Sequential Numerical Integration in Nonlinear State Space Models for Microeconometric Panel Data

Author(s): Florian Heiss

Source: *Journal of Applied Econometrics*, Vol. 23, No. 3 (Apr., 2008), pp. 373-389
Published by: Wiley
Stable URL: http://www.jstor.org/stable/25144553
Accessed: 25-06-2018 17:01 UTC

JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **23**: 373–389 (2008)
Published online 28 February 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/jae.993

SEQUENTIAL NUMERICAL INTEGRATION IN NONLINEAR STATE SPACE MODELS FOR MICROECONOMETRIC PANEL DATA

FLORIAN HEISS*
University of Munich, Department of Economics, Ludwigstr. 28 RG, 80539 Munich, Germany

SUMMARY

This paper discusses the estimation of a class of nonlinear state space models including nonlinear panel data models with autoregressive error components. A health economics example illustrates the usefulness of such models. For the approximation of the likelihood function, nonlinear filtering algorithms developed in the time-series literature are considered. Because of the relatively simple structure of these models, a straightforward algorithm based on sequential Gaussian quadrature is suggested. It performs very well both in the empirical application and a Monte Carlo study for ordered logit and binary probit models with an AR(1) error component. Copyright © 2008 John Wiley & Sons, Ltd.

Received 28 June 2006; Revised 27 May 2007

1. INTRODUCTION

Many microeconometric models, especially limited dependent variable (LDV) models, are inherently nonlinear. This nonlinearity complicates the analysis of panel data models, for a general discussion see for example Chamberlain (1984). In applied microeconometric research, the time-series dimension of the panel data is therefore often modeled in a very simple way to keep the computations tractable. A leading example are random effects models, where the error term is split into an i.i.d. and a time-constant part.

This paper discusses the approximation of the likelihood function for a class of models which can be represented in a simple state space framework. An important example of such models are LDV models which replace the assumption of a time-constant random effect with a random process of unobserved influences. The discussed algorithms can be divided into two classes. One approach is to numerically integrate out the full sequence of unobserved states. Simulation is a natural way to do this.

Another approach is to use a nonlinear version of the Kalman filter. Outside of economics, these nonlinear filtering approaches are widely studied e.g. in engineering (Doucet, De Freitas, and Gordon 2001). In the econometric time-series literature, they have been discussed e.g. by Danielsson and Richard (1993), Tanizaki and Mariano (1994), Shephard and Pitt (1997), Durbin and Koopman (2002), and Fernández-Villaverde and Rubio-Ramírez (2006). For a survey of these methods, see Durbin and Koopman (2001) and Tanizaki (2003). The models for which these algorithms are usually applied have a much more complicated structure than the type of

* Correspondence to: Florian Heiss, Department of Economics, University of Munich, Ludwigstr. 28 RG, 80539 Munich, Germany. E-mail: mail@florian-heiss.de

models considered in this paper, see for example Fernández-Villaverde and Rubio-Ramírez (2005), Koopman and Lucas (2005), and Bauwens and Hautsch (2006). The relatively simple structure of the models allows to express the likelihood function as a product of univariate integrals. Therefore, a nonlinear filtering algorithm is suggested which is based on the sequential application of Gaussian quadrature rules.

As an example for the type of models considered here, a health economics application is presented. For studying the evolution of health over time, the literature has so far focused on first-order Markov chain and random effects models. Contoyannis, Jones, and Rice (2004) thoroughly discuss these models and their estimation. It is argued that a model with an AR(1) error term is theoretically more convincing. Furthermore it is more parsimonious and captures the observed intertemporal correlation pattern much better.

Different algorithms are implemented for the illustrative health model. While all converge to the same results as the computational effort is increased, the speed of this convergence differs notably. The nonlinear filtering algorithms are much more accurate with a given level of computational costs than the methods based on joint simulation. The sequential Gaussian quadrature (SGQ) approach clearly performs best in these exercises.

A Monte Carlo study confirms that the relative advantages of the algorithms persist for different parameterizations of the same model. Finally, Monte Carlo results are presented for a panel probit model for which the specialized Geweke-Hajivassilou-Keane (GHK) simulator is available. This algorithm performs better than partially analytic simulators but worse than the nonlinear filtering approaches, in particular than SGQ.

The paper is structured as follows. Section 2 specifies the class of state space models and discusses approaches to the approximation of their likelihood functions. An example application is presented in Section 3. The accuracy of the likelihood approximation and the estimated parameters is compared for different algorithms. Monte Carlo simulations are used to study the determinants of the approximation accuracy in different model settings for ordered logit and binary probit models with an AR(1) error component in Section 4. Section 5 summarizes and concludes.

2. MICROECONOMETRIC STATE SPACE MODELS

Consider a limited dependent variables (LDV) panel data model with a random effect

$$\begin{aligned} y_{it}^* &= \mathbf{x}_{it}\boldsymbol{\beta} + a_i + e_{it} \\ y_{it} &= G(y_{it}^*) \end{aligned} \tag{1}$$

where $i = 1, \ldots, N$ indicates cross-sectional units (say individuals) and $t = 1, \ldots, T$ indicates the time dimension of the data. The latent variable y_{it}^* is driven by a vector of explanatory variables $\mathbf{x}_{it}$, a time-constant individual random effect a_i and an i.i.d. error term e_{it}. The latter two random variables are assumed to be mutually independent and independent of $\mathbf{x}_{it}$ and to have a parametric distribution. The observed dependent variable y_{it} is a parametric function of this latent variable.

This paper considers generalizations of this class of models in which the time-constant random effect a_i is replaced with a stochastic process $\mathbf{a}_{i,1:T} = [a_{it} : t = 1, \ldots, T]$. These models can be interpreted in the state space framework with a_{it} representing latent states. Section 2.1 defines the structure of and requirements on these models and sections 2.2 and 2.3 discuss the numerical approximation of the likelihood function.

J. Appl. Econ. **23**: 373–389 (2008)
DOI: 10.1002/jae

2.1. Model Specification

Suppose a sequence of dependent variables $\mathbf{y}_{it}$ is observed a number of times $t = 1, \ldots, T$ for a number of cross-sectional units $i = 1, \ldots, N$. In the following discussion, assume that T is the same number for each cross-sectional unit so that we are dealing with a balanced panel. This is merely for notational convenience—unbalanced panels can easily be dealt with if the individual number of observations is random or modeled jointly.

While the vectors of dependent variables $\mathbf{y}_{it}$ are one-dimensional in many applications, the more general case does not create any complications neither in the notation nor in the analysis. The vector of dependent variables may consist of discrete, continuous, or both types of random variables. Let $\mathbf{x}_i$ denote a vector of explanatory variables which may contain time-varying strictly exogenous variables. In this case, $\mathbf{x}_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}]$ collects all time-specific values.

Unobserved heterogeneity creates a correlation of the dependent variables over time, conditional on $\mathbf{x}_i$. It is captured by a sequence of scalar latent 'states' a_{it} which are correlated over time, but not necessarily perfectly. In this sense, random effects are included as the special case in which $\mathbf{a}_{i,1:T} = [a_{it} : t = 1, \ldots, T]$ has a degenerate joint distribution with $a_{it} = a_i$ for all $t = 1, \ldots, T$. In the more general case, it could for example follow an AR(1) process.

Assume that the marginal distribution of a_{it} is known up to a finite set of parameters included in the general parameter vector $\boldsymbol{\theta}$. For notational simplicity, assume that these marginal distributions are identical for all $i = 1, \ldots, N$ and $t = 1, \ldots, T$ and denote its p.d.f. conditional on the exogenous covariates as $f(a_{it}|\mathbf{x}_i)$. For identification of the model, it will in many cases be necessary to assume independence of $\mathbf{x}_i$ analogous to random effects models.

As noted before, states are allowed to be dependent over time. For notational and analytical convenience, assume that they are first-order Markov. Also assume that there is no feedback from the sequence of dependent variables $\mathbf{y}_{i,1:T} = [\mathbf{y}_{it} : t = 1, \ldots, T]$. Therefore, the conditional p.d.f. can be written as

$$f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}, \mathbf{a}_{i,1:t-1}) = f(a_{it}|\mathbf{x}_i, a_{i,t-1}) \tag{2}$$

This conditional distribution is assumed to be known up to parameters.

Just like in a typical random-effects model, $\mathbf{x}_i$ and a_{it} do not perfectly determine the dependent variable due to i.i.d. error terms. Let $P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}, \mathbf{a}_{i,1:T})$ represent the joint probability mass (or density) of $\mathbf{y}_{it}$ conditional on the explanatory variables, the sequence of states, and past values of the dependent variable. Make the conditional independence assumption

$$P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}, \mathbf{a}_{i,1:T}) = P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it}) \quad \forall i = 1, \ldots, N, t = 1, \ldots, T \tag{3}$$

Conditional on $\mathbf{x}_i$ and the contemporaneous value of the latent state a_{it}, the outcome probability of $\mathbf{y}_{it}$ is independent of both past and future values of the state process $\mathbf{a}_{i,1:T}$ and lagged dependent variables. The latter assumption avoids the usual initial value problems which could be dealt with using common approaches, see Heckman (1981), Wooldridge (2005), and Honoré and Tamer (2006). Under this assumption, all contemporaneous correlation of $\mathbf{y}_{it}$ conditional on $\mathbf{x}_i$ is generated by the correlation of latent states $\mathbf{a}_{i,1:T}$.

Assume that this conditional probability $P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})$ is a known parametric function. Section 3 provides an example how an ordered logit model with an AR(1) error component is represented in this more general model specification.

DOI: 10.1002/jae

2.2. Approximation of the Likelihood Contributions by Joint Simulation

Suppose we are interested in likelihood-based estimation such as maximum likelihood or Bayesian analysis. With $P(\mathbf{y}_{i,1:t}|\mathbf{x}_i;\boldsymbol{\theta})$ denoting the joint probability mass (or probability density) of $\mathbf{y}_{i,1:t} = [\mathbf{y}_{is} : s = 1, \ldots, t]$ conditional on $\mathbf{x}_i$, and collecting all model parameters in the vector $\boldsymbol{\theta}$, the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} P(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta}) \tag{4}$$

A marginalization of the conditional outcome probabilities specified in (3) with respect to $\mathbf{a}_{i,1:T}$ is generally not feasible, so the expressions $P(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta})$ have to be approximated numerically.

The simplest approach is to numerically integrate out the full latent process $\mathbf{a}_{i,1:T}$

$$P(\mathbf{y}_{i,1:T}|\mathbf{x}_i) = \int \cdots \int P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{a}_{i,1:T}) f(\mathbf{a}_{i,1:T}|\mathbf{x}_i) da_{i1} \cdots da_{iT} \tag{5}$$

By (3), $P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{a}_{i,1:T}) = \prod_{t=1}^{T} P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})$. So all terms in this equation are known by assumption, but the integral does in general not have an analytic solution. The typical approach in applied microeconometrics to these kinds of problems is simulation: Generate a number of R draws $[\mathbf{a}_{i,1:T}^r : r = 1, \ldots, R]$ from the joint distribution $f(\mathbf{a}_{i,1:T}|\mathbf{x}_i)$. The simulated probability is equal to

$$\tilde{P}^{\mathrm{SIM}}(\mathbf{y}_{i,1:T}|\mathbf{x}_i) = \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T} P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it}^r) \tag{6}$$

Pseudo-maximum likelihood estimators of $\boldsymbol{\theta}$ using $\tilde{P}^{\mathrm{SIM}}(\mathbf{y}_{i,1:T}|\mathbf{x}_i)$ instead of its true value are under weak regularity conditions consistent (in N) if the number of replications R rises with N (Hajivassiliou and Ruud 1994).

It has been shown in various cases that with a given number of replications R, the accuracy of the simulated probabilities and estimators based on them improves if instead of (pseudo-)random draws antithetic or quasi-random draws are used. A further general approach to improve the computational efficiency of simulation estimators is the use of importance sampling instead of drawing from the joint distribution $f(\mathbf{a}_{i,1:T}|\mathbf{x}_i)$, see for example Richard and Zhang (2005).

Instead of simulation, deterministic numerical integration methods can be used to approximate analytically infeasible integrals. While Gaussian quadrature is known to work effectively in univariate integration problems, the integral in (5) is T-dimensional even though a_{it} is assumed to be one-dimensional. The well-known product rule extension of Gaussian quadrature to multiple dimensions suffers from exponentially rising computational costs as the number of dimensions increases.

Heiss and Winschel (2006) use a different general approach of extending Gaussian quadrature to multiple dimensions. Instead of approximating the integrand by a multivariate polynomial with a bound on the maximal exponent, this method of integration on sparse grids (SGI) approximates it by a complete polynomial.[1] This leads to much slower increase of the computational costs with a small decrease of approximation accuracy.

[1] In two dimensions, a product rule of order 2 would approximate the integrand with a polynomial including terms of x_1, x_2, x_1^2, x_2^2, x_1x_2, $x_1^2x_2$, $x_1x_2^2$, and $x_1^2x_2^2$. A complete polynomial bounds the sum of exponents, so the higher-order terms $x_1^2x_2$, $x_1x_2^2$, and $x_1^2x_2^2$ would be ignored.

DOI: 10.1002/jae

The higher the time-series dimension of the data, the worse can all methods of integrating out the full sequence of latent states be expected to work. This is extremely problematic for the Gaussian integration based on the product rule. But also for SGI, the computational burden rises with the dimensions of integration. While the asymptotic (in R) properties of simulation estimators do not depend on the dimension, the accuracy given a finite number of replications often does, see e.g. Lee (1997) for Monte Carlo results for simulation estimators.

2.3. Nonlinear Filtering

Nonlinear filter techniques separate the full integral in (5) into a sequence of lower-dimensional integrals using the structure of the model. These approaches can be interpreted as a generalization of the Kalman filter to nonlinear models with possibly nonnormal disturbances.

The general idea how to decompose the integral in (5) for the model structure described in Section 2.1 is as follows. By the rules of conditioning, the probabilities of interest can in general be written as

$$P(\mathbf{y}_{i,1:T}|\mathbf{x}_i) = P(\mathbf{y}_{i1}|\mathbf{x}_i)\prod_{t=2}^{T} P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}) \tag{7}$$

The initial probability can be expressed in a straightforward way as the univariate integral

$$P(\mathbf{y}_{i1}|\mathbf{x}_i) = \int P(\mathbf{y}_{i1}|\mathbf{x}_i, a_{i1}) f(a_{i1}) da_{i1} \tag{8}$$

By (3), the outcome probabilities conditional on past values in (7) can be written as

$$P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}) = \int P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it}) f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}) da_{it} \tag{9}$$

This equation reflects the model assumption that all dependence of $\mathbf{y}_{it}$ conditional on $\mathbf{x}_i$ is induced by the presence of the latent state process of a_{it}. Now all these probabilities are expressed as univariate integrals. The problem with their numerical approximation is that the conditional distribution $f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ is a complicated function of past realizations $\mathbf{y}_{i,1:t-1}$.

For general nonlinear state space models, there are various approaches to this problem, for an overview see Durbin and Koopman (2001) and Tanizaki (2003). Typical applications come from engineering, finance or macroeconomics. These models are typically much more involved than the class of models considered here. The most important difference is that the state space and therefore the integrals in (9) in this paper is restricted to be one-dimensional which motivates the use of Gaussian quadrature for its approximation as discussed below.

The nonlinear particle filter (NPF) makes random draws from the conditional distribution $f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ and approximates the integrals in (9) by simulation. The draws are obtained sequentially using a resampling algorithm, see Doucet, De Freitas, and Gordon (2001) and for an application to econometric time series models Fernández-Villaverde and Rubio-Ramírez (2006). Because of the discrete nature of this resampling step, the resulting approximated likelihood is not smooth in the model parameters. This impedes gradient-based maximization algorithms for the likelihood function.

Other algorithms use some sort of importance sampling where values for a_{it} are not drawn from their conditional distribution but from some other distribution and a weighting scheme

DOI: 10.1002/jae

is sequentially updated to capture the conditioning on the data and the difference between the conditional and the sampling distribution. Tanizaki and Mariano (1994) use such a sequential importance sampling (SIS) algorithm for nonlinear state space models but provide no general rule for the best choice of the sampling distribution. For the efficient choice of a sampling distribution, see for example DeJong, Dharmarajan, Liesenfeld, and Richard (2006), Durbin and Koopman (2002) and Shephard and Pitt (1997).

As noted above, the types of models this paper is concerned with has a one-dimensional state space, for example an AR(1) error component. Gaussian quadrature is known to work well for many univariate numerical integration problems—for the random effects probit model, see Butler and Moffit (1982). The algorithm prescribes a set of R nodes $[z^r : r = 1, \ldots, R]$ and corresponding weights $[w^r : r = 1, \ldots, R]$ for a general univariate integration problem of the form $\int g(z)w(z)dz$ which depend on the weighting function $w(z)$. The approximation is then given as $\sum_{r=1}^{R} w^r g(z^r)$. It will be the exact solution of the integral if $g(z)$ is a polynomial of order $2R - 1$ or less. If the integrand is reasonably smooth and can therefore be closely approximated by a polynomial, the approximation can be expected to be very accurate.

The sequential quadrature algorithm suggested for the integration problems in (9) uses a reweighting scheme similar to sequential importance sampling algorithms. Define

$$q_{it}(a_{it}) = \frac{f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})}{f(a_{it}|\mathbf{x}_i)} \tag{10}$$

and rewrite (9) as

$$P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}) = \int q_{it}(a_{it})P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})f(a_{it}|\mathbf{x}_i)da_{it} \tag{11}$$

This corresponds to an importance sampling approach with the "proposal density" $f(a_{it}|\mathbf{x}_i)$. As long as $q_{it}(a_{it})P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})$ is known and smooth in a_{it}, Gaussian quadrature can be applied with a rule appropriate for the known p.d.f. $f(a_{it}|\mathbf{x}_i)$.

Approximating (11) with Gaussian quadrature requires $f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ to be evaluated for all quadrature nodes $a_{it} \in \{a^1, \ldots, a^R\}$. These values can be constructed in a sequential fashion. Start with $t = 1$ and denote $f(a_{i1}|\mathbf{x}_i, \mathbf{y}_{i,1:0}) = f(a_{i1}|\mathbf{x}_i)$ and $P(\mathbf{y}_{i1}|\mathbf{x}_i, \mathbf{y}_{i,1:0}) = P(\mathbf{y}_{i1}|\mathbf{x}_i)$. For $t = 1$, $f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ is known and $P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ can be approximated. With this initialization, suppose this is true for some t and all $a_{it} \in \{a^1, \ldots, a^R\}$ and derive $f(a_{i,t+1}|\mathbf{x}_i, \mathbf{y}_{i,1:t})$ in two steps so that $P(\mathbf{y}_{i,t+1}|\mathbf{x}_i, \mathbf{y}_{i,1:t})$ can be approximated.

First, note that Bayes' rule and the model assumption (3) imply

$$f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t}) = f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})\frac{P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})}{P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})} \tag{12}$$

This gives an expression of the state density conditional on past and current values of the dependent variable. For the next recursion step, we need the density of the state in the next period conditional on this information. The conditional independence assumption (2) implies

$$f(a_{i,t+1}|\mathbf{x}_i, \mathbf{y}_{i,1:t}) = \int f(a_{i,t+1}|\mathbf{x}_i, a_{it})f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t})da_{it} \tag{13}$$

DOI: 10.1002/jae

Combining this result with (12) and (10) yields

$$f(a_{i,t+1}|\mathbf{x}_i, \mathbf{y}_{i,1:t}) = \int q_{it}(a_{it}) f(a_{i,t+1}|\mathbf{x}_i, a_{it}) \frac{P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})}{P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})} f(a_{it}|\mathbf{x}_i) da_{it} \tag{14}$$

This univariate integral can again be approximated for each value of $a_{i,t+1} \in \{a^1, \ldots, a^R\}$ by Gaussian quadrature with a rule appropriate for $f(a_{it}|\mathbf{x}_i)$. The term $q_{it}(a_{it})$ as well as $P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})$ are know from the previous step, $P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ has been approximated in the previous step and both $f(a_{i,t+1}|\mathbf{x}_i, a_{it})$ and $f(a_{it}|\mathbf{x}_i)$ are known by assumption.

Whether this algorithm works well depends on how well the integrands in (11) and (14) can be approximated by a low-order polynomial. Since this depends on the model structure and parameters, this method is implemented and compared to alternative approaches for an empirical example in Section 3 and in Monte-Carlo studies in Section 4.

3. EMPIRICAL EXAMPLE

3.1. Background and Data

One of the most frequently studied measures of individual health is the self-rated health status (SRHS). Many surveys include it as the answer to a question like "Would you say your health is excellent, very good, good, fair, or poor?". Despite its obvious subjectiveness, it has been found a useful and powerful measure. It maps the high-dimensional and complex concept of health into one dimension using individual perceptions and judgments. It is also a very powerful predictor of objective events such as mortality.

The data used for the empirical example are from the Health and Retirement Study (HRS).[2] The HRS contains data on different cohorts of elderly Americans. The sample consists of all cohorts with the only restriction that they are at least 50 years old at the time of the first interview. This includes 25,353 respondents with up to 6 observations over time each. A total of 102,233 observations are available.

Table I shows the distribution of SRHS in the sample. As the tabulations conditional on the previous response indicate, SRHS is highly correlated over time. In the literature, this correlation is almost exclusively modeled as time-constant unobserved heterogeneity and/or state dependence of SRHS. Contoyannis, Jones, and Rice (2004) discuss and compare these approaches.

Table II gives an impression on the intertemporal correlation pattern over a longer period of time. It shows the results of an ordered logit regression of SRHS in wave 6 on a typical set of covariates plus lagged values of SRHS.[3] The two most interesting results are that the coefficients of all lags (i) are all highly significantly different from zero and (ii) get smaller the further away the respective observation is from wave 6.

A random effects model would imply equal predictive power of all lags which contradicts observation (ii). A first-order Markov chain model in which the outcome is explained by a lagged

[2] The HRS is sponsored by the National Institute of Aging (grant number NIA U01AG009740) and conducted by the University of Michigan. For the analyses presented here, I use the RAND HRS Data File (Version E) which is a user-friendly data set produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration. See http://www.rand.org/labor/aging/dataprod for details.

[3] Note that this is obviously only done for respondents with six observations. Due to the sampling scheme, this holds only for the original HRS cohort born between 1931 and 1941.

DOI: 10.1002/jae

Table I. Distribution of SRHS

	poor	fair	good	very good	excellent
Frequency	10,099	19,579	30,811	27,665	14,079
Percent	9.9	19.2	30.1	27.1	13.8
By previous SRHS [%]:					
poor	56.9	30.5	9.5	2.4	0.8
fair	16.6	48.0	26.5	7.3	1.6
good	4.4	19.1	49.9	22.4	4.2
very good	1.8	6.7	27.7	50.6	13.3
excellent	1.0	3.2	12.9	33.9	48.9

Table II. Ordered Logit of SRHS in wave 6 on past SRHS

age	−0.0188	(0.006)**
female	0.0823	(0.046)$^{+}$
high school	0.1861	(0.065)**
some college	0.1762	(0.076)*
college degree+	0.3308	(0.078)**
nonwhite	−0.1198	(0.063)$^{+}$
SRHS wave 5	0.9847	(0.039)**
SRHS wave 4	0.5175	(0.038)**
SRHS wave 3	0.3251	(0.036)**
SRHS wave 2	0.2034	(0.035)**
SRHS wave 1	0.2390	(0.032)**
Observations	7173	
Log likelihood	−7663.4	

Robust SE in parentheses, $^{+}$: $p < 0.10$, *: $p < 0.05$, **: $p < 0.01$

dependent variable would imply no additional predictive power of waves 1 through 4 once wave 5 is controlled for. This contradicts observation (i). A combination of a time-constant random effect (RE) with a first-order Markov chain model would imply predictive power of all waves with wave 5 having a higher predictive power than waves 1 through 4. But a Wald test of the hypothesis of equal predictive power of the earlier four waves is clearly rejected (test statistic $\overset{a}{\sim} \chi^2_3 = 40.05$).

These findings can be interpreted as an indication that the models typically used for modeling SRHS in panels such as Contoyannis, Jones, and Rice (2004) are not capable of capturing the correlation pattern found in the data.[4] An obvious strategy to capture the correlation pattern better would be to combine a higher-order Markov chain model of state dependence with a RE specification. But this would aggravate the initial values problem already present in the first-order Markov chain model with RE.

In a structural model for SRHS, state-dependence of SRHS is also not very convincing. While for example in a model of labor force participation lagged outcomes can causally affect today's outcome, this is unlikely for this application: Which of the five SRHS categories a respondent ticks in a survey won't affect future health. In a model with state dependence and RE, the coefficients determining the state dependence can be interpreted to capture the diminishing predictive power of

[4] Note that these models typically do not specify the lagged values as the 5-point scale SRHS measure but as four dummy variables. This does not change the conclusions from Table II but only makes the results harder to read.

DOI: 10.1002/jae

higher lags evident in Table II in a reduced-form fashion. But a structurally more plausible model would be one in which SRHS depends on actual health and this underlying variable follows some random process over time with decreasing correlation. For the discussion of such issues, also see Heiss, Börsch-Supan, Hurd, and Wise (2007).

3.2. Model Structure

Let y_{it}^* denote a latent variable which represents actual health. It is modeled as a function of covariates $\mathbf{x}_{it}$, an unobserved stochastic process a_{it} and an i.i.d. error term e_{it}. For simplicity, consider the linear specification

$$y_{it}^* = \mathbf{x}_{it}\boldsymbol{\beta} + a_{it} + e_{it} \tag{15}$$

Assume that SRHS $y_{it} \in \{1, \ldots, 5\}$ is generated by the standard ordered response model

$$y_{it} = j \Leftrightarrow \alpha_{j-1} \leq y_{it}^* < \alpha_j \quad \text{with} \quad 1 \leq j \leq 5 \tag{16}$$

where $\alpha_0 = -\infty$, $\alpha_5 = \infty$, and α_1 through α_4 are unknown model parameters.

In order to derive a parametric expression of conditional outcome probabilities, assume that the i.i.d. error terms e_{it} are logistic. They may represent transitory health problems like a cold, general mood at the time the survey was completed or general measurement errors. This parametric assumption leads to a standard ordered logit specification except that the latent process a_{it} is present. With $\Lambda(\cdot)$ representing the logistic c.d.f., the conditional outcome probabilities in (3) can in this model be written as

$$P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it}) = \Lambda(\alpha_{y_{it}} - \mathbf{x}_{it}\boldsymbol{\beta} - a_{it}) - \Lambda(\alpha_{y_{it}-1} - \mathbf{x}_{it}\boldsymbol{\beta} - a_{it}) \tag{17}$$

To complete the model, the joint distribution of the state space a_{it} has to be specified. Assume independence of $\mathbf{x}_i$ and a normal stationary AR(1) process. With $\phi(\cdot; \mu, \sigma^2)$ denoting the normal p.d.f. with mean μ and variance σ^2, the marginal distribution is

$$f(a_{it}|\mathbf{x}_i) = \phi(a_{it}; 0, \sigma^2) \tag{18}$$

Assume the AR(1) structure

$$a_{it} = \rho a_{i,t-1} + u_{it} \tag{19}$$

where the innovations u_{it} are i.i.d. normal with zero mean and variance $(1 - \rho^2)\sigma^2$. The correlation parameter $-1 \leq \rho \leq 1$ is another model parameter. This leads to a conditional distribution corresponding to (2) of

$$f(a_{it}|\mathbf{x}_i, a_{i,t-1}) = \phi(a_{it}; \rho a_{i,t-1}, (1 - \rho^2)\sigma^2) \tag{20}$$

This completes the model definition discussed in general in Section 2.1 with a parameter vector $\boldsymbol{\theta} = [\boldsymbol{\beta}, \alpha_1, \ldots, \alpha_4, \sigma, \rho]$.

A standard ordered logit model follows in the special case $\sigma = 0$ and a random effects ordered logit model follows in the case $\rho = 1$. The correlation between a_{it} and a_{is} conditional on the

DOI: 10.1002/jae

covariates $\mathbf{x}_i$ is $\rho^{|t-s|}$. With $0 < \rho < 1$, it can explain the significant but decreasing predictive power of lagged dependent variables in Table II.

3.3. Implementation and Results

For this model, different algorithms for evaluating the likelihood function are implemented and compared. The results labeled *Random Simulation* are based on a simulation of the full sequence of the latent state space as discussed in Section 5. The draws are based on a standard random number generator.[5] *Antithetic Simulation* uses the same algorithm but bases the draws on Modified Latin Hypercube Sequences (MLHS) instead of random draws. These are straightforward to implement and work effectively in the context of likelihood approximation, see Hess, Train, and Polak (2006) for details. *Sparse grids integration* denotes numerical integration on sparse grids for the integration of the whole sequence. The algorithm for generating nodes and weights is given in Heiss and Winschel (2006).[6]

The implemented filtering algorithms are a *Nonlinear particle filter* (Fernández-Villaverde and Rubio-Ramírez 2006) with MLHS for the initial state and innovations to improve the performance. A *Sequential importance sampling* algorithm similar to Tanizaki and Mariano (1994) is implemented except that a fixed grid of nodes instead of antithetic draws are used. For the univariate state space, this proved to be most successful. Finally, *Sequential Gaussian quadrature* denotes the algorithm discussed in detail in Section 2.3.

Each algorithm is implemented for a different number of nodes R at which all likelihood contributions are to be evaluated. As $R \to \infty$, all methods should converge to the true likelihood. The question is how fast they do and what computational costs each algorithm needs to achieve accurate results.

Figure 1 shows the approximated log likelihood value at a fixed parameter vector for the different algorithms with the numbers of calculations of the conditional outcome probabilities R on the abscissa—note the logarithmic scaling. As expected, all algorithms converge to the same value, but the speed of convergence differs markedly. Random simulation of the whole sequence converges the slowest and even with 5000 replications, there is still a notable difference to the limiting value. For small R, the approximation is severely biased downwards. This is due to the fact that while outcome probabilities are simulated without bias, the concave log transformation creates downward bias by Jensen's inequality. Antithetic simulation with MLHS performs better requiring roughly half as many evaluations to achieve a comparable accuracy. Sparse grids integration converges notably faster.

Coming to the sequential algorithms, the nonlinear particle filter performs better than the joint algorithms with $R < 100$ but it is still far away from the limiting value and converges slower than sparse grids integration with a higher number $R > 100$. Compared to this, the sequential importance sampling algorithm with a fixed grid of nodes is very successful. With $R = 200$ the results are hardly different from the limiting value. The fastest algorithm is sequential Gaussian quadrature. With only $R = 20$ replications, the results are nearly indistinguishable from the limit

[5] A vector $\mathbf{z}^r_{i,1:T}$ is drawn from an independent standard normal distribution. The states $\mathbf{a}_{i,1:T}$ are jointly normal with zero mean and a covariance matrix Σ. A draw from this distribution is obtained by $\mathbf{a}^r_{i,1:T} = L\mathbf{z}^r_{i,1:T}$, where and L is the Choleski decomposition of Σ such that $LL' = \Sigma$.

[6] Code in Matlab and Stata for generating nodes and weights for integration on sparse grids can be requested from the author.

DOI: 10.1002/jae

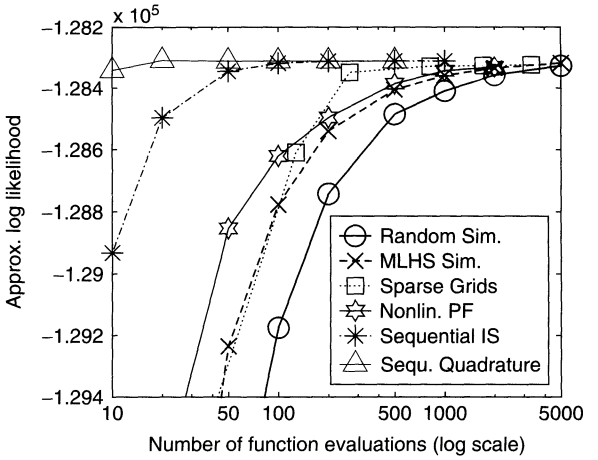

Figure 1. Approximate log likelihood at fixed parameter vector

$R \to \infty$. It should be noted that for models with a higher-dimensional state space it is likely that the advantage of the sequential algorithms would remain but that their simulation-based versions become more powerful relative to SGQ.

The number of evaluations of the conditional outcome probabilities R is not the only determinant of computational costs. A comparison of the total run time of the algorithms obviously depends on how efficient the algorithms are coded. In the implementations used for this paper, the main difference to comparing by R is that the nonlinear particle filter is relatively slow because the resampling is costly. Details can be requested from the author.

The ultimate goal of the approximated likelihood functions is to base parameter estimation on them. Intuitively, a better approximation of the likelihood function leads to better estimates based on it. As seen above, the sequential Gaussian quadrature algorithm seems to perform very well with only $R = 20$ function evaluations. To be on the safe side, this algorithm with $R = 50$ is declared as a "reference algorithm". Table III shows the ML estimates obtained by approximations using this algorithm. Notably, the estimated standard deviation of the latent state process σ is large compared to the standard logistic i.i.d. error term e_{it} which has a standard deviation normalized to $\pi/\sqrt{3} \approx 1.82$. The correlation parameter ρ is large but highly significantly smaller than unity.

Figure 2 shows the estimates of these two most interesting parameters that drive the intertemporal correlation pattern using the different algorithms and number of evaluations R. Note that the particle filter was not used for estimation since, as noted above, its approximated likelihood function is not smooth in the parameters so that gradient-based optimization does not work for this algorithm. The qualitative picture is the same as for the likelihood values. While with 20 evaluations, the sequential Gaussian quadrature algorithm has reached its limiting value, the other methods need considerably more computations with random simulation performing worst. The estimated standard deviation σ seems to be downward biased with the simulation methods, while the correlation parameter is upward biased.

DOI: 10.1002/jae

Table III. Parameter estimates (sequ. quadrature with $R = 50$)

age splines: 50+	−0.1069	(0.0048)**
age splines: 60+	0.0624	(0.0076)**
age splines: 70+	−0.0485	(0.0082)**
age splines: 80+	−0.0074	(0.0122)
age splines: 90+	0.0912	(0.0306)**
female	0.0828	(0.0360)**
nonwhite	−1.0119	(0.0465)**
high school	1.2658	(0.0380)**
some college	1.8584	(0.0474)**
college degree+	2.7922	(0.0509)**
Latent states a_{it}: SD σ	2.8764	(0.0276)**
Latent states a_{it}: corr. ρ	0.9439	(0.0128)**
Individuals	25,353	
Observations	102,233	
Log likelihood	−128,311.0	

Robust SE in parentheses, **: $p < 0.01$

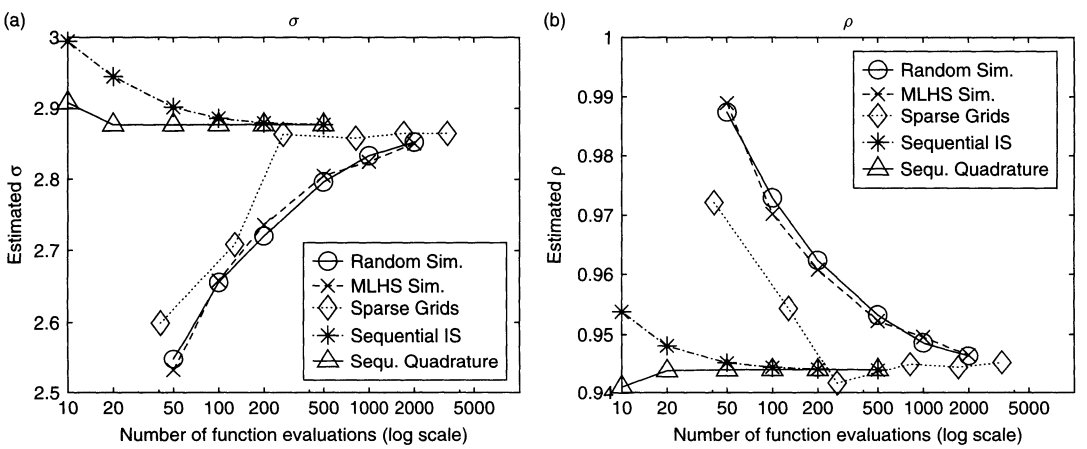

Figure 2. Results: Estimated Parameters σ and ρ

4. MONTE CARLO SIMULATIONS

4.1. Ordered Logit Models with an AR(1) Error Component

In order to check the sensitivity of the approximation algorithms with respect to the data generating process, this section reports results of a small Monte Carlo analysis. The models are also ordered logits with different parameterizations. All results use artificial panel data with $N = 1000$ cross-sectional units. The time-series dimension varies between specifications. The model structure corresponds to the empirical example. The explanatory variables x_{it} are generated from a univariate AR(1) process with a marginal standard normal distribution and a correlation over time of 0.5. The slope parameter is set to $\beta = 1$. The AR(1) error component a_{it} is normally distributed with a standard deviation of σ and a correlation over time of ρ, both of which are varied between specifications. The error term e_{it} is i.i.d. logistic. The observed dependent variable is an ordinal

DOI: 10.1002/jae

variable with five outcomes, where the cut points which translate the latent variable into the outcomes are set such that all outcomes are roughly equally populated.

The parameters of specification 1 are $T = 10$, $\sigma = 1$, and $\rho = 0.5$. Each of these are then separately set to a higher and a lower number to study their impact on the approximation performance, leading to a total of 7 model specifications. These values are given in Table IV.

For each of these model specifications, 100 artificial data sets are generated. Using each of these data sets, the log likelihood value at the true parameters is approximated by different algorithms with different levels of accuracy. Besides SGQ, SIS, random simulation and MLHS simulation, a quasi Monte Carlo simulation based on Halton sequences is applied.

The upper part of Table V shows their average over the 100 replications for these approximation exercises. All methods converge to the same value as the accuracy is increased. But while the simulation algorithms still have not completely converged with 5000 nodes, SGQ delivers the limiting value (with an accuracy of four digits) already with 11 nodes. The lower part of Table V shows approximation errors. The root mean squared errors are calculated relative to SGQ with 151 nodes for the respective data set. The conclusions from these results are the same: While the error of the simulation algorithms converge toward zero as the number of nodes increase, it is still at least 0.5 even with 5000 nodes. This might be accurate enough for maximum likelihood estimation but imposes large computational costs. SGQ delivers better results even with only five nodes and with 51 nodes, the error is smaller than 10^{-11} which is indistinguishable from roundoff errors with double precision computations.

Table VI shows these root mean squared approximation errors for the other model specifications. The broad picture is the same for all specifications. As it turns out, specification 5 with a high

Table IV. Monte-Carlo Model specifications

Specification number	1	2	3	4	5	6	7
T	10	5	30	10	10	10	10
σ	1	1	1	.2	4	1	1
ρ	.5	.5	.5	.5	.5	.1	.9

Table V. Monte Carlo Results Specification 1: $T = 10$, $\sigma = 1.0$, $\rho = 0.5$

Sequential Algorithms			Joint Simulation			
nodes	SGQ	SIS	nodes	MLHS	Halton	Random
Average log likelihood:						
5	−14975.9365	−14984.3503	50	−15018.6150	−15010.9704	−15039.2498
11	−14975.9244	−14977.5606	100	−14999.4423	−14996.2652	−15010.3644
25	−14975.9244	−14976.2011	1000	−14978.7097	−14977.1020	−14979.7530
51	−14975.9244	−14975.9731	2000	−14977.2547	−14976.4254	−14977.8923
101	−14975.9244	−14975.9282	5000	−14976.5179	−14976.0043	−14976.6189
Root mean squared error: (taking SGQ with 151 nodes as the true value)						
5	7.031×10^{-2}	$9.053 \times 10^{+0}$	50	$4.035 \times 10^{+1}$	$3.058 \times 10^{+1}$	$6.044 \times 10^{+1}$
11	1.007×10^{-4}	$2.059 \times 10^{+0}$	100	$2.044 \times 10^{+1}$	$2.013 \times 10^{+1}$	$3.054 \times 10^{+1}$
25	3.045×10^{-9}	9.034×10^{-1}	1000	$3.043 \times 10^{+0}$	$1.097 \times 10^{+0}$	$4.061 \times 10^{+0}$
51	3.071×10^{-12}	4.049×10^{-1}	2000	$2.005 \times 10^{+0}$	$1.008 \times 10^{+0}$	$2.080 \times 10^{+0}$
101	3.018×10^{-12}	2.031×10^{-1}	5000	$1.034 \times 10^{+0}$	5.055×10^{-1}	$1.049 \times 10^{+0}$

DOI: 10.1002/jae

Table VI. Root mean squared errors of the likelihood approximation

Sequential Algorithms			Joint Simulation			
nodes	SGQ	SIS	nodes	MLHS	Halton	Random
Specification 2: $T = 5$, $\sigma = 1.0$, $\rho = 0.5$, Average log likelihood: −7471.10						
5	5.028×10^{-2}	$5.006 \times 10^{+0}$	50	$8.073 \times 10^{+0}$	$4.076 \times 10^{+0}$	$2.004 \times 10^{+1}$
11	8.040×10^{-5}	$1.052 \times 10^{+0}$	100	$5.006 \times 10^{+0}$	$2.007 \times 10^{+0}$	$1.009 \times 10^{+1}$
25	2.049×10^{-9}	5.070×10^{-1}	1000	9.075×10^{-1}	2.064×10^{-1}	$1.076 \times 10^{+0}$
51	1.073×10^{-12}	2.073×10^{-1}	2000	7.043×10^{-1}	1.044×10^{-1}	$1.013 \times 10^{+0}$
101	1.038×10^{-12}	1.039×10^{-1}	5000	4.001×10^{-1}	6.051×10^{-2}	7.002×10^{-1}
Specification 3: $T = 30$, $\sigma = 1.0$, $\rho = 0.5$, Average log likelihood: −44896.06						
5	1.027×10^{-1}	$2.087 \times 10^{+1}$	50	$5.061 \times 10^{+2}$	$1.003 \times 10^{+3}$	$6.010 \times 10^{+2}$
11	2.033×10^{-4}	$6.072 \times 10^{+0}$	100	$3.080 \times 10^{+2}$	$5.029 \times 10^{+2}$	$4.015 \times 10^{+2}$
25	7.017×10^{-9}	$1.098 \times 10^{+0}$	1000	$9.067 \times 10^{+1}$	$1.026 \times 10^{+2}$	$9.093 \times 10^{+1}$
51	1.018×10^{-11}	8.077×10^{-1}	2000	$6.008 \times 10^{+1}$	$7.032 \times 10^{+1}$	$6.022 \times 10^{+1}$
101	1.004×10^{-11}	4.042×10^{-1}	5000	$3.025 \times 10^{+1}$	$3.052 \times 10^{+1}$	$3.024 \times 10^{+1}$
Specification 4: $T = 10$, $\sigma = 0.2$, $\rho = 0.5$, Average log likelihood: −14659.03						
5	4.071×10^{-4}	3.002×10^{-1}	50	5.069×10^{-1}	8.027×10^{-1}	$2.006 \times 10^{+0}$
11	1.090×10^{-9}	1.049×10^{-1}	100	3.043×10^{-1}	7.004×10^{-1}	$1.049 \times 10^{+0}$
25	1.086×10^{-12}	6.100×10^{-2}	1000	7.028×10^{-2}	6.022×10^{-2}	3.048×10^{-1}
51	1.068×10^{-12}	3.056×10^{-2}	2000	6.053×10^{-2}	3.057×10^{-2}	2.080×10^{-1}
101	1.091×10^{-12}	1.085×10^{-2}	5000	3.085×10^{-2}	1.078×10^{-2}	1.042×10^{-1}
Specification 5: $T = 10$, $\sigma = 4.0$, $\rho = 0.5$, Average log likelihood: −15227.06						
5	$2.061 \times 10^{+2}$	$2.067 \times 10^{+1}$	50	$2.087 \times 10^{+3}$	$2.061 \times 10^{+3}$	$2.097 \times 10^{+3}$
11	$1.019 \times 10^{+1}$	$4.041 \times 10^{+0}$	100	$2.008 \times 10^{+3}$	$2.001 \times 10^{+3}$	$2.015 \times 10^{+3}$
25	2.024×10^{-1}	$1.028 \times 10^{+0}$	1000	$6.022 \times 10^{+2}$	$5.056 \times 10^{+2}$	$6.022 \times 10^{+2}$
51	9.062×10^{-3}	5.048×10^{-1}	2000	$4.008 \times 10^{+2}$	$3.072 \times 10^{+2}$	$4.009 \times 10^{+2}$
101	9.064×10^{-5}	2.061×10^{-1}	5000	$2.025 \times 10^{+2}$	$2.004 \times 10^{+2}$	$2.022 \times 10^{+2}$
Specification 6: $T = 10$, $\sigma = 1.0$, $\rho = 0.1$, Average log likelihood: −14965.60						
5	4.010×10^{-2}	$4.021 \times 10^{+0}$	50	$4.086 \times 10^{+1}$	$4.051 \times 10^{+1}$	$7.011 \times 10^{+1}$
11	5.080×10^{-5}	$1.017 \times 10^{+0}$	100	$2.076 \times 10^{+1}$	$2.030 \times 10^{+1}$	$3.099 \times 10^{+1}$
25	2.004×10^{-9}	4.027×10^{-1}	1000	$3.087 \times 10^{+0}$	$2.017 \times 10^{+0}$	$5.047 \times 10^{+0}$
51	3.031×10^{-12}	1.092×10^{-1}	2000	$2.019 \times 10^{+0}$	$1.012 \times 10^{+0}$	$3.020 \times 10^{+0}$
101	2.059×10^{-12}	9.004×10^{-2}	5000	$1.042 \times 10^{+0}$	5.035×10^{-1}	$1.063 \times 10^{+0}$
Specification 7: $T = 10$, $\sigma = 1.0$, $\rho = 0.9$, Average log likelihood: −14712.88						
5	$3.036 \times 10^{+1}$	$1.052 \times 10^{+1}$	50	$2.016 \times 10^{+1}$	$1.059 \times 10^{+1}$	$3.072 \times 10^{+1}$
11	$1.069 \times 10^{+0}$	$4.082 \times 10^{+0}$	100	$1.015 \times 10^{+1}$	$7.046 \times 10^{+0}$	$1.099 \times 10^{+1}$
25	1.012×10^{-2}	$1.074 \times 10^{+0}$	1000	$1.077 \times 10^{+0}$	9.004×10^{-1}	$2.035 \times 10^{+0}$
51	7.086×10^{-7}	8.021×10^{-1}	2000	$1.017 \times 10^{+0}$	4.062×10^{-1}	$1.069 \times 10^{+0}$
101	3.009×10^{-12}	4.027×10^{-1}	5000	7.021×10^{-1}	2.013×10^{-1}	8.096×10^{-1}

variance of the AR(1) term is the most problematic for the likelihood approximation. SGQ needs 51 nodes to achieve an error of less than 10^{-2}. The simulation methods look pretty worthless in this case. Even with 5000 nodes, their error remains larger than 10^2. As a comparison between specifications 2 and 3 shows, the length of the time-series affects the simulation algorithms more than SGQ. On the other hand, a high value of ρ makes SGQ but not simulation less efficient. But even with $\rho = .9$ (and as the empirical application showed with $\rho = .94$), SGQ with a sufficient number of nodes is by far the most accurate method. With $\rho = 1$, the model becomes a random effects model and SGQ becomes a standard quadrature algorithm for random effects models.

DOI: 10.1002/jae

4.2. A Panel Probit Model

For the special case of panel probit models, the Geweke-Hajivassiliou-Keane (GHK) simulator is the leading algorithm for the likelihood approximation. It makes use of the joint normality of the compound error term $a_{it} + e_{it}$ in these models and samples from its distribution conditional on the data, see Börsch-Supan and Hajivassiliou (1993) and Keane (1994).

Table VII shows Monte Carlo results for a probit model with an AR(1) error component. The model structure is the same as the one for the ordered logit model except that the i.i.d. error term e_{it} is standard normal and that the observed dependent variable is binary. The parameters are the same as for specification 1. The methods include the same SGQ and random simulation algorithms as used above and two GHK simulators—one based on Halton sequences and one using a standard random number generator. As previous research like Hajivassilou, McFadden, and Ruud (1996) and Geweke, Keane, and Runkle (1997) suggests, the GHK algorithm performs well and much better than the partial analytic simulation algorithm which samples from a_{it} alone. But also for this model, SGQ is clearly the most accurate algorithm.

5. CONCLUSIONS

This paper deals with a certain class of nonlinear panel data models including limited dependent variable models with AR(1) error components. In an example application, such a generalization of random effects models proves to be useful. The likelihood function for these models is generally infeasible analytically, so different approaches to its numerical approximation are discussed.

The models can be expressed in a state space form, so nonlinear Kalman filter algorithms available for such models provide a natural approach besides the simulation of the latent process. The state space in these models are simple compared to the models these algorithms are typically applied to. A sequential Gaussian quadrature algorithm is suggested which makes use of this fact.

For the estimation of this illustrative model, this algorithm performs very well and clearly better than the other algorithms compared to it. Monte Carlo evidence suggests that this result is robust to a variety of alternative parameterizations of a similar model. It is also more powerful than the GHK simulator for a panel probit model analyzed in another set of Monte Carlo experiments.

Table VII. Monte Carlos Results Probit model with $T = 10$, $\sigma = 1$, $\rho = 0.5$

nodes	SGQ	Joint Simulation			
		nodes	GHK (Halton)	GHK (Random)	Random
Average log likelihood:					
5	−5631.6677	50	−5632.0445	−5633.4226	−5688.5083
11	−5631.6717	100	−5631.8543	−5632.4109	−5661.8823
25	−5631.6718	1000	−5631.6798	−5631.8364	−5633.3628
51	−5631.6718	2000	−5631.6753	−5631.7693	−5632.4744
101	−5631.6718	5000	−5631.6737	−5631.7141	−5631.7924
Root mean squared error: (taking SGQ with 151 nodes as the true value)					
5	1.068×10^{-1}	50	$1.009 \times 10^{+0}$	$2.051 \times 10^{+0}$	$5.078 \times 10^{+1}$
11	2.060×10^{-4}	100	6.065×10^{-1}	$1.043 \times 10^{+0}$	$3.012 \times 10^{+1}$
25	2.003×10^{-10}	1000	7.034×10^{-2}	4.040×10^{-1}	$2.059 \times 10^{+0}$
51	1.097×10^{-12}	2000	4.032×10^{-2}	3.010×10^{-1}	$1.038 \times 10^{+0}$
101	1.024×10^{-12}	5000	1.086×10^{-2}	1.075×10^{-1}	6.080×10^{-1}

DOI: 10.1002/jae

ACKNOWLEDGEMENTS

The author would like to thank Axel Börsch-Supan, Dan McFadden, Paul Ruud, Arthur van Soest, Viktor Winschel, Joachim Winter, and three anonymous referees for valuable discussion, comments and suggestions.

REFERENCES

Bauwens L, Hautsch N. 2006. Stochastic conditional intensity processes. *Journal of Financial Econometrics* **4**: 450–493.

Börsch-Supan A, Hajivassiliou V. 1993. Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics* **58**: 347–368.

Butler JS, Moffit R. 1982. A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica* **50**: 761–764.

Chamberlain G. 1984. Panel data. In *Handbook of Econometrics Vol. 2*, Griliches Z, Intriligator MD, eds., Elsevier. 1247–1318.

Contoyannis P, Jones AM, Rice N. 2004. The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics* **19**: 473–503.

Danielsson J, Richard JF. 1993. Accelerated Gaussian importance sampler with application to dynamic latent variable models. *Journal of Applied Econometrics* **8**: S153–S173.

DeJong DN, Dharmarajan H, Liesenfeld R, Richard JF. 2006. Efficient likelihood analysis and filtering for state-space representations. Department of Economics, University of Pittsburgh.

Doucet A, De Freitas N, Gordon N (eds.). 2001. *Sequential Monte Carlo Methods in Practice*. Springer: New York.

Durbin J, Koopman SJ. 2001. *Time Series Analysis by State Space Methods*. Oxford University Press.

Durbin J, Koopman SJ. 2002. A simple and efficient simulation smoother for state space time series analysis. *Biometrika* **89**: 603–616.

Fernández-Villaverde J, Rubio-Ramírez JF. 2005. Estimating dynamic equilibrium economies: Linear versus nonlinear likelihood. *Journal of Applied Econometrics* **20**: 891–910.

Fernández-Villaverde J, Rubio-Ramírez JF. 2006. Estimating macroeconomic models: A likelihood approach. NBER Technical Working Paper 321.

Geweke JF, Keane MP, Runkle DE. 1997. Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics* **80**: 125–165.

Hajivassiliou V, McFadden D, Ruud P. 1996. Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. *Journal of Econometrics* **72**: 85–134.

Hajivassiliou VA, Ruud PA. 1994. Classical estimation methods for LDV models using simulation. In *Handbook of Econometrics Vol. 4*, Engle RF, McFadden DL, eds., Elsevier. 2383–2441.

Heckman JJ. 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete time - discrete data stochastic process. In *Structural Analysis of Discrete Data and Econometric Applications*, Manski CF, McFadden D, eds., MIT Press: Cambridge, Mass. 179–195.

Heiss F, Börsch-Supan A, Hurd M, Wise D. 2007. Pathways to disability: Predicting health trajectories. In *Health in Older Ages: The Causes and Consequences of Declining Disability Among the Elderly*, Wise D, ed., University of Chicago Press. Forthcoming.

Heiss F, Winschel V. 2006. Estimation with numerical integration on sparse grids. Department of Economics Discussion paper No. 2006-15, University of Munich, http://econpapers.repec.org/paper/lmumuenec/916.htm.

Hess S, Train KE, Polak JW. 2006. On the use of a modified latin hypercube sampling (MLHS) method in the estimation of a mixed logit model for vehicle choice. *Transportation Research Part B* **40**: 147–163.

Honoré BE, Tamer E. 2006. Bounds on parameters in panel dynamic discrete choice models. *Econometrica* **74**: 611–629.

Keane MP. 1994. A computationally practical simulation estimator for panel data. *Econometrica* **62**: 95–116.

Koopman SJ, Lucas A. 2005. Business and default cycles for credit risk. *Journal of Applied Econometrics* **20**: 311–323.

Lee LF. 1997. Simulated maximum likelihood estimation of dynamic discrete choice statistical models: Some Monte Carlo results. *Journal of Econometrics* **82**: 1–35.
Richard JF, Zhang W. 2005. Efficient high-dimensional importance sampling. Working Paper, University of Pittsburgh, Dept. of Economics.
Shephard N, Pitt MK. 1997. Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**: 653–667.
Tanizaki H. 2003. Nonlinear and non-Gaussian state-space modeling with Monte Carlo techniques: A survey and comparative study. In *Handbook of Statistics*, Shanbhag D, Rao C, eds., Elsevier: vol. 21. 871–929.
Tanizaki H, Mariano RS. 1994. Prediction, filtering and smoothing in non-linear and non-normal cases using Monte Carlo integration. *Journal of Applied Econometrics* **9**: 163–179.
Wooldridge JM. 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* **20**: 39–54.

J. Appl. Econ. **23**: 373–389 (2008)
DOI: 10.1002/jae