

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

Phone: +49(621)2180-6291, Fax: +49(621)2180-3954

`florian.heiss@lrz.uni-muenchen.de`

February 15, 2007

This paper discusses the estimation of nonlinear panel data models with autoregressive error components and a broader class formulated in the framework of nonlinear state space models. An example from health economics shows that such a model is more plausible and parsimonious and captures the correlation pattern better than the commonly applied random effects and Markov chain models. For the approximation of the likelihood function, nonlinear filtering algorithms developed in the time-series literature are considered. For the relatively simple structure of these models, a straightforward algorithm based on sequential Gaussian quadrature is suggested. It performs well both in the empirical application and a Monte Carlo study for an ordered logit and a probit model with an AR(1) error component.

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

1 Introduction

Panel data provide repeated observations on the same individuals, firms, or other units over time. This allows the identification of a much richer set of effects in a more general setting than pure cross-sectional data. Many microeconomic models, especially limited dependent variable models, are inherently nonlinear. This nonlinearity complicates the analysis of panel data models, for a general discussion see for example Chamberlain (1984). In applied microeconomic research, many nonlinear panel data models specify unobserved heterogeneity as time-constant individual effects and/or state dependence as a low-order Markov model by including lagged dependent variables as regressors. While the estimation of these models is fairly straightforward, they impose a quite inflexible dynamic structure on the data.

As an example, 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 approaches and their estimation. I argue that an ordered logit 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.

The widespread application of such models is hampered by the computational difficulties encountered in their estimation. This paper discusses these problems and different solutions for a class of models which includes limited dependent variable models with AR(1) errors but is much more general. It is formulated in a state space framework. This approach has a long tradition in linear time series models, see Hamilton (1994). The increase in computational power makes it also feasible for general nonlinear models which generated increased interest in the econometric time series literature, see for example Fernández-Villaverde and Rubio-Ramírez (2005), (Koopman and Lucas 2005), and (Bauwens and Hautsch 2006).

The computational problem in evaluating the likelihood function of such models is that the unobserved state process has to be integrated out. With continuously distributed states, these integrals have to be approximated numerically and their dimension is typically proportional to the time-series dimension of the data. Unlike time series models, this data dimension is usually moderate for microeconomic analyses and asymptotic arguments are applied to the cross-sectional dimension. This makes it feasible to approximate the full multidimensional integral for example by Monte Carlo simulation or by numerical integration.

For time-series models, various attempts have been made to break up the full integral into a sequence of lower-dimensional integrals in the spirit of the Kalman filter. Outside of economics, these nonlinear filtering approaches are widely studied e.g. in engineering (Doucet, De Freitas and Gordon, eds 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).

Also applications with a moderate time-series dimension can profit from nonlinear filtering techniques. Loosely speaking does each reduction of dimensionality help for each method of numerical integration. Given the features of microeconomic models discussed here with a moderate time series dimension and a univariate latent state space, it is then argued that a straightforward approach using sequential Gaussian quadrature can be expected to perform well.

Different algorithms are then 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. Finally, a Monte Carlo study compares the approximation accuracy of various algorithms for an ordered logit model with different parameters of the data generating process and a panel probit model.

The paper is structured as follows. Section 2 discusses the structure of the state-space model and approaches to its likelihood approximation. Section 3 presents an example application and gives results for the accuracy of the likelihood approximation and the estimated parameters. Section 4 presents Monte Carlo simulations 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. Section 5 concludes.

2 Microeconomic State-Space Models

This paper discusses estimation for a set of micro-econometric models that can be represented in a simple state-space framework. We start out by defining the structure of and requirements on these models in section 2.1 and discuss the numerical approximation of the likelihood function in sections 2.2 through 2.4.

2.1 Model Specification

Suppose a sequence of dependent variables is observed over time for a number N of cross-sectional units, say individuals. All random variables involved in the model are assumed to be independent across cross-sectional units. Let T be the number of observations over time (“waves”) for each cross-sectional unit. 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.

The vectors of random variables \mathbf{Y}_{it} for $i = 1, \dots, N$ and $t = 1, \dots, T$ represent dependent variables of individual i in wave t . In many applications, they are one-dimensional, but I allow for the more general case since this 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. They are modeled conditional on exogenous

variables \mathbf{x}_i , unobserved states which are correlated over time a_{it} and unobserved i.i.d. error terms \mathbf{e}_{it} . The model is specified as

$$\mathbf{Y}_{it} = g(\mathbf{x}_i, a_{it}, \mathbf{e}_{it}; \boldsymbol{\theta}), \quad (1)$$

where $g(\cdot)$ is a general parametric function. The vectors \mathbf{x}_i contain time-constant and time-varying strictly exogenous variables. In the latter case, $\mathbf{x}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}]$ collects all time-specific values. The random variables (“states”) a_{it} are allowed to be continuously distributed and correlated over time in a relatively flexible way discussed in detail below. In this paper, a_{it} is assumed to be a scalar random variable. Generalizations to models with a higher-dimensional state-space are conceptually straightforward but affect the numerical approximation discussed below. The i.i.d. error terms \mathbf{e}_{it} may reflect measurement errors and/or transitory influences on \mathbf{Y}_{it} . For notational simplicity, all model parameters are collected in the vector $\boldsymbol{\theta}$.

This model can be viewed as a generalization of a random effects model. In this case, $\mathbf{a}_{i,1:T} = [a_{it} : t = 1, \dots, T]$ has a degenerate joint distribution with $a_{it} = a_i$ for all $t = 1, \dots, T$. In the more general case, it could for example represent an AR(1) component of the error term.

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 $\mathbf{Y}_{i,1:t} = [\mathbf{Y}_{is} : s = 1, \dots, t]$ conditional on \mathbf{x}_i , evaluated at the observed values $\mathbf{y}_{i,1:t}$, the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N P(\mathbf{y}_{i,1:T}|\mathbf{x}_i; \boldsymbol{\theta}) \quad (2)$$

The computational problem in evaluating this expression mainly arises due to the presence of the latent states a_{it} in the model. Before discussing this problem and solutions in detail, the class of models is restricted in the following way. For convenience of presentation, it is understood that all expressions depend on the parameter vector $\boldsymbol{\theta}$ which is in the following left out of the notation.

Measurement

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 $\mathbf{Y}_{i,1:t}$ conditional on \mathbf{x}_i , $\mathbf{a}_{i,1:T}$, and past values of y_{it} , evaluated at the observed values \mathbf{y}_{it} .

Make the following conditional independence assumption.

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

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 with the usual 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 sequence of latent states $\mathbf{a}_{i,1:T}$ which is correlated over time.

Assume that the i.i.d. error terms \mathbf{e}_{it} can easily be integrated out of the model so that $P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})$ is a known parametric function. It might for example follow from a typical limited dependent variable (LDV) model specification in (1) in which the unknown a_{it} enter as additional regressors analogous to LDV models with random effects.

States

For the sequence of latent states $\mathbf{a}_{i,1:T}$, assume that the marginal distribution of a_{it} is known up to a finite set of parameters included in the general parameter vector $\boldsymbol{\theta}$. In the following, these states are treated as vectors of continuous random variables. The only real difference with discrete or mixed distributions is that the numerical analysis would be less difficult. For notational simplicity assume that these marginal distributions are the same for all $i = 1, \dots, N$ and $t = 1, \dots, T$ 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 $\mathbf{y}_{i,1:T}$. Therefore, for the conditional p.d.f.

$$f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:T}, \mathbf{a}_{i,1:t-1}) = f(a_{it}|\mathbf{x}_i, a_{i,t-1}) \quad (4)$$

This conditional distribution is again assumed to be known up to parameters. This structure allows to write the joint p.d.f. of $\mathbf{a}_{i,1:T}$ as

$$f(\mathbf{a}_{i,1:T}|\mathbf{x}_i) = f(a_{i1}|\mathbf{x}_i) \prod_{t=2}^T f(a_{it}|\mathbf{x}_i, a_{i,t-1}) \quad (5)$$

Section 3 provides an example how an ordered logit model with an AR(1) error component is represented in this general model specification.

2.2 Approximation of the Likelihood Contributions by Joint Simulation

For the evaluation of the likelihood function (2), the probabilities $P(\mathbf{y}_{i,1:T}|\mathbf{x}_i)$ have to be evaluated. Because of the presence of the latent process $\mathbf{a}_{i,1:T}$ in the conditional outcome probabilities specified in (3), this expression can in general not be evaluated directly. The simplest approach to approximation is to numerically integrate out the full latent process $\mathbf{a}_{i,1:T}$:

$$P(\mathbf{y}_{i,1:T}|\mathbf{x}_i) = \int \cdots \int P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{a}_{i,1:T}) f(\mathbf{a}_{i,1:T}|\mathbf{x}_i) da_{i1} \cdots da_{iT} \quad (6)$$

All terms in this equation are known by assumption. The integral does in general not have an analytic solution. The typical approach in micro-econometrics to these kinds of problems is simulation: Generate a number of R draws $[\mathbf{a}_{i,1:T}^r : r = 1, \dots, R]$ from the joint distribution $f(\mathbf{a}_{i,1:T}|\mathbf{x}_i)$.¹ The simulated probability is equal to

$$\tilde{P}^{\text{SIM}}(\mathbf{y}_{i,1:T}|\mathbf{x}_i) = \frac{1}{R} \sum_{r=1}^R P(\mathbf{y}_{i,1:T}|\mathbf{x}_i; \mathbf{a}_{i,1:T}^r). \quad (7)$$

¹If $[a_{i1}, \dots, a_{iT}]$ are jointly normal with zero mean and a general covariance matrix Σ , these draws can be obtained as $\mathbf{a}_{i,1:T}^r = L\mathbf{z}_{i,1:T}^r$, where $\mathbf{z}_{i,1:T}^r$ is a vector of independent draws from a standard normal distribution and L is the Choleski decomposition of Σ such that $LL' = \Sigma$.

Pseudo-maximum likelihood estimators of θ using $\tilde{P}^{\text{SIM}}(\mathbf{y}_{i,1:T}|\mathbf{x}_i)$ instead of its true value is 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 Richard and Zhang (2005).

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). The GHK simulator has been shown to work effectively for probit model for example by Hajivassiliou, McFadden and Ruud (1996).

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 (6) is T -dimensional even if a_{it} is one-dimensional and a multiple thereof otherwise. The well-known product rule extension of Gaussian quadrature to multiple dimensions suffers from exponentially rising computational costs as the number of dimensions increases. Even in 3 or 4 dimensions, this “curse of dimensionality” makes this approach computationally inefficient. In higher dimensions, it quickly becomes infeasible even on modern computers.

Heiss and Winschel (2006) use a different approach of extending Gaussian quadrature to multiple dimensions for integration problems such as (6). 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.² This leads to much slower increase of the computational costs with a small decrease of approximation accuracy.

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 so 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 (6) 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. Compared to time series models in engineering, finance or macroeconomics for which nonlinear filters are usually discussed, the typical microeconomic panel data model has a low-dimensional state space and a short time-series dimension.

The general idea how to decompose the integral in (6) for the model structure described in section 2.1 is the following. For a simplification of notation, denote $P(\mathbf{y}_{i1}|\mathbf{x}_i, \mathbf{y}_{i,1:0}) = P(\mathbf{y}_{i1}|\mathbf{x}_i)$. By the rules of conditioning, the probabilities of interest can then in general be written as

$$P(\mathbf{y}_{i,1:T}|\mathbf{x}_i) = \prod_{t=1}^T P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}). \quad (8)$$

²In 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.

Each of these terms $P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ are now approximated separately by some $\tilde{P}(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ and the approximated individual likelihood contributions are

$$\tilde{P}^{\text{SEQ}}(\mathbf{y}_{i,1:T}|\mathbf{x}_i) = \prod_{t=1}^T \tilde{P}(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}). \quad (9)$$

Note that the expressions $P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ are nontrivial to calculate. They are complicated functions of past values because of the presence of the unobserved sequence of latent states a_{it} .

The structure of the model allows to obtain these approximations in a sequential fashion. The outcome probabilities conditional on past values can be written as

$$P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}) = \int P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it}) f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}) da_{it} \quad (10)$$

by (3). 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} . The problem of this equation is that the conditional distribution $f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ is again a complicated function of past realizations $\mathbf{y}_{i,1:t-1}$.

For $t = 1$, the conditional densities are simply equal to the initial distribution $f(a_{i1}|\mathbf{x}_i)$ which is known by the model specification. For $t > 1$, they can be expressed in a recursive fashion. Suppose that the conditional density $f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ is known so that $P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ can be calculated by (10). Then, density for the next wave $t + 1$ can be derived as follows. First, note that by the conditional independence assumption (4),

$$f(a_{it}, a_{i,t+1}|\mathbf{x}_i, \mathbf{y}_{i,1:t}) = f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t})f(a_{i,t+1}|\mathbf{x}_i, a_{it}). \quad (11)$$

A marginalization with respect to a_{it} leads to the wanted expression of the conditional distribution for $t + 1$:

$$f(a_{i,t+1}|\mathbf{x}_i, \mathbf{y}_{i,1:t}) = \int f(a_{i,t+1}|\mathbf{x}_i, a_{it}) f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t}) da_{it}. \quad (12)$$

The first term in this integral is known by the model specification, the second term can be expressed in terms of known functions. Bayes' rule and the model assumption (3) imply

$$f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t}) = f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}) \frac{P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})}{P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})}. \quad (13)$$

A combination of (12) and (13) results in an expression for $f(a_{i,t+1}|\mathbf{x}_i, \mathbf{y}_{i,1:t})$ as a function of terms which are either known by the model specification ($P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})$ and $f(a_{i,t+1}|\mathbf{x}_i, a_{it})$) or from the previous recursion step ($P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ and $f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$).

The computational problem lies in the fact that the integrals in (10) and (12) generally do not have an analytic solution. There are various approaches to numerical approximations, see (Durbin and Koopman 2001) and Tanizaki (2003).

For a simulation approximation of the likelihood function, draws from the conditional distribution $f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ must be obtained. A direct way to do this is the nonlinear particle filter (NPF), see Doucet et al., eds (2001) and for an application to econometric time series models Fernández-Villaverde and Rubio-Ramírez (2005). The conditioning on the data is introduced by resampling from a set of draws (“particles”) where each particle is drawn with an appropriate probability depending on the data. 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. In the empirical illustration below, the NPF is therefore only used to approximate the likelihood function for a given set of parameters but not for estimation.

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 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 using an approximation of the model, see Durbin and Koopman (1997, 2002) and Shephard and Pitt (1997).

2.4 Sequential Gaussian Quadrature

The types of models this paper is concerned with has a one-dimensional state space, for example an AR(1) error component. This makes it feasible and possibly efficient to construct

a nonlinear Kalman filter based on sequential Gaussian quadrature. Gaussian quadrature prescribes a set of R nodes $[z_r : r = 1, \dots, R]$ and corresponding weights $[w_r : r = 1, \dots, R]$ for a general 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. Gaussian quadrature has long been used for univariate integration problems such as random effects models, see for example Butler and Moffit (1982).

A problem with applying Gaussian quadrature directly to the integrals in (10) and (12) is that the natural weight functions $f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ have no closed-form expression and therefore the appropriate nodes and weights cannot be derived. Therefore, a reformulation of the integrals very much in the spirit of the sequential importance sampling algorithm is used here. Define a ‘‘proposal density’’ for which a Gaussian quadrature rule is known. For simplicity, here the marginal distribution $f(a_{it}|\mathbf{x}_i)$ is used. Define the ratio of the two densities as

$$q_{it}(a_{it}) = \frac{f(a_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})}{f(a_{it}|\mathbf{x}_i)}. \quad (14)$$

With this definition, rewrite (10) as

$$P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}) = \int q_{it}(a_{it}) P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it}) f(a_{it}|\mathbf{x}_i) da_{it} \quad (15)$$

and combine (12) and (13) to obtain

$$f(a_{i,t+1}|\mathbf{x}_i, \mathbf{y}_{i,1:t}) = \int q_{it}(a_{it}) f(a_{i,t+1}|\mathbf{x}_i, a_{it}) \frac{P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})}{P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})} f(a_{it}|\mathbf{x}_i) da_{it}. \quad (16)$$

Using their definition in (14), this gives a recursion of the ‘‘importance weights’’:

$$q_{i,t+1}(a_{i,t+1}) = \int q_{it}(a_{it}) \frac{f(a_{i,t+1}|\mathbf{x}_i, a_{it})}{f(a_{i,t+1}|\mathbf{x}_i)} \frac{P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it})}{P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})} f(a_{it}|\mathbf{x}_i) da_{it}. \quad (17)$$

Both integrals in (15) and (17) can now be sequentially approximated by Gaussian quadrature with quadrature nodes $[a^r : r = 1, \dots, R]$ and weights $[w^r : r = 1, \dots, R]$ ap-

proprate for $f(a_{it}|\mathbf{x}_i)$. Initialize $q_{i1}^r = 1$ for all $r = 1, \dots, R$ and for all $t = 1, \dots, T$ do the following calculations:

1. Approximate $P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})$ as

$$\tilde{P}(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1}) = \sum_{r=1}^R q_{it}^r P(\mathbf{y}_{it}|\mathbf{x}_i, a^r) w^r \quad (18)$$

2. For all $s = 1, \dots, R$, approximate $q_{i,t+1}(a^s)$ as

$$q_{i,t+1}^{*s} = \sum_{r=1}^R \frac{f(a^s|\mathbf{x}_i, a^r)}{f(a^s|\mathbf{x}_i)} \frac{q_{it}^r P(\mathbf{y}_{it}|\mathbf{x}_i, a^r)}{\tilde{P}(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{y}_{i,1:t-1})} w^r. \quad (19)$$

3 Example: An Ordered Logit Model of Health with an AR(1) Error Term

3.1 Background and Data

One of the most frequently studied measures of individual health is the self-rated health status (SRHS). The Health and Retirement Study (HRS) asks all respondents “Would you say your health is excellent, very good, good, fair, or poor?”. It is included in many other surveys with a similar wording. 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.

Panel data analyses of SRHS are not only useful because unobserved heterogeneity of health itself, but also heterogeneity of reporting SRHS *given* health can be accounted for. Table 1 shows the distribution of SRHS in the sample. As the tabulations conditional on the previous response indicate, SRHS is highly correlated over time. Here, we focus on the question how to model this correlation. In the literature, this correlation is almost exclusively modeled as time-constant unobserved heterogeneity and/or state dependence of SRHS. Contoyannis et al. (2004) discuss and compare these approaches. I will argue

that a model with correlated error terms is both more plausible and fits the correlation pattern better.

The data used for the empirical example is from the Health and Retirement Study (HRS).³ The HRS contains data on different cohorts of elderly Americans. I use a sample of all cohorts with the only restriction that they are at least 50 years old at the time of the first interview. The sample includes 25,353 respondents with up to 6 observations over time each. A total of 102,233 observations are available.

Table 2 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.⁴ 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 would imply no additional predictive power of waves 1 through 4 once wave 5 is controlled for which 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 $\overset{a}{\sim} \chi_3^2 = 40.05$).

I interpret these findings as an indication that the models typically used for modeling SRHS in panels such as Contoyannis et al. (2004) are not capable of capturing the corre-

³The 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/dataproduct> for details.

⁴Note 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.

lation pattern found in the data.⁵ 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, state-dependence of SRHS is actually 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. So in a model with state dependence and RE, the coefficients determining the state dependence can be interpreted to capture the diminishing predictive power of higher lags evident in Table 2 in a reduced-form fashion. But a structurally more plausible model would be one in which SRHS depends on current health and this underlying variable follows some random process over time with decreasing correlation. I suggest a model with an AR(1) error term.

3.2 Model Structure

Let Y_{it}^* denote a latent variable which represents a continuous measure of 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

$$Y_{it}^* = \mathbf{x}_{it}\boldsymbol{\beta} + a_{it} + e_{it}. \quad (20)$$

Assume that SRHS $Y_{it} \in \{1, \dots, 5\}$ is generated by a standard ordered response model.

$$Y_{it} = j \Leftrightarrow \alpha_{j-1} \leq Y_{it}^* < \alpha_j \quad \text{with } 1 \leq j \leq 5, \quad (21)$$

where $\alpha_0 = -\infty$, $\alpha_5 = \infty$, and α_1 through α_4 are unknown model parameters. In the general notation of section 2.1, equations (20) and (21) correspond to the specification of the model in (1).

⁵Note 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 2 but only makes the results harder to read.

In order to derive a parametric expression of conditional outcome probabilities, assume that the i.i.d. error terms e_{it} are i.i.d. with a logistic distribution. 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

$$P(\mathbf{y}_{it}|\mathbf{x}_i, a_{it}) = \Lambda(\alpha_{y_{it}} - \mathbf{x}_{it}\boldsymbol{\beta} - a_{it}) - \Lambda(\alpha_{y_{it-1}} - \mathbf{x}_{it}\boldsymbol{\beta} - a_{it}). \quad (22)$$

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 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

$$f(a_{it}|\mathbf{x}_i) = \phi(a_{it}; 0, \sigma^2). \quad (23)$$

Assume the AR(1) structure

$$a_{it} = \rho a_{i,t-1} + u_{it} \quad (24)$$

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 (4) of

$$f(a_{it}|\mathbf{x}_i, a_{i,t-1}) = \phi(a_{it}; \rho a_{i,t-1}, (1 - \rho^2)\sigma^2). \quad (25)$$

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

A standard ordered logit model follows in the special case $\sigma = 0$ and a standard random effects ordered logit model follows in the case $\rho = 1$. The correlation between Y_{it} and Y_{is} conditional on the 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 2.

3.3 Likelihood approximation and estimation performance

For the SRHS model described in section 3, this section implements and compares the following algorithms for evaluating the likelihood function discussed above:

- **Random Simulation:** Simulation of the full sequence of the latent state space using a standard random number generator for drawing from the joint distribution.
- **Antithetic Simulation:** The same algorithm but using 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:** Numerical integration on sparse grids. The algorithm for generating nodes and weights is given in Heiss and Winschel (2006).⁶
- **Nonlinear particle filter:** the standard nonlinear particle filter (Fernández-Villaverde and Rubio-Ramírez 2006) is implemented with MLHS for the initial state and innovations to improve the performance.
- **Sequential importance sampling:** The algorithm of ? except that a fixed grid of nodes instead of antithetic draws are used. For the univariate state space, this proved to be most successful.
- **Sequential Gaussian quadrature** as detailed in section 2.4. Pseudo-Code is shown in the appendix.

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

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

needs to achieve accurate results. First, the approximate values of the log likelihood function at a fixed parameter vector is calculated and compared to its limiting value. Second, each method is used to estimate the parameter vector by maximizing the approximated likelihood function.

Figure 1 shows the approximated log likelihood value at a fixed parameter vector for the different algorithms with the numbers of calculations R on the abscissa – note the logarithmic scaling. As expected, all algorithms converge to the same value, but the speed of convergence differs dramatically. 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 practically indistinguishable from the limit $R \rightarrow \infty$.

The number of evaluations of the conditional outcome probabilities R is not the only determinant of computational costs. The additional required calculations are

- The random simulation, antithetic simulation, and nonlinear particle filter require a large number of random numbers. With $R = 5000$, $102233 * 5000 \approx 500$ million random numbers have to be generated. With 8 bytes storage for each number, this corresponds to roughly 4GB. This is beyond the RAM capacity of most modern per-

sonal computers, so the numbers have to be sequentially generated for each likelihood evaluation.

- The resampling step of the nonlinear particle filter is computationally costly.
- The updating of the importance weights of the sequential importance sampling and Gaussian quadrature algorithms is not too expensive with low R but rises quadratically with R since it involves the calculation of R weighted sums over R elements.

To provide a different comparison, Figure 2 shows the same results as Table 1 but with the total time the implemented methods needed for each likelihood evaluation instead of the number of function evaluations. The methods were implemented in Matlab and runs on a Pentium 4 PC with 3GHz. Of course, these results depend on how efficiently the different algorithms were coded. While the author tried to do a good job for all algorithms, these results should not be interpreted too literally. As can be seen, the long run times for the resampling step make the nonlinear particle filter less competitive than when just considering the number of evaluations R . The random simulation now performs better than the MLHS because the generation of random numbers is considerably faster. The sequential Gaussian quadrature run very fast, so its advantage over the other methods is at least as pronounced as in Figure 1.

The ultimate goal of the approximated likelihood functions is to base parameter estimation on them. Intuitively, a better approximation of the likelihood function *ceteris paribus* leads to better estimates based on it. For the different algorithms and accuracy levels discussed above, the model parameters can for example be estimated by maximization of the approximated likelihood. 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 3 shows the QML estimates obtained by 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 3 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.⁷

A measure of overall deviations of the estimated parameters from their limiting values is shown in Figure 4. It shows the LR test statistic for the null hypothesis that all parameters are equal to the estimates obtained by the different algorithms, where the statistic is calculated for the “reference algorithm”. The broad message of this graph is the same as from the previous figures: all methods converge to the parameters obtained by the “reference algorithm” so that the test statistic approaches zero. Sequential quadrature does so faster than the other algorithms and much faster than the brute force simulation methods.

3.4 Implied Correlation Pattern

How well is this simple and parsimonious model able to capture the intertemporal correlation patterns in the data? Table 4 approaches this question. In the first column, the

⁷This may be due to the downward bias of the simulated log likelihood function discussed above. This bias due to the log transformation tends to be larger if the simulated probabilities are noisier which depends in turn on the number of draws but also on the variance parameter. With a smaller value of σ , the downward bias of the simulated log likelihood decreases so that it peaks at a lower value of σ than the actual log likelihood function.

parameter estimates from the descriptive regression on covariates and five lags of SRHS are repeated from Table 2. The other three columns show results from simulated data sets. Given the SRHS model, the parameter estimates and the actual covariates in the data, 100 different data set were simulated. The same descriptive regression is then repeated for each of the simulated data sets. The table shows the mean and the 95% confidence interval over the 100 repetitions.

Most of the parameters are well in line with the original estimates. The coefficients of lagged SRHS are all highly significant and decrease over time. Only the coefficient of the most recent lag (SRHS wave 5) is significantly higher in the original estimates than with the simulated data. This might be an indication that the model specification can be refined, for example the AR(1) process is too simple. Or the correlation parameters differ across the population so that some interactions might help. Remember that the model is estimated on the full sample, whereas the results in Table 4 are obviously for the subsample with six observations. This subsample only contains the original (relatively young) HRS cohort. Overall, the model does a very good job in replicating the intertemporal correlation pattern – much better than e.g. a first-order markov chain model with a random effect which would imply equal coefficients for SRHS in wave 1 through 4.

The general model structure discussed in section 2.1 would also easily allow more elaborate models. An obvious extension would be to add mortality to the measurement model. This would allow a straightforward and model-consistent treatment of the obvious dynamic selection effect through mortality. Another straightforward generalization would be to specify the unobserved state process in continuous time. The only required change would be in the transition equation (24). This would allow to easily take care of the fact that the time between surveys, and therefore probably also the correlation between adjacent measures, differ considerably in the HRS. For the discussion of such issues, see Heiss, Börsch-Supan, Hurd and Wise (2006).

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 presents results of a small Monte Carlo study. 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 latent variable is

$$Y_{it}^* = x_{it}\beta + a_{it} + e_{it} \quad \forall i = 1, \dots, N, \quad t = 1, \dots, T$$

where x_{it} generated from an AR(1) process with a marginal standard normal distribution and a correlation over time of 0.5. The parameter 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 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 5.

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 6 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 6 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 larger computational costs. SGQ delivers better results even with only five nodes and with 51 nodes, the error is smaller than 10^{-11} . With calculations being done with double precision, this is indistinguishable from rounding errors.

Table 7 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 variance of the AR(1) term is the most problematic for the likelihood approximation. SGQ needs 51 nodes to achieve an errors of less than 10^{-2} . The simulation methods look pretty worthless in this case. Even with 5000 nodes, there error remains larger than 10^2 . As a comparison between specifications 2 and 3 show, 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 fixed effects.

4.2 A Panel Probit Model

Finally, Table 8 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 a GHK simulators – one based on Halton sequences and one using a standard random number generator. As

discussed above, GHK is a specialized algorithm which makes use of the joint normality of and samples from the compound error term $a_{it} + e_{it}$. As previous research like Hajivassiliou et al. (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 the numerical approximation of the likelihood for a certain class of nonlinear panel data models including limited dependent variable models with AR(1) error terms. The computational difficulties arise because the likelihood function involves multiple integrals without analytic solutions. While methods for multiple numerical integration are available, their accuracy decreases with a rising dimensionality if the computational effort is held constant. Equivalently, the computational costs for a given accuracy increase with a rising dimensionality.

The paper discusses how these models allow to split the multiple integrals into several integrals with lower dimensions using nonlinear filtering algorithms. Since these integrals are approximated accurately with relatively low computational costs, the overall approximation can be expected to perform better than the “brute force” approach to approximate the joint integral. In the models discussed here, the state space is one-dimensional, allowing a straightforward and potentially highly accurate numerical integration by Gaussian quadrature. An nonlinear Kalman filter algorithm based on this approach is presented.

In an application, the panel data modeling of self-rated health status is discussed. It is argued that an ordered logit model with an AR(1) error term is more plausible than the typically specified random effects and/or first-order Markov models. It is also more parsimonious and yet captures the observed intertemporal correlation pattern better. For the estimation of this illustrative model, different algorithms are implemented and their approximation accuracy is compared. The sequential algorithms work much better than the

joint simulation. In this application with a univariate state space, the sequential Gaussian quadrature approach clearly performs best. Finally, a Monte Carlo study confirms the favorable performance of the proposed sequential quadrature approach for various data generating processes in the class of models discussed here.

Appendix: Pseudo-Code for sequential Gaussian quadrature

The sequential Gaussian quadrature algorithm was discussed in section 2.4. In the following, pseudo-code for the implementation of the SRHS model is presented for the convenience of the reader.⁸

1. Preparations:

- Fix a number of replications R .
- Obtain R nodes and weights for Gaussian quadrature and store them in the $R \times 1$ vectors \mathbf{a} and \mathbf{w} , respectively.
- For the updating, relative densities are required often. Since they do not change, calculate them once and reuse them every time: Generate a $R \times R$ “transition matrix” \mathbf{m} , where

$$\mathbf{m}(\mathbf{r}, \mathbf{c}) = \frac{\phi(\mathbf{a}(\mathbf{r}); \text{rho} \cdot \mathbf{a}(\mathbf{c}), (1 - \text{rho}^2) \cdot \text{sigma}^2)}{\phi(\mathbf{a}(\mathbf{r}); 0, \text{sigma}^2)} \text{ represents } \frac{f(\mathbf{a}(\mathbf{r})|\mathbf{x}_i, \mathbf{a}(\mathbf{c}))}{f(\mathbf{a}(\mathbf{r})|\mathbf{x}_i)}.$$

2. For each cross-sectional unit $i = 1, \dots, N$

- a) Initialize the vector \mathbf{q} as a $R \times 1$ vector of ones.
- b) For each wave $t = 1, \dots, T$ (T may differ across i):
 - Calculate the $R \times 1$ vector of weighted conditional probabilities \mathbf{qp} as $\mathbf{qp}(\mathbf{r}) = \mathbf{q}(\mathbf{r}) \cdot P(y(i, t)|\mathbf{x}(i, t), \text{sigma} \cdot \mathbf{a}(\mathbf{r}))$.
 - Approximate the likelihood contribution according to (18) as $L(i, t) = \mathbf{qp}'\mathbf{w}$.
 - Update the importance weights according to (19) as $\mathbf{q} = \frac{1}{L(i, t)}(\mathbf{m} * \mathbf{qp})'\mathbf{w}$ with “*” denoting matrix multiplication.

⁸All calculations for this paper were done in Matlab. The actual code can be requested from the author.

References

- Bauwens, Luc, Hautsch, Nikolaus 2006. Stochastic conditional intensity processes. *Journal of Financial Econometrics* **4**(3): 450–493.
- Börsch-Supan, Axel, Hajivassiliou, Vassilis 1993. Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics* **58**: 347–368.
- Butler, J. S., Moffit, Robert 1982. A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica* **50**(3): 761–764.
- Chamberlain, Gary 1984. Panel data. In *Handbook of Econometrics*, ed. Zvi Griliches and Michael D. Intriligator, vol. II. Amsterdam, New-York: Elsevier pp. 1247–1318.
- Contoyannis, Paul, Jones, Andrew M, Rice, Nigel 2004. The dynamics of health in the british household panel survey. *Journal of Applied Econometrics* **19**: 473 – 503. Mimeo, University of York.
- Danielsson, Jon, Richard, Jean-François 1993. Accelerated gaussian importance sampler with application to dynamic latent variable models. *Journal of Applied Econometrics* **8**: S153–S173.
- Doucet, Arnaud, Nando De Freitas, Neil Gordon, eds 2001 *Sequential Monte Carlo Methods in Practice*. New York: Springer Verlag.
- Durbin, James, Koopman, Siem Jan 1997. Monte carlo maximum likelihood estimation for non-gaussian state space models. *Biometrika* **84**(3): 669–684.
- 2001 *Time Series Analysis by State Space Methods*. Oxford University Press.
- 2002. A simple and efficient simulation smoother for state space time series analysis. *Biometrika* **89**(3): 603–616.

- Fernández-Villaverde, Jesús, Rubio-Ramírez, Juan F. 2005. Estimating dynamic equilibrium economies: Linear versus nonlinear likelihood. *Journal of Applied Econometrics* **20**: 891–910.
- 2006. Estimating macroeconomic models: A likelihood approach. NBER Technical Working Paper 321.
- Geweke, John F., Keane, Michael P., Runkle, David E. 1997. Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics* **80**: 125–165.
- Hajivassiliou, Vassilis A., Ruud, Paul A. 1994. Classical estimation methods for LDV models using simulation. In *Handbook of Econometrics Vol. IV*, ed. Robert F. Engle and Daniel L. McFadden. New-York: Elsevier pp. 2383–2441.
- Hajivassiliou, Vassilis, McFadden, Daniel, Ruud, Paul 1996. Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. *Journal of Econometrics* **72**: 85–134.
- Hamilton, James D. 1994. State-space models. In *Handbook of Econometrics Volume 4*, ed. R. Engle and D. McFadden. North-Holland chapter 50.
- Heckman, James J. 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*, ed. Charles F. Manski and Daniel McFadden. Cambridge, Mass.: MIT Press pp. 179–195.
- Heiss, Florian, Börsch-Supan, Axel, Hurd, Michael, Wise, David 2006. Pathways to disability: Predicting health trajectories. In *Perspectives on the Economics of Aging*, ed. David Wise. University of Chicago Press. forthcoming.
- Heiss, Florian, Winschel, Viktor 2006. Estimation with numerical integration on sparse grids. Technical Report, Department of Economics Discussion paper No. 2006-15, University of Munich. <http://econpapers.repec.org/paper/lmumuenec/916.htm>.

- Hess, Stephane, Train, Kenneth E., Polak, John W. 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, Bo E., Tamer, Elie 2006. Bounds on parameters in panel dynamic discrete choice models. *Econometrica* **74**: 611–629.
- Keane, Michael P. 1994. A computationally practical simulation estimator for panel data. *Econometrica* **62**(1): 95–116.
- Koopman, Siem Jan, Lucas, André 2005. Business and default cycles for credit risk. *Journal of Applied Econometrics* **20**(2): 311–323.
- Lee, Lung-Fei 1997. Simulated maximum likelihood estimation of dynamic discrete choice statistical models: Some monte carlo results. *Journal of Econometrics* **82**: 1–35.
- Richard, Jean-François, Zhang, Wei 2005. Efficient high-dimensional importance sampling. Technical Report, Working Paper, University of Pittsburgh, Dept. of Economics.
- Shephard, Neil, Pitt, Michael K. 1997. Likelihood analysis of non-gaussian measurement time series. *Biometrika* **84**: 653–667.
- Tanizaki, H., Mariano, R.S. 1994. Prediction, filtering and smoothing in non-linear and non-normal cases using monte carlo integration. *Journal of Applied Econometrics* **9**(2): 163–179.
- Tanizaki, Hisashi 1999. Nonlinear and nonnormal filter using importance sampling: Antithetic monte-carlo integration. *Communications in Statistics, Simulation and Computation* **28**: 463–486.
- 2003. Nonlinear and non-gaussian state-space modeling with monte carlo techniques: A survey and comparative study. In *Handbook of Statistics*, ed. D.N. Shanbhag and C.R. Rao, vol. 21. Elsevier pp. 871–929.

Wooldridge, Jeffrey M. 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* **20**: 39–54.

Table 1: 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 2: 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, + : $p < 0.10$, * : $p < 0.05$, ** : $p < 0.01$

Table 3: 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, ** : $p < 0.01$

Table 4: Correlation patterns: ordered logit on original and simulated data

	original	simulated data		
	data	mean	2.5%	97.5%
age	-0.0188	-0.0036	-0.0140	0.0083
female	0.0823	0.0214	-0.0726	0.1138
high school	0.1861	0.1590	0.0294	0.2853
some college	0.1762	0.2310	0.0938	0.4135
college degree+	0.3308	0.3642	0.2209	0.5177
nonwhite	-0.1198	-0.1286	-0.2601	-0.0196
SRHS wave 5	0.9847	0.7832	0.7206	0.8481
SRHS wave 4	0.5175	0.4874	0.4255	0.5567
SRHS wave 3	0.3251	0.3161	0.2470	0.3795
SRHS wave 2	0.2034	0.2112	0.1394	0.2804
SRHS wave 1	0.2390	0.1676	0.0937	0.2206

Table 5: 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 6: Monte Carlos 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}	$1.049 \times 10^{+0}$

Table 7: 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}

(continued on next page...)

(...continued from previous page)

Sequential Algorithms			Joint Simulation			
nodes	SGQ	SIS	nodes	MLHS	Halton	Random
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}

Table 8: Monte Carlos Results Probit model with $T = 10, \sigma = 1, \rho = 0.5$

Joint Simulation					
nodes	SGQ	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}

Figure 1: Approximate log likelihood at fixed parameter vector

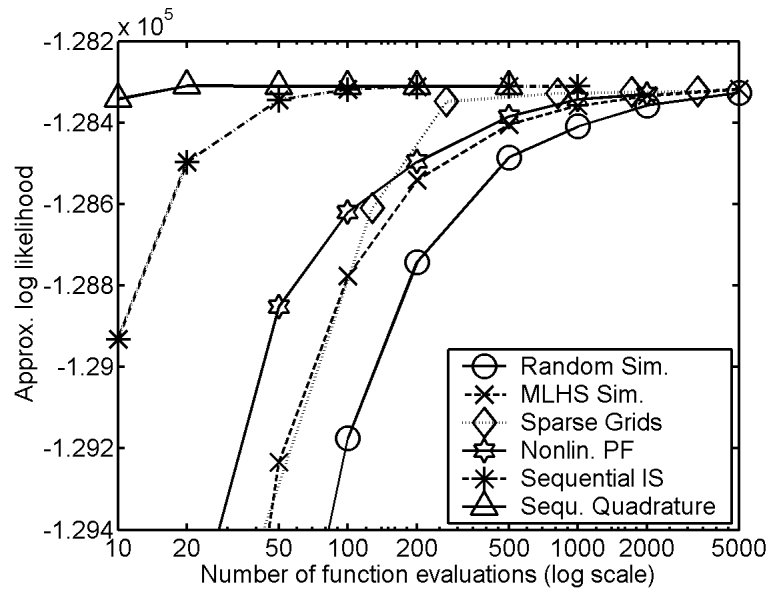


Figure 3: Results: Estimated Parameters σ and ρ

(a) σ

(b) ρ

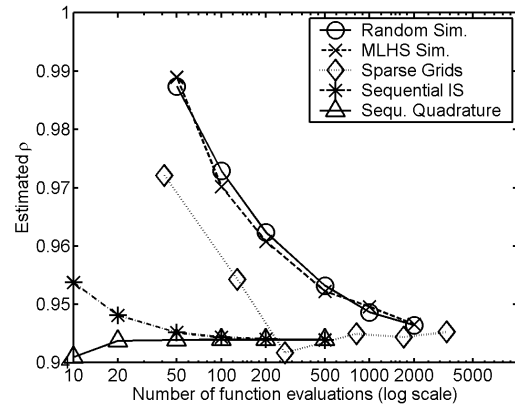
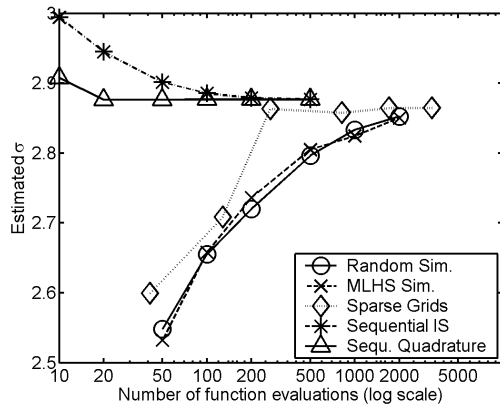


Figure 4: Results: LR statistic for estimated parameters

