Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Divide and Conquer: Recursive Likelihood Function Integration for Hidden Markov Models with Continuous Latent Variables

Gregor Reich

informs
http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Vol. 66, No. 6, November–December 2018, pp. 1457–1470
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Divide and Conquer: Recursive Likelihood Function Integration for Hidden Markov Models with Continuous Latent Variables

Gregor Reich[a]

[a] Department of Business Administration, University of Zurich, 8044 Zurich, Switzerland
Contact: gregor.reich@uzh.ch, http://orcid.org/0000-0002-1665-6835 (GR)

Received: February 6, 2015
Revised: June 13, 2016; November 7, 2017; February 7, 2018
Accepted: March 2, 2018
Published Online in Articles in Advance: November 23, 2018

Subject Classification: economics: econometrics; dynamic programming: applications; statistics: estimation
Area of Review: Computational Economics

https://doi.org/10.1287/opre.2018.1750

Abstract. This paper develops a method to efficiently estimate hidden Markov models with continuous latent variables using maximum likelihood estimation. To evaluate the (marginal) likelihood function, I decompose the integral over the unobserved state variables into a series of lower dimensional integrals, and recursively approximate them using numerical quadrature and interpolation. I show that this procedure has very favorable numerical properties: First, the computational complexity grows linearly in the number of periods, making the integration over hundreds and thousands of periods feasible. Second, I prove that the numerical error accumulates sublinearly in the number of time periods integrated, so the total error can be well controlled for a very large number of periods using, for example, Gaussian quadrature and Chebyshev polynomials. I apply this method to the bus engine replacement model of Rust [*Econometrica* 55(5): 999–1033] to verify the accuracy and speed of the procedure in both actual and simulated data sets.

Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2018.1750.

Keywords: hidden Markov models • maximum likelihood estimation • numerical integration • interpolation

1. Introduction

This paper develops a method to efficiently estimate hidden Markov models with continuous latent variables using maximum likelihood estimation (MLE). To evaluate the (marginal) likelihood function, I decompose the integral over the unobserved state variables into a series of lower dimensional integrals and recursively approximate them using numerical quadrature and interpolation. I show that this procedure has very favorable numerical properties: First, the computational complexity grows linearly in the number of periods, making the integration over hundreds and thousands of periods feasible. Second, I prove that the numerical error accumulates sublinearly in the number of time periods integrated, so the total error can be well controlled for a very large number of periods using, for example, Gaussian quadrature and Chebyshev polynomials. I apply this method to the bus engine replacement model of Rust (1987) to verify the accuracy and speed of the procedure in both actual and simulated data sets.

An important application of hidden Markov models within economics are the dynamic discrete choice models (DDCMs). Although plenty of other uses exist—inside and outside of economics (see, for example, the classic textbook of Elliott et al. 2008)—the focus of this paper's application is DDCMs of economic decision making, which has become a popular tool in the last three decades: First, many (individual) economic decisions we actually can observe are, in fact, discrete in nature, for example, the choice of a brand or medical treatment. Second, the underlying utility maximization problem of the agents is often dynamic in nature: decisions made today not only influence today's payoffs, but also influence future decisions and payoffs. By capturing these key facts, DDCMs have a wide range of uses; for pioneering papers see, for example, Miller (1984), Wolpin (1984), Pakes (1986), and Rust (1987). For extensive surveys, see Aguirregabiria and Mira (2010) and Keane et al. (2011).

The majority of contributions to the literature on the estimation of DDCMs make strong distributional assumptions about the errors and other unobserved state variables. Probably most prominent is the assumption of independently and identically extreme value type I ($EV1$ iid) or Gumbel distributed errors (see, e.g., Walck 1996); obviously implied by the $EV1$ iid assumption but usually stated explicitly by a conditional independence assumption (CI), the errors are assumed to be serially uncorrelated. However, there exists a wide consensus that these assumptions are not made based on the existence of much empirical evidence, but rather for numerical tractability: serial independence—alongside other distributional assumptions—induce closed-form solutions to potentially high-dimensional integrals that arise in the solution to the dynamic optimization problem and in the choice probabilities in the likelihood function. These closed-form solutions go back to the work of McFadden (1974, 1981) and Rust (1987). If, however, no closed-form solutions exist, it is common understanding that the likelihood function is hard to compute: "the likelihood function for a DDCM can be thought of as an integral over latent variables (the unobserved state variables). If the unobservables are serially correlated, computing this integral is very hard" (Norets 2009, p. 1665).

This conclusion follows from the fact that the integral over serially correlated errors really has dimensionality equal to the time horizon of the data times the number of choices with a serially correlated error component attached to it.

Although relaxing the $EV1$ error assumption has attracted some attention—for example, Larsen et al. (2012) test the statistical significance of allowing for more general distributions in the Rust (1987) model—several papers have developed integrated methods to estimate models without the CI assumption, thus allowing for a general notion of serially correlated unobserved state variables. Among those are the expectation–maximization algorithm based on the conditional choice probability estimation of Arcidiacono and Miller (2011), the particle filter method of Blevins (2016), and the Markov chain Monte Carlo (MC) approaches of Norets (2009, 2012). Apart from those, several papers use Monte Carlo integration to directly approach the integration over the unobserved state variables; among them are the simulation and interpolation method of Keane and Wolpin (1994); the patent model of Pakes (1986), which is considered one of the pioneering DDC models; and the application of Gaussian quadrature and interpolation as discussed in Stinebrickner (2000). The application of MC is motivated by the fact that the variance of the MC estimate of the integral does not depend on the dimension of the integral and, thus, not on the time horizon of the data. However, MC creates a source of stochastic error or "noise" in the estimation problem, and thus, standard likelihood maximization techniques are likely to fail. Therefore, methods to "average out" the sampling error, such as the simulated method of moments (SMM; McFadden 1989), have been developed; see Eisenhauer et al. (2015) for a comparison of SMM and MLE.

The approach followed in this paper is quite different by identifying and exploiting the structure that is present in the integral over the unobserved state variables in the (marginal) likelihood function: Given the serial dependence of the unobserved state variables in Markov, the time structure allows the high-dimensional integral over the time horizon to be decomposed and rewritten as a sequence of low-dimensional integrals. Then, I can recursively approximate this sequence to high accuracy, using highly efficient approximation schemes for low-dimensional integrals, such as Gaussian quadrature, and interpolate this approximation to iterate over the time dimension.

Although it is straightforward to see that the computational complexity of computing this integral is linear in the time dimension, one of the main contributions of this paper is the analysis of the numerical properties of this method, which I call "recursive likelihood function integration" (RLI). First, I prove that the numerical error in the RLI method accumulates sublinearly in the number of time periods integrated. Note that linear error growth can easily be controlled by choosing efficient numerical quadrature and interpolation methods. Second, the convergence rate of the method is derived in terms of the convergence rates of the particular quadrature and the interpolation methods used. Finally, I formulate generic assumptions on the continuous-state hidden Markov models that make the convergence results of the RLI method applicable.

Recursive computation of the likelihood function for serially correlated unobserved Markov states is not a new idea in general. However, to the best of my knowledge, its application has been limited to discrete state spaces so far and, therefore, did neither require numerical quadrature nor function approximation; see, for example, Cosslett and Lee (1985) for the estimation of models with Markov regime switching. Rather, the "integration" of the likelihood function really constitutes a finite sum, which can be easily computed up to machine precision. Recently, Connault (2016) applied this idea to compute the likelihood function of dynamic discrete choice models with discrete unobserved state variables.[1]

Although the focus of the paper is to approximate the likelihood function of hidden Markov models, solving the DDC model usually also requires substantial numerical work unless a two-step estimator in the sense of Hotz and Miller (1993) is used (also see Arcidiacono and Ellickson (2011) for a recent study on two-step estimators) or unless the solution of the model is combined with its estimation in a Bayesian framework as done by Imai et al. (2009) and Norets (2009). Several approaches to value function approximation have been proposed (see, for example, Rust 1996, Judd 1998, Cai and Judd 2013), and to stay flexible and generic, I use interpolation over an adaptively refined grid as proposed by Grüne and Semmler (2004); for the computation of the expectation over the value, I use Gaussian quadrature as was first proposed and successfully implemented in the context of DDCMs with serially correlated unobserved state variables by Stinebrickner (2000). Finally, I solve the maximum likelihood problem using a nested fixed-point algorithm (NFXP; Rust 1987), which is interconnected with the grid refinement process of the expected value function approximation.

As an application, I estimate the bus engine replacement model of Rust (1987) with serially correlated errors. One motivation for serial correlation in this model is a test for misspecification from the original paper, which leads to the following conclusion: "for groups 1, 2, and 3 and the combined groups 1–4 there is strong evidence that (CI) does not hold. The reason for rejection in the latter cases may be due to the presence of 'fixed-effects' heterogeneity which induces serial correlation in the error terms" (Rust 1987, p. 1027).

Testing for statistical significance of serially correlated errors, I find that in some subsamples of the original data set I can reject serially uncorrelated errors. Also, the parameter estimates vary substantially; their relative sizes, however, are rather stable.

The remainder of this paper is organized as follows: Section 2 first presents a motivating example by extending the bus engine replacement model of Rust (1987) to feature serially correlated errors (Section 2.1). Second, its solution and estimation procedure is discussed and a simple version of the recursive likelihood function integration algorithm is derived (Section 2.2). Finally, the model is estimated and tested for serial correlation in the errors (Section 2.3). The method developed to estimate the motivating example is then formalized in Section 3: First, rigorous definitions and assumptions on the integration and interpolation problems are introduced (Sections 3.1 and 3.2). Second, the recursive likelihood function integration method is analyzed with respect to accumulation of numerical error and convergence speed (Section 3.3) as well as its applicability to general continuous-state hidden Markov models (Section 3.4). Section 4 concludes.

2. A Motivating Example

2.1. The Bus Engine Replacement Model of Rust (1987)

In the bus engine replacement model of Rust (1987), an agent repeatedly makes decisions about the maintenance of a fleet of buses: Each period, the agent observes the state of each of the buses, including mileage, damage, signs of wear, etc. Based on these observations, the agent decides whether to do regular maintenance work only or a general overhaul; the latter is usually referred to as a replacement of the engine. The engine replacement causes a fixed cost of *RC* plus some random component, and the cost of regular maintenance is a function $c(\cdot)$ that is increasing in the current mileage state, plus some random component.

Formally, the agent faces single-period costs (or negative utility) for each individual bus,

$$u_\theta(i, x_t) + \varepsilon_t(i), \quad u_\theta(i, x_t) = \begin{cases} -RC & \text{if } i = 1 \\ -c(x_t, \theta_1) & \text{if } i = 0, \end{cases} \tag{1}$$

where i is the decision variable with $i = 1$ indicating engine replacement and $i = 0$ regular maintenance; $\varepsilon_t = (\varepsilon(0), \varepsilon(1))$ is a random utility component (or cost if negative as in this application) that is observed by the agent for all possible choices before making the actual decision; and x_t is the mileage of the individual bus at time t, which is reset to zero after an engine replacement. The replacement cost RC as well as the cost function parameter θ_1 are both parameters to be estimated. The maintenance cost function is assumed to be of the form $c(x_t, \theta_1) = \theta_1 \cdot x_t$. From the econometrician's point of view, mileage at the time of decision and the decision itself are observable for each bus and each time period. The random utility component, however, is only observable to the agent but not to the econometrician; consequently, it is often referred to as the unobserved state variable.

For the agent, the decision problem is how long to run a bus with regular maintenance only, with increasing costs induced by increasing mileage, and when to replace its engine, thus facing the one-time replacement cost but, at the same time, reducing the maintenance costs in the future because mileage is reset to zero. Assuming that the agent behaves dynamically optimally, the Bellman equation defines the value per bus as a function of its mileage state and the random utility components

$$\begin{aligned} V_\theta(x_t, \varepsilon_t) = \max_{i \in \{0,1\}} \{ & u_\theta(i, x_t) + \varepsilon_t(i) \\ & + \beta \mathbb{E}[V_\theta(x_{t+1}, \varepsilon_{t+1}) | i, x_t, \varepsilon_t; \theta] \}. \end{aligned} \tag{2}$$

The conditional expected continuation value in (2) is defined by

$$\begin{aligned} & \mathbb{E}[V_\theta(x_{t+1}, \varepsilon_{t+1}) \,|\, i, x_t, \varepsilon_t; \theta] \\ & \quad = \int V_\theta(x_{t+1}, \varepsilon_{t+1}) p_{x\varepsilon}(x_{t+1}, \varepsilon_{t+1} | i, x_t, \varepsilon_t; \theta) d(x_{t+1}, \varepsilon_{t+1}), \end{aligned} \tag{3}$$

with subscript θ denoting the dependence of the value function on the parameter values RC and θ_1 and where the integration limits are ignored for better readability; $p_{x\varepsilon}$ is the conditional joint density function of the state variable process.

The original model makes the following conditional independence assumption regarding the joint distribution of the state variables:

$$p_{x\varepsilon}(x_{t+1}, \varepsilon_{t+1} | i, x_t, \varepsilon_t; \theta) = \tilde{p}_{\varepsilon|x}(\varepsilon_{t+1} | x_{t+1}; \theta) p_x(x_{t+1} | i, x_t; \theta). \tag{4}$$

Assumption (4) ensures that (i) the mileage state transition is—conditional on the decision i—independent of the random utility component and (ii) that the random utility components are serially uncorrelated. If the CI assumption holds and if, moreover, the random utility components $\varepsilon(i)$ are distributed extreme value type I ($EV1$) iid, the integral in (3) has a closed-form solution. However, to allow for serial correlation in ε while keeping (i), I assume

$$p_{x\varepsilon}(x_{t+1},\varepsilon_{t+1}|i,x_t,\varepsilon_t;\theta)=p_{\varepsilon|i}(\varepsilon_{t+1}|i,\varepsilon_t;\theta)p_x(x_{t+1}|i,x_t;\theta). \tag{5}$$

A choice for serial correlation in the unobserved state variables that is frequently used in the literature is the $AR(1)$ process. More specifically, I define

$$\begin{aligned}\varepsilon_t(0)&=\rho\varepsilon_{t-1}(i_{t-1})+\tilde{\varepsilon}_t(0), &\tilde{\varepsilon}_t(0)\text{ iid}\\ \varepsilon_t(1)&=\tilde{\varepsilon}_t(1), &\tilde{\varepsilon}_t(1)\text{ iid,}\end{aligned} \tag{6}$$

and $p_{\tilde{\varepsilon}}(\cdot)$ as the probability density function of the innovations $\tilde{\varepsilon}_t(i)$ with zero mean. Note that ρ is an additional parameter of the estimation; furthermore, I assume that $\varepsilon_0(i)$ is distributed with density $p_{\tilde{\varepsilon}}(\cdot)$. Thus, I only assume the random utility component of regular maintenance to be serially correlated. It is important to note that definition (6) nests the original model for $\rho=0$ and, the density function $p_{\tilde{\varepsilon}}(\cdot)$ being extreme value type 1, $EV1$. Moreover, I consider two variants for each density function: the first variant uses the "standard" form of the distribution—like the standard normal distribution with mean zero and variance one—whereas the second normalizes the distribution of the innovation $\tilde{\varepsilon}$ such that the resulting $AR(1)$ process has zero mean and constant variance (i.e., its variance is independent of ρ), which is achieved by setting the location and scale parameters accordingly (as follows).

Given that mileage state x_t and decision i_t are observable for all buses but random utility components ε_t are not, the aim is to estimate this model's parameter $\theta=\{\theta_1,RC,\rho\}$, given the data $\{x_t,i_t\}_{t=0}^T$, by maximum likelihood estimation.

2.2. Maximum Likelihood Estimation with Unobserved Serial Correlation

This subsection develops a numerical method to estimate the bus engine replacement model with serially correlated unobserved state variables from the previous subsection. In particular, I motivate and develop a recursive method to integrate out the serially correlated state variables in the computation of the likelihood function, therefore approximating it to high accuracy using highly efficient deterministic quadrature and interpolation rules for low and medium dimensions. I refer to this method as *recursive likelihood function integration*. Note that the detailed strategies for the likelihood maximization, the solution of the model with the new state variable, and the integration of model solution and estimation are deferred to the e-companion EC.2.3. A rigorous assessment of the numerical properties of the RLI method and its applicability to the estimation of general dynamic Markov models with unobserved serially correlated states are formally derived in Section 3.

The *marginal* likelihood function of observing a particular history of state transition and maintenance decisions for one individual bus derives as follows:

$$\begin{aligned}L_T(\theta)&\equiv P_{xi}(\{x_t,i_t\}_{t=1}^T|\{x_0,i_0\};\theta)\\ &=\int\ldots\int p_\varepsilon(\varepsilon_0;\theta)P_{xi\varepsilon}(\{x_t,i_t,\varepsilon_t\}_{t=1}^T|\{x_0,i_0,\varepsilon_0\};\theta)\\ &\quad\cdot d\varepsilon_0 d\varepsilon_1\ldots d\varepsilon_T,\end{aligned} \tag{7}$$

where the integration limits are ignored for better readability but are each from minus infinity to plus infinity in this example.

The likelihood function of the full panel computes as the product of the likelihood functions of the individual buses because the state variables are assumed to be independently distributed across buses. Incorporating the assumption that all state transitions are Markov, I can factorize the likelihood of observing a particular time series as

$$\begin{aligned}&P_{xi\varepsilon}(\{x_t,i_t,\varepsilon_t\}_{t=1}^T|\{x_0,i_0,\varepsilon_0\};\theta)\\ &\quad=\prod_{t=2}^T p_{xi\varepsilon}(x_t,i_t,\varepsilon_t|x_{t-1},i_{t-1},\varepsilon_{t-1};\theta).\end{aligned} \tag{8}$$

I can further decompose the joint transition probability density in (8), using the fact that, given x_t and ε_t, i_t is independent of i_{t-1}, ε_{t-1}, and x_{t-1} as well as incorporating Equation (5):

$$\begin{aligned}&p_{xi\varepsilon}(x_t,i_t,\varepsilon_t|x_{t-1},i_{t-1},\varepsilon_{t-1};\theta)\\ &\quad=p_{i|x\varepsilon}(i_t|x_t,\varepsilon_t;\theta)p_{\varepsilon|i}(\varepsilon_t|i_{t-1},\varepsilon_{t-1};\theta)p_x(x_t|x_{t-1},i_{t-1};\theta).\end{aligned}$$

For notational simplicity, I define

$$m_{it}\equiv u_\theta(i,x_t)+\beta\mathbb{E}[V_\theta(x_{t+1},\varepsilon_{t+1})|i,x_t,\varepsilon_t;\theta].$$

Although $p_{\varepsilon|i}(\varepsilon_t|i_{t-1},\varepsilon_{t-1},\theta)$ is determined by (6) and $p_x(x_t|x_{t-1},i_{t-1})$ is estimated independently (and, therefore, omitted from now on), the density function of the conditional decision probability, $p_{i|x\varepsilon}(i_t|x_t,\varepsilon_t,\theta)$, is given by

$$\begin{aligned}p_{i|x\varepsilon}(1|x_t,\varepsilon_t(0),\varepsilon_t(1);\theta)&=\mathbb{1}(m_{1t}+\varepsilon_t(1)>m_{0t}+\varepsilon_t(0))\\ p_{i|x\varepsilon}(0|x_t,\varepsilon_t(0),\varepsilon_t(1);\theta)&=\mathbb{1}(m_{1t}+\varepsilon_t(1)\le m_{0t}+\varepsilon_t(0)),\end{aligned} \tag{9}$$

where $\mathbb{1}(\cdot)$ is the index function that is equal to one if its argument is true and zero otherwise; note that the conditional decision probabilities are actually degenerate because—loosely speaking—there is no randomness left given ε_t.

Finally, exploiting the Markov structure for the integration and dropping parameter dependence for better readability, I can write the likelihood function (7) as

$$\begin{aligned} &L_T(\theta)\\ &= \int\ldots\int p_\varepsilon(\varepsilon_0;\theta)\prod\nolimits_{t=1}^{T-1} p_{i|x\varepsilon}(i_t|x_t,\varepsilon_t;\theta)p_{\varepsilon|i}(\varepsilon_t|i_{t-1},\varepsilon_{t-1};\theta)\\ &\left(\int p_{i|x\varepsilon}(i_T|x_T,\varepsilon_T;\theta)p_{\varepsilon|i}(\varepsilon_T|i_{T-1},\varepsilon_{T-1};\theta)d\varepsilon_T\right)d\varepsilon_0\ldots d\varepsilon_{T-1}. \end{aligned} \tag{10}$$

To numerically approximate (10), I define the following recurrence relation:

$$g_t(\varepsilon)=\begin{cases}1 & t>T\\ \int p_{i|x\varepsilon}(i_t|x_t,\varepsilon';\theta)p_{\varepsilon|i}(\varepsilon'|i_{t-1},\varepsilon;\theta)g_{t+1}(\varepsilon')d\varepsilon' & \text{otherwise.}\end{cases} \tag{11}$$

Now, given $g_{t+1}(\cdot)$, I can numerically approximate the function $g_t(\cdot)$ using both numerical integration and function approximation. Because $g_t(\cdot)$ is known to be unity for $t>T$, I can use backward iteration starting from $g_T(\cdot)$ to solve for $g_0(\cdot)$, which is the approximation of the likelihood function $L_T(\theta)$. Note that this procedure is somewhat similar to solving for the value function of a finite-horizon, discrete-time dynamic programming problem by backward iteration. Algorithm 1 proposes a simple implementation of the procedure but is generic with respect to both the numerical integration scheme and the function approximation schemes as long as the latter depend on function evaluations only.

Algorithm 1. Computation of the likelihood function (10) by recursive likelihood function integration (RLI)

```
$\Gamma \leftarrow$ initialize ε-grid with D elements
$\hat{g}(\cdot) \leftarrow$ initialize interpolant $\{(e,\tilde{g}_e)\}_{e\in\Gamma}$ to 1
for $t = T,\ldots,1$ do
    for $e \in \Gamma$ do
        $\tilde{g}_e \leftarrow \int p_{i|x\varepsilon}(i_t|x_t,\varepsilon';\theta)p_{\varepsilon|i}(\varepsilon'|i_{t-1},\varepsilon;\theta)\hat{g}(\varepsilon')d\varepsilon'$
    end for
    $\hat{g}(\cdot) \leftarrow$ construct interpolant $\{(e,\tilde{g}_e)\}_{e\in\Gamma}$
end for
$L_{T\theta} \leftarrow \int p_\varepsilon(\varepsilon_0;\theta)\hat{g}(\varepsilon')d\varepsilon'$
return $L_{T\theta}$
```

Note that each integral over ε_t is generally still N-dimensional. Thus, the procedure decomposes the $T\cdot N$-dimensional integral of (7) to an N-dimensional integration that is repeated $D\cdot T$ times, where D is the number of nodes used for the approximation of $g_t(\cdot)$. Because the computational complexity of deterministic numerical integration is exponential in the number of dimensions in the worst case, this reduction is highly desirable even for large D because it enters the complexity of the overall algorithm linearly:

$$O(\exp(T\cdot N)) \gg O(D\cdot T\exp(N)),$$

where the complexity notation for "work" w reads as $w = O(f(T)) \Leftrightarrow \exists k<\infty : w \leq k\cdot f(T)$.

Given that serial correlation is only allowed in some dimensions but not all, I can potentially replace parts of the integral in (11) by a closed-form solution; this is particularly the case if the cumulative distribution of those unobserved state variables that are not serially correlated does have a closed form. Recall that the integration over ε_t is really N-dimensional, thus two-dimensional in the model under consideration:

$$\begin{aligned}\iint & p_{\varepsilon(0)|i}(\varepsilon_t(0)|i_{t-1},\varepsilon_{t-1}(0);\theta)p_{\varepsilon(1)}(\varepsilon_t(1);\theta)\\ &\cdot p_{i|x\varepsilon}(i_t|x_t,\varepsilon_t(0),\varepsilon_t(1);\theta)d\varepsilon_t(1)d\varepsilon_t(0).\end{aligned}$$

Using (9), I can write the integral over $\varepsilon_t(1)$ in terms of its cumulative distribution function $F_{\varepsilon(1)}(\cdot;\theta)$,

$$\begin{aligned}&\int_{-\infty}^{\infty} \mathbb{1}(\varepsilon_t(1) > m_{0t} - m_{1t} + \varepsilon_t(0))p_{\varepsilon(1)}(\varepsilon_t(1);\theta)d\varepsilon_t(1)\\ &= \int_{m_{0t}-m_{1t}+\varepsilon_t(0)}^{\infty} p_{\varepsilon(1)}(\varepsilon_t(1);\theta)d\varepsilon_t(1)\\ &= 1 - F_{\varepsilon(1)}(m_{0t} - m_{1t} + \varepsilon_t(0);\theta),\end{aligned}$$

which no longer involves numerical quadrature if an analytical formula for $F_{\varepsilon(1)}$ exists.

2.3. Estimation Results for the Bus Engine Replacement Model

In this subsection, I present estimation results for the bus engine replacement model of Rust (1987) featuring a serially correlated, unobserved random utility component as specified in Section 2.1; additionally, in the e-companion to this paper (cf. Section EC.2.2), I present an extensive Monte Carlo study with simulated data sets to assess the question to what extent the algorithm is able to reproduce the parameters of a distribution with known parameters and what estimator variance can be expected from various data set sizes.

The original data set of Rust (1987) consists of monthly odometer readings and engine replacement decisions for a fleet of 162 buses, subdivided into eight groups depending on their manufacturer and model. Because buses are heterogeneous across groups, it is common to create different subsamples to estimate the parameters of model (1); I follow the literature by estimating three subsamples separately, consisting of groups $\{1,2,3\}$, $\{1,2,3,4\}$, and $\{4\}$, and I discretize mileage in "bins" of 5,000 miles each. The highest possible mileage state is 90 (which corresponds to 450,000 miles and is far from ever being reached in the data), formally $x \in X = \{1,\dots,90\}$. I assume the mileage transition to follow a Markov process (conditional on the replacement decision), for which I estimate the parameters independently. I parameterize the discount factor by $\beta = 0.9999$ as in the original paper.

Table 1 presents the estimation results using the original data set of Rust (1987), again for both extreme value type I and normally distributed innovations $\tilde{\varepsilon}$ and for each distribution family with and without normalization of the innovation distribution. Note that the $EV1$ distribution has a location and a scale parameter, μ and $\beta > 0$, respectively; the mean of a random variable with $EV1(\mu,\beta)$ distribution is $\mu + \gamma\beta$, and its variance is $\beta^2\pi^2/6$, where $\gamma \approx 0.5772$ is the Euler–Mascheroni constant. In particular, an $AR(1)$ process with "standard" $EV1$ innovations with density $EV1(-\gamma,1)$ will have mean zero and variance $\pi^2/6(1-\rho^2)^{-1}$; however, with normalized innovations, that is, if innovations are distributed according to the density $EV1(-\gamma\sqrt{1-\rho^2},\ \sqrt{1-\rho^2})$, the corresponding mean and variance will be zero and $\pi^2/6$, respectively. For normal innovations, the $AR(1)$ process without normalization, that is, with $N(0,1)$ innovations, will have zero mean and variance $(1-\rho^2)^{-1}$ whereas the normalized version with $N(0,1-\rho^2)$ innovations has mean zero and variance one.

For the $EV1$ case, I observe that, although the parameter estimates in the presence of serial correlation are substantially different from the estimates without serial correlation, the ratio of engine replacement cost to the regular maintenance cost parameter is relatively stable; thus, the trade-off for the decision maker has not changed much quantitatively. This result holds true for both innovation specifications, that is, with and without normalization, although the values of the parameters in the normalized version are somewhere in between the values without serial correlation and the values with serial correlation but no normalization. Moreover, the relative costs and the corresponding likelihood function values are almost identical for the two specifications with serial correlation. Performing a likelihood ratio test to compute the statistical significance of the quantitative changes induced by the introduction of serial correlation, I find that only on the largest subsample of the data set (bus groups 1–4) can I reject the hypothesis of no serial correlation at a reasonable significance level.

The case of normally distributed $\tilde{\varepsilon}$ yields similar results with two notable differences: First, not only do the cost parameter values change substantially, but also the ratios and the implied trade-off for the decision maker (i.e., the bus fleet manager). However, at the same time, carrying out a likelihood ratio test, I cannot reject the hypothesis of no serial correlation at a reasonable significance level for any of the subsamples in the normal case. Second, the density normalization has very little influence on the parameter estimates in the normal case in contrast to the $EV1$ case where the difference is substantial.

I interpret the change of the ratio of the cost parameters in this particular model as follows (as an example, I assume the ratio in the restricted model to be larger than in the unrestricted one): If I ignore serial correlation, the relative costs of regular maintenance are underestimated. Consequently, using the true relative costs in a model without serial correlation, I would predict more (or, equivalently, earlier) engine replacement than I find in the data. Thus, allowing for serial correlation explains why I do not observe more frequent engine replacement given the high (true) relative costs of regular maintenance. Conversely, in a model with serial correlation, but based on the biased relative costs estimates, I would predict the buses to run for too long without engine replacement.

Assessing the question of the statistical significance of the estimates from the original data set is difficult though. First, as I demonstrate in the Monte Carlo study in the e-companion (Section EC.2.2), my experiments with artificial data sets indicate that the results are rarely significant for small samples even if the true model features serial correlation as defined by (6). Consequently, given the number of buses in the original data set, p-values as for groups 1–4 with extreme value distributed $\tilde{\varepsilon}(i)$ are not what I can generally expect. Second, I still cannot conclude that the serial correlation I found in the data is really coming from an unobserved source as different bus groups are pooled together for two of the three subsamples, thus creating a heterogeneous sample that is treated as homogeneous by the model. Consequently, as long as I do not find the serial correlation *within* one single bus group to be significant, these estimations have to be taken with a grain of salt.

3. Recursive Likelihood Function Integration

This section derives the numerical properties of the recursive likelihood function integration method as outlined in the previous section, including error analysis, convergence rates, and the necessary properties of the model for the theoretical results to be applicable.

The section is structured as follows: First, to obtain a unified nomenclature, the concepts of numerical quadrature and interpolation are introduced (Sections 3.1 and 3.2, respectively); particular quadrature and interpolation methods are briefly presented and analyzed for their applicability in the RLI context in the

Table 1. Estimation Results for the Model of Rust (1987) with Serially Correlated Errors

	Bus groups 1–3				Bus groups 1–4				Bus group 4			
	Rust	Baseline	Standardized	Normalized	Rust	Baseline	Standardized	Normalized	Rust	Baseline	Standardized	Normalized
							$\tilde{\varepsilon} \sim EV1(\mu, \beta)$					
RC	11.7270	11.7266	25.0624	18.1397	9.7558	9.7560	27.0368	18.1927	10.0750	10.0749	22.0634	15.7403
	(2.602)	(1.928)	(10.127)	(10.859)	(1.227)	(0.898)	(16.939)	(6.724)	(1.582)	(1.351)	(13.873)	(9.452)
θ_1	4.8259	4.8257	9.8531	7.1312	2.6275	2.6276	7.4151	4.9894	2.2930	2.2929	4.8132	3.4330
	(1.792)	(1.366)	(4.564)	(5.591)	(0.618)	(0.469)	(5.551)	(2.424)	(0.639)	(0.554)	(3.738)	(2.610)
RC/θ_1	2.4300	2.4300	2.5396	2.5437	3.7130	3.7128	3.6462	3.6463	4.3938	4.3935	4.5840	4.5850
ρ	—	—	0.6899	0.6898	—	—	0.7396	0.7396	—	—	0.7001	0.7000
			(0.098)	(0.168)			(0.112)	(0.091)			(0.134)	(0.189)
$-L$	2,708.366	2,708.366	2,707.777	2,707.777	6,055.250	6,055.250	6,053.340	6,053.340	3,304.155	3,304.156	3,303.912	3,303.912
p (LR)			0.2903	0.2903			0.0506	0.0506			0.4848	0.4848
							$\tilde{\varepsilon} \sim N(0, \sigma)$					
RC		7.0372	13.8320	13.3909		6.0018	18.8660	17.5170		6.0747	10.8680	10.8111
		(1.029)	(2.736)	(0.552)		(0.481)	(2.671)	(2.777)		(0.758)	(0.948)	(4.056)
θ_1		2.5406	5.3814	5.4492		1.3990	5.2595	5.0840		1.1829	2.2881	2.4086
		(0.732)	(1.316)	(0.192)		(0.263)	(0.940)	(0.816)		(0.327)	(0.319)	(1.099)
RC/θ_1		2.7700	2.5717	2.4574		4.2900	3.5870	3.4455		5.1354	4.7497	4.4886
ρ		—	0.5203	0.5117		—	0.6680	0.6510		—	0.4887	0.4920
			(0.086)	(0.012)			(0.042)	(0.049)			(0.032)	(0.164)
$-L$		2,707.901	2,707.832	2,707.817		6,054.082	6,053.683	6,053.649		3,303.919	3,303.899	3,303.889
p (LR)			0.7103	0.6819			0.3717	0.3521			0.8446	0.8065

Notes. Estimation results for different subsamples of the original data set with the innovation distribution being extreme value type 1 $EV1(\mu, \beta)$ (top) and normal $N(0, \sigma)$ (bottom). The "Rust" columns quote the results of the original paper (table IX of Rust 1987), and as in Rust (1987), θ_1 is rescaled by 10^3; the "baseline" columns refer to the model without serial correlation; the "standardized" columns refer to innovation densities without normalization, that is, $EV1(-\gamma, 1)$ and $N(0, 1)$, whereas the "normalized" columns refer to normalized innovation densities, that is, $EV1(-\gamma\sqrt{1-\rho^2}, \sqrt{1-\rho^2})$ and $N(0, 1-\rho^2)$. $-L$ is the *negative* value of the log-likelihood function at the solution; p (LR) is the p-value of the likelihood ratio test with $H_0: \rho = 0; \beta = .9999$. For the estimation of the standard errors, the inverse of the negative Hessian of the likelihood function at its maximum is used: $(-H(\hat{\theta}|\{x_t, i_t\}_{t=0}^{T}))^{-1}$.

e-companion, Sections EC.1.2 and EC.1.3. As a result, the convergence speed for *parametric* integration is derived in dependence of the convergence speed of the quadrature and interpolation methods. Second, a *recursive* version of parametric integration is defined and analyzed for error propagation (Section 3.3); again, the convergence speed of the numerical approximation by recursively applying quadrature and interpolation is derived in dependence of the convergence speed of the quadrature and interpolation methods. The subsection is again accompanied by several examples, partly in the e-companion: a comparison of the method to Monte Carlo integration, confirming the theoretically derived error and convergence behavior (Section EC.1.4), as well as a convergence analysis for the application to the model of Rust (1987) from Section 2 (Section EC.2.1). Finally, the scope of applicability of the method to integrate out serially correlated unobserved state variables to obtain the marginal likelihood function is analyzed (Subsection 3.4).

3.1. Numerical Quadrature

Definition 1 (Kernel Integral). Given a function $f:\mathbb{R}^m \supseteq D \rightarrow \mathbb{R}$, the integral of f against a nonnegative and bounded *kernel* or *weighting function* $q:\mathbb{R}^m \supseteq D \rightarrow [0,a]$ with $a \in \mathbb{R}_+$ is denoted by

$$I_f = \int_D f(x)q(x)dx, \tag{12}$$

where $\int_D q(x)dx \le 1$; note that only finite integrals are considered; that is, $I_f < \infty$.

Note that the multiplication of the integrand f by a weighting function q in Definition 1 is without loss of generality (w.l.o.g.) because the kernel can be chosen as unity. However, it is generally needed to cover integrals over unbounded domains (e.g., $D = \mathbb{R}^m$).

Definition 2 (Approximation by Quadrature). $\hat{I}_f$ is an *approximation* of I_f by (numerical) quadrature if

$$I_f = \hat{I}_f + \epsilon_f^Q, \quad |\epsilon_f^Q| \ll 1,$$

where ϵ_f^Q is the *approximation error*.

In the context of this paper, I limit my attention to numerical quadrature rules that comply with the following definition:

Definition 3 (Quadrature Rule). A *quadrature rule* is any systematic choice of nodes and weights $\{(x_i,\omega_i)\}_{i=1}^{n^Q}$ such that I_f is approximated by

$$\hat{I}_f = \sum_{i=1}^{n_Q} \omega_i f(x_i), \tag{13}$$

and where $\{(x_i,\omega_i)\}_{i=1}^{n^Q}$ depend deterministically on n^Q but not on f.

The e-companion discusses two popular quadrature rules, namely Simpson integration and Gaussian quadrature along with their convergence properties in Section EC.1.2.

Note that Definition 3 rules out two popular approaches to numerical integration, namely Monte Carlo integration (nondeterministic) and adaptive integration methods (dependence on integrand).

Definition 4 (Convergence of Quadrature Rule). Given $f \in C^i$, the quadrature rule converges at rate s_Q if

$$|\epsilon_f^Q| = O(n_Q^{-s_Q}) \quad \Leftrightarrow \quad \forall f \in C^i : \exists k < \infty : |\epsilon_f^Q| \le k n_Q^{-s_Q}.$$

Note that by C^i, I refer to the space of functions that are i times continuously differentiable and for which the ith derivative is, moreover, bounded. Because functions in this context can be multivariate, this includes all partial derivatives and must hold for all dimensions. Moreover, because the integral (12) can be multivariate, the convergence rate as defined in Definition 4 is the total rate over *all* dimensions.

Definition 5 (Parametric Kernel Integral). Given a function $f:\mathbb{R}^{m_x} \times \mathbb{R}^{m_y} \supseteq D \times E \rightarrow \mathbb{R}$, the *parametric* kernel integral of f is denoted by the function

$$I_f(y) = \int_D f(x,y)q(x)dx, \tag{14}$$

where $I_f:\mathbb{R}^{m_y} \rightarrow \mathbb{R}$. Its approximation $\hat{I}_f(y)$ is given by

$$I_f(y) = \hat{I}_f(y) + \epsilon_f^Q(y),$$

where $\epsilon_f^Q(y)$ is the approximation error in dependence of the parameter y.

Definition 6 (Parametric Form). Given a function $f:\mathbb{R}^{m_x} \times \mathbb{R}^{m_y} \supseteq D \times E \rightarrow \mathbb{R}$, $f_{\bar{y}}(x) \equiv f(x,\bar{y}) = f(x,y)|_{y=\bar{y}}$ and $f_{\bar{x}}(y) \equiv f(\bar{x},y) = f(x,y)|_{x=\bar{x}}$ denote the *parametric form* of f with respect to y and x, respectively.

Loosely speaking, the parametric form of f fixes one of its two (potentially multivariate) arguments.

Definition 7 (Uniform Convergence of Parametric Integration). A quadrature rule for parametric integration as in Definition 5 converges uniformly if, for all integrands $f \in C^i$,

$$\lim_{n_Q \rightarrow \infty} \sup_{y \in E} |\epsilon_f^Q(y)| = 0.$$

Moreover, it converges at rate s_Q if

$$\|\epsilon_f^Q\|_\infty \equiv \sup_{y \in E} |\epsilon_f^Q(y)| = O(n_Q^{-s_Q}).$$

Remark 1 (Preservation of Smoothness). Note that the approximation of $I_f(y)$ by $\hat{I}_f(y)$ through a quadrature rule as defined by Definition 3 preserves smoothness in y because $\hat{I}_f(y)$ is just a weighted sum of functions in y:

$$\hat{I}_f(y) = \sum_{i=1}^{n_Q} \omega_i f_{\tilde{x}_i}(y).$$

Therefore, $f \in C^i \Rightarrow \hat{I}_f \in C^i$. However, the smoothness is not necessarily preserved by other numerical integration methods, such as adaptive quadrature or Monte Carlo integration, because the x_i either depend (nonsmoothly) on f or are nondeterministic; also note that, if only finite integrals and convergent quadrature methods are considered, the potentially infinite sum (as $n_Q \to \infty$) is always bounded.

3.2. Interpolation

Definition 8 (Interpolation). Given a function $f: \mathbb{R}^m \supseteq E \to \mathbb{R}$, n_I pairs of argument values and the corresponding function values, $\{(y_i, f(y_i))\}_{i=1}^{n_I}$, and an Ansatz function $\phi(y; \mathbf{a})$ with n_I parameters $\mathbf{a} = (a_1, \dots, a_{n_I})$, $\mathcal{I}_f \equiv \phi(\cdot\,; \mathbf{a})$ is called the interpolant of f if the parameters $\mathbf{a}$ are chosen such that

$$\phi(y_i; \mathbf{a}) = f(y_i) \forall i. \tag{15}$$

The corresponding interpolation error $\epsilon_f^I(y)$ is defined by

$$f(y) = \mathcal{I}_f(y) + \epsilon_f^I(y)$$

and is zero at the interpolation nodes, $\epsilon_f^I(y_i) = 0$.

The e-companion discusses three popular interpolation schemes, namely Chebyshev polynomials, piecewise linear interpolation, and cubic splines along with their convergence properties in Section EC.1.3.

Definition 9 (Uniform Convergence for Interpolation). An interpolation scheme converges uniformly if, for all functions $f \in C^i$,

$$\lim_{n_I \to \infty} \sup_{y \in E} |\epsilon_f^I(y)| = 0.$$

Moreover, it converges at rate s_I if

$$\|\epsilon_f^I\|_\infty \equiv \sup_{y \in E} |\epsilon_f^I(y)| = O(n_I^{-s_I}).$$

Because the interpolant in (15) can be multivariate, the convergence rate as defined in Definition 9 is the total rate over *all* dimensions.

Definition 10 (Preservation of Smoothness). The interpolation scheme preserves smoothness up to order j if

$$f \in C^i \Rightarrow \mathcal{I}_f \in C^j.$$

It fully preserves smoothness if, moreover, $i = j$.

Definition 11 (Approximation of Parametric Integration). Consider a parametric integration problem as in Definition 5; its approximation by quadrature and interpolation is defined as

$$I_f(y) = \hat{I}_f(y) + \epsilon_f^Q(y) = \mathcal{I}_{\hat{I}_f}(y) + \underbrace{\epsilon^I_{\hat{I}_f}(y) + \epsilon_f^Q(y)}_{\equiv \epsilon_{I_f}(y)}.$$

where $\epsilon_{I_f}(y)$ denotes the overall approximation error.

Note that although ϵ_f^Q depends on n_Q only, $\epsilon^I_{\hat{I}_f}$ depends on n_I and also on n_Q in a potentially nonmonotone way through $\hat{I}_f$; consequently, the proofs of the convergence rate results that follow will have to account for function *sequences* in multiple dimensions.

Proposition 1 (Convergence of Parametric Integration). *Consider the approximation of a parametric integration problem by quadrature and interpolation as in Definition 11, where $f \in C^i$, and quadrature and interpolation methods that converge uniformly in the sense of Definitions 7 and 9, respectively. Then, the approximation of the parametric integration problem converges uniformly as n_Q and n_I tend to infinity:*

$$\lim_{n_Q, n_I \to \infty} \|\epsilon_{I_f}\|_\infty = 0. \tag{16}$$

Moreover, if the quadrature and interpolation methods converge at rates s_Q and s_I, respectively, and if the number of quadrature and interpolation nodes are chosen as $n_Q = n^\theta$ and $n_I = n^{(1-\theta)}$ with $\theta \in (0, 1)$, then the approximation of the parametric integration problem in terms of total integrand evaluations, $n = n_Q n_I$, converges uniformly at rate $s = \min\{s_Q\theta, s_I(1-\theta)\}$:

$$\|\epsilon_{I_f}\|_\infty = O(n^{-\min\{s_Q\theta, s_I(1-\theta)\}}). \tag{17}$$

All proofs are deferred to the e-companion (Section EC.1.1).

Remark 2 (Asymptotic Convergence Rate). Note that Proposition 1 constitutes an "asymptotic" convergence rate, which might be observed only as n_Q—and, therefore, n—approaches infinity. However, in the numerical examples we present here and in the e-companion (Section EC.1.4), the result manifests itself already for small n_Q.

Corollary 1 (Optimal Node Distribution). *Given the assumptions of Proposition 1 and known convergence rates of the respective methods, optimal convergence of the parametric integration can be obtained by* minimizing *the error bound (17) with respect to θ:*

$$\frac{s_I}{s_Q + s_I} = \arg\min_\theta n^{-\min\{s_Q\theta, s_I(1-\theta)\}}.$$

Example 1 (Optimal Node Distribution). If $s_Q = s_I$, it is optimal to balance quadrature and interpolation nodes by choosing $n_Q = n_I = \sqrt{n}$.

Example 2 (Convergence Rate in Limit Cases). If the convergence rate of either quadrature or interpolation is very much higher than the other one, formally if $s_Q \ll s_I$ or $s_Q \gg s_I$, the optimal θ tends to one or zero, respectively. In the limiting case, however, the convergence rate of the recursive parametric integration turns out to be $s = \min\{s_Q, s_I\}$.

3.3. Recursive Parametric Integration.

Definition 12 (Recursive Parametric Integral). Given a function $f: \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{N} \supseteq D \times D \times \mathbb{N} \to \mathbb{R}$, and a kernel $q: \mathbb{R}^m \supseteq D \to [0, a], a \in \mathbb{R}_+$ in the sense of Definition 1, the *recursive parametric integral* of f of order $T < \infty$ is denoted by the function

$$L_f^T = \int \dots \int_{D^T} \prod_{t=1}^{T} f_t(x_t, x_{t-1}) q_t(x_t) dx_1, \dots, dx_T, \tag{18}$$

where f_t and q_t are parametric forms of f and q, respectively, in the sense of Definition 6: $f_t(x_t, x_{t-1}) \equiv f(x_t, x_{t-1}, t)$ and $q_t(x_t) \equiv q(x_t, t)$ with $\forall t: \int_D q(x_t, t) dx_t \leq 1$ and x_0 is given. Its approximation $\hat{L}_f^T$ is defined by

$$L_f^T = \hat{L}_f^T + \epsilon_{L_f^T}$$

where $\epsilon_{L_f^T} \ll 1$ is the approximation error.

As in Definition 1, the role of the kernel is w.l.o.g.

Proposition 2 (Convergence of Recursive Parametric Integration). *Given a recursive parametric integral L_f^T as in Definition 12 with the restricted integrand $f: \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{N} \supseteq D \times D \times (1, \dots, \bar{T}) \to [0,1]$, and $f_t, q_t \in C^i, t = 1, \dots, \bar{T}$ with T bounded by $T \leq \bar{T} < \infty$, consider its approximation by recursive application of the parametric integral approximation using quadrature and interpolation as in Definition 11, where the quadrature and the interpolation methods converge uniformly in the sense of Definitions 7 and 9, respectively, and where the interpolation method is sufficiently smoothness preserving as by Definition 10. Then*

1. *for fixed T, Trecursive parametric integration converges; that is, the maximum error of the approximation of the recursive integral, $\hat{L}_f^T$, vanishes as $n_Q = n_{Q,t}$ and $n_I = n_{I,t}$ tend to infinity:*

$$\lim_{n_Q, n_I \to \infty} |\epsilon_{L_f^T}| = 0.$$

2. *fixing the number of quadrature and interpolation nodes for each iteration to n_Q and n_I, respectively, the maximum approximation error of the recursive integral as a function of T is bounded linearly:*

$$|\epsilon_{L_f^T}| = O(T).$$

3. *if the quadrature and interpolation method converge at rates s_Q and s_I, respectively, the convergence rate of the overall approximation error $|\epsilon_{L_f^T}|$ in terms of total integrand evaluations, n, is given by*

$$|\epsilon_{L_f^T}| = O\left(T\left(\frac{n}{T}\right)^{-\min\{s_Q\theta, s_I(1-\theta)\}}\right),$$

where $n_{Q,t} = (n/\mathrm{T})^{\theta}$ and $n_{I,t} = (n/\mathrm{T})^{1-\theta}$ with $\theta \in (0,1)$.

Remark 3 (Role of Assumptions and Restrictions in Proposition 2). I list a couple of remarks on the role and limitations of the assumptions and restrictions necessary to prove Proposition 2:

Boundedness of the Integrand. Although the proof of Proposition 2 requires the integrand to map to $[0, 1]$, for practical purposes, the image space really only needs to be bounded (which is required by the definition of C^i anyway): Then, because integration is a linear operator, the integrand can be (linearly) transformed to comply with the restrictions and integrated recursively, and finally, the integral is transformed back. Note that the same argument can be applied for kernels that integrate to more than one (cf. Definition 1).

Smoothness Preservation of Interpolation. Note that although the proof of Proposition 2 requires the interpolation method to be smoothness preserving as in Definition 10, it does not require the degree of smoothness preservation to be such that the *maximum* possible convergence rate will be attained; rather, different degrees of smoothness preservation are generally required for Parts 1 and 2 (together) or 3 to hold. For example, it is well known that every continuous and bounded function over a compact interval is a uniform limit of piecewise linear continuous functions, and interpolation using piecewise linear continuous functions preserves continuity. At the same time, the trapezoidal rule for numerical integration converges for every Riemann integrable function, that is, for all bounded and continuous functions over a compact domain, and the compactness of the domain for the interpolation assures uniformity of convergence of the parametric integration problem. Therefore, Parts 1 and 2 of Proposition 2 still assure convergence and linear error growth.

Boundedness of Time Horizon. The assumption is that $T \leq \bar{T} < \infty$ has no practical relevance as $\bar{T}$ can be chosen arbitrarily large as long as it is finite. Rather, it is technically needed to bound the sum of "period-wise" errors independently of T (cf. proof of Proposition 2, Equations (EC.7), (EC.14), and (EC.19)).

Corollary 2 (Sublinear Error Bound). *Given the approximation problem of Proposition 1 with the corresponding* *assumptions, suppose that the computational effort for each iteration implied by n_Q and n_I is fixed, and $T \le \bar{T} < \infty$. Then there exists an error bound for $|\epsilon_{L_f^T}|$ that grows sublinearly in T.*

Remark 4 (Sharpness of Error Bound). Note that the error bounds of Proposition 2 and Corollary 2 are still not sharp because, by taking absolute errors, they ignore sign changes in $\epsilon_{I_{\tilde{f}_t}}$, which potentially make the errors cancel or "average out" (cf. proof of Proposition 2, Equation (EC.8)). Rather, (sub)linear error growth is derived from the boundedness of the integrand by one in this paper. However, because the sign changes and their effect on the overall error depend on the function $\epsilon_{I_{\tilde{f}_t}}$ and, therefore, on f itself, they are much harder to quantify ex ante.

I conclude the section with a couple of examples that demonstrate the different interpretations of the theoretical results and a comparison with Monte Carlo integration; more examples, in particular for error growth in T, and convergence results for the application to the model of Rust (1987) can be found in the e-companion, Sections EC.1.4 and EC.2.1.

Example 3 (Increasing Accuracy). Suppose the level of accuracy for $\hat{L}_f^T$ shall be increased by a factor of i; one wants to know how many more integrand evaluations are necessary to obtain this level of accuracy. Therefore, the following equation needs to be solved for j:

$$T\left(\frac{n}{T}\right)^{-s} = iT\left(\frac{jn}{T}\right)^{-s},$$

where $s = \min\{s_Q\theta, s_I(1-\theta)\}$, yielding $j = i^{\frac{1}{s}}$. For example, to double the level of accuracy ($i = 2$) for a given time horizon T when approximating a one-dimensional integrand from C^4 using a cubic spline and Simpson integration ($s_Q = s_I = 4$, assuming that the third and fourth derivatives are approximated well enough to preserve smoothness) with equally many quadrature and interpolation nodes ($\theta = 0.5$), one needs to increase the total number of integrand evaluations by a factor of $2^{\frac{1}{2}} \approx 1.4$. In contrast, to double the average level of accuracy using Monte Carlo integration, four times more integrand evaluations are needed.

Example 4 (Convergence Rate Comparison). This example compares the rates of convergence for different approximation methods for the recursive parametric integral from Definition 12 for a particular integrand, $f: \mathbb{R}^2 \to [0,1], f(x_t, x_{t-1}) = \Phi((0.5x_{t-1} + x_t + 4)^2)$, where Φ is the cdf of the standard normal distribution, and the kernel is chosen as its pdf.

The methods employed in this example are Gauss–Hermite quadrature and Chebyshev polynomial interpolation using $2n_Q^{GH} = n_I^{Cheb}$ (see Examples EC.2 and EC.4), the compound Simpson rule and cubic spline interpolation using $n_Q^{Simp} = n_I^{CS}$ (see Examples EC.1 and EC.6), and Monte Carlo integration (see Example EC.3); the number of quadrature and interpolation nodes are chosen to match the number of integrand evaluations in the MC integration: $n = Tn_Qn_I = Tn_{MC}$ with $T = 100$.

Figure 1 (a) depicts the error of the two recursive versions and MC integration as a function of integrand evaluations (normalized by T and averaged over 50 runs for MC integration); because both axes are on a log scale, convergence rates can be read from the slope of the (log) error as a (linear) function of the (log) number of nodes or draws (and, thus, the number of integrand evaluations). The example confirms the result from Proposition 2, Part 3; given the convergence rates of Gauss quadrature and polynomial interpolation are exponential at best, the rates of Simpson integration and cubic splines are both four, and the standard deviation of Monte Carlo estimates reduces at rate 1/2. Note that although the cubic spline is not smoothness preserving in the sense of Definition 10, the convergence rates predicted by Proposition 2 can still be observed.

Figure 1(b) plots the run time in seconds needed to achieve a particular degree of numerical accuracy (again both in log terms). Because MC integration is naturally faster for a comparable total number of integrand evaluations (because no interpolant creation and evaluation is needed) but has a far slower convergence, only by putting the run time of the integration into relation with the numerical accuracy achieved within a particular amount of time does one obtain a realistic perception of the true computational efficiency of the methods. In this particular example, it is obvious that, although MC integration is far more efficient to obtain a rough estimate of the integral, the recursive method is way more efficient to obtain accurate approximations of the integral, which can be highly beneficial when put into an optimization context (see Section 3.4).[2]

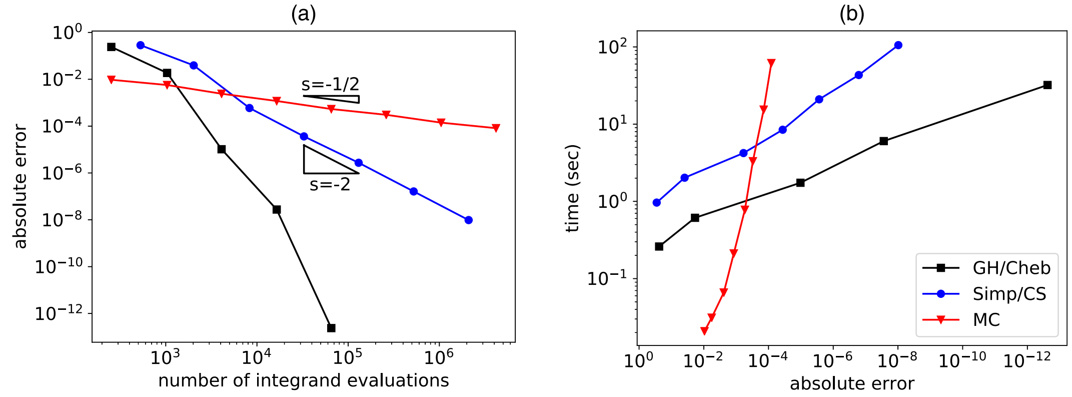
3.4. Recursive Likelihood Function Integration

This subsection derives the conditions under which the results for recursive parametric integration apply to maximum likelihood estimation.

Definition 13 (Model). Suppose the model under consideration predicts observations according to the joint probability density function

$$P_{xi\varepsilon}(\{i_t, x_t, \varepsilon_t\}_{t=1}^T | \{i_0, x_0, \varepsilon_0\}; \theta), \tag{19}$$

Figure 1. (Color online) Comparison of Numerical Approximation Error and Running Times



Note. Numerical approximation error as a function of the total number of integrand evaluations (a), and running times (in seconds) as a function of numerical accuracy (b) for recursive parametric integration using Gauss–Hermite integration with Chebyshev interpolation ("GH/Cheb," square) and Simpson integration and cubic spline interpolation ("Simp/CS," circle), respectively, compared with Monte Carlo integration ("MC," triangle).

where i_t is the dependent variable observed at time t ("outcome"); x_t and ε_t are the observable and the unobservable parts of the independent variables at time t, respectively; and θ is a vector of parameters of the model.

Note that i_t, x_t, and ε_t in Definition 13 are generally vector valued.

Assumption 1 (Markov Property). *The model in Definition 13 is Markov:*

$$P_{xi\varepsilon}(\{i_t, x_t, \varepsilon_t\}_{t=1}^{T} | \{i_0, x_0, \varepsilon_0\}; \theta) = \prod_{t=1}^{T} p_{xi\varepsilon}(\{i_t, x_t, \varepsilon_t\} | \{i_{t-1}, x_{t-1}, \varepsilon_{t-1}\}; \theta).$$

In this paper, I only consider Markov models of order 1, but the RLI method generalizes to higher orders.

Rewriting the transition probability density function using conditional density functions yields

$$p_{xi\varepsilon}(\{i_t, x_t, \varepsilon_t\} | \{i_{t-1}, x_{t-1}, \varepsilon_{t-1}\}; \theta) = p_{xi|\varepsilon}(i_t, x_t | i_{t-1}, x_{t-1}, \varepsilon_{t-1}, \varepsilon_t; \theta) p_{\varepsilon|i}(\varepsilon_t | i_{t-1}, x_{t-1}, \varepsilon_{t-1}; \theta).$$

Assumption 2 (Smoothness of Conditional Density Functions). *The conditional density functions $p_{xi|\varepsilon}(i_t, x_t | i_{t-1}, x_{t-1}, \varepsilon_{t-1}, \varepsilon_t; \theta)$ and $p_{\varepsilon|i}(\varepsilon_t | i_{t-1}, x_{t-1}, \varepsilon_{t-1}; \theta)$ are in C^i with respect to ε_t and ε_{t-1}.*

Assumption 3 (Boundedness of Conditional Density Function for Observed Variables). *The conditional density function of the observed variables, $p_{xi|\varepsilon}(i_t, x_t | i_{t-1}, x_{t-1}, \varepsilon_{t-1}, \varepsilon_t; \theta)$, is bounded by one.*

Because the process ε_t is unobserved, the likelihood function of model (19) cannot be computed directly. However, the *marginal* likelihood function computes as the marginalization with respect to $\{\varepsilon_t\}_{t=0}^{T}$.[3]

Definition 14 (Marginal Likelihood Function). The *marginal* likelihood function of model (19) reads as

$$L_T(\theta) \equiv P_{xi}(\{i_t, x_t\}_{t=0}^{T} | \{i_0, x_0\}; \theta) \tag{20}$$

$$= \int_{\underline{\varepsilon}_0}^{\bar{\varepsilon}_0} \cdots \int_{\underline{\varepsilon}_T}^{\bar{\varepsilon}_T} p_{\varepsilon}(\varepsilon_0; \theta) \cdot \prod_{t=1}^{T} p_{xi|\varepsilon}(i_t, x_t | i_{t-1}, x_{t-1}, \varepsilon_{t-1}, \varepsilon_t; \theta) \cdot p_{\varepsilon|i}(\varepsilon_t | i_{t-1}, x_{t-1}, \varepsilon_{t-1}; \theta) d\varepsilon_0 \ldots d\varepsilon_T. \tag{21}$$

Note that because ε_t in (21) can be vector valued, the integrals over each ε_t form potentially multidimensional integrals themselves.

To allow for random variables with infinite support (without truncation), the integral in the marginal likelihood function (20) has to be transformed into a kernel integral in the sense of Definition 1. Therefore, I require the following assumption:

Assumption 4 (Change of Variable). *There exists an invertible change of variable*

$$\varepsilon_t = \varphi(\tilde{\varepsilon}_t, \varepsilon_{t-1}, i_{t-1}, x_{t-1}; \theta)$$

such that

$$p_{\varepsilon|i}(\varphi(\tilde{\varepsilon}_t, \varepsilon_{t-1}, i_{t-1}, x_{t-1}; \theta) | i_{t-1}, x_{t-1}, \varepsilon_{t-1}; \theta) = q_t(\tilde{\varepsilon}_t; \theta) \tag{22}$$

which is in C^{i+1} with respect to $\tilde{\varepsilon}_t$ and ε_{t-1} and where

$$\int_{\varphi^{-1}(\underline{\varepsilon}_t, \varepsilon_{t-1}, i_{t-1}, x_{t-1}; \theta)}^{\varphi^{-1}(\bar{\varepsilon}_t, \varepsilon_{t-1}, i_{t-1}, x_{t-1}; \theta)} q_t(\tilde{\varepsilon}_t; \theta) d\tilde{\varepsilon}_t \leq 1 \tag{23}$$

$$\varphi'(\tilde{\varepsilon}_t, \varepsilon_{t-1}, i_{t-1}, x_{t-1}; \theta) \equiv \frac{\partial}{\partial \tilde{\varepsilon}_t} \varphi(\tilde{\varepsilon}_t, \varepsilon_{t-1}, i_{t-1}, x_{t-1}; \theta) \leq 1. \tag{24}$$

Example 5 (*AR*(1) Process). The change of variable φ in Assumption 4 often coincides with the "functional form" of the process ε_t. A simple but practically very important example for a particular process and the corresponding change of variable is the $AR(1)$ process, $\varepsilon_t = \rho\varepsilon_{t-1} + \tilde{\varepsilon}_t$, where $\tilde{\varepsilon}_t$ is white noise, distributed identically and independently according to the density function $q(\cdot)$, and $\varepsilon_t = 0$ for $t \le 0$. The corresponding change of variable that fulfills (22) is

$$\varphi(\tilde{\varepsilon}_t, \varepsilon_{t-1}, i_{t-1}, x_{t-1}; \theta) = \varphi(\tilde{\varepsilon}_t, \varepsilon_{t-1}; \theta) = \rho\varepsilon_{t-1} + \tilde{\varepsilon}_t$$

$$\varphi'(\tilde{\varepsilon}_t, \varepsilon_{t-1}, i_{t-1}, x_{t-1}; \theta) = 1$$

$$\varphi^{-1}(\varepsilon_t, \varepsilon_{t-1}, i_{t-1}, x_{t-1}; \theta) = \varepsilon_t - \rho\varepsilon_{t-1}.$$

Proposition 3 (Recursive Likelihood Function Integration). *Given a marginal likelihood function $L_T(\theta)$ as in Definition 14, consider its approximation by recursive application of the parametric integral approximation using quadrature and interpolation as in Definition 11, where the quadrature and the interpolation methods converge uniformly at rates s_Q and s_I in the sense of Definitions 7 and 9 for sufficiently smooth integrands, respectively, and where the interpolation method is, moreover, smoothness preserving as by Definition 10. Under Assumptions 1–4, the results of Proposition 2 as well as Corollary 2 apply.*

Remark 5 (Role of Assumptions in Proposition 3). I list a couple of remarks on the role and limitations of the assumptions necessary for Proposition 3 to apply.

Continuity of Density Functions. Assumption 2 is not always fulfilled by default and needs some care in model design; for example, in discrete-choice models it can happen that that the *conditional* choice probability is binary and, thus, degenerate: $p_{i|x\varepsilon}(i_t|x_t, \varepsilon_t, \theta) \in \{0, 1\}$, which is not even continuous; however, there exist ways to avoid this kind of problem, such as introducing smooth (uncorrelated) errors, etc.

Unit Bound on Density Function. Although the unit bound restriction of Assumption 3 is technically needed to prove Proposition 3, it is rarely restricting in the practice of maximum likelihood estimation because even if it fails to hold, any monotone transformation of the likelihood function—such as rescaling $p_{xi|\varepsilon}$—will preserve the location of its maximum. The same holds true for the unit bound on the kernel integral and the corresponding change of variable in Equations (23) and (24). Alternatively, the rescaling can be done within the numerical integration as noted in Remark 3.

Change of Variable. Assumption 4 is w.l.o.g. because the use of the kernel integral as in Definition 1 is itself w.l.o.g.; numerically, even nontrivial kernels can always be made part of the integrand f, speaking in terms of Definition 1. However, if expectations over random variables with infinite support are integrated, either the kernel integral has to be used, or the domain of integration has to be truncated. Even if, however, the change of variables is necessary, the unit boundedness in Equations (23) and (24) is also w.l.o.g as argued in Remark 3.

4. Conclusion

This paper develops a method to efficiently approximate the (marginal) likelihood function of continuous-state hidden Markov models. More precisely, I decompose the integral over the unobserved state variables in the likelihood function into a series of lower dimensional integrals and successively approximate them using lower dimensional quadrature rules and interpolation between the time steps. I call this procedure recursive likelihood function integration, and I provide rigorous error and convergence analysis of the new method as well as assumptions on the model for the theoretical results to be applicable.

The key conclusions of the analysis are as follows. First, the computational complexity grows linearly in the number of periods, which makes the integration over hundreds and thousands of periods feasible. Second, I prove that the numerical error accumulates sublinearly in the number of time periods integrated. Consequently, using highly efficient and fast-converging numerical quadrature and interpolation methods for low and medium dimensions, such as Gaussian quadrature and Chebyshev polynomials, the numerical error can be well controlled even for very large numbers of periods.

As an application, I apply this method to the bus engine replacement model of Rust (1987), featuring serially correlated errors and using the original data set, finding barely any serial correlation. Also, the parameter estimates vary substantially compared with the case of serially uncorrelated errors. Second, but deferred to the e-companion, I verify the RLI algorithm's ability to recover the parameters of the same model in an extensive Monte Carlo study with simulated data sets, finding that the method is indeed able to recover the parameters used for the simulation, particularly in the case of the serial correlation parameter, which is recovered to very high precision.

Acknowledgments

The author is indebted to his advisor Karl Schmedders, as well as to Ken Judd, John Rust, and Che-Lin Su for support and guidance in this project. The author thanks Greg Crawford, Philipp Eisenhauer, Katharina Erhardt, Dennis Kristensen, János Mayer, Bertel Schjerning, Ole Wilms, the handling associate editor at *Operations Research*, and two anonymous referees, as well as seminar audiences at the University of Chicago, Hoover Institution Stanford University, University of Zurich, and the 68th European Meeting of the Econometric Society for helpful comments and suggestions. The author also thanks Dave Brooks for editorial comments on the manuscript.

This paper was first submitted on February 6, 2015 as "Divide and Conquer: Recursive Likelihood Function Integration for Dynamic Discrete Choice Models with Serially Correlated Unobserved State Variables;" earlier versions of this paper circulated as "Divide and Conquer: A New Approach to Dynamic Discrete Choice with Serial Correlation."

Endnotes

[1] This is not to be confused with the recursive maximum likelihood estimation (RMLE) algorithm of Kay (1983) for the estimation of *AR* processes, which allows one to recursively update maximum likelihood estimates to higher order *AR* models.

[2] All examples in this section are written in Python (partially using Numpy and Scipy) without any parallelization. All computations are carried out on a 2012 laptop with one four-core Intel "Core i7 Ivy Bridge" processor running at 2.6 GHz and 16 GB RAM.

[3] I use the term "marginal likelihood function" in this context in the frequentist's sense, in that the unobserved random variables ε_t can be thought of as *nuisance parameters* with a distribution attached to them, which allows to integrate them out (instead of being optimized over).

References

Aguirregabiria V, Mira P (2010) Dynamic discrete choice structural models: A survey. *J. Econometrics* 156(1):38–67.

Arcidiacono P, Ellickson PB (2011) Practical methods for estimation of dynamic discrete choice models. *Annual Rev. Econom.* 3(1): 363–394.

Arcidiacono P, Miller RA (2011) Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica J. Econometric Soc.* 79(6):1823–1867.

Blevins JR (2016) Sequential Monte Carlo methods for estimating dynamic Microeconomic models. *J. Appl. Econometrics* 31(5): 773–804.

Cai Y, Judd KL (2013) Advances in numerical dynamic programming and new applications. Schmedders K, Judd KL, eds. *Handbook of Computational Economics* (Elsevier, Amsterdam), 479–516.

Connault B (2016) Hidden rust models. Working paper, University of Pennsylvania, Philadelphia.

Cosslett SR, Lee LF (1985) Serial correlation in latent discrete variable models. *J. Econometrics* 27(1):79–97.

Eisenhauer P, Heckman JJ, Mosso S (2015) Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments. *Internat. Econom. Rev.* 56(2):331–357.

Elliott RJ, Aggoun L, Moore JB (2008) *Hidden Markov Models: Estimation and Control*, Vol. 29 (Springer, New York).

Grüne L, Semmler W (2004) Using dynamic programming with adaptive grid scheme for optimal control problems in economics. *J. Econom. Dynam. Control* 28(12):2427–2456.

Hotz VJ, Miller RA (1993) Conditional choice probabilities and the estimation of dynamic models. *Rev. Econom. Stud.* 60(3): 497–529.

Imai S, Jain N, Ching A (2009) Bayesian estimation of dynamic discrete choice models. *Econometrica J. Econometric Soc.* 77(6):1865–1899.

Judd KL (1998) *Numerical Methods in Economics* (The MIT Press, Cambridge, MA).

Kay SM (1983) Recursive maximum likelihood estimation of autoregressive processes. *IEEE Trans. Acoust. Speech Signal Process.* 31(1):56–65.

Keane MP, Wolpin KI (1994) The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence. *Rev. Econom. Statist.* 76(4):648–672.

Keane MP, Todd PE, Wolpin KI (2011) The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications. Ashenfelter O, Card D, eds. *Handbook of Labor Economics* (Elsevier, Amsterdam), 331–461.

Larsen BJ, Oswald F, Reich G, Wunderli D (2012) A test of the extreme value type I assumption in the bus engine replacement model. *Econom. Lett.* 116(2):213–216.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed. *Frontiers in Econometrics* (Academic Press, Cambridge, MA), 105–142.

McFadden D (1981) Econometric models for probabilistic choice. Manski CF, McFadden D, eds. *Structural Analysis of Discrete Data with Econometric Applications* (The MIT Press, Cambridge, MA), 198–272.

McFadden D (1989) A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica J. Econometric Soc.* 57(5):995–1026.

Miller RA (1984) Job matching and occupational choice. *J. Political Econom.* 92(6):1086–1120.

Norets A (2009) Inference in dynamic discrete choice models with serially correlated unobserved state variables. *Econometrica J. Econometric Soc.* 77(5):1665–1682.

Norets A (2012) Estimation of dynamic discrete choice models using artificial neural network approximations. *Econometric Rev.* 31(1): 84–106.

Pakes A (1986) Patents as options: Some estimates of the value of holding European patent stocks. *Econometrica J. Econometric Soc.* 54(4):755–784.

Rust J (1987) Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica J. Econometric Soc.* 55(5): 999–1033.

Rust J (1988) Maximum likelihood estimation of discrete control processes. *SIAM J. Control Optim.* 26(5):1006–1024.

Rust J (1996) Numerical dynamic programming in economics. Amman HM, Kendrick DA, Rust J, eds. *Handbook of Computational Economics* (Elsevier, Amsterdam), 619–729.

Stinebrickner TR (2000) Serially correlated variables in dynamic, discrete choice models. *J. Appl. Econometrics* 15(6):595–624.

Walck C (1996) Hand-book on statistical distributions for experimentalists. Technical report, University of Stockholm, Stockholm.

Wolpin KI (1984) An estimable dynamic stochastic model of fertility and child mortality. *J. Political Econom.* 92(5):852–874.

Gregor Reich is a postdoctoral fellow at the Hoover Institution, Stanford University, and the Department of Business Administration, University of Zurich. His research interests are in solution and estimation methods for structural economic models.