1} dx$ then gives

$$\log \prod_{k=m}^{n} |I + \sigma_k^2 M H_k| < c\beta^{-1}(m^\beta - n^\beta). \tag{22}$$

Lemma 3 can be applied to Eq. 22 to demonstrate that

$$\sum_{m=1}^{n-1} |\sigma_m^2| \prod_{k=m+1}^{n} |I + \sigma_k^2 M H_k| = O(1).$$

Lemma 4 can then be applied, with $w_{m,n-1} = |\sigma_m^2| \prod_{k=m+1}^{n} |I + \sigma_k^2 M H_k|$ and $b_n = E[\eta_n]$. This gives

$$E\Big[\sum_{m=1}^{n-1} \Big\{ \prod_{k=m+1}^{n} (I + \sigma_k^2 M H_k) \Big\} \sigma_m^2 \eta_m\Big] \to 0. \tag{23}$$

A very similar argument, replacing $w_{m,n-1}$ by $|\sigma_m^2| \prod_{k=m+1}^n |I+\sigma_k^2 MH_k|^2$ and b_n by $\mathrm{Var}(\sigma_n^2 \eta_n)$, allows the use of Lemma 4 to give

$$\mathrm{Var}\Big[\sum_{m=1}^{n-1}\Big\{\prod_{k=m+1}^{n}(I+\sigma_k^2 MH_k)\Big\}\sigma_m^2\eta_m\Big] \to 0. \tag{24}$$

In addition, Eq. 22 implies that

$$\prod_{k=1}^{n}(I+\sigma_k^2 MH_k)(\hat{\theta}^{(1)}-\hat{\theta}) \to 0. \tag{25}$$

Eq. 23, Eq. 24 and Eq. 25 imply convergence in probability for Eq. 21, which completes the proof. □

Lemma 2. *If A is a negative definite matrix with $|A| < 1$ and with largest eigenvalue $\pi < 0$ then $\log|I+A| < \pi$.*

Proof. Let u be an arbitrary vector with $|u| = 1$.

$$\begin{aligned}\log|I+A| &= \log(\sup_u |u'(I+A)u|)\\ &= \log(\sup_u |1+u'Au|)\end{aligned}$$

By hypothesis $u'Au > -1$, and so $\sup_u |1+u'Au| = 1+\sup_u u'Au$. Therefore,

$$\log|I+A| \;=\; \log(1+\sup_u u'Au) = \log(1+\pi) < \pi,$$

where we use the inequality $\log(1+\pi) < \pi$ for $\pi \in (-1,0)$. □

Lemma 3. *If $c > 0$ and $0 < \beta < 1$ then*

$$\sum_{m=1}^{n}\exp\{c(m^\beta - n^\beta)\}m^{\beta-1} = O(1). \tag{26}$$

Proof. We write the sum in Eq. 26 as

$$n^\beta \frac{1}{n}\sum_{m=1}^{n}\exp\{-c(1-(m/n)^\beta)n^\beta\} \times \Big(\frac{m}{n}\Big)^{\beta-1}. \tag{27}$$

As $n \to \infty$, Eq. 27 can be compared to the integral

$$n^\beta \int_0^1 \exp\{-c(1-x^\beta)n^\beta\}x^{\beta-1}\,dx.$$

This can be analyzed in two parts. Firstly,

$$\begin{aligned}n^\beta \int_0^{1/2}\exp\{-c(1-x^\beta)n^\beta\}x^{\beta-1}\,dx &< n^\beta \int_0^{1/2}\exp\{-(1-(1/2)^\beta)cn^\beta\}x^{\beta-1}\,dx\\ &= n^\beta \exp\{-(1-(1/2)^\beta)cn^\beta\}(1/2)^\beta/\beta \to 0. \quad (28)\end{aligned}$$

For the second part, change variable to $y = (1 - x^\beta)$ and proceed as follows:

$$\begin{aligned} n^\beta \int_{1/2}^{1} \exp\{-c(1-x^\beta)n^\beta\} x^{\beta-1}\, dx &= n^\beta \int_0^{1-(1/2)^\beta} \exp\{-cyn^\beta\}\beta x^{2(\beta-1)}\, dy \\ &< n^\beta 2^{2(1-\beta)} \int_0^\infty \exp\{-cyn^\beta\}\, dy = 2^{2(1-\beta)}/c. \quad (29) \end{aligned}$$

Eq. 28 and Eq. 29 together yield the required result. □

Lemma 4. *Suppose $b_n \to 0$ and $\sum_{m=1}^n |w_{m,n}| < C$ with $w_{m,n} \to 0$ as $n \to \infty$ for each m. Then $\sum_{m=1}^n b_n w_{m,n} \to 0$.*

Proof. b_n is bounded, say $|b_n| < K$. For $\epsilon > 0$, $\exists n_0 : |b_n| < \epsilon\ \forall n > n_0$. Also, $\exists n_1 : |w_{m,n}| < \epsilon/n_0$ whenever $m \leq n_0$ and $n > n_1$. Then, for $n > \max(n_0, n_1)$, $|\sum_{m=1}^{n_0} b_n w_{m,n}| < K\epsilon$ and $|\sum_{m=n_0+1}^{n} b_n w_{m,n}| < C\epsilon$. Thus, $|\sum_{m=1}^{n} b_n w_{m,n}| < (K + C)\epsilon$. □

S6 Standard errors and confidence intervals

The Fisher information can be estimated by

$$\hat{\mathcal{I}}_{ij} = \sum_{t=1}^{T} \frac{\partial}{\partial \theta_i} \log f(y_t | y_{1:t-1}, \hat{\theta}) \frac{\partial}{\partial \theta_j} \log f(y_t | y_{1:t-1}, \hat{\theta}) \quad (30)$$

leading to corresponding standard errors $\mathrm{SE}(\hat{\theta}_i) = [\hat{\mathcal{I}}^{-1/2}]_{ii}$. Procedure 4 details how this was implemented in this article.

Procedure 4. (Standard errors)

1. Let $\hat{\theta}$ be the sample mean of the (vector) estimates $\{\hat{\theta}_k, k = 1, \ldots, K\}$ from K replications of Procedure 1. Calculate $\ell_{t,ij} = \log f(y_t | y_{1:t-1}, \hat{\theta} + h_{ij}\delta_i)$ for $1 \leq i \leq m$ and $1 \leq j \leq q$, where δ_i is a vector of zeros with a one in the i^{th} position. $\{h_{ij}\}$ can be the offsets used for a sliced likelihood diagnostic plot. Alternatively, one can use $q = 2$ with $h_{i1} = 0$ and $h_{i2} = h\sqrt{\Phi_{ii}}$, where Φ is the sample covariance matrix of $\{\hat{\theta}_k\}$. The constant h is chosen by trial and error, and Φ gives the relative scale of the uncertainty in the components of θ.

2. Regress $\ell_{t,ij}$ on h_{ij} for each i, giving rise to regression coefficients $\dot{\ell}_{t,i}$ with variance estimates $\hat{\mathrm{Var}}(\dot{\ell}_{t,i})$.

3. Estimate the Fisher information by $\hat{I}_{ij} = \sum_t \dot{\ell}_{t,i}\dot{\ell}_{t,j}$ and estimate the derivative of the log likelihood at $\hat{\theta}$ by $\dot{\ell}_i = \sum_{t=1}^T \dot{\ell}_{t,i}$.

Procedure 4 step 2 calculates numerical derivatives, averaging over a neighborhood given by $\{h_{ij}\}$. If $\{h_{ij}\}$ are too small, the Monte Carlo error in the likelihood evaluation will dominate the numerical derivative. Since $E[\dot{\ell}_{t,i}] \approx \partial\ell/\partial\theta_i$, $\sum_{t=1}^T E[\dot{\ell}_{t,i}^2] \approx \sum_{t=1}^T \{(\partial\ell/\partial\theta_i)^2 + \mathrm{Var}(\dot{\ell}_{t,i})\}$. Thus the bias of $\hat{I}_{ii}$ as an estimator of I_{ii} is approximately $\sum_{t=1}^T \hat{\mathrm{Var}}(\dot{\ell}_{t,i})$. We monitor this quantity and trust the estimate $\hat{I}_{ii}$ only if $\hat{I}_{ii} \gg \sum_{t=1}^T \hat{\mathrm{Var}}(\dot{\ell}_{t,i})$. Otherwise, either the neighborhood used to calculate the numerical derivative or the Monte Carlo sample size must be increased. There could be some advantage in calculating the numerical derivatives in the directions of the eigenvectors of Φ, with the eigenvalues giving the appropriate scaling in each direction. We prefer not to do this, since K is not necessarily large compared to m. In particular, if $K \leq m$ then Φ is singular.

Note that one can use $\tilde{\theta} = \hat{\theta} + \hat{I}^{-1}\dot{\ell}$ as a possibly improved parameter estimate, based on a quadratic approximation to the local likelihood surface (35). However, $\tilde{\theta}$ involves the potentially inaccurate Monte Carlo derivative estimates that MIF carefully avoids, and in our experience $\hat{\theta}$ is more reliable for the situation arising in this article.

Standard errors are usually interpreted in the context of a normal approximation for the MLE: one is invited to think of $\hat{\theta}_i \pm 2\,\mathrm{SE}(\hat{\theta}_i)$ as an approximate 95% confidence interval. A more accurate confidence interval comes from the profile log likelihood (34). Profile likelihoods can be calculated using MIF, but at considerably more computational expense than the SEs from Procedure 4. If θ is partitioned into two components ζ and η, of dimensions d_ζ and d_η respectively, then the profile log likelihood of η is defined by $\ell_{(p)}(\eta) = \sup_\zeta \ell(\zeta, \eta)$. An approximate 95% confidence interval for η is given by $\{\eta : 2[\ell_{(p)}(\hat{\eta}) - \ell_{(p)}(\eta)] < \chi^2_{0.95}(d_\eta)\}$ where $\chi^2_{0.95}(d_\eta)$ is the 0.95 quantile of a χ^2 random variable on d_η degrees of freedom, and $\hat{\eta} = \operatorname{argmax} \ell_{(p)}(\eta)$.