Essays on Specification and Estimation of Latent Variable Models

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von
Florian Heiß

im Sommersemester 2005

Referent:	Prof. Axel Börsch-Supan, Ph.D.
Korreferent:	Prof. Arthur van Soest, Ph.D.
Dekan:	Prof. Dr. Ulrich Falk
Tag der mündlichen Prüfung:	06. Juli 2005

Acknowledgements

I am indebted to many people for their invaluable help with writing this thesis. First and foremost, I would like to thank my supervisor Axel Börsch-Supan for his encouragement and support. During the last years, he provided me with an excellent work environment and boosted my interest in and understanding of research in economics.

In addition, my coauthors on health-related work Michael Hurd and David Wise had a significant impact on my thinking and writing about these topics. Besides our joint work, I benefitted from endless discussions with Viktor Winschel on numerics, statistics, philosophy and all the rest. With detailed discussion, comments and suggestions on the thesis and related work, Melanie Lührmann, Joachim Winter, Alexander Ludwig, Hendrik Jürges, and many others substantially improved the result.

Thanks also go to all current and former colleagues and friends at the Mannheim Research Institute for the Economics of Aging (MEA). Each of them – explicitly including Isabella Nohe, Petra Worms-Lickteig and Brunhild Griesbach – contributed to the unique atmosphere at MEA which not only helped to work on this thesis but also to enjoy it.

I gratefully acknowledge financial support from the Land Baden-Württemberg, the Gesamtverband der Deutschen Versicherungswirtschaft (GDV), the National Institute on Aging (NIA), and the Deutsches Institut für Altersvorsorge (DIA).

Last but not least, I would like to thank my parents for their endless support not only during the time I worked on this dissertation.

Contents

List of Tables

List of Figures

1 Introduction

The field of econometrics combines economic theory, statistical methods, and data. Only the combination of all three ingredients allows to understand economic phenomena and derive sound policy recommendations. Theory can provide a set of competing models based on alternative assumptions which might lead to conflicting conclusions. On pure theoretical grounds, the choice between them can merely be based on a subjective assessment of plausibility. Furthermore, economic theories involve a set of parameters whose magnitude is unknown in reality. Only by confronting the alternative models with data, it is possible to judge them on a more objective basis and to quantify their implications.

On the other hand, it is impossible to draw conclusions from data alone. They only provide measures like correlations between different quantities of interest. An underlying theory is needed to interpret them in a structural way. With data from classical experiments, the link between theory and data is straightforward. The researcher varies quantities of interest while holding all other influences constant. Obviously, the differences of the observed outcomes of the experiment are caused by the changes of the experimental design which are controlled by the researcher. Experiments are of limited use for studying many phenomena relevant for economic analyses, although for specific questions, this approach provided valuable insights into individual decision making. Typically, econometricians have to rely on data generated "in the real world". Since the world is complex, economic theory is essential for the structural interpretation of the data.

In order to unite economic theory with data, statistical methods are needed. The complexity of human behavior and economic systems prohibits the identification of deterministic relationships between quantities of interest. An important reason for randomness is the impossibility to observe all relevant influences. Models are always simplifications of the reality and statistics allows to explicitly consider unaccounted influences.

The broad range of ingredients and goals of econometrics is also reflected in the sources of inspiration for innovations in this field. Advances in economic theory or statistical methods can motivate new developments as well as the availability of different data sources or simply the demand for policy evaluations. In the early days of econometrics, analyses were mainly based on aggregate macroeconomic data. In the 1960s, researchers realized that the insights into economic relationships this approach allows are limited. For example individual heterogeneity cannot be considered. At the same time, data on individuals, firms and other units started to become available and the computing power increased so that they could be analyzed. This lead to a fast advancement of microeconometric models and methods.

One of the main advantages of microeconometric analysis is that theories on individual decision making can be connected to data on individual decisions. This allows a much more direct and valid inference about the underlying mechanisms and the consideration of individual heterogeneity. In the year 2000, the nobel prize in economics was shared by two pioneers of this literature. Daniel McFadden developed the main idea and specific approaches to identify individual behavior based on economic models of utility maximization with data on observed decisions. As he put it in his prize lecture, this exemplary combination of economic theory and statistical methods

> "[...] has been successful because it emphasized empirical tractability and could address a broad array of policy questions within a framework that allowed results to be linked back to the economic theory of consumer behavior".

This dissertation contributes four essays to the broad literature on microeconometric modeling and the computational problems that arise in the implementation of such models. Chapter 2 revisits a classical model for a specific class of problems. The nested logit model is concerned with situations in which the researcher is faced with observations on individual choices among a finite set of mutually exclusive alternatives. Examples for these situations include brand or travel mode choices. It models these decisions based on McFadden's concept of utility maximization and allows subsets of the alternatives to be similar in an unobserved way. The nested logit model has a long tradition in the econometric literature and is still one of the major tools used for empirical analyses in many areas.

The implementation of this model caused severe confusion. Parallel to the original model, which is based on a utility maximization model with a certain specification of the stochastic components, an alternative specification of outcome probabilities emerged. It was originally implemented in commercial statistical packages such as LIMDEP and Stata and documented as "the nested logit model". In empirical applications, researchers commonly used these implementations, believing that they applied the original model specification. Chapter 2 discusses the differences between these two specifications and provides examples for the bias of interpretation and inference if they are ignored. It furthermore describes an implementation of the original model in Stata.

After the power of micro-level data analysis had been fully appreciated, the use of repeated observations on a number of individuals or other units over time attracted a lot of attention. This panel data structure has invaluable advantages over single observations for each individual. By observing changes over time, individual heterogeneity can be accounted for in a much richer way and dynamic aspects can be considered. James Heckman, the second nobel prize laureate in the year 2000, advanced among other important areas the analysis of heterogeneity and dynamics in microeconometric analyses. In his prize lecture he noted that

> “Different assumptions about the sources of unobserved heterogeneity have a profound effect on the estimation and economic interpretation of empirical evidence, in evaluating programs in place and in using the data to forecast new policies and assess the effect of transporting existing policies to new environments”.

Chapter 3 of this dissertation discusses different modeling strategies for panel data. It was motivated during work of the author on a larger project on the complex interrelationships between health, socioeconomic status, and living arrangements. In a first paper on this project, Heiss, Hurd, and Börsch-Supan (2005) (HHB) study the joint evolution of these conditions using panel data from a large survey of elderly Americans, the health and retirement study (HRS). We analyze the individual trajectories of measures of each of these conditions. The econometric model structure used in this paper is relatively simple since the goal was to provide a broad overview of the relevant variables and their dependencies. In order to understand the causal paths underlying the findings of HHB, a more careful analysis of the sources of intertemporal relationships is required as pointed out in the quotation above.

In general, applied econometric models often make use of the panel structure in a very limited way, although James Heckman already provided a quite general discussion in the early 1980s. Heterogeneity not captured by observed covariates is typically treated as a latent variable that is constant over time. Some models, including those discussed in HHB, consider causal dependence between outcomes over time. More elaborate models are typically developed for specific models only.

A discussion of more flexible modeling approaches for panel data in a general setting is given in chapter 3 of this dissertation. The models are formulated in a state space framework. In this approach, the model is separated into a part that specifies the evolution of a set of latent variables over time and another part that models the connection between this state space and the observed outcomes. Using this strategy, the formulation and implementation of a rich set of flexible and yet plausible and parsimonious models is straightforward.

In addition to the general approach, chapter 3 considers special issues. These include a specification of the evolution of latent variables in continuous time and the joint modeling of multiple dependent variables. The correction of problems caused by the systematic loss of individuals over time known as panel attrition is discussed as a special application of joint models. These ideas are applied to study the evolution of individual health. Health itself is modeled as a latent variable for which the answers to survey questions give indications. A simple model for this process captures the data much better than traditional approaches like random effects or causal models. Mortality obviously affects the study of health evolution, since the least healthy individuals die at younger ages. Using a joint model of health and mortality, this can be easily accounted for. The chapter also demonstrates the effects of ignoring this fact.

Based on the findings and methodological suggestions of chapter 3, Heiss, Börsch-Supan, Hurd, and Wise (2005) study more carefully the evolution of health, disability, and mortality. The state space approach suggested in chapter 3 proves to be useful for this enterprise. It will also help to get back to the issue of a joint analysis of health, socioeconomic status, and living arrangements in future research. The state space approach is well suited for this task, since it allows straightforward specifications of joint models with flexible correlation structures both across time and over different outcomes.

Flexible econometric models like those discussed in chapter 3 are not easily estimated. In recent years, advances in econometric research have been triggered by the technical progress leading to a surge of computational power. This development permits analyses of such models. Previously, the specification of econometric models has been confined to a class for which the required computational effort was limited. The ongoing increase in computing resources lead to the development of simulation-based estimation methods which hardly restrict the specification of models. As Kenneth Train puts it in the introduction of his textbook on these methods (Train 2003):

> “The researcher is therefore freed from previous constraints on model specification – constraints that reflected mathematical convenience rather than the economic reality of the situation. This new flexibility is a tremendous boon to research.”

However, even with the computing power researchers have typically access to today, simulation-based estimation still poses a challenge in many situations. The more computing power is available, the higher is the flexibility of the implemented models, which again creates higher computational costs. The results in chapter 3 were based on simulation estimation which allowed the flexible model specification. However, some of them required several days to run on a modern PC. Improving on simulation methods is therefore an active area of research in statistics and econometrics. Chapters 4 and 5 are concerned with improvements of and alternatives to simulation-based estimation. The approaches for different problems in both chapters succeed strikingly well in mitigating the computational burden and allow to save about 99% of computing time in the presented examples.

For the estimation of econometric models, certain measures such as outcome probabilities have to be evaluated. The model often provides these as functions of random variables only. Taking the expectation over these variables corresponds to the integration over their distribution. These integrals pose the challenge to specification and/or computation. The classical approach is to concentrate on models for which they have a known explicit solution, which holds only in very special cases. The alternative is to resort to computationally intensive numerical methods.

There are different approaches to numerical integration. In practice, econometricians typically use simulation techniques. They are relatively easy to implement and the speed of their convergence to the true value does not depend on the dimension of the integral. However, they also have disadvantages over deterministic approaches such as Gaussian quadrature. They suffer from simulation noise and both their accuracy and convergence rate are worse, at least in low dimensions. While one-dimensional Gaussian quadrature is known to be efficient, the usual extension to higher dimensions requires computational costs that rise exponentially with the number of dimensions and thereby quickly becomes prohibitive.

Chapter 4 discusses this problem for a class of panel data models including the important case of general nonlinear models with AR(1) error processes, a special case of the models described in chapter 3. It is argued that and shown how it is possible to split the high-dimensional integral into several lower-dimensional integrals in these cases. Sequential numerical integration of each of these integrals is much easier than integration of the original high-dimensional integral. Several approaches for implementation are discussed and compared. The suggested method uses Gaussian quadrature with importance weights.

For an application taken from chapter 3, the sequential quadrature and a sequential simulation-based approach are compared to the simulation of the whole integral, which is the strategy currently used in the literature. The suggested approach needs less computational costs by a factor of more than 100 to reach the same accuracy. In practice, computations are done in less than one hour compared to several days.

Chapter 5 is based on joint work with Viktor Winschel and discusses cases in which multiple integrals cannot be split into several lower-dimensional parts or in which these parts are still relatively high-dimensional. It suggests estimation based on an integration rule that extends Gaussian quadrature in a more careful way than the widely known product rule. As opposed to the latter, it does not suffer from exponentially growing computational costs, a problem known as the “curse of dimensionality” of numerical integration. In high dimensions, computational requirements of the proposed approach are dramatically lower without losing the numerical advantages of Gaussian quadrature.

The method is easy to implement. Given nodes and weights, existing simulation code has only to be adjusted by replacing raw with weighted means. Software for the generation of the appropriate nodes and weights is available from the author. The chapter presents extensive Monte Carlo experiments for a widely used discrete choice model, the random parameters or mixed logit model. The suggested approach strikingly outperforms traditional simulation-based estimators, again saving computational effort by a factor of about 100.

To summarize, this dissertation contributes four essays to the broad literature on microeconometric modeling and the computational problems that arise in the implementation of such models. Chapter 2 clarifies on the specification and presents an implementation of the more classical and widely used nested logit model. The other three essays discuss specification and estimation in the light of advanced numerical methods. They allow to fully exploit the potential of latent variable models in microeconometrics by freeing the researcher from constraints in the model specification. Chapter 3 demonstrates this advantage and presents flexible and yet parsimonious modeling strategies for nonlinear panel data. The numerical methodology required for the estimation of such and many other models is advanced in chapters 4 and 5. They suggest to replace the commonly used simulation techniques with deterministic integration rules by a separation of the complex task into several less demanding tasks if this is possible (chapter 4) or by an efficient multidimensional extension of Gaussian quadrature in the general case (chapter 5). Both approaches dramatically mitigate the computational burden and thereby provide even more flexibility and convenience to the applied researcher.

2 The Nested Logit Model: Clarification and Implementation

2.1 Introduction

The nested logit model has become an important tool for the empirical analysis of discrete outcomes. It is attractive since it relaxes the strong assumptions of the multinomial (or conditional) logit model. At the same time, it is computationally straightforward and fast compared to the multinomial probit, mixed logit, or other even more flexible models due to the existence of a closed-form expression for the likelihood function.

There is some confusion about the specification of the outcome probabilities in nested logit models. Two substantially different formulas and many minor variations of them are presented and used in the empirical literature and in textbooks. Many researches are neither aware of this issue nor which version is actually implemented by the software they use. This obscures the interpretation of their results. This problem has been previously discussed by Hensher and Greene (2000), Hunt (2000), Koppelman and Wen (1998), and Louviere, Hensher, and Swait (2000, section 6.5). This chapter provides a comparison of both approaches in line with this literature. It argues and shows in numerous examples that one of these specifications is preferable in most situations. The package `nlogit` of Stata 7.0 does not implement this specification. Therefore the package `nlogitrum` is presented, which does.

The remainder of this chapter is organized as follows: section 2.2 introduces basic concepts of discrete choice and random utility maximization (RUM) models and discusses the conditional logit model as the most straightforward example. Section 2.3 presents one version of the nested logit model, the so-called RUMNL model. It can directly be derived from a RUM model. Section 2.4 introduces the other variant, that is implemented as `nlogit` in Stata 7.0. It is shown that this model is more difficult to interpret and might imply counterintuitive and undesired restrictions. This is often overlooked by applied researchers. Section 2.5 compares both models in special cases of nesting structures. The Stata implementation of the preferred RUMNL model is introduced in section 2.6 and section 2.7 concludes.

2.2 Fundamental Concepts

Discrete Choice models are used to make statistical inferences in the case of discrete dependent variables. This chapter deals with a special class of discrete choice models for which there are more than two possible outcomes and the outcomes cannot be sensibly ordered. A classical example is the travel mode choice. This chapter uses a well-known data set on this topic to provide empirical examples. Among others, Greene (2000, example 19.18), Hunt (2000), and Louviere, Hensher, and Swait (2000, section 6.4) present nested logit estimates based on them. The data contain 210 non-business travelers between Sydney, Canberra, and Melbourne. They had four travel modes alternatives: car, train, bus, and plane.

Section 2.2.1 presents the concept of random utility maximization (RUM) models. Different types of variables can enter RUM models of discrete choice. Since this will be important for the following discussion, section 2.2.2 characterizes these variable types and the specification of their coefficients. Section 2.2.3 presents the RUM interpretation of the well-known conditional logit model and first estimates.

2.2.1 Random Utility Maximization Models

Econometricians often interpret discrete choice models in terms of underlying structural models of behavior, called random utility maximization (RUM) models. They assign a utility level U_{ij} to each alternative $j = 1, \ldots, J$ for each decision maker $i = 1, \ldots, I$. The decision makers are assumed to choose the alternative from which they derive the highest utility.

The utilities are determined by a large number of characteristics of the decision maker and the alternatives. The researchers have information on some of those determinants, but not on all. This is reflected by splitting the utilities into a deterministic part V_{ij} and a stochastic part ε_{ij}:

$$U_{ij} = V_{ij} + \varepsilon_{ij}. \tag{2.1}$$

The probability P_{ij} that individual i chooses some alternative j is equal to the probability of U_{ij} being the largest of all $U_{i1}, \ldots U_{iJ}$. With $y_i \in \{1 \ldots J\}$ denoting the alternative that decision maker i chooses, this probability is

$$\begin{aligned} P_{ij} = \Pr(y_i = j) &= \Pr(U_{ij} > U_{ik} \quad \forall\, k = 1, \ldots, J : k \neq j) \\ &= \Pr(\varepsilon_{ik} - \varepsilon_{ij} \leq V_{ij} - V_{ik} \quad \forall\, k = 1, \ldots, J : k \neq j). \end{aligned} \tag{2.2}$$

Given the deterministic parts of the utility functions $V_{i1}, \ldots, V_{iJ}$, this probability will depend on the assumptions on the distribution of the (differences of) the stochastic error terms $\varepsilon_{i1}, \ldots \varepsilon_{iJ}$. For some distributions, there exists a closed-form solution for this expression. The most prominent examples are the conditional logit model discussed in section 2.2.3 and the random utility version of the nested logit model discussed in section 2.3.2.

A look at equation (2.2) reveals two interesting properties of the RUM outcome probabilities: They are based on utility *differences* only. The addition of a constant to all utilities does not change the outcome probabilities. In addition to that, the scale of utility is not identified: Multiplying *each* of the utilities $U_{i1}, \ldots, U_{iJ}$ by a constant factor does not change the probabilities. So RUM models have to normalize the utilities.

2.2.2 Types of variables and coefficients

The deterministic utility components V_{ij} may consist of different types of determinants. Alternative-specific constants α_j for all but one (the reference) alternative should enter the model. They capture choice probabilities *relative to the reference alternative* that cannot be attributed to the other explanatory variables. In addition, individual-specific and/or alternative-specific variables may enter the utilities.

Individual-specific variables describe characteristics of the decision maker. These variables may influence the relative attractiveness of the alternatives. Prominent examples are socio-economic variables like income or age. They are collected in a vector $\mathbf{z}_i$ for each decision maker $i = 1, \ldots, I$. A parameter vector $\boldsymbol{\gamma}_j$ for each alternative j is associated with the individual-specific variables. Since only utility *differences* are relevant for the choice, the parameters for one (the reference) alternative have to be normalized to zero for identification.[1] The other parameters can be estimated freely. They represent the effect of the individual-specific variables on the utility of the respective alternatives *relative* to the reference alternative. In the travel mode choice example, the respondents were asked about their household income. The individual-specific variable $\texttt{inc}_i$ represents the income of individual i in ten thousand dollars.

Alternative-specific variables vary both over individuals and alternatives. A prominent example is the price in models of brand choice. In the travel mode choice data, there is a variable $\texttt{time}_{ij}$ that represents the time (in hours) that individual i would need for the trip with travel mode j. These variables will be collected in a vector $\mathbf{x}_{ij}$ for each decision maker $i = 1, \ldots, I$ and each alternative $j = 1, \ldots, J$. They may enter the utilities in two different ways. Since the variation over alternatives provides additional ground for identification, a sepa-

[1] Of course any other value can be chosen for normalization. The normalization to zero simplifies the interpretation of the other parameters.

rate parameter for *each* alternative is statistically identified. In the travel mode choice example, spending one hour in the own car might be associated with a lower disutility than spending one hour in the bus. This would be reflected in a larger $\beta_{\texttt{bus}}$ than $\beta_{\texttt{car}}$ in absolute value.

Including all these variables, the deterministic part of the utility V_{ij} can in general be written as

$$V_{ij} = \alpha_j + \mathbf{x}'_{ij}\boldsymbol{\beta}_j + \mathbf{z}'_i\boldsymbol{\gamma}_j. \tag{2.3}$$

On the other hand, researchers often want to estimate a joint coefficient $\boldsymbol{\beta}$ for all alternatives. This is possible because of the variation of $\mathbf{x}_{ij}$ over the alternatives. These variables will be called generic variables and their coefficients will be restricted as

$$\boldsymbol{\beta}_j = \boldsymbol{\beta} \quad \forall j = 1, \dots, J. \tag{2.4}$$

With this specification, the joint parameter β of travel time in our example may be interpreted as the value of time in terms of utility. If price is included as a generic variable, its parameter is often used to rescale the utility in dollar terms. Whether or not generic variables enter the model will affect the discussion of the nested logit model below.

2.2.3 Multinomial/Conditional/McFadden's Logit Model

The multinomial logit (MNL) and conditional logit (CL) models are probably the most widely used tools for analyzing discrete dependent variables. The terminology is not consistent in the literature, but this chapter refers to the MNL model as a special case of a CL model in which all explanatory variables are individual-specific. Such a model is implemented in Stata as `mlogit`, see [R] **mlogit**. The more general conditional logit model is implemented as the `clogit` command, see [R] **clogit**. The same model without the interpretation in terms of an underlying RUM model is often referred to as multinomial logistic regression. In the following, this chapter will discuss the most general CL model.

Consider a RUM model as described in section 2.2.1. The CL model assumes that the error terms $\varepsilon_{i1}, \dots \varepsilon_{iJ}$ are i.i.d. as Extreme Value Type I. This distribution has a variance of $\sigma^2 = \frac{\pi^2}{6}$ which implicitly sets the scale of the utilities. (McFadden 1974) shows that under these assumptions, the resulting probability P_{ij}^{CL} that individual $i = 1, \dots, I$ chooses some alternative $j = 1, \dots, J$ has a straightforward analytical solution:

$$P_{ij}^{\mathrm{CL}} = \frac{\mathrm{e}^{V_{ij}}}{\sum_{k=1}^{J} \mathrm{e}^{V_{ik}}}. \tag{2.5}$$

Table 2.1 shows estimation results for two CL models of the travel mode choice example. Both consider `income` and `time` as explanatory variables and

Table 2.1: Conditional Logit estimates

Model		(A)		(B)	
		Coef.	z	Coef.	z
const×	car	-4.122	-4.09	-3.886	-3.97
	bus	-2.614	-2.33	-2.678	-2.68
	train	-1.153	-1.14	-1.523	-1.60
hinc×	car	-0.209	-1.66	-0.201	-1.60
	bus	-0.454	-3.00	-0.457	-3.02
	train	-0.680	-4.92	-0.678	-4.93
time				-0.600	-8.29
time×	air	-3.364	-7.92	-2.754	-7.43
	car	-0.572	-7.58		
	bus	-0.609	-6.92		
	train	-0.639	-8.02		
Log likelihood		-201.34		-202.19	

define the outcome `air` as the reference outcome — i.e. $\alpha_{\texttt{air}}$ and $\gamma_{\texttt{air}}$ are normalized to zero. The deterministic parts of the utility in equation (2.3) are therefore

$$\begin{array}{llllll} V_{i,\texttt{air}} & = & & \beta_{\texttt{air}} \cdot \texttt{time}_{i,\texttt{air}}, & & \\ V_{i,\texttt{car}} & = \alpha_{\texttt{car}} & + & \beta_{\texttt{car}} \cdot \texttt{time}_{i,\texttt{car}} & +\gamma_{\texttt{car}} & \cdot\texttt{inc}_i, \\ V_{i,\texttt{bus}} & = \alpha_{\texttt{bus}} & + & \beta_{\texttt{bus}} \cdot \texttt{time}_{i,\texttt{bus}} & +\gamma_{\texttt{bus}} & \cdot\texttt{inc}_i, \text{ and} \\ V_{i,\texttt{train}} & = \alpha_{\texttt{train}} & + & \beta_{\texttt{train}} \cdot \texttt{time}_{i,\texttt{train}} & +\gamma_{\texttt{train}} & \cdot\texttt{inc}_i \end{array} \tag{2.6}$$

for both models.

Model A allows for different time parameters β_j for all alternatives. The estimate of all three γ_j alternatives is negative. This implies that higher income c.p. decreases the probability to chose any other travel mode rather than fly. The relative magnitude can also be interpreted: the order of the coefficients corresponds to the order of the marginal effects of the choice probabilities. All `time` parameters are highly significantly negative. This implies that the time spent for the trip is associated with a disutility and that the probability to choose any travel mode decreases when it gets slower.

As the results from model A indicate, the `time` parameters for the alternatives train, bus, and car are very similar. A test of the hypothesis that they are actually equal cannot be rejected. So it makes sense to impose equality, that is to specify `time` as a generic variable. This has two advantages. It improves the efficiency of the estimates and allows an interpretation of the coefficient as the implicit value of time in terms of utility. But $\beta_{\texttt{air}}$ is significantly higher in absolute value than the other parameters. So model B specifies `time` as a generic variable and additionally includes an interaction for the `air` alternative.[2] As expected, the log likelihood value decreases relative to the unconstrained model A, but this decrease is insignificant. The marginal effects and elasticities do not change significantly either.

The CL/MNL model is widely used because of its convenient form of the choice probabilities and due to its globally concave likelihood function that makes maximum likelihood estimation straightforward. But it imposes strong restrictions on the distribution of the error terms. Most notably, they are assumed to be independently distributed. Note that these terms capture all unobserved determinants of the choices. If two alternatives are similar, it is plausible to assume that their errors are positively correlated. In our example, if there are unobserved individual characteristics that affect the utility of both public transportation modes `bus` and `train` similarly, the error terms of those alternatives are correlated. This is ruled out by the CL model. If the assumption of independent error terms is violated, the CL parameter estimates are biased.

2.3 Nested Logit Models I: RUMNL

The basic idea of nested multinomial logit (NMNL) models is to extend the CL model in order to allow groups of alternatives to be similar to each other in an unobserved way, that is to have correlated error terms. The general approach of NMNL models is introduced in section 2.3.1. Section 2.3.2 presents a NMNL model that is derived from a RUM model and therefore called RUMNL model in this chapter. Finally, section 2.3.3 extends the CL example for this model. A Stata implementation of the RUMNL model is introduced later in this chapter, see section 2.6.

[2] There may be different reasons for the unequal parameter. Either the disutility of spending time in the plane is higher than for the other travel modes, or time actually enters the utility nonlinearly (the mean travel time by air is obviously significantly lower than the time for the other alternatives), or the situations in which people choose to fly differs in that time is more crucial. The reason for this will not be further explored at this place since this chapter is not really about travel mode choice and adding nonlinear terms etc. complicates the model unnecessarily for the purpose of demonstration.

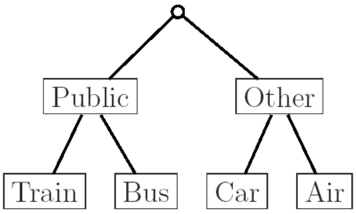
2.3.1 General Approach

The researcher partitions the choice set into M subsets ('nests') $B_m, m = 1, \ldots, M$.[3] So each alternative belongs to exactly one nest. Denote the nest to which alternative $j = 1, \ldots, J$ belongs as $B(j)$:

$$B(j) = \{B_m : j \in B_m, \quad m = 1, \ldots M\} \tag{2.7}$$

For the travel mode example, one possible nesting structure is depicted in figure 2.1. The number of nests is $M = 2$. The public transportation modes (train and bus) share the nest $B_{\texttt{public}} = \{\texttt{bus, train}\}$ and the other modes (air and car) share the nest $B_{\texttt{other}} = \{\texttt{car, air}\}$. In our notation, $B(\texttt{bus})$ is equivalent to $B_{\texttt{public}}$ just as $B(\texttt{train})$ is. This notation will help in formulating the choice probabilities below.

Figure 2.1: Nesting structure for models C through H



In order to develop an intuitive expression for the choice probabilities, it helps to decompose them into two parts. The probability of individual i choosing alternative j, $\Pr(y_i = j)$ is equal to the product of the probability $\Pr(y_i \in B(j))$ to choose some alternative in nest $B(j)$ and the conditional probability to choose exactly alternative j *given* some alternative in the same nest $B(j)$ is chosen $\Pr(y_i = j | y_i \in B(j))$:

$$P_j = \Pr(y = j) = \Pr(y = j | y \in B(j)) \cdot \Pr(y \in B(j)), \tag{2.8}$$

where the individual subscript i is dropped from now on for the sake of a more concise notation. In our example the probability to take the bus $\Pr(y = \texttt{bus})$ is equal to the probability to choose public transportation $\Pr(y \in B(\texttt{bus}))$ times the conditional probability to take the bus *given* a public transportation mode is

[3] This can be generalized to various nesting levels in a straightforward way by grouping the alternatives within such a nest in sub-nests and so on, but this chapter will concentrate on the simplest case of only one nesting level.

chosen $\Pr(y = \texttt{bus}|y \in B(\texttt{bus}))$. Note that this decomposition is valid in general by the rules of conditioning. But it is especially useful for thinking about the nested logit model.

2.3.2 Nested Logit as a RUM Model

The NMNL model can be derived from a RUM model just as the CL model. Consider a RUM model as described in 2.2.1. The CL model assumes that the error terms $\varepsilon_{i1}, \ldots \varepsilon_{iJ}$ are i.i.d. as Extreme Value Type I. Instead, the RUMNL model assumes a generalized version of this distribution. This special form of the generalized extreme value (GEV) distribution extends the Extreme Value Type I distribution by allowing the alternatives within a nest to have mutually correlated error terms.

For each nest $m = 1, \ldots, M$, the joint distribution of the error terms has an additional parameter τ_m that represents a measure of the mutual correlation of the error terms of all alternatives within this nest. Actually, this chapter specifies τ_m to be equal to $\sqrt{1-\rho_m}$ with ρ_m representing the correlation coefficient. So it is an *inverse* measure of the correlation. Therefore, it is often called dissimilarity parameter.[4] The *marginal* distribution of each error term is again Extreme Value Type I.

The RUMNL conditional choice probability to choose alternative j *given* some alternative in its nest is chosen $\Pr(y = j|y \in B(j))$ corresponds to a simple CL model for the choice between the alternatives in nest $B(j)$. The utilities are rescaled by the inverse of the dissimilarity parameter $\tau(j)$ of this nest:

$$\Pr(y = j|y \in B(j)) = \frac{\mathrm{e}^{\frac{1}{\tau(j)}V_j}}{\sum_{k \in B(j)} \mathrm{e}^{\frac{1}{\tau(j)}V_k}}, \tag{2.9}$$

The most intuitive explanation is based on the consideration of the implicit scaling in the logit model. As seen in section 2.2.1, the RUM choice probabilities depend on the utility *differences*. As noted above, the CL model implicitly scales all utilities such that the error terms have a variance of $\sigma^2 = \frac{\pi^2}{6}$. Since they are assumed to be independent in the CL model, their differences have a variance of $2\sigma^2$. But the RUMNL error terms within a nest are positively correlated. The higher the correlation between the error terms, the lower is the variance of these differences. With the relationship between the dissimilarity parameter τ_m and the coefficient of correlation ρ_m presented above, it is straightforward to show that the variance of the difference is $2\sigma^2\tau_m^2$. By normalizing the utilities by the

[4] Other equivalent parameterizations are used in the literature. For example, McFadden (1981) replaces τ_m with $\sigma_m = 1 - \tau_m$ and Louviere, Hensher, and Swait (2000) replace τ_m with $\mu_m = 1/\tau_m$.

factor $\frac{1}{\tau_m}$, the variance of this normalized difference becomes $2\sigma^2$. Without this normalization, the utilities in each nest would be scaled by a different factor and would therefore not be comparable across nests.

The denominator in equation (2.9) represents a (rescaled) measure of the attractiveness of the nest $B(j)$. The log of this expression for each nest m is called inclusive value IV_m. It corresponds to the expected value of the utility individual i obtains from the alternatives in nest m:

$$IV_m = \ln \sum_{k \in B_m} \mathrm{e}^{\frac{1}{\tau_m} V_k}. \tag{2.10}$$

The probability $\Pr(y \in B(j))$ to choose some alternative from nest k is again a CL probability for the choice between the nests. The scaled back inclusive values take the role of the deterministic parts of the utilities:

$$\Pr(y \in B(j)) = \frac{\mathrm{e}^{\tau(j) IV(j)}}{\sum_{m=1}^{M} \mathrm{e}^{\tau_m IV_m}}. \tag{2.11}$$

Because of the way the dissimilarity parameters enter this equation, they are also called IV parameters.

Nested Logit models can be estimated sequentially. First estimate a submodel for each nest according to equation (2.9). Then calculate the inclusive values defined in equation (2.10) and estimate a model for the choice of a nest shown in equation (2.11). See, among others, Train (2003) for a discussion of this sequential estimation and the necessary decomposition of the the explanatory variable into nest- and alternative-specific variables. Alternatively, all these equations can be plugged into equation (2.8). In this way, the marginal choice probability for alternative j can be obtained as

$$P_j^{\mathrm{RNL}} = \frac{\mathrm{e}^{\frac{1}{\tau(j)} V_j}}{\mathrm{e}^{IV(j)}} \cdot \frac{\mathrm{e}^{\tau(j) IV(j)}}{\sum_{m=1}^{M} \mathrm{e}^{\tau_m IV_m}}. \tag{2.12}$$

This probability is the full information likelihood contribution.

The CL model follows in the special case of $\tau_m = 1, \forall m = 1, ..., M$. This can be easily checked: the nests merely partition the choice set, so $\sum_{m=1}^{M} \mathrm{e}^{IV_m} = \sum_{k=1}^{J} \mathrm{e}^{V_k}$ must hold in this case. The RUMNL model is consistent with RUM if all τ_m lie in the unit interval.[5] For an introduction to this model also see Train (2003) and Maddala (1983).

[5] This condition can be relaxed for *local* consistency with RUM, see Börsch-Supan (1990).

2.3.3 Examples

Table 2.2 shows estimation results for two RUMNL models of the travel mode choice example with the nesting structure depicted in figure 2.1. Model C corresponds to a RUMNL version of the CL model A. The log likelihood value increases considerably by allowing the IV parameters to diverge from unity. A likelihood ratio test clearly rejects the CL model that implicitly restricts the IV parameters to unity. The IV parameter $\tau_{\texttt{public}}$ is within the unit interval and corresponds to a correlation of the two error terms of about .71. The IV parameter $\tau_{\texttt{other}}$ is clearly above 1. This implies that this model is inconsistent with RUM. This will be ignored for now and discussed in section 2.5.2.

Table 2.2: RUMNL estimates

Model		(C)		(D)	
		Coef.	z	Coef.	z
const×	car	-5.751	-1.60	-6.383	-2.24
	bus	-2.499	-0.76	-2.782	-1.03
	train	-1.253	-0.39	-1.786	-0.66
hinc×	car	-0.354	-0.90	-0.362	-0.93
	bus	-0.556	-1.94	-0.554	-1.93
	train	-0.827	-2.90	-0.831	-2.91
time				-1.301	-5.6
time×	air	-7.027	-5.49	-5.878	-5.54
	car	-1.325	-5.12		
	bus	-1.281	-5.37		
	train	-1.305	-5.54		
τ public		0.539	3.69	0.545	3.79
τ other		4.879	3.58	4.801	3.84
Log likelihood		-165.12		-165.26	

The other parameters tend to be larger in the RUMNL model than in the CL model. They cannot be compared however since the scaling differs across the models. One can either compare ratios of coefficients or calculate statistics such as the estimated marginal effects or elasticities of the choice probabilities with respect to the explanatory variables. The interpretation within the RUMNL model is equivalent to the interpretation in the CL model A. Model D in table 2.2 shows a RUMNL model with time entering as a generic variable analogous to the CL model B. Again, the interpretation remains the same. The generic restrictions of model D cannot be rejected by a likelihood ratio test.

2.4 Nested Logit Models II: NNNL

This section discusses a variant of the nested logit model. It will be called non-normalized nested logit (NNNL) model for reasons that are explained below. This is the model that is presented as the nested logit model for example presented by Greene (2000, section 19.7.4). It is also the model implemented in Stata 7.0 by the command `nlogit`, see [R] **nlogit**.

2.4.1 Structure of the Model

A latent variable $\widetilde{V}_j$ similar to the deterministic part of the utility in a RUM model is defined as a linear combination of the explanatory variables:

$$\widetilde{V}_j = \widetilde{\alpha}_j + \mathbf{x}_j' \widetilde{\boldsymbol{\beta}}_j + \mathbf{z}' \widetilde{\boldsymbol{\gamma}}_j. \tag{2.13}$$

If alternative-specific variables enter the model as generic variables, that is with a common coefficient $\widetilde{\boldsymbol{\beta}}_j$ for all alternatives, analogous restrictions to equation (2.4) are imposed:

$$\widetilde{\boldsymbol{\beta}}_j = \widetilde{\boldsymbol{\beta}} \quad \forall j = 1, \ldots, J. \tag{2.14}$$

The reason for adding the tilde to the V and the parameters is that the variable V_j is reserved to represent deterministic utility parts in this chapter and as will be explained below, this linear combination $\widetilde{V}_j$ may not be interpreted in this way.

With the inclusive value for any nest m defined as

$$\widetilde{IV}_m = \ln \sum_{k \in B_m} \mathrm{e}^{\widetilde{V}_k}, \tag{2.15}$$

the choice probabilities of the NNNL model are

$$P_j^{\mathrm{NNL}} = \frac{\mathrm{e}^{\widetilde{V}_j}}{\mathrm{e}^{\widetilde{IV}(j)}} \cdot \frac{\mathrm{e}^{\tau(j)\widetilde{IV}(j)}}{\sum_{m=1}^{M} \mathrm{e}^{\tau_m \widetilde{IV}_m}}. \tag{2.16}$$

Comparing these equations to equations (2.10) and (2.12), the relevant difference is that the deterministic utilities are not scaled by the inverse of the IV parameter in the conditional probability within the nest, $\frac{\mathrm{e}^{\widetilde{V}_j}}{\mathrm{e}^{\widetilde{IV}(j)}}$. This is the reason for the calling this model non-normalized nested logit (NNNL) model. As argued in section 2.3.2, this implies different scaling of the utilities across nests. In consequence, the interpretation of this model as a RUM model with the deterministic utility defined as $\widetilde{V}_j$ is challenged. This can be confirmed formally by considering what happens in a RUM model when the utility of each alternative is increased by some value a. According to section 2.2.1, the RUM choice probabilities do not change. Now have a closer look at equation (2.15). Adding the constant a to every $\widetilde{V}_j$ *does* alter the NNNL choice probabilities.

As a result, this model is not based on a RUM model with the deterministic parts of the utilities defined as $\widetilde{V}_j$ as was noted by Hensher and Greene (2000), Hunt (2000), Koppelman and Wen (1998), and Louviere, Hensher, and Swait (2000, section 6.5). But the next section argues that it *can* be interpreted in RUM terms with other deterministic utilities.

2.4.2 Interpretation of the NNNL as a RUM Model

As a result of the discussion above, the parameters $\widetilde{\alpha}_j, \widetilde{\boldsymbol{\beta}}_j$, and $\widetilde{\boldsymbol{\gamma}}_j$ of a NNNL model may not be interpreted as the structural parameters of an underlying RUM model as many researchers tend to do. But how can the parameters be interpreted? A reformulation of the NNNL model that is motivated from the insights of section 2.4.1 helps to answer this question. Suppose the deterministic part of the utility is not defined as $\widetilde{V}_j$ but as a scaled version V_j^{NNL} of it:

$$V_j^{\text{NNL}} = \tau(j)\widetilde{V}_j = \tau(j)\left(\widetilde{\alpha}_j + \mathbf{x}_j'\widetilde{\boldsymbol{\beta}}_j + \mathbf{z}'\widetilde{\boldsymbol{\gamma}}_j\right), \tag{2.17}$$

where $\tau(j)$ is the IV parameter of the nest to which alternative j belongs. Adding the constant a to every V_j^{NNL} means adding $\frac{a}{\tau(j)}$ to $\widetilde{V}_j$ and the inclusive value $\widetilde{IV}(j)$. As can be easily seen from equation (2.16), this leaves the choice probabilities unchanged.

If $\widetilde{V}_j$ in equations (2.15) and (2.16) are replaced with the equivalent term $\frac{a}{\tau(j)}V_j^{\text{NNL}}$, the equations become equivalent to the RUMNL equations (2.10) and (2.12). So the difference between the NNNL and the RUMNL model boils down to the in the specification of the utilities. While the RUMNL model directly considers the deterministic utilities and their parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, the NNNL model specifies utility according to equation (2.17).

A researcher with access to NNNL software but not to RUMNL software can apply a NNNL model and deduce the implicit RUM assumptions and parameters according to equation (2.17). Depending on the nesting structure and the presence of generic variables, this can be more or less straightforward and more or less sensible. In some cases, the NNNL parameters “only” have to be rescaled to recover the RUM parameters. In other cases, the NNNL model implicitly imposes restrictions that are usually undesired and unnoticed by researchers and readers of their work. The next sections identify these cases in order to illustrate the the theoretical arguments and to provide a guideline of how to interpret NNNL results.

2.4.3 Example 1: Alternative-specific coefficients only

In many applications, no generic variables enter the model. This case will turn out to be the least problematic for NNNL estimation in the sense that no implicit restrictions are imposed and the utility parameters can be recovered easily from the estimates.

The NNNL utility from equation (2.17) can be rewritten as

$$\begin{aligned} V_j^{\text{NNL}} &= \tau(j)\left(\widetilde{\alpha}_j + \mathbf{x}_j'\widetilde{\boldsymbol{\beta}}_j + \mathbf{z}'\widetilde{\boldsymbol{\gamma}}_j\right) \\ &= \tau(j)\widetilde{\alpha}_j + \mathbf{x}_j'(\tau(j)\widetilde{\boldsymbol{\beta}}_j) + \mathbf{z}'(\tau(j)\widetilde{\boldsymbol{\gamma}}_j). \end{aligned} \tag{2.18}$$

So with

$$\begin{aligned} \alpha_j &:= \tau(j)\widetilde{\alpha}_j, \\ \boldsymbol{\beta}_j &:= \tau(j)\widetilde{\boldsymbol{\beta}}_j, \quad \text{and} \\ \boldsymbol{\gamma}_j &:= \tau(j)\widetilde{\boldsymbol{\gamma}}_j, \end{aligned} \tag{2.19}$$

the utility simplifies to the equivalent of the RUMNL specification (2.3):

$$V_j^{\text{NNL}} = V_j = \alpha_j + \mathbf{x}_j'\boldsymbol{\beta}_j + \mathbf{z}'\boldsymbol{\gamma}_j. \tag{2.20}$$

So assume you have estimated a nested logit model using NNNL software such as the `nlogit` command of Stata. The estimates of $\widetilde{\alpha}_j$, $\widetilde{\boldsymbol{\beta}}_j$, and $\widetilde{\boldsymbol{\gamma}}_j$ do not directly have a structural interpretation in terms of a RUM model. But the underlying parameters α_j, $\boldsymbol{\beta}_j$, and $\boldsymbol{\gamma}_j$ can be recovered according to equation (2.19).

As an illustration, table 2.3 shows the results for a NNNL (model E) that corresponds to the RUMNL model C. Both models are equivalent in terms of the log likelihood and the implied marginal effects and elasticities. The estimated IV parameters are also identical. But the other parameter estimates differ. For example, the RUMNL estimate for the structural parameter of 'hinc×train' is $\widehat{\gamma}_{\text{hinc}\times\text{train}} = -0.474$. It can be recovered from the NNNL estimates by multiplying the estimated coefficient $\widehat{\widetilde{\gamma}}_{\text{hinc}\times\text{train}} = -0.879$ with the estimated IV parameter of the respective nest $\widehat{\tau}_{\text{public}} = 0.539$ as can be easily verified: $-0.474 = -0.879 \times 0.539$. Table 2.3 does these calculations for each of the coefficients. The third column shows the scaling factors which corresponds to the respective estimated IV parameter. The products of these factors and

Table 2.3: NNNL estimates without generic variables

		Model (E)		Recovering RUM params		
		Coef.	z	Factor	Par.	z*
const×	car	-1.179	-1.29	4.879	-5.751	1.60
	bus	-4.635	-0.73	0.539	-2.499	0.77
	train	-2.323	-0.38	0.539	-1.253	0.40
hinc×	car	-0.072	-0.90	4.879	-0.354	0.90
	bus	-1.031	-1.82	0.539	-0.556	1.94
	train	-1.534	-2.48	0.539	-0.827	2.90
time×	air	-1.440	-3.63	4.879	-7.027	5.49
	car	-0.272	-5.03	4.879	-1.325	5.12
	bus	-2.376	-4.92	0.539	-1.281	5.37
	train	-2.420	-4.87	0.539	-1.305	5.54
τ public		0.539	3.69			
τ other		4.879	3.58			
Log likelihood		-165.12				

*: Wald test of H_0: Rescaled parameter = 0. Shown is the square root of the test statistic $\overset{a}{\sim} N(0,1)$

the estimated NNNL coefficient can be found in the fourth column of table 2.3. Their equality to the RUMNL parameters can be easily verified by a comparison with the RUMNL results in the first column of table 2.2.

Note that the NNNL parameters cannot be interpreted in terms of RUM directly. The relative size of the coefficients does not have any meaning before they are rescaled. The scaling also has to be taken into account when testing hypotheses based on the parameters. For example the presented asymptotic t-statistic for $\widetilde{\gamma}_{\text{hinc}\times\text{train}}$ for the NNNL model does not correspond to the respective test for the RUMNL parameter $\gamma_{\text{hinc}\times\text{train}}$. The tests of the RUMNL parameters can be reproduced from the NNNL estimates. The appropriate null hypothesis $\text{H}_0 : \widetilde{\gamma}_{\text{hinc}\times\text{train}} \times \tau_{\text{public}} = 0$ can for example be tested using a Wald test. The respective test statistics for all parameters are shown in the fifth column of table 2.3. They are equivalent to the asymptotic t-statistics of the RUMNL model (C).[6]

[6] The test statistic for the Wald test is $\overset{a}{\sim} \chi^2_1$. The displayed value is the square root of this statistic which is $\overset{a}{\sim} N(0,1)$ by the properties of the χ^2 distribution with one d.f.

So in the case without generic variables, the NNNL and RUMNL models are equivalent. But while the RUMNL model directly estimates the parameters of interest, the estimated coefficients from NNNL have to be rescaled before they can be interpreted. This rescaling has also be taken into account when testing hypotheses. For example, the asymptotic t-statistics from the output of `nlogit` do not correspond to tests of intrinsically interesting hypotheses.

2.4.4 Example 2: Inclusion of Generic Variables

As discussed above, the researcher may often want to constrain the coefficients $\boldsymbol{\beta}_j$ of alternative-specific variables to be equal for each alternative. This constraint (2.4): $\boldsymbol{\beta}_j = \boldsymbol{\beta} \quad \forall j = 1, \dots, J$ could easily be imposed for the CL model B and for the RUMNL model D. However, the corresponding constraints on the NNNL parameters according to equation (2.14) are *not* equivalent. Instead of equal RUM parameters, they impose equal *scaled* RUM parameters:

$$\begin{aligned} & \widetilde{\boldsymbol{\beta}}_j = \widetilde{\boldsymbol{\beta}} \quad \forall j = 1, \dots, J \\ \Leftrightarrow \quad & \frac{1}{\tau(j)} \boldsymbol{\beta}_j = \widetilde{\boldsymbol{\beta}} \quad \forall j = 1, \dots, J \\ \Leftrightarrow \quad & \boldsymbol{\beta}_j = \tau(j) \widetilde{\boldsymbol{\beta}} \quad \forall j = 1, \dots, J \end{aligned} \tag{2.21}$$

The structural parameters $\boldsymbol{\beta}_j$ are not restricted to be equal across alternatives. Instead, they are constrained to be proportional to the IV parameters of their nest. The author of this chapter cannot think of a RUM model for which these constraints could make any sense. Why should the travel time in our example be associated with more disutility for travel modes that happen to share a nest with relatively dissimilar alternatives?

Table 2.4 shows NNNL estimates with `time` specified as a ‘generic’ variable in the sense of equation (2.21). A comparison of models F and D illustrates that the NNNL model does not give the same estimates as the RUMNL model in this case. In particular, the log likelihood values differ. While the corresponding RUMNL model D shows very different IV parameters for both nests (0.55 vs. 4.80), the estimates of the NNNL IV parameters from model F are relatively similar for both nests (2.54 vs. 2.64). With the intuition developed so far, this can be readily interpreted. In the RUMNL model, the IV parameters solely capture the (dis)similarity of the alternatives within the corresponding nest. While the public transportation modes appear to be quite similar, the other modes are not. This is reflected in the RUMNL estimates. The IV parameters in the NNNL model capture another effect: the relative importance of travel time for the alternatives within the nest. The diverging IV parameters that are in accordance with the dissimilarity would imply that travel time is much more

Table 2.4: NNNL estimates with generic variables

Model	(F) NNNL		(G) NNNL		(H) RUMNL	
	Coef.	z	Coef.	z	Coef.	z
const×car	-2.325	-2.56	-2.556	-3.01	-6.645	-3.26
bus	-2.364	-2.87	-2.398	-3.03	-6.235	-2.88
train	-1.319	-1.73	-1.358	-1.86	-3.531	-1.89
hinc× car	-0.138	-1.34	-0.150	-1.47	-0.390	-1.47
bus	-0.196	-1.56	-0.191	-1.54	-0.497	-1.64
train	-0.352	-3.18	-0.349	-3.24	-0.907	-3.68
time	-0.460	-6.75	-0.456	-6.73	-1.185	-5.64
time× air	-1.988	-5.39	-2.079	-6.04	-5.405	-5.46
τ public	2.535	4.29	2.600	4.41	2.600	4.41
τ other	2.638	4.36	2.600	4.41	2.600	4.41
Log likelihood	-194.01		-194.29		-194.29	

important for the car alternative than for the public transportation modes. This is not the case as is obvious from the previous results. So both effects that are captured by the same NNNL IV parameters are not in line with each other.

The 'generic' specification for the NNNL model implies a counterintuitive restriction that can hardly be motivated from a RUM model. As a result, specifications like model F should be avoided. RUM Models like model D can in general not be estimated with NNNL software like Stata's `nlogit` command if generic variables are present. There are exceptions some of which are discussed in the next section.

2.5 Special Nesting Structures

Section 2.4.4 argued, that the specification of NNNL models with generic variables can in general imply implausible binding constraints. This section discusses special cases for which this is not be true.

2.5.1 Equal IV Parameters across all Nests

If one is willing to assume a priori that the dissimilarity parameters of all nests in a nesting level have the same value, the scaling problem of the NNNL model disappears. The restrictions (2.21) imply essentially the same as the generic restrictions in a RUMNL model according to equation 2.4. The presence of the generic variable does not distort the estimates of the NNNL model, since its parameter is forced to be scaled equally in each nest.

Table 2.4 shows results for a NNNL and a RUMNL model that differ from the previous ones in that their IV parameters are constrained to be equal. The RUMNL parameters (model H) can be deduced from the NNNL estimates by multiplying them with the joint IV parameter. For example the estimated RUMNL income coefficient for the train alternative is $\widehat{\gamma}_{\text{hinc}\times\text{train}} = -0.907$. It can be recovered from the NNNL estimates as $\widehat{\widehat{\gamma}}_{\text{hinc}\times\text{train}} \times \widehat{\tau}_{\text{public}} = -0.349 \times 2.600$.

The problem with this constraint is that it cannot be tested with NNNL estimates, because the unconstrained model F is misspecified. In contrast, both RUMNL specifications are valid and a comparison of the log likelihood values of models D and H clearly shows that this constraint is rejected by the data.

2.5.2 Degenerate nests

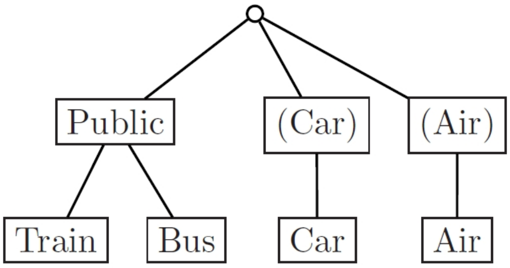
If a nest contains only one alternative, it is called a degenerate nest. The dissimilarity parameter of degenerate nests is not defined in the RUMNL model. This can be easily seen from equations (2.10) and (2.12). Since the degenerate nest $B(j)$ only contains alternative j, its inclusive value (2.10) simplifies to $IV(j) = \frac{1}{\tau(j)}V_j$. The dissimilarity parameter $\tau(j)$ cancels out of the choice probability (2.12). This is intuitive since the concept of (dis)similarity does not make sense with only one alternative.

In the NNNL model however, the dissimilarity parameter of degenerate nests does not vanish from the choice probability and may be statistically identified. As discussed above, the identification in general comes from two sources: the dissimilarity and the relative importance of the 'generic' variables in the respective nest. Like in the RUMNL model, the former source disappears in degenerate nests. But the latter source may be present if generic variables enter the model. Without 'generic' variables, the dissimilarity parameters are not jointly identified with the other parameters. So they can be constrained to any nonzero value. The only effect of choosing this value is that the respective parameters are scaled accordingly as discussed in section 2.4.3.

If at least one 'generic' variable is included in the NNNL model, the IV parameter of degenerate nests may be identified along with the other model parameters. This identification comes from the restriction of equally *scaled* parameters $\frac{1}{\tau(j)}\boldsymbol{\beta}_j$ across alternatives and nests and the parameters only constitute this scaling. A conventional approach to restrict the IV parameter to be equal to unity does not result in a model that is consistent with the underlying RUM model.

This is demonstrated with the estimates shown in table 2.5. The fact that the estimated dissimilarity parameter of the nest 'other' in table 2.2 is substantially larger than 1 indicates that the alternatives 'air' and 'car' should not share a nest. Therefore, the nesting structure is modified by splitting this nest into two degenerate nests. The resulting nesting structure is depicted in figure 2.2. In models I and J shown in table 2.5, the variable 'time' purely enters as a generic variable. The dissimilarity parameters of the degenerate nests 'air' and 'car' are not identified from the RUMNL model I. As argued above, they cancel out in the likelihood function. In contrast, all IV parameters are identified in the NNNL model J. It has two more free parameters than the RUMNL model and a substantially higher likelihood value.

Figure 2.2: Nesting structure for models I through L



However, these IV parameters do not have anything to do with (dis)similarity. They simply relax the constraint of equal scaling of the generic variable coefficient across nests. To demonstrate this, models K and L shown in table 2.5 do the same explicitly by estimating a separate 'time' coefficient for each nest. As a result, the IV parameters of the degenerate nests are not jointly identified with the other parameters of the corresponding nests in the NNNL model and have to be constrained to any nonzero number. Both models result in the same log likelihood value and the parameters are equivalent if the NNNL parameters are rescaled with the value of the corresponding IV parameter. The results are also equivalent to model J. This supports the assertion that the IV parameters in model J do nothing more than relax the constraint of equal scaling.

So if there is only one generic variable present in the model, the NNNL estimate of the IV parameter can be interpreted in a straightforward way, although this is probably not the way the researcher intends to interpret IV parameters.

Table 2.5: Degenerate Nests

Model		(I)		(J)		(K)		(L)	
		RUMNL		NNNL		RUMNL		NNNL	
		Coef.	z	Coef.	z	Coef.	z	Coef.	z
const×	car	1.140	1.97	-19.400	-2.74	-3.613	-3.83	-3.613	-3.83
	bus	3.206	6.17	-7.283	-1.48	-1.433	-1.56	-7.283	-1.48
	train	3.371	6.19	-5.130	-1.07	-1.010	-1.11	-5.130	-1.07
hinc×	car	-0.011	-0.10	-0.695	-1.09	-0.130	-1.09	-0.130	-1.09
	bus	-0.451	-4.31	-2.328	-2.74	-0.458	-3.81	-2.328	-2.74
	train	-0.505	-4.83	-3.013	-3.13	-0.593	-4.86	-3.013	-3.13
time		-0.165	-3.79	-2.319	-4.66				
time×	public					-0.456	-6.17	-2.319	-4.66
	air					-2.654	-6.73	-2.654	-6.73
	car					-0.432	-6.11	-0.432	-6.11
τ public		0.073	2.96	0.197	3.78	0.197	3.78	0.197	3.78
τ air		—	—*	1.144	3.86	—	—*	1	—**
τ car		—	—*	0.186	3.74	—	—*	1	—**
Log likelihood		-212.45		-182.57		-182.57		-182.57	

*: Parameter not defined.

**: Parameter normalized to 1.

It is much more direct to explicitly relax the specification of generic variables. If there is more than one generic variable, the interpretation becomes more obscure. Then the NNNL specification imposes the restriction that the coefficients of all generic variables differ proportionally across nests. Greene (2000, example 19.18) presents a model in which this problem appears. It is a NNNL model based on the data used in this chapter. In addition to 'time', the generic variable 'cost' is included. As a result, the estimates have no clear interpretation. The RUMNL avoids the danger of misspecification and misinterpretation.

2.5.3 Dummy nests

There is a way to 'trick' NNNL software into estimating a RUM consistent nested logit model with generic variables and without imposing equality of dissimilarity parameters. Koppelman and Wen (1998) propose to add degenerate dummy nests and constrain their IV parameters appropriately. This can most easily be explained by an example.

Figure 2.3 shows the nesting structure for the travel mode choice example according to figure 2.1 with appropriate dummy nests added. For each alternative, such a degenerate nest is specified. The corresponding IV parameters θ_1 through θ_4 are shown next to each nest along with the respective constraint. The two 'public' alternatives each have a degenerate dummy nest whose IV parameters are constrained to be equal to the IV parameter of the 'other' nest. Intuitively, their parameters are first scaled by $1/\tau_{\texttt{public}}$. Then the additional dummy nest scales them by $1/\tau_{\texttt{other}}$. For the two 'other' alternatives, this works accordingly. As a result, the parameters of *all* alternatives are scaled by $\frac{1}{\tau_{\texttt{public}} \cdot \tau_{\texttt{other}}}$. While τ_1 and τ_2 can be allowed to differ, this does not translate into different scaling across nests.

Figure 2.3: Nesting structure with dummy nests

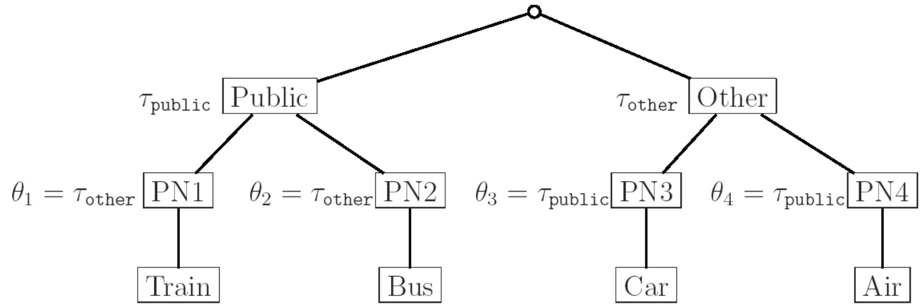

$\tau_{\texttt{public}}$
Public
$\tau_{\texttt{other}}$
Other
$\theta_1 = \tau_{\texttt{other}}$
PN1
$\theta_2 = \tau_{\texttt{other}}$
PN2
$\theta_3 = \tau_{\texttt{public}}$
PN3
$\theta_4 = \tau_{\texttt{public}}$
PN4
Train
Bus
Car
Air

Table 2.6 shows the results from a specification according to this strategy. Model M is identical to model D. It could not be reproduced by NNNL since it contains generic variables and the IV parameters are allowed to differ between nests. Model N is a NNNL model with the dummy nests added as described above (NNNL-DN). As can be seen, this specification mimics the RUMNL model except for the scaling of the parameters for the explanatory variables. The structural coefficients can be recovered from these estimates by multiplying the estimated coefficients by both estimated IV parameters. For example, the coefficient for 'hinc×train is -0.831. it can be calculated from the NNNL-DN estimates as -0.831 = -0.317 × 4.801 × 0.545.

Depending on the original nesting structure, a large number of dummy nests may be needed for this strategy. This complicates both the specification and the

Table 2.6: Dummy nests

Model		(M)=(D)		(N)	
		RUMNL		NNNL-DN	
		Coef.	z	Coef.	z
const×car		-6.383	-2.24	-2.438	-1.51
	bus	-2.782	-1.03	-1.063	-0.86
	train	-1.786	-0.66	-0.682	-0.59
hinc×	car	-0.362	-0.93	-0.138	-0.92
	bus	-0.554	-1.93	-0.212	-1.58
	train	-0.831	-2.91	-0.317	-1.98
time		-1.301	-5.60	-0.497	-3.08
time×	air	-5.878	-5.54	-2.245	-2.69
τ public		0.545	3.79	0.545	3.79
τ other		4.801	3.84	4.801	3.84
Log likelihood		-165.257		-165.26	

estimation.[7] This strategy therefore seems to be a real alternative to RUMNL only for researchers who just have access to a NNNL implementation.

2.6 Stata implementation of RUMNL

The NNNL model is available for Stata 7.0 users as the `nlogit` command. As argued in this chapter, the RUMNL model is preferable in most situations. This section introduces the command `nlogitrum.ado` that implements the RUMNL model. It was used to produce all RUMNL estimates in this chapter. Furthermore, the command `nlogitdn.ado` is described. It adds dummy nests to any specified nesting structure as discussed in section 2.5.3.

2.6.1 Installation

The implementation of `nlogitrum.ado` was presented in Heiss (2002) in the Stata Journal. The package including the implementation as well as the data and code for the examples used in this paper is therefore available for automatic installation. From Stata, the command `net search nlogitrum` will provide

[7] The command `nlogitdn`, introduced in section 2.6.5, automates the generation of dummy nests and appropriate constraints.

instructions. Alternatively, the package is available for manual download at `http://www.stata-journal.com/software/sj2-3/st0017/`.

2.6.2 Data setup

The data setup for `nlogitrum` is equivalent to `nlogit`. That is, a set of categorical variables *altsetvarB* [... *altsetvar2 altsetvar1*] is generated using `nlogitgen`. The tree structure can be visualized using `nlogittree`. For a thorough description see [R] **nlogit**.

2.6.3 Syntax

```
nlogitrum depvar indepvars [weight] [if exp] [in range] , group(varname)
    nests(altsetvarB [ ...  altsetvar2 altsetvar1]) [ notree nolabel clogit
    level(#) nolog robust ivconstraints(string) constraints(numlist)
    maximize_options ]
```

The syntax is similar to that of `nlogit` with one major difference. `nlogit` insists on explanatory variables for each nesting level and `nlogitrum` only allows explanatory variables to directly enter the conditional probabilities of the alternatives. There are three reasons for this change. The first reason is that in many cases it is hard to find a variable that is specific to a nest instead of an alternative. So one often ends up throwing nonsense variables into the specification of nest-specific explanatory variables and constraining their coefficients to zero. The second reason is that for the RUMNL model, it does not make a difference at all if a nest-specific variable is specified for a nest or for all alternatives within the nest. The third reason is that it greatly simplifies the syntax and makes it equivalent to the syntax of `clogit` except for the additional options.

The option `d1` of `nlogit` does not exist for `nlogitrum`. The current version uses the `ml` method `d0`.

2.6.4 Predictions

The syntax for `predict` after `nlogitrum` is nearly identical to the syntax after `nlogit` estimation. The only difference is that the options `xbb` and `xbb#` are replaced by the option `xb`, since the linear prediction can only be sensibly defined for the bottom level (the alternatives).

2.6.5 Generating dummy nests: nlogitdn

The command nlogitdn is a wrapper for nlogit. Its syntax is equivalent to the nlogit syntax. nlogitdn analyzes the specified nesting structure, adds appropriate dummy nests and constraints to the specification as discussed in section 2.5.3, and calls nlogit. It was used for the estimation of model N in table 2.6.

2.6.6 Examples

In order to help the reader become accustomed to the syntax, the commands used to produce the example models A through N are listed below. Most variable names should be self-explanatory. The variable grp identifies the observations and the variable travel identifies the alternatives and takes the values 0 for air, 1 for train, 2 for bus, and 3 for car. The variable mode is the 0/1 coded dependent variable. For most NMNL models, the nesting structure is depicted in figure 2.1. The respective variable type was generated using nlogitgen. For the models I through L, the nesting structure according to figure 2.2 was generated with the variable typedeg:

```
. nlogitgen type = travel(public: 1 | 2, other: 0 | 3 )
new variable type is generated with 2 groups
lb_type:
           1 public
           2 other
. nlogitgen typedeg = travel(public: 1 | 2, air: 0, car: 3)
new variable typedeg is generated with 3 groups
lb_typedeg:
           1 public
           2 air
           3 car
```

Since no variables enter the models on the level of the nests, the nonsense variables nothing1 and nothing2 were generated. The constraints that show up in the nlogit commands constrain their coefficients to zero. The models themselves were estimated using the following commands:

```
. * Model A:
. clogit mode asc_* hinc_* time_*, group(grp)
. * Model B:
. clogit mode asc_* hinc_* time time_air, group(grp)
. * Model C:
. nlogitrum mode asc_* hinc_* time_*, group(grp) nests(travel type)
. * Model D:
. * Model E:
. nlogit mode (travel = asc_* hinc_* time_* )(type=nothing1), group(grp) const(
> 1) d1
. *Model F:
. nlogit mode (travel = asc_* hinc_* time time_air )(type=nothing1), group(grp)
>  const(1)
```

```
. * Model G:
. nlogit mode (travel = asc_* hinc_* time time_air)(type=nothing1), group(grp)
> const(1) ivc(other=public)
. * Model H:
. nlogitrum mode asc_* hinc_* time time_air, group(grp) nests(travel type) ivc(
> other=public)
. * Model I:
. nlogitrum mode asc_* hinc_* time , group(grp) nests(travel typedeg) ivc(air=3
> .14159, car=3.14159)
. * Model J:
. nlogit mode (travel = asc_* hinc_* time )(typedeg=nothing2), group(grp) const
> (2)
. * Model K:
. nlogitrum mode asc_* hinc_* timepublic time_air time_car, group(grp) nests(tr
> avel typedeg) ivc(air=3.14159, car=3.14159)
. * Model L:
. nlogit mode (travel = asc_* hinc_* timepublic time_air time_car)(typedeg=noth
> ing2), group(grp) const(2) ivc(air=1, car=1)
. * Model M = Model E
. * Model N:
. nlogitdn mode (travel = asc_* hinc_* time time_air)(type=nothing1), group(grp
> ) const(1)
```

Note that the IV parameters of ‘air’ and ‘car’ in models I and K do not actually exist as discussed in section 2.5.2. Since the algorithm does not realize this beforehand, these parameters have to be restricted to an arbitrary nonzero number (in the examples, 3.14159 was chosen to illustrate the arbitrariness).

2.7 Conclusions

The name ‘nested logit’ has been given to different models. This chapter argues and demonstrates that the seemingly slight difference in the specification of the outcome probabilities can lead to substantially different results and interpretations thereof. So researchers using a nested logit model (and the readers of their results) should be aware of the actual variant used.

One of these variants (called RUMNL in this chapter) is derived from a random utility maximization (RUM) model that is prevalent in econometrics. The estimated coefficients can be readily interpreted and simple tests like asymptotic t-tests directly test hypotheses of interest. This holds irrespective of the type of included explanatory variables and specified nesting structure.

The alternative (called NNNL in this chapter) implies a varying scaling of the underlying utilities across alternatives. Depending on the model specification, it can give equivalent results to those of RUMNL and the structural parameters can be recovered. But in order to do so, the estimated coefficients have to be rescaled and this also has to be kept in mind for hypothesis tests. This is the case if only alternative-specific parameters enter the model. If generic variables (variables with a common coefficient across alternatives) are present, the NNNL model places restrictions on the parameters that are often counterintuitive and undesired. The reason is that the inclusive value parameters in this case not only constitute the (dis)similarities of the alternatives, but also the different scaling of the generic variable coefficients across nests.

Stata 7.0 comes with an implementation of the NNNL model. This chapter introduces the Stata package `nlogitrum` that implements the preferred RUMNL model.

3 State Space Approaches to Microeconometric Panel Data Modeling: Health Trajectories with Initial and Dynamic Selection through Mortality

3.1 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 microeconometric models, especially limited dependent variable models, are inherently nonlinear. This nonlinearity complicates the analysis of panel data models, see Chamberlain (1984). Heckman (1981b) discusses a general setup for nonlinear panel data models in the context of binary choice models. In applied research, the vast majority of nonlinear panel data models specify unobserved heterogeneity as time-constant fixed or random effects and/or state dependence as a low-order Markov model.

State space models separate the model into the specification of a latent state process and a measurement model which connects it to the observed outcomes. 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. State-space models also provide a general and intuitive basis to formulate microeconometric panel data models. Commonly used approaches like random effects models are directly nested within a more flexible and potentially more plausible specification.

In this chapter, I discuss the general model structure and simulation-based estimation of state space models. Furthermore, special topics are covered. These include the specification of the state process in continuous time and joint modeling of multiple dependent variables. As a special case of simultaneous models, panel attrition can be modeled jointly with the variable of interest to allow for selectivity correction. To analyze quantitative results of the estimated model, the simulation of conditional trajectories is discussed.

These ideas are then applied to modeling the evolution of self-reported health (SRH). Panel data models usually applied in this literature include Markov chain and random effects models. I show that a simple and parsimonious state space model captures the data much better. It is based on a simple process of latent health in continuous time that generates the SRH answers in the surveys. Furthermore, I discuss the problem of selectivity caused by mortality both in the initial sample and by panel attrition. With a joint model of SRH and mortality, I show how these biases can be corrected and demonstrate the selectivity effects.

The chapter is structured as follows. Section 3.2 presents the general model structure and the requirements on the specification and estimation of state space models. It also presents a discussion of topics like continuous time modeling of the state space and selectivity correction in a simultaneous model. In section 3.3, an empirical application for self-reported health is presented. Different model specifications are implemented and tested against each other. Section 3.4 discusses the problem of selection through mortality and presents a joint model of health and mortality. The effects of different specifications are demonstrated in simulation exercises. Section 3.5 concludes.

3.2 Nonlinear State-Space Models for Panel Data

3.2.1 Model Structure

A large share of microeconometric models involve sets of unobserved random variables that are not mutually independent conditional on observable covariates. Examples include panel data models with unobserved heterogeneity, sample selection models with selection on unobservables, and simultaneous equation models. State space models specify separate models for the unobserved random variables (state space) and the connection with the observed variables (measurement). Let $\mathbf{y}_{it}$ with $i = 1, ..., N$ and $t = 1, ..., T$ denote the observed dependent variable of individual i at wave t in $N \times T$ dimensional panel data. In general, $\mathbf{y}_{it}$ may be a vector if different outcomes are modeled simultaneously.

Let $\mathbf{x}_i$ denote a vector of observed strictly exogenous covariates. These can be constant or or varying over time. In the latter case, $\mathbf{x}_i$ collects all time-specific values. In addition, the model is formulated in terms of unobserved random variables ("states") $\mathbf{u}_{it}$ and possibly i.i.d. random shocks $\mathbf{e}_{it}$. In general, these can be random vectors if more than one state variable is defined. The complete model consists of two parts:

State Space

The state space model specifies the joint distribution of the states $\mathbf{u}_{it}$ conditional on the covariates. Let $\mathbf{u}_{it}$ be continuously distributed and denote the vector of individual state sequences as $\mathbf{u}_{i,1:T} = [\mathbf{u}_{i1}, ..., \mathbf{u}_{iT}]$. Assume that its joint p.d.f. conditional on the covariates $f(\mathbf{u}_{i,1:T}|\mathbf{x}_i; \boldsymbol{\theta})$ is a function known up to a finite-dimensional parameter vector $\boldsymbol{\theta}$. Furthermore, assume that vectors of random numbers can be drawn from this distribution.

Measurement

The measurement model specifies the data generating process *conditional on* the latent states. So $\mathbf{u}_{i,1:T}$ can be treated like the observed covariates $\mathbf{x}_i$. Let $\mathbf{y}_{i,1:T} = [\mathbf{y}_{i1}, ..., \mathbf{y}_{iT}]$ denote the vector of all individual outcomes. The most important assumption needed for the further analysis is that the joint distribution of $\mathbf{y}_{i,1:T}$ conditional on $\mathbf{x}_i$ and $\mathbf{u}_{i,1:T}$ is known up to a finite number of parameters. Discrete and continuous dependent variables are in the following treated jointly. With a slight misuse of terminology, in the following I will refer to $P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta})$ as the corresponding probability for $\mathbf{y}_{i,1:T}$ conditional on $\mathbf{x}_i$ and $\mathbf{u}_{i,1:T}$.

A special case of this class of models can be called *contemporaneous* state space models. Conditional on $\mathbf{u}_{it}$, the outcome $\mathbf{y}_{it}$ is assumed to be independent of $\mathbf{y}_{is}$ and $\mathbf{u}_{is}$ for all $s \neq t$. It follows that

$$P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) = \prod_{t=1}^{T} P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{u}_{it}; \boldsymbol{\theta}). \tag{3.1}$$

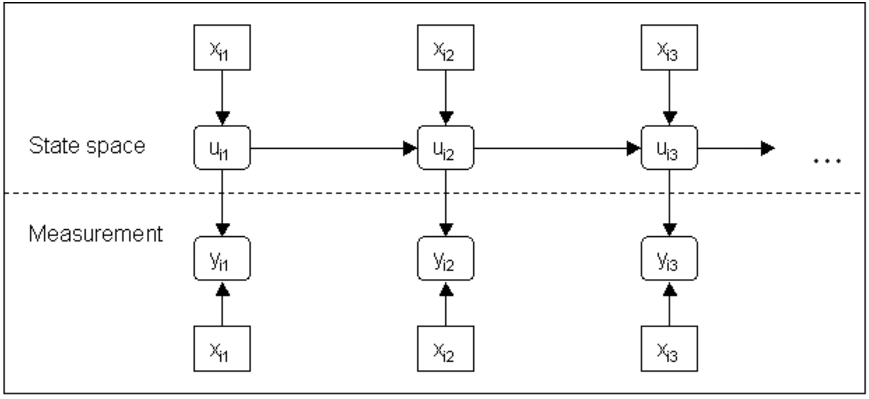
A graphical illustration of such a model is presented in Figure 3.1. The unobserved states possibly depend on covariates and are dependent over time. Conditional on $\mathbf{u}_{it}$, the outcomes $\mathbf{y}_{it}$ are serially independent. But without this conditioning, the serial correlation of $\mathbf{u}_{it}$ induces serial correlation of $\mathbf{y}_{it}$ conditional on $\mathbf{x}_i$.

A simple example of a binary panel probit model with a one-dimensional AR(1) error term may help to clarify the approach. The measurement model is a straightforward probit model with $\mathbf{x}_i = [\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}]$ and $\mathbf{u}_{i,1:T} = [u_{i1}, ..., u_{iT}]$ as explanatory variables. Conditional on those, the outcomes are assumed to be independent. With the parameter vector $\boldsymbol{\theta} = [\boldsymbol{\beta}, \sigma, \rho]$, the probit specification implies the measurement model

$$P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) = \prod_{t=1}^{T} \Phi\left(\mathbf{x}_{it}\boldsymbol{\beta} + u_{it}\right)^{y_{it}} \left(1 - \Phi\left(\mathbf{x}_{it}\boldsymbol{\beta} + u_{it}\right)\right)^{1-y_{it}}, \tag{3.2}$$

where Φ denotes the standard normal c.d.f. Let the states u_{it} be specified as a Gaussian AR(1) process that is independent of $\mathbf{x}_i$. The marginal distribution of each u_{it} is normal with zero mean and variance σ^2. Conditional on the past values

Figure 3.1: Contemporaneous state space models



$u_{i1}, ..., u_{i,t-1}$, the distribution of u_{it} is normal with mean $\rho u_{i,t-1}$ and variance $(1-\rho^2)\sigma^2$, where $|\rho| < 1$. The joint density of $\mathbf{u}_{i,1:T}$ is therefore

$$f(\mathbf{u}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta}) = \frac{1}{\sigma}\,\phi\left(\frac{u_{it}}{\sigma}\right)\prod_{t=2}^{T}\frac{1}{\sigma\sqrt{1-\rho^2}}\,\phi\left(\frac{u_{it}-\rho u_{i,t-1}}{\sigma\sqrt{1-\rho^2}}\right). \tag{3.3}$$

3.2.2 Estimation

In this chapter, only maximum likelihood estimation is discussed. For Bayesian analyses, the main problem of the likelihood evaluation is analogous, so the discussion can be easily applied. Other methods like GMM suffer from similar computational problems. Assume that the random variables involved in the model are independent across cross-sectional units. The likelihood function for the general panel data state space model can then be written as

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y},\mathbf{x}) = \prod_{i=1}^{N}\ell(\boldsymbol{\theta}|\mathbf{y}_{i,1:T},\mathbf{x}_i), \tag{3.4}$$

where $\mathbf{y} = [\mathbf{y}_{i,1:T} : i = 1,\ldots,N]$ and $\mathbf{x} = [\mathbf{x}_i : i = 1,\ldots,N]$. The likelihood contributions represent the outcome probabilities, interpreted as a function of the parameters: $\ell(\boldsymbol{\theta}|\mathbf{y}_{i,1:T},\mathbf{x}_i) = P(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta})$. These likelihood contributions cannot in general be evaluated explicitly since the model does not provide $P(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta})$. However, it can be expressed as an expectation of the known $P(\mathbf{y}_{i,1:T}|\mathbf{x}_i,\mathbf{u}_{i,1:T};\boldsymbol{\theta})$ over the conditional state distribution:

$$P(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta}) = \int P(\mathbf{y}_{i,1:T}|\mathbf{x}_i,\mathbf{u}_{i,1:T};\boldsymbol{\theta})\,f(\mathbf{u}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta})\,d\mathbf{u}_{i,1:T}. \tag{3.5}$$

This multi-dimensional integral cannot be expressed in closed form except for very special cases. Instead, it can be approximated numerically by different methods. It is well known that Monte Carlo integration is a flexible solution to such problems. For the implementation, a number R of draws from the joint distribution characterized by $f(\mathbf{u}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta})$ has to be generated. For many specifications, this is straightforward. Geweke (1996), Hajivassiliou and Ruud (1994) and Train (2003) provide general discussions on univariate and multivariate random number generation. For the AR(1) example from section 3.2.1, this can be done as follows. For each $r = 1, ..., R$, a random number u_{i1}^r is drawn from the univariate normal distribution of the initial state u_{i1}. Then sequentially for each $t = 2, ..., T$, a random number u_{it}^r is drawn from the conditional distribution of u_{it} given $u_{1,t-1} = u_{1,t-1}^r$. The resulting vector $\mathbf{u}_{i,1:T}^r = [u_{i1}^r, ..., u_{iT}^r]$ is a draw from the joint distribution $f(\mathbf{u}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta})$. Alternatively, standard sampling methods from multivariate normal distributions based on Choleski factorization of the covariance matrix can be used.

Given these draws, the simulated likelihood contribution is calculated as

$$\tilde{P}(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta}) = \frac{1}{R}\sum_{r=1}^{R} P(\mathbf{y}_{i,1:T}|\mathbf{x}_i,\mathbf{u}_{i,1:T}^r;\boldsymbol{\theta}). \tag{3.6}$$

Under mild regularity conditions, the simulated likelihood contribution $\tilde{P}(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta})$ is an unbiased estimate of $P(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta})$ and converges to its true but unknown value almost surely as $R \to \infty$ by a law of large numbers, see for example Geweke (1996).

For maximum likelihood estimation however, the log likelihood function involves taking the logarithm of the likelihood contributions. As a result of Jensen's inequality, $\log(\tilde{P}(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta}))$ is biased downwards for $\log(P(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta}))$ with a given finite number of draws R. But with $R \to \infty$, the consistency carries over to the log likelihood. As $N \to \infty$, the maximum simulated likelihood estimator based on $\log(\tilde{P}(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta}))$ is consistent if R rises with N. It is asymptotically equivalent to the infeasible maximum likelihood estimator based on $\log(\tilde{P}(\mathbf{y}_{i,1:T}|\mathbf{x}_i;\boldsymbol{\theta}))$ if R rises fast enough so that $R/\sqrt{N} \to \infty$ as $N \to \infty$. See for example Hajivassiliou and Ruud (1994) for a more detailed discussion of maximum simulated likelihood estimation. Similar methods based on simulating moments (McFadden 1989) or scores (Hajivassiliou and McFadden 1998) are discussed in the literature.

This dissertation also discusses alternative approaches to estimation. Instead of simulation, the integral in equation 3.5 can be approximated by deterministic integration. Gaussian quadrature and related methods are inherently defined for one-dimensional integrals. For these problems, they are known to work efficiently (Butler and Moffit 1982). But even if the state space (that is each $\mathbf{u}_{it}$) is one-dimensional, integration is over all periods and therefore T-dimensional. The well-known product rule of extending quadrature methods to higher dimensions requires computational costs that rise exponentially with the number of dimensions and is therefore infeasible for dimensions larger than 4 or 5.

Chapter 5 proposes and alternative way to extend one-dimensional quadrature to higher dimensions. For a different application, it is shown to strikingly outperform Monte Carlo integration in terms of accuracy and computational effort. An alternative is to reformulate equation 3.5 such that the high-dimensional integral can be split into several lower-dimensional integrals. Chapter 4 discusses this for state space models that have a contemporaneous structure as defined above and for which the state space has a markov-property such that only a limited number of lagged values have independent predictive power for $\mathbf{u}_{it}$. In this chapter, I concentrate on Monte Carlo integration as described above.

3.2.3 State Space in Continuous Time

In the measurement model, a finite number of values of the latent states $\mathbf{u}_{i1}, ..., \mathbf{u}_{iT}$ affect the observed outcomes. Since for the estimation of the model only draws from their joint distribution is needed, the original formulation of the state process can easily be defined in continuous time. In the empirical application discussed in section 3.3, the states constitute the latent health of the respondents. Since the time lag between the interviews in the panel data vary considerably, a formulation of the health evolution in continuous time is more natural than a model of changes from wave to wave.

Let $\mathbf{u}_i(\tau)$ denote such a process, where τ denotes continuous calendar time. Assume that the measurement model specifies $\mathbf{y}_{it}$ to depend on $\mathbf{u}_{it} := \mathbf{u}_i(\tau_t)$, where τ_t may constitute the calendar time at which the corresponding survey was conducted. So $\mathbf{u}_{i,1:T}$ corresponds to a sample from the continuous-time process $\mathbf{u}_i(\tau)$ at certain points in time $\tau_1, ..., \tau_T$. In this chapter, these points in time are treated as exogenous. A joint model of endogenous sampling times may be interesting for certain application but will be left for future research.

As long as the joint distribution of $\mathbf{u}_{i,1:T}$ can be evaluated from the properties of their underlying process $\mathbf{u}_i(\tau)$, this specification does not lead to additional problems for the parameter estimation. Introducing an example for a specification that is also used in the application in section 3.3 may help to clarify the approach. Assume for simplicity a one-dimensional state space. The latent states evolve according to an Ornstein-Uhlenbeck process with zero mean. It corresponds to the continuous-time analogue to an AR(1) process in discrete time. The marginal distribution of $u_i(\tau)$ is normal with zero mean and variance σ^2. Conditional on a previous realization $u_i(\tau - \Delta)$, $u_i(\tau)$ is normally distributed with mean $\rho^{\Delta} u_i(\tau - \Delta)$ and variance $(1 - \rho^{2\Delta})\sigma^2$, where $|\rho| < 1$. The correlation between $u_i(\tau)$ and $u_i(\tau - \Delta)$ is ρ^{Δ}. With $\mathbf{u}_{it} = \mathbf{u}_i(\tau_t)$, the joint distribution of $\mathbf{u}_{i,1:T}$ is normal with zero mean and covariance matrix

$$\sigma^2 \begin{bmatrix} 1 & \rho^{\tau_2-\tau_1} & \rho^{\tau_3-\tau_1} & \cdots & \rho^{\tau_T-\tau_1} \\ \rho^{\tau_2-\tau_1} & 1 & \rho^{\tau_3-\tau_2} & \cdots & \rho^{\tau_T-\tau_2} \\ \rho^{\tau_3-\tau_1} & \rho^{\tau_3-\tau_2} & 1 & \cdots & \rho^{\tau_T-\tau_3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{\tau_T-\tau_1} & \rho^{\tau_T-\tau_2} & \rho^{\tau_T-\tau_3} & \cdots & 1 \end{bmatrix}.$$

From this multivariate normal distribution, random numbers can be as easily generated as from the distribution implied by an AR(1) process as discussed above.

3.2.4 Sample Selection and Panel Attrition

The general model structure discussed in section 3.2.1 allows a model of multiple dependent variables that depend on the same set of unobserved state variables. This allows a specification of both a contemporaneous and intertemporal correlation structure between different outcome variables conditional on the covariates $\mathbf{x}_i$. As a special case, a random variable that represents selection into the sample can be modeled. This allows a straightforward model with selection on observables *and* states – a special form of selection on unobservables that explicitly takes advantage of the panel structure. Since attrition is observed in the data, this selection process can be estimated. This also allows to correct for initial sample selection if both sources of selection are driven by the same mechanism.

For the application of state space models to health, selection through mortality is the most obvious source of selectivity. Let $\mathbf{y}_{it} = [y_{it}^h, y_{it}^m]$, where y_{it}^h denotes a measure of interest such as self-reported health and y_{it}^m denotes an indicator of mortality. So y_{it}^h is observed if $y_{it}^m = 0$. Assume that conditional on $\mathbf{x}_i$ and $\mathbf{u}_{i,1:T}$, these two outcomes are independent and independent over time. This would correspond to a classical selection on observables specification if $\mathbf{u}_{i,1:T}$ were observed. The measurement model for the observed outcomes can now be written as

$$P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) = \prod_{t=1}^{T} P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) \tag{3.7}$$

with

$$P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) = \begin{cases} P(y_{it}^m|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) P(y_{it}^h|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) & \text{if } y_{it}^m = 0 \\ P(y_{it}^m|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) & \text{if } y_{it}^m = 1 \end{cases} \tag{3.8}$$

If the initial sample were representative for the whole population, this joint model of health and mortality could be easily estimated since it corresponds to the general setup discussed above. A further complication arises in the mortality example since the selection process has been at work prior to the initial survey wave. The initial sample consists of respondents who survived up to the first interview only, so $y_{i1}^m = 0$ for all i. This is obviously a selected sample. Because of the conditional independence assumption, the joint outcome probability taking into account this initial selectivity can be written as

$$P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{u}_{i,1:T}, y_{i1}^m = 0; \boldsymbol{\theta}) = P(y_{i1}^h|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) \prod_{t=2}^{T} P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}), \quad (3.9)$$

where $P(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta})$ is defined as above. Given the sequence $\mathbf{u}_{i,1:T}$, this does not create additional problems. For the likelihood evaluation, this expression has to be integrated over the appropriate distribution of $\mathbf{u}_{i,1:T}$. The conditional independence assumption implies $P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{u}_{i,1:T}, y_{i1}^m = 0; \boldsymbol{\theta}) = P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta})$, so

$$P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, y_{i1}^m = 0; \boldsymbol{\theta}) = \int P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta})\, f(\mathbf{u}_{i,1:T}|\mathbf{x}_i, y_{i1}^m = 0; \boldsymbol{\theta})\, d\mathbf{u}_{i,1:T}. \quad (3.10)$$

The conditioning on $y_{i1}^m = 0$ shows up in the conditional distribution of $\mathbf{u}_{i,1:T}$. Clearly, the conditional density $f(\mathbf{u}_{i,1:T}|\mathbf{x}_i, y_{i1}^m = 0; \boldsymbol{\theta})$ is not equal to the density $f(\mathbf{u}_{i,1:T}|\mathbf{x}_i; \boldsymbol{\theta})$ which is given by the state space specification. The difference is driven by the fact that only respondents with a favorable latent state sequence survive and can be sampled in the initial survey.

One strategy to approximate the integral in equation 3.10 is importance sampling. Rewrite

$$\begin{aligned} &P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, y_{i1}^m = 0; \boldsymbol{\theta}) \\ &= \int P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) \frac{f(\mathbf{u}_{i,1:T}|\mathbf{x}_i, y_{i1}^m = 0; \boldsymbol{\theta})}{f(\mathbf{u}_{i,1:T}|\mathbf{x}_i; \boldsymbol{\theta})}\, f(\mathbf{u}_{i,1:T}|\mathbf{x}_i; \boldsymbol{\theta})\, d\mathbf{u}_{i,1:T}. \end{aligned} \quad (3.11)$$

The importance sampling factor can be written by Bayes' rule as

$$q_i(\mathbf{u}_{i,1:T}) = \frac{f(\mathbf{u}_{i,1:T}|\mathbf{x}_i, y_{i1}^m = 0; \boldsymbol{\theta})}{f(\mathbf{u}_{i,1:T}|\mathbf{x}_i; \boldsymbol{\theta})} = \frac{P(y_{i1}^m = 0|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta})}{P(y_{i1}^m = 0|\mathbf{x}_i; \boldsymbol{\theta})} \quad (3.12)$$

As long as $P(y_{i1}^m = 0|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta})$ can be evaluated, an importance sampling algorithm based on equation 3.11 is the following:

1. Draw R sequences $\mathbf{u}_{i,1:T}^1, ..., \mathbf{u}_{i,1:T}^R$ from the joint distribution $f(\mathbf{u}_{i,1:T}|\mathbf{x}_i; \boldsymbol{\theta})$
2. For each $r = 1, ..., R$ evaluate $Q_i^r = P(y_{i1}^m = 0|\mathbf{x}_i, \mathbf{u}_{i,1:T}^r; \boldsymbol{\theta})$ and $P_i^r = P(\mathbf{y}_{i,1:T}|\mathbf{x}_i, \mathbf{u}_{i,1:T}^r; \boldsymbol{\theta})$
3. The simulated likelihood contribution is $\tilde{P}_i = \frac{\sum_{r=1}^R P_i^r Q_i^r}{\sum_{r=1}^R Q_i^r}$

3.2.5 Simulating Conditional Trajectories

Using the same idea of importance sampling as for initial sample selection solution discussed above, outcome probabilities conditional on any other outcomes can be evaluated. This is particularly useful for the simulation of conditional trajectories to study implied features of the estimated models as demonstrated in the application in section 3.3. In the health application, an example is the probability that a respondent reports poor health at some age conditional on SRH in another wave and/or survival to yet another age.

Let $\mathbf{y}_i^A$ denote a subset of the individual outcome sequence $\mathbf{y}_{i,1:T}$. The researcher is interested in the outcome probability of $\mathbf{y}_i^A$ conditional on another subset $\mathbf{y}_i^B$ of the individual outcome sequence. By conditional independence and Bayes' rule, the conditional outcome probability can be written equivalently to equations 3.11 and 3.12 as

$$P(\mathbf{y}_i^A|\mathbf{x}_i, \mathbf{y}_i^B; \boldsymbol{\theta}) = \int P(\mathbf{y}_i^A|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) \frac{P(\mathbf{y}_i^B|\mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta})}{P(\mathbf{y}_i^B|\mathbf{x}_i; \boldsymbol{\theta})} f(\mathbf{u}_{i,1:T}|\mathbf{x}_i; \boldsymbol{\theta}) \, d\mathbf{u}_{i,1:T}. \tag{3.13}$$

The importance sampling approximation of this integral is equivalent to that for initial selectivity correction. Each draw $\mathbf{u}_{i,1:T}^1, ..., \mathbf{u}_{i,1:T}^R$ from the joint state distribution is assigned an importance sampling weight $q_i^r = \frac{P(\mathbf{y}_i^B|\mathbf{x}_i, \mathbf{u}_{i,1:T}^r; \boldsymbol{\theta})}{\sum_{s=1}^R P(\mathbf{y}_i^B|\mathbf{x}_i, \mathbf{u}_{i,1:T}^s; \boldsymbol{\theta})}$. The simulated conditional outcome probability is the weighted mean $\tilde{P}(\mathbf{y}_i^A|\mathbf{x}_i, \mathbf{y}_i^B; \boldsymbol{\theta}) = \sum_{r=1}^R q_i^r P(\mathbf{y}_i^A|\mathbf{x}_i, \mathbf{u}_{i,1:T}^r; \boldsymbol{\theta})$.

3.3 Models of Latent and Self-Reported Health

3.3.1 Data and Descriptive Evidence

This section discusses modeling strategies to study the evolution of individual health over time. The application is based on panel data from the Health and Retirement Study (HRS) which is sponsored by the National Institute of Aging (NIA) and conducted by the University of Michigan. The data version used is the RAND HRS Data File (Version D). It was developed by the RAND Center for the Study of Aging with funding from the National Institute on Aging (NIA) and the Social Security Administration (SSA).

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. This applies to 25,499 respondents. After excluding respondents with missing information on essential variables, all analyses are based on a sample of 25,453 respondents. For those, a total of 118,674 observations are available (excluding observations after the death of a respondent).

In 4,423 of those cases, the respondent was "lost" from the sample and not even the mortality status could be ascertained. Those cases are dropped from the sample. The death of 5,414 respondents is observed during the time covered by the panel data.

The HRS provides information on a large number of various aspects of health such as self-reported health, prevalence and incidence of certain health conditions and functional limitations. In this chapter, I concentrate on a frequently studied measure, the self-reported health (SRH). The wording of this question in the HRS is "Would you say your health is excellent, very good, good, fair, or poor?". This health measure has the advantage that it is very general in the sense that all aspects of health are included, weighted by the individual perception. Although the answer to this question is obviously subjective, it contains considerable objective information. For example, it is a strong predictor for mortality, even when controlling for a large number of relevant covariates. I come back to a discussion of the objectiveness of this measure below. In 5,587 cases in the sample, the respondent is known to be alive but does not provide a SRH answer due to survey or item nonresponse (predominantly the former). This leaves a total of 103,250 observations of SRH.

The answers to the SRH question are highly persistent. Table 3.1 shows the transition rates from the answers in one wave to the answers in the next wave for all waves and all respondents that provide an answer in both adjacent waves. About half of the respondents give exactly the same answer in two adjacent waves and another 30-40% change the answer by only one "unit".

Table 3.1: Health transitions

Previous health			Current health				
	Obs.	%	poor	fair	good	v. good	exc.
poor	6,293	8.3	**57.0**	30.6	9.3	2.4	0.8
fair	13,704	18.1	16.6	**48.0**	26.5	7.2	1.6
good	22,915	30.2	4.3	19.1	**50.1**	22.4	4.1
very good	21,537	28.4	1.7	6.6	27.5	**51.0**	13.2
excellent	11,467	15.1	1.0	3.1	12.7	33.8	**49.4**
Total	75,916	100.0	9.7	19.3	30.4	27.8	12.8

There are various explanations for the high persistence of health over time that is illustrated in Table 3.1. The classical discussion differentiates between observed and unobserved heterogeneity and true state dependence. True state dependence is interpreted as the fact that the observed outcome in one wave structurally affects the outcome of the next wave. A prominent example is labor force participation: Not working in one period lowers the own market wage due to signaling and human capital depreciation and thereby the probability of working in the next period. Many studies of SRH specify a Markov chain model in which current health is modeled to depend on lagged SRH outcomes, see for example Contoyannis, Jones, and Rice (2003). The coefficient of these lagged outcome variables is usually found to be highly significant. But the interpretation as causal state-dependence is implausible in SRH studies. The answer a respondent gives to the SRH question surely has no structural effect on the true health status in the next wave. Typically, unobserved heterogeneity is modeled as time-constant fixed or random effect. Hernandez-Quevedo, Jones, and Rice (2004) for example present a random effects ordered probit model for the study of SRH.

Table 3.2 provides a more detailed look at the structure of intertemporal correlation of SRH in the original HRS subsample. It shows the results of a simple logit regression of poor or fair SRH (the "bad state") in wave 5 on the number and timing of bad states in the prior waves and can be seen as a condensed description of the runs patterns. There are no other explanatory variables included, but the picture does not change qualitatively when controlled for observed heterogeneity. Results from such regressions can be requested from the author. If time-constant heterogeneity drives the persistence of SRH, the number of previous bad states should be the relevant statistic to predict wave 5 SRH. The results show that in fact the higher the number of previous bad states, the higher is the probability of another bad state in wave 5. But in addition, the timing of those previous bad states plays a significant role. The closer to wave 5 these bad states occur, the higher is their predictive power for the status in wave 5. This rules out time-constant unobserved heterogeneity as the only source of the intertemporal correlation.

If the results are instead interpreted as state dependence in form of a k^{th} order Markov chain, only the first k lags should have predictive power. The results show significant predictive power of all previous waves and additionally the number of bad states. Obviously, this could be modeled as a saturated 4^{th} order Markov chain. But in this case, only the last wave is available for analyses since the have to be conditioned on. Furthermore, there is no reason to believe that if there were yet an earlier wave available, it would have no predictive power as would be required by a 4^{th} order Markov chain. Since *conditional on the number of bad states* the predictive power of wave 1 and 2 cannot be statistically distinguished, a strategy could be to model SRH as a third-order Markov chain with unobserved heterogeneity. But this would create the problem of initial conditions for the first three waves (Heckman 1981a). When moving from the binary indicator measure of health to the full 5-point scale, the number of nuisance parameters would also become very large.

Table 3.2: Structure of intertemporal correlation of SRH

Logit estimates, dependent variable: poor/fair health in wave 5				
	Odds Ratio	Std. Err.	z	P
# bad states=0	1.00		(reference)	
# bad states=1	4.06	0.50	11.29	0.00
# bad states=2	8.40	1.61	11.11	0.00
# bad states=3	11.92	3.17	9.33	0.00
# bad states=4	39.24	13.18	10.93	0.00
status w1 = bad	1.00		(reference)	
status w2 = bad	1.06	0.15	0.43	0.67
status w3 = bad	1.44	0.17	2.99	0.00
status w4 = bad	2.24	0.26	6.85	0.00
Number of obs	LL	LR chi2(7)	Pseudo R2	Prob ¿ chi2
8771.00	-3049.04	3698.25	0.38	0.00

The question is whether there is a possibility to formulate a model that captures the structure of intertemporal correlation observed in Table 3.2 in a parsimonious and plausible way. This section departs from classical models of state dependence and unobserved heterogeneity and develops a state space model that has both aspects. Health itself is interpreted as a latent state variable with unobserved heterogeneity. This latent variable follows a stochastic process that is correlated over time like an AR(1) process so *it* can be interpreted to be state dependent. The observed SRH measure is merely an indicator of the current health status and *therefore* correlated over time. In a parsimonious way, this strategy produces the pattern observed in Table 3.2. Health is correlated over the whole life. But the correlation between SRH in two points in time diminishes the further they are apart since shocks in the unobserved health status accumulate.

3.3.2 A State-Space Model of Health and SRH

In this section, a simple state space model is suggested in which health is modeled as a latent state that evolves over time and self-reported health is driven by this "true" health together with contemporaneous shocks or measurement errors. Independent and random effects models are discussed as a special case of this model with parametric restrictions.

Objective health is modeled as a one-dimensional continuous state process $u_{i1}, ..., u_{iT}$. Assume that given the current health state, SRH is generated by an ordered logit model. In the notation of the general model, $y^h_{it} \in \{1, ..., 5\}$ where

1 denotes "poor" and 5 denotes "excellent" SRH. The measurement model can be formulated as

$$y^h_{it} = j \quad \Leftrightarrow \quad \alpha_{j-1} \leq u_{it} + e_{it} < \alpha_j \qquad \forall j = 1,..,5, \tag{3.14}$$

where the SRH answer is driven by the health stock u_{it} and temporary shocks e_{it} which are specified as i.i.d. logistic random variables. These variables may be interpreted as transitory health problems like a cold or random misclassifications due to the current mood of the respondents or other factors (Crossley and Kennedy 2002). If the combined expression $u_{it} + e_{it}$ has a value between α_{j-1} and α_j, then the respondent gives answer $y^h_{it} = j$. The parameters α_1 through α_4 are estimated, $\alpha_0 = -\infty$ and $\alpha_5 = \infty$. This corresponds to a usual ordered logit model with the unobserved health component as an explanatory variable. Its coefficient is normalized to unity to scale the latent health state. The conditional outcome probabilities for each wave are simply

$$P(y^h_{it} = j|\mathbf{x}_i, \mathbf{u}_{it}, \boldsymbol{\theta}) = \Lambda(\alpha_j - u_{it}) - \Lambda(\alpha_{j-1} - u_{it}), \tag{3.15}$$

where Λ denotes the logistic c.d.f.

Health itself is unobserved and represented by a one-dimensional state space. For each point in time, u_{it} denotes the continuous health status. It is assumed to be additively separable into two components:

$$u_{it} = \mu_{it} + a_{it}. \tag{3.16}$$

The "deterministic" part μ_{it} is modeled as a parametric function of explanatory variables $\mathbf{x}_{it}$ such as age. Throughout this chapter, the assumption of a linear specification is maintained, but a generalization to any parametric function would be straightforward:

$$\mu_t = \mathbf{x}_t\boldsymbol{\beta}. \tag{3.17}$$

The "stochastic" part a_{it} is modeled as a latent process over time. Throughout this chapter I assume that the marginal distribution of each a_{it} is normal with zero mean and a variance of σ^2 and independent of the explanatory variables:

$$a_{it} \sim \mathcal{N}(0, \sigma^2) \tag{3.18}$$

This assumption can easily be replaced with a different parametric distribution or for example by allowing for heteroscedasticity.

Different assumptions on the correlation over time are specified and tested. As a starting point, an independent model is implemented. In the special case of $\sigma^2 = 0$, the unobserved states have a degenerate distribution with $a_{it} = 0$ for all $t = 1, ..., T$. The health process and therefore the outcomes are independent conditional on the exogenous variables. Alternatively, a random effects model is specified by assuming that the "stochastic" health part is constant over time: $a_{it} = a_i$ for all $t = 1, ..., T$. If an ordered probit model of SHR is specified as the measurement model, a random effects ordered probit model similar to the specification of Hernandez-Quevedo, Jones, and Rice (2004) is the result.

As already seen in Table 3.2, the random effects specification is unlikely to capture the structure of intertemporal correlation of SRH. In a third specification, I allow correlations of latent health at two different points in time which depends on the time gap and an unknown correlation parameter ρ. The time gap between two interviews varies considerably in the HRS. The correlation of latent health in one and in the next wave plausibly depends on this gap. The Ornstein-Uhlenbeck (OU) process already discussed in section 3.2.3 allows a straightforward and natural approach to capture these effects. It is an equivalent to an AR(1) process in continuous time.

The full set of model parameters for this model is $\boldsymbol{\theta} = [\boldsymbol{\beta}, \alpha_1, ..., \alpha_4, \sigma, \rho]$. The random effects model follows as a special case with $\rho = 1$ and the i.i.d. specification restricts $\sigma = 0$. For these three specifications of the latent health stock process, the ordered logit model of SRH is estimated. As explanatory variables, only linear splines of age were included in the health stock equation to capture the general deterioration of health in a flexible but straightforward way.

Table 3.3 provides an overview over the results. Model 1 is the i.i.d. specification. It corresponds to a simple ordered logit model with 5 parameters for the age splines and 4 parameters for the cut points. Model 2 uses the RE specification. It corresponds to a random effects ordered logit model. In addition to the parameters in Model 1, the variance of the (normally distributed) random effect is estimated. Its estimate is highly significantly different from zero and its magnitude is large compared to the i.i.d. error term in the SRH equation (7 vs. $\pi^2/3 = 3.3$). Consequently, the i.i.d. specification of Model 1 is clearly rejected by a LR test (test statistic = 49922.2 with 1 degree of freedom) . This variance also leads to a high correlation of SRH over time in the model. Since the random effect is constant over time, the correlation of a_{it} over time is restricted to 1.

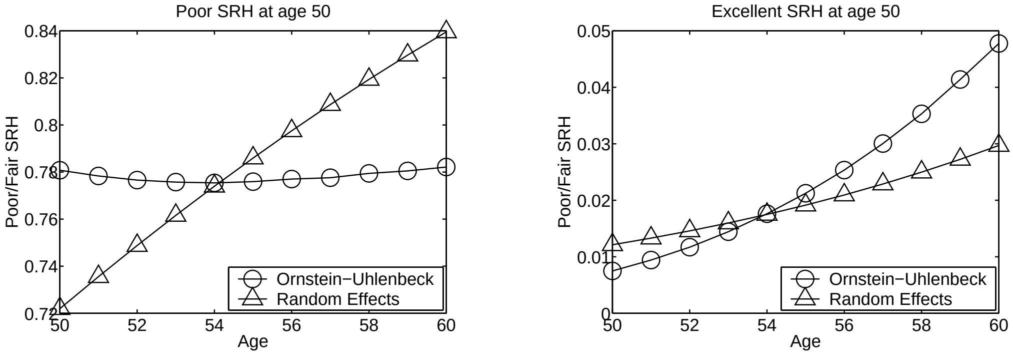
Model 3 relaxes the assumption of perfectly correlated unobserved health with the Ornstein-Uhlenbeck specification. The estimated correlation between health at one point in time and health 1 year later is $\rho = .96$. So the correlation with health 10 years later is $\rho^{10} = .69$. The correlation is very high but significantly smaller than one. Consequently an LR test clearly rejects Model 2 against this specification (test statistic = 621.5 with 1 degree of freedom).

The RE and the OU specification imply different patterns of intertemporal correlation. Figure 3.2 illustrates these differences by showing the path of predicted probabilities to report poor or fair SRH for a respondent who once reported poor (left figure) or excellent (right figure) SRH at age 50. In the RE

Table 3.3: Results: Different specifications of latent health

	Model 1: i.i.d.		Model 2: RE		Model 3: OU	
	Estimate	z-Stat	Estimate	z-Stat	Estimate	z-Stat
Latent Health:						
variance	0.0000	(restr.)	6.9925	122.71	9.1261	73.82
corr. (1year)	—	(n.a.)	1.0000	(restr.)	0.9630	1922.50
# individuals	25,453		25,453		25,453	
# observations	103,250		103,250		103,250	
# parameters	9		10		11	
Log-Likelihood	-156,165.4		-131,204.3		-130,893.5	

Figure 3.2: Random effects vs. Ornstein-Uhlenbeck



specification, the unobserved health component is constant over time. The slope of the corresponding lines is completely driven by the age gradient of health. In the OU specification, unobserved health is highly but not perfectly correlated over time. In addition to the age gradient, the corresponding lines are driven by a regression to the mean effect, since the predictive power of health at age 50 decreases over time. The higher estimated health variance compensates for this so that the "average" predictive power has a similar level. Loosely speaking, this "average" predictive power identifies the variance of the random effect, since the average respondent is observed for about eight years. The parameter ρ of the OU model is identified by the decreasing predictive power which was already documented in Table 3.2 and drives the regression to the mean which generates the different slopes in Figure 3.2.

As discussed above, another way to account for the correlation of SRH over time is a Markov chain model in which current outcome probabilities are conditioned on previous outcomes. This strategy has the advantage that it can be computationally simpler since lagged dependent variables can be added just as any other explanatory variable and a straightforward ordered logit model can be estimated using standard software. But as argued above, a dependence of the latent health variable is much more plausible than this direct specification of the observed outcome. In addition, the conditioning on lagged dependent variables is problematic for two reasons: For the initial observation, there is obviously no information on the lagged outcome. This leads to the initial conditions problem (Heckman 1981a) if outcomes are not conditionally independent. In addition, missing information for example due to item nonresponse creates a similar problem. Even if these values are missing at random so that the use as a dependent variable is not problematic, the information is additionally missing for the next observation as an explanatory variable. With conditionally dependent observations, the researcher has to resort to computationally intensive methods so that the advantage of this approach disappears.

In order to test the proposed modeling approach, I compare the results of a model with the i.i.d. specification (Model 4) and an OU model (Model 6) corresponding to Models 1 and 3, respectively, to a first-order Markov chain model (Model 5) in Table 3.4. This model is an ordered logit model with the full set of 4 lagged dependent variables in addition to the age splines as explanatory variables. Since it requires observed lagged dependent variables, I constrain the sample in all three models and condition on the initial observations. For the OU model, this is done using the general importance sampling method that is also used for initial selectivity correction. Because of the high persistence of SRH already mentioned, the first-order markov chain model fits the data much better in terms of the likelihood than the i.i.d. specification. But although the OU specification is more parsimonious, it clearly outperforms the first-order markov model in terms of the likelihood. Compared to the discussed alternatives, the state space model with an OU-process for the latent health process succeeds very well in capturing the intertemporal correlation structure of the data in a parsimonious way.

Table 3.4: Results: First-order Markov chain vs. state space model

	Model 4: i.i.d.	Model 5: Markov	Model 6: OU
# individuals	22,787	22,787	22,787
# observations	75,915	75,915	75,915
# parameters	9	13	11
Log-Likelihood	-114,276.0	-92,141.6	-89,362.0

3.4 Joint Models of Health and Mortality

Obviously health and mortality are related. Table 3.5 shows mortality rates from wave to wave in the HRS data. Overall, about six percent of the respondents die between two waves. Mortality strikingly differs by previous SRH: While only 1.7 percent of respondents who report excellent health at one wave die before the next wave, more than 20 percent of those who report poor health do.

Table 3.5: Wave to wave mortality (percent)

total	by previous SRH				
	poor	fair	good	v. good	excel.
5.96	20.6	9.4	4.4	2.3	1.7

3.4.1 Model Structure

With a joint model of health and mortality, both initial selectivity and panel attrition due to mortality can be treated in a consistent way. I therefore add a measurement model for mortality risk depending on the same latent health stock that drives SRH. Let $\tau_1, ..., \tau_T$ denote the times of interview in continuous time where τ_T is the time of death if the respondent dies during the period covered by the panel data. The mortality hazard rate at these points in time is specified as

$$\lambda_i(\tau_t) = \lambda_0 e^{\delta u_{it}}. \tag{3.19}$$

For the survival probability between two waves, the the whole path of hazard rates and therefore health is relevant. Note that the health stock changes over time due to the change of explanatory variables and the stochastic process of its unobserved path. The former effects can be easily dealt with. The OU specification of latent health is based on continuous time, but a hazard rate model with continuously changing hazard of this form has no closed-form expression. See Yashin and Manton (1997) for a general discussion of this problem. The simulation-based estimation suggested above relies on a finite number of relevant states over time to allow sampling from their joint distribution. I suggest a simple approximation of the continuous-time problem in discrete time. For all points in time between τ_t and τ_{t+1}, the hazard rate is based on the linear interpolation

$$\tilde{\lambda}_i(\tau) = \lambda_0 e^{\delta \tilde{u}_i(\tau)} \qquad \text{with } \tilde{u}_i(\tau) = \tfrac{\tau-\tau_t}{\tau_{t+1}-\tau_t} u_{it} + \tfrac{\tau_{t+1}-\tau}{\tau_{t+1}-\tau_t} u_{i,t+1}. \tag{3.20}$$

Since the estimated correlation of the health process over time is very high, the variance of $u_i(\tau)$ conditional on u_{it} and $u_{i,t+1}$ is very low and by far the most variation is created by variations of u_{it} and $u_{i,t+1}$. So the linear interpolation should be unproblematic.

With the interpolated hazard rate, the survival probabilities can easily be derived. Let y_{it}^m for $t = 1, \dots, T$ denote the binary random variable that corresponds to survival to τ_t given survival to τ_{t-1}. With $\delta \neq 0$, the conditional probability of survival is

$$P(y_{it}^m = 0 | \mathbf{x}_i, \mathbf{u}_{i,1:T}; \boldsymbol{\theta}) = \exp\left(-(\tau_t - \tau_{t-1}) \frac{\lambda_i(\tau_t) - \lambda_i(\tau_{t-1})}{\delta(u_{it} - u_{i,t-1})} \right). \tag{3.21}$$

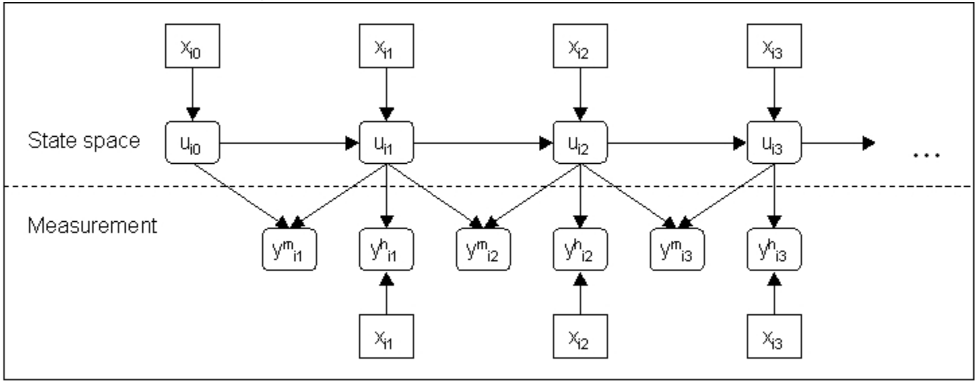
The panel attrition problem discussed so far is similar to the dynamic selection effect (Cameron and Heckman 1998) discuss in a random effects model for the sequential decision to continue schooling. While in their application the assumption of initially independent observed and unobserved heterogeneity is plausible, it is intrinsically inconsistent in our application if the initial time is defined as the first interview. Since the HRS initially samples respondents aged 50 or older, there has been substantial selectivity prior to the first wave for respondents who are sampled at higher ages. This implies that the age of the initial interview is positively correlated with the health trajectory prior to the first wave which in turn is correlated with current and future health. This effect leads to biased inferences if it is ignored in the analysis and only panel attrition after wave 1 is modeled. The depreciation of health over time is underestimated since it is obscured by the selection effect. The same is true for other covariates. For example if the selection mechanism works identically for men and women and men are unhealthier and thereby have a higher mortality risk, the observed cross-sectional health differences between gender are obscured by the differential selection effect.

With a joint model of health and mortality, it is straightforward to account for these effects, especially with the OU specification of latent health. Note that the fact of being interviewed implies conditioning on the outcome that the individual survived up to the age of the first interview. This can be used to correct the likelihood by the method for conditioning on outcomes discussed in section 3.2.4. The specification of the state space has to be slightly modified. Instead of assuming i.i.d. initial health at wave 1, I assume i.i.d. health at age 50 which is added as an artificial "wave" 0, so $\mathbf{u}_{i,0:T} = [u_{i0}, \dots, u_{iT}]$.

In addition, the measurement specification of SRH is modified by including age splines as explanatory variables in the ordered logit model. The parameters of these variables are identified jointly with the parameters in the latent health equation. The latter capture the age path of mortality risk, whereas the former capture the age path of SRH. Figure 3.3 depicts the structure of the joint model of SRH and mortality with a latent health process that drives both outcomes.

The observed-data likelihood of the joint model for SRH health and mortality can be evaluated according to equation 3.12.

Figure 3.3: Structure of the joint model



3.4.2 Comparison of Different Specifications

Analogous to Table 3.3, Table 3.6 gives an overview over results from joint models of SRH and mortality. Model 7 is a simple i.i.d. specification. SRH and mortality are modeled as independent ordered logit and hazard rate models, respectively. The model contains 5 parameters for the age splines in each of the outcome equations, a constant in the mortality equation, and four cut point parameters in the ordered logit specification. Model 8 involves an unobserved latent health state modeled as an OU process that drives panel attrition through mortality. This additionally requires three parameters: Its variance and correlation over time and its parameter in the mortality equation. The parameter for the SRH equation is normalized to unity – this identifies the variance.

Model 8 assumes i.i.d. initial health in wave 1. As argued above, independence from initial age is an implausible restriction. Model 9 replaces it with the assumption of i.i.d. health at age 50 and conditions on survival to the age of the first interview. This specification is also statistically superior in terms of the log likelihood. Both OU models nest the i.i.d. Model 7 which is clearly rejected by LR tests.

Table 3.6: Joint models: Different specifications of latent health

	Model 7: i.i.d.	Model 8: OU	Model 9: OU
initial distribution	n.a.	wave 1	age 50
# individuals	25,453	25,453	25,453
# observations	103,250	103,250	103,250
# parameters	15	18	18
Log-Likelihood	-177,281.1	-150,781.6	-150,438.0

3.4.3 Results and Simulations

Table 3.7 shows the parameter estimates of the preferred Model 9. Since they are not easy to interpret quantitatively, I show a number of simulation results that highlight the most important and interesting features of the fitted model. Unless otherwise indicated, all simulations presented in this section are based on this most general Model 9. Most of the simulations require some form of conditioning. The method for doing so was discussed in section 3.2.5.

Table 3.7: Model 9: Parameter estimates

		Latent Health		SRH		Mortality	
Std. Dev.	σ	3.241	(0.0153)				
Correlation	ρ	0.982	(0.0006)				
Constant		0	(restr.)	0	(restr.)	-6.888	(0.0918)
Age splines:	50+	-0.591	(0.0425)	0.461	(0.0434)	0	(restr.)
	60+	0.301	(0.0573)	-0.250	(0.0573)	0	(restr.)
	70+	-0.013	(0.0443)	-0.069	(0.0431)	0	(restr.)
	80+	-0.243	(0.0390)	0.184	(0.0374)	0	(restr.)
	90+	-0.097	(0.0501)	0.033	(0.0496)	0	(restr.)
Latent health				1	(restr.)	-0.330	(0.0057)
Cut points				(4 param.)			

Standard Errors in parentheses

The marginal density of the unobserved health process a_{it} is normal with mean zero and estimated standard deviation of $\sigma = 3.2$. Its distribution is shown in Figure 3.4 as the "unconditional" density $f(a_{it})$. The other two densities are conditional on SRH for a 50 years old respondent. These are by Bayes' rule simply the rescaled densities

$$f(a_{it}|y_{it}^h = j) = f(a_{it}) \frac{P(y_{it}^h = j|a_{it})}{P(y_{it}^h = j)}. \tag{3.22}$$

The enormous differences between health states illustrate the importance of the latent health for SRH.

Figure 3.4: Unobserved health by SRH at age 50

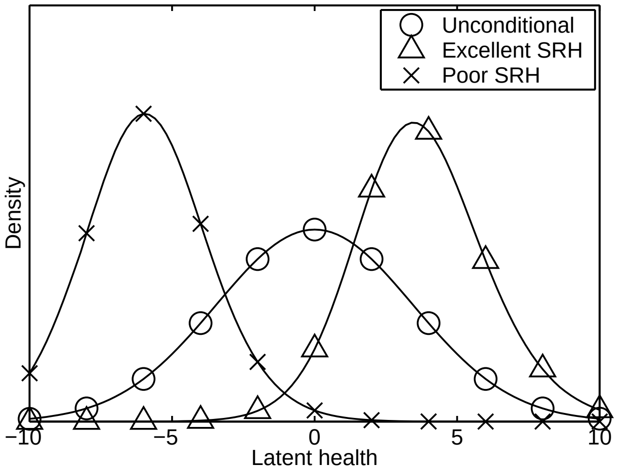
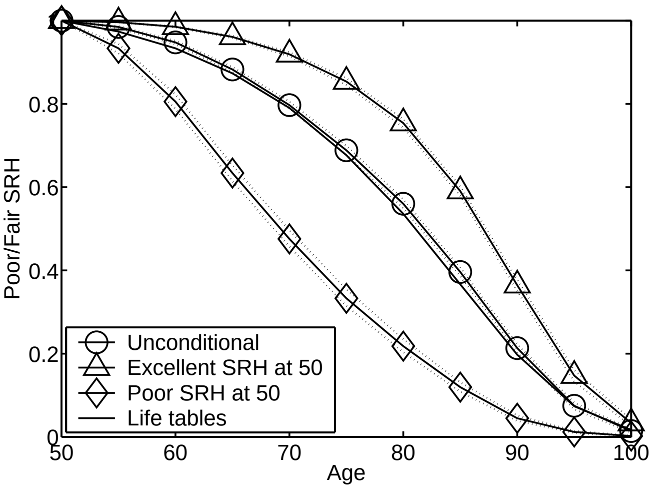


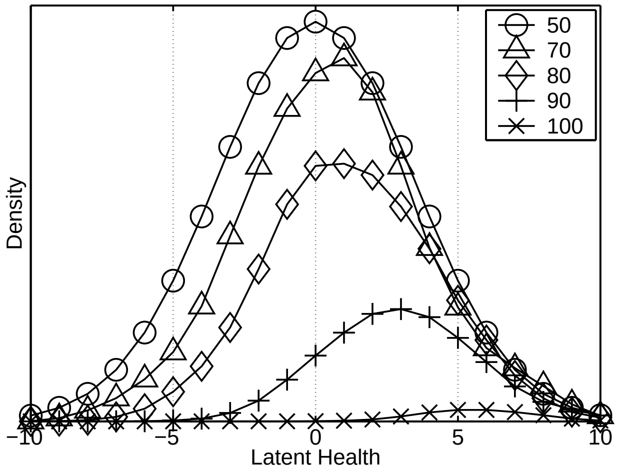
Figure 3.5: Survival probabilities



Since latent health also drives the mortality risk, SRH predicts mortality by signaling the conditional distribution of latent health as depicted in Figure 3.4. Figure 3.5 shows the simulated survival probabilities unconditional and conditional on SRH at age 50. The dotted lines for the simulated series indicate 95% confidence bands. Self-reported health and mortality are strongly connected. The simulations are based on the method of importance sampling to evaluate conditional outcome probabilities as discussed in section 3.2.5. The differences are striking. For example at age 70, only 48.5 percent of those respondents who report poor health at age 50 are alive, whereas 91.6 percent of those reporting excellent health survive.

Figure 3.5 also shows the survival rates conditional on survival to age 50 from the life tables for 1997 (National Center for Health Statistics: National Vital Statistics Report, Vol. 47, No. 28). The simulated unconditional survival probabilities tend to be slightly higher than the numbers from the life tables. This might be due to the fact that the HRS samples only individuals who are initially non-institutionalized. But the differences are small and both lines agree very well.

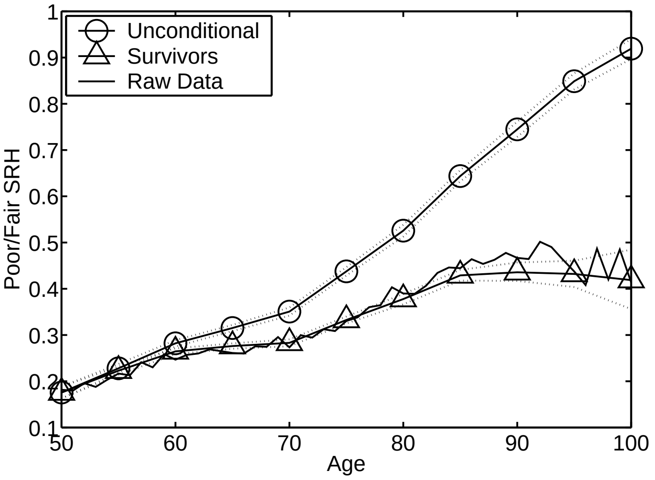
Figure 3.6: Densities of unobserved health of the surviving by age



Because of the strong effect of latent health on mortality and the high intertemporal correlation of latent health, the surviving population is a selected sample. Figure 3.6 illustrates this selection effect of mortality. It shows the distribution of the unobserved health state of the surviving population by age. The weighted kernel density estimates are scaled by survival probabilities so that the curves integrate to the share of surviving population at the respective age.

This shows the selection effect of mortality: Those at the lower end of the health distribution die earlier and therefore drop out of the sample.

Figure 3.7: Poor/fair SRH: Unconditional and surviving population



The selection effect of mortality on the distribution of latent health shown in Figure 3.6 directly translates to the trajectories of SRH. Figure 3.7 shows the simulated paths of the probability to report “poor” or “fair” health by age. The line labeled “unconditional” corresponds to the underlying true deterioration of health in the course of aging. This would be observed directly in the data if no deaths would occur or if health and mortality were independent. Because unhealthy people have a much higher mortality risk, the cross-sectional morbidity of the surviving population is much lower for the surviving population at higher ages. Its path is labeled “survivors” and condition on survival up to the respective age. It corresponds to the expected risk of poor or fair SRH given the selection-driven distribution of latent health that is shown in Figure 3.6. The simulated health path of the survivors tracks the cross-sectional means quite closely.

Each SRH or mortality observation at some age affects the complete predicted health path. In Figure 3.8, the whole health trajectories conditional on survival through at least age 80 and 90 are shown. This information has a considerable impact on the whole path of health because of the high intertemporal persistence of latent health. For example while the risk of poor or fair SRH for the whole population at age 50 is 17.5 percent, of those who are alive at age 80 or 90 only 9.6 or 6.3 percent were in poor or fair health at that age, respectively.

Figure 3.8: Poor/fair SRH by survival to higher ages

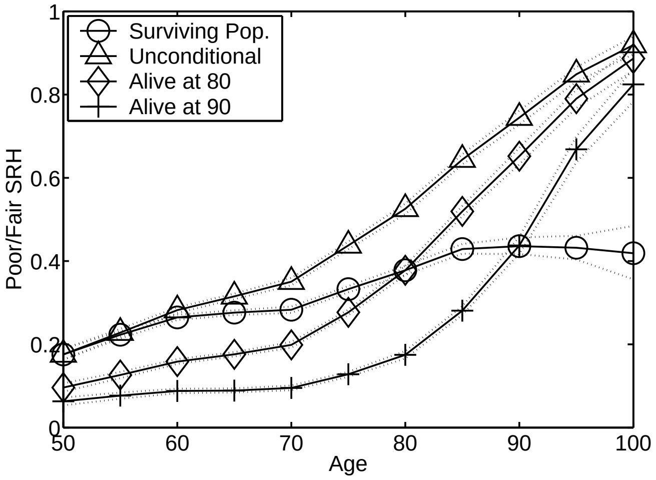
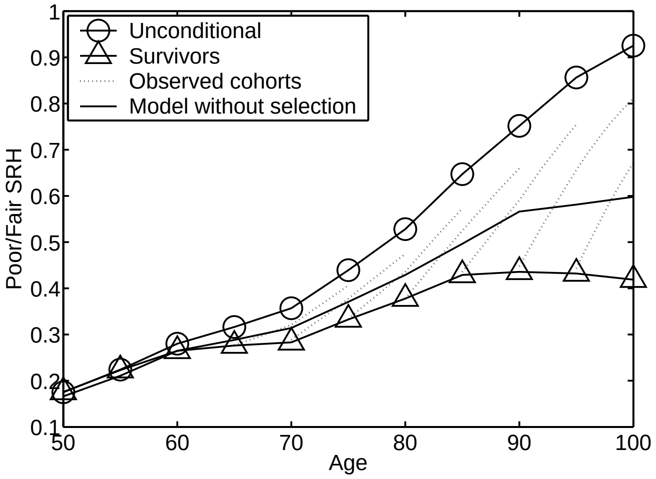
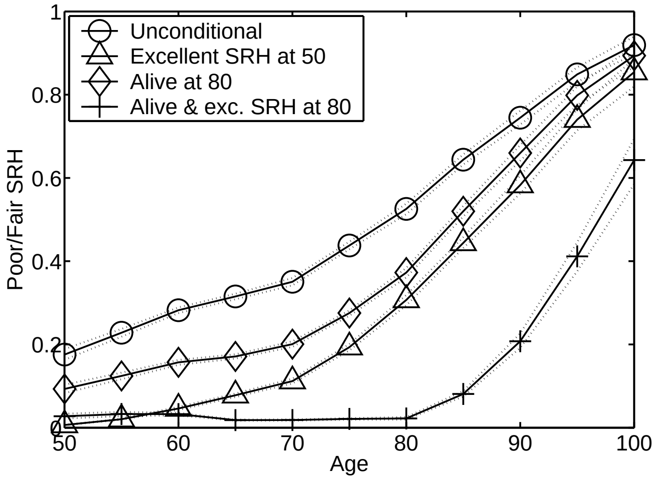


Figure 3.9: Poor/fair SRH: Cohorts and selection



The effects of ignoring differential mortality before and after the first wave can be seen in Figure 3.9. It shows the “unconditional” and “survivors” trajectories already known from Figure 3.7. In the initial wave, each respondent is drawn from the surviving population at the respective age. For different initial ages, the dotted lines correspond to the true evolution of health over the next ten years which corresponds to the maximum time each respondent is tracked in the data. Model 3 which ignores selection by mortality identifies these individual paths but falsely assumes that the individual health distribution is independent of age. As a result, the implied trajectory labeled "model without selection" lies between the unconditional and the survivors trajectory and has no direct interpretation.

Figure 3.10: Poor/fair SRH paths by survival and SRH



Finally, Figure 3.10 shows the SRH trajectories conditional on survival and SRH at ages 50 and 80. Both survival and excellent SRH are "positive" information so that the conditional paths of poor or fair SRH are all below the unconditional path. Health is highly persistent. While conditioning on SRH at some age has the biggest impact on nearby ages, the whole path is affected significantly. Excellent SRH at age 80 provides more extreme information on unobserved health a_{it} than at age 50 since the age gradient of latent health makes excellent SRH less likely.

3.4.4 Differences by Sex

In the models discussed so far, age was the only covariate. Of course, controlling for more covariates is straightforward. In this section, sex is considered in addition to age. The goal is to discuss specification and interpretation of the results. A dummy variable for female respondents and an interaction term with age is entered into the latent health equation. These two variables also enter the SRH equation. The parameters in the former equation therefore capture mortality differences between male and female respondents, the latter capture response differences.

Table 3.8: Model 10: Parameter estimates

		Latent Health		SRH		Mortality	
Std. Dev.	σ	3.237	(0.0784)				
Correlation	ρ	0.982	(0.0009)				
Constant		0	(restr.)	0	(restr.)	-6.779	(0.0957)
Age splines:	50+	-0.614	(0.0242)	0.467	(0.0250)	0	(restr.)
	60+	0.320	(0.0399)	-0.265	(0.0388)	0	(restr.)
	70+	-0.036	(0.0418)	-0.048	(0.0409)	0	(restr.)
	80+	-0.245	(0.0383)	0.182	(0.0371)	0	(restr.)
	90+	-0.115	(0.0500)	0.051	(0.0486)	0	(restr.)
Female		0.969	(0.2804)	-1.315	(0.2445)		
Female * age		0.031	(0.0096)	-0.003	(0.0086)		
Latent health				1	(restr.)	-0.334	(0.0063)
Cut points				(4 param.)			
Log likelihood:		-150277.1					
LR against Model 9:		321.9018 (4 d.f.)					

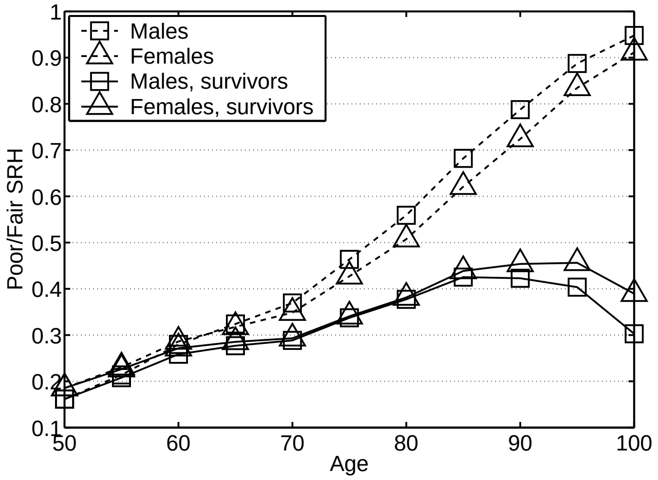
Standard Errors in parentheses

Table 3.8 shows the parameter estimates for the specification of Model 9, enriched with these covariates. Women for example are in better "objective health" identified by mortality risk. But *given* objective health, they report worse SRH. The total effect of gender on SRH can be calculated by adding the two coefficients which leads to a slightly worse SRH for women at higher ages.

Figure 3.11 shows the true morbidity paths and the morbidity of the surviving population by gender. At age 50, both start at about the same self-reported morbidity. It increases more rapidly for males. But at the same time, the selection due to mortality is stronger for men. As a result, SRH of the surviving population is about the same or even better for the male population in the cross-sectional data. So while observed SRH is worse for females at high ages in cross-sectional data, this can be explained by differential mortality alone. The true paths are the opposite.

Given these results, it is still unclear why women report worse SRH than men with the same health identified by mortality risk. This effect can be seen in the negative coefficient for females in the SRH equation. Women have a much lower mortality risk but still report about the same or even worse SRH than men. There may be different reasons for that. Health is a complex and multidimensional concept and SRH represents a summary measure, capturing all kinds of different health conditions and problems. Case and Paxson (2004)

Figure 3.11: Poor/fair SRH paths by gender

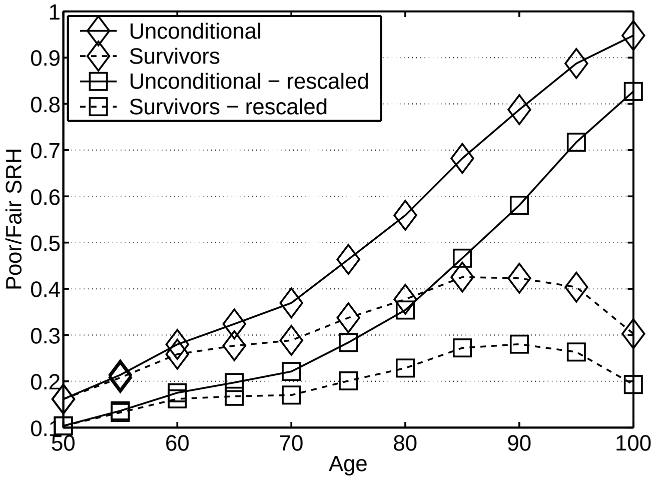


argue that an important reason for this "puzzle" might be a different prevalence of chronic conditions. Women suffer more often from conditions such as arthritis that have a large effect on well-being but only a minor effect on mortality risk, whereas men suffer more often from "deathly" conditions such as cardiovascular diseases.

A different explanation is that given the same objective health, the response process for SRH differs between males and females. For example, women might be more willing to admit or simply more aware of being in a bad health status than men. In a sense, Model 10 provides the opportunity to adjust for those different response scales. The coefficients of the sex variables in the latent health equation capture differential mortality risk. The coefficients in the SRH equation capture the response differences given latent health and thereby mortality risk. This does not strictly correspond to response scales since as argued above mortality risk and self-reported health are different functions of the complex and multidimensional concept of health.

Figure 3.12 shows simulations of the risk of poor or fair SRH for females. The lines labeled "rescaled" represents the simulated path for females adjusted for their differential responses. These lines show the trajectories for women *if they would give the same answer as men with the same mortality risk.*

Figure 3.12: Poor/fair SRH: Females with male response scales



3.5 Summary and Conclusions

This chapter discusses the use of state space models for microeconometric panel data analyses. It has been shown that this class of models provides a very flexible approach to formulate complex model structures in a straightforward and parsimonious way. The drawback is that for the analysis one has to resort to numerical integration techniques which are computationally intensive. But with modern computers, the computational burden is not prohibitive anymore and the technological progress can be expected to attenuate it further.

After a general discussion of requirements on the model structure, special issues have been discussed. Latent state processes in continuous time can easily be combined with observations at certain points in time such as survey interviews. Panel attrition is merely a special case of joint models for several dependent variables. The panel dimension directly allows to model sample selection on latent states which effectively uses the time series dimension of the panel data to identify selectivity.

In an application, a simple and parsimonious model of a latent health state process fits the data of self-reported health (SRH) much better than the Markov chain or random effects models frequently applied to this and similar measures. Furthermore, it was argued and the results show that the evolution of individual health cannot be studied in an isolated model of SRH and thereby implicitly conditional on survival. Various simulations demonstrate the strong intertemporal correlation of health and the strong link between SRH, the underlying health process and mortality.

Because of differential mortality, the surviving population is a selected sample and this selection systematically varies by age and sociodemographic characteristics. Ignoring this selectivity leads to biased results. For example the age gradient of health is underestimated severely since it is confounded by the survival of the healthiest. While at the average, elderly females tend to report worse SRH than males, this can be solely explained by a stronger selection due to a higher mortality of males.

4 Sequential Likelihood Evaluation for Nonlinear Panel Data Models

4.1 Introduction

This chapter discusses the estimation for a certain class of panel data models. It includes limited dependent or generally nonlinear models with an AR(1) error term. More general cases for simultaneous models that are driven by the same error process or multiple processes with certain properties are also covered.

In the straightforward case of a univariate error process, the likelihood function involves T-dimensional integrals, where T represents the time-series dimension of the panel data. The usual approach in the econometric literature to numerically approximate these integrals is Monte Carlo simulation of the full integral. While the dimension of the integral T does not affect the asymptotic properties of the simulation approximation as the number of replications R rises, the accuracy of the approximation worsens as T rises with a given R. Lee (1997) provides Monte Carlo evidence on this effect for panel probit models.

For applications with a large time series dimension, various attempts have been made to break up the T-dimensional integral into T one-dimensional integrals to circumvent this problem. Doucet, De Freitas, and Gordon (2001) provide a discussion of the basics and various topics of such sequential Monte Carlo methods. These approaches can be seen as generalizations of the Kalman filter, which is appropriate for linear models with normally distributed error processes. For a discussion of these approaches to econometric time series models and panel data models with large time-series dimension see for example Fernández-Villaverde and Rubio-Ramírez (2004), Tanizaki and Mariano (1994), and Zhang and Lee (2004).

In the typical panel data used in microeconometric applications, the time dimension is small enough for joint Monte Carlo simulation. However, the computational cost of this approach can be high or even prohibitive if R is chosen such that the approximation is reasonably precise. It may therefore pay off to follow a similar approach for these applications. I present methods discussed in the literature and suggest a method that is based on sequential Gaussian quadrature of T one-dimensional integrals in the case of univariate error processes. This approach is straightforward to implement and allows similarly powerful numerical integration as in random effects models such as in Butler and Moffit (1982). In an application to an ordered logit model of health with an AR(1) error process, I show that this method clearly outperforms other methods and requires as little as 10 to 20 function evaluations to achieve a better precision than full Monte Carlo simulation with 2000 antithetic random draws.

The chapter is organized as follows. Section 4.2 discusses the class of models for which the approaches are appropriate. Section 4.3 presents the problem of and known solutions to the approximation of the likelihood function and introduces the method of sequential Gaussian quadrature. Section 4.4 presents an application and compares the performance of different algorithms. Section 4.5 concludes and the appendix section 4.6 presents proofs for the results used in section 4.3.

4.2 Model Specification

The class of panel data models discussed in this chapter is relatively rich. Suppose a sequence of dependent variables is observed over time for a number N of cross-sectional units such as individuals, households, or firms. The 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.

The vectors $\mathbf{Y}_{it}$ for $i = 1, ..., N$ and $t = 1, \ldots, T$ contain the dependent random variables for the corresponding wave. In many applications, it is 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. Let $\mathbf{y}_{it}$ denote the observed outcomes that represent realizations of $\mathbf{Y}_{it}$. It is modeled as a function of unknown parameters $\boldsymbol{\theta}$, a vector of exogenous variables $\mathbf{x}_i$ and two random error vectors $\mathbf{a}_{it}$ and $\mathbf{e}_{it}$. The vectors of exogenous variables are allowed to vary over time. In this case, $\mathbf{x}_i$ collects all time-specific values. The measurement model discussed below specifies which part of $\mathbf{x}_i$ affects the dependent variables at which point in time. The random errors $\mathbf{e}_{it}$ are assumed to be independent over time and might represent measurement errors or contemporaneous shocks. The D-dimensional vector $\mathbf{a}_{it}$ denotes unobserved variables that are allowed to be dependent over time ("states"). The measurement model specifies the data generating process *given* values of the unobserved variables:

$$\mathbf{y}_{it} = g(\mathbf{x}_i, \mathbf{a}_{it}, \mathbf{e}_{it}, \boldsymbol{\theta}) \tag{4.1}$$

To simplify the presentation, let the random variables $\mathbf{Y}_{it}$ be discrete. In the discrete case, the conditional outcome probabilities or the probability mass function of $\mathbf{Y}_{it}$ at $\mathbf{y}_{it}$ are of interest. The discussion can be directly translated for continuous or mixed distributions, in which case the corresponding probability densities are relevant. Define the conditional outcome probabilities

$$\begin{aligned} P_{it}(\mathbf{y}_{it}) &= \Pr\left(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{x}_i, \boldsymbol{\theta}\right) \\ \text{and } P_{it}(\mathbf{y}_{it}|\mathbf{a}) &= \Pr\left(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{a}_{it} = \mathbf{a}, \mathbf{x}_i, \boldsymbol{\theta}\right) \end{aligned} \tag{4.2}$$

Assumption 1 (Measurement Model) *The conditional outcome probabilities $P_{it}(\mathbf{y}_{it}|\mathbf{a})$ are smooth functions of $\mathbf{a}$ which are known up to a finite set of parameters.*

In many microeconometric applications, $g(\mathbf{x}_i, \mathbf{a}_{it}, \mathbf{e}_t, \boldsymbol{\theta})$ is a known parametric function and $\mathbf{e}_t$ can be integrated out. In these cases, assumption 1 follows directly. Examples are panel binary or ordered discrete choice models and multinomial or nested logit models with $\mathbf{x}_i$ and $\mathbf{a}_{it}$ as explanatory variables. In other cases such as the multinomial probit model, $P_{it}(\mathbf{y}_{it}|\mathbf{a})$ might have to be computed using numerical approximation techniques.

Assumption 2 (Conditional Independence) *Conditional on $\mathbf{x}_i$ and $\mathbf{a}_{it}$, the outcome probabilities are independent of all other outcomes and states:*

$$\Pr\left(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{a}_{it} = \mathbf{a}, \mathbf{a}_{is}, \mathbf{y}_{is}, \mathbf{x}_i, \boldsymbol{\theta}\right) = P_{it}(\mathbf{y}_{it}|\mathbf{a}) \quad \forall s = 1, ..., T \neq t.$$

For the distribution of the latent states, make the following assumptions.

Assumption 3 (Marginal distribution) *The states $\mathbf{a}_{it}$ conditional on $\mathbf{x}_i$ and $\boldsymbol{\theta}$ are identically distributed with a known parametric p.d.f. $f(\mathbf{a})$.*

Assumption 4 (Transition) *Conditional on $\mathbf{a}_{i,t-1}$, the states $\mathbf{a}_{it}$ are identically distributed with a known parametric p.d.f. $f_c(\mathbf{a}|\mathbf{a}_{i,t-1})$. The states are first-order Markov such that the distribution of $\mathbf{a}_{it}$ conditional on $\mathbf{a}_{i,t-1}$ is independent of $\mathbf{a}_{i,t-s}$ with $s > 1$.*

To clarify these assumptions, consider the following example which corresponds to the example application presented in section 4.4. It is an ordered logit model with a stationary AR(1) state process and independent error terms. Let $y_{it} \in \{1, ..., J\}$ denote the realization of a random variable with a finite set of ordered outcomes.

$$\begin{array}{rrcl} \text{Measurement:} & y_{it}^* & = & \mathbf{x}_{it}\boldsymbol{\beta} + a_{it} + e_{it} \\ & y_{it} & = & y \Leftrightarrow \alpha_y \leq y_{it}^* < \alpha_{y+1} \\ \text{States:} & a_{i1} & \sim & \text{i.i.d. } \mathcal{N}\left(0, \sigma^2\right), \quad a_{it} = \rho a_{i,t-1} + u_{it} \\ & e_{it} & \sim & \text{i.i.d. logistic,} \quad u_{it} \sim \text{i.i.d. } \mathcal{N}\left(0, (1-\rho^2)\sigma^2\right), \end{array}$$

where $|\rho| < 1$. In the notation introduced above, the ordered logit specification implies

$$P_{it}(y_{it}|a_{it}) = \Lambda\left(\alpha_{y_{it}+1} - \mathbf{x}_{it}\boldsymbol{\beta} - a_{it}\right) - \Lambda\left(\alpha_{y_{it}} - \mathbf{x}_{it}\boldsymbol{\beta} - a_{it}\right),$$

where Λ is the logistic c.d.f., $\alpha_1 = -\infty$, $\alpha_{J+1} = \infty$, and α_2 through α_J are unknown parameters. This satisfies assumptions 1 and 2. The states a_{it} are specified as a one-dimensional normally distributed AR(1) process with

$$f(a) = \frac{1}{\sigma}\phi\left(\frac{a}{\sigma}\right) \quad \text{and} \quad f_c(a|a') = \frac{1}{\sqrt{1-\rho^2}\sigma}\phi\left(\frac{a-\rho a'}{\sqrt{1-\rho^2}\sigma}\right),$$

where ϕ is the standard normal p.d.f. This state process clearly satisfies assumptions 3 and 4. The vector of model parameters in this application is $[\boldsymbol{\beta}, \sigma, \rho, \alpha_2, ..., \alpha_J]$.

4.3 Evaluation of the Likelihood Contributions

For the estimation of the model parameters with maximum likelihood or Bayesian analyses, the likelihood functions have to be evaluated. For other methods such as GMM, similar statistics are needed. By independence over cross-sectional units, the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \ell_i, \tag{4.3}$$

where the likelihood contributions ℓ_i are the joint outcome probabilities of individual i. Let $\mathbf{Y}_{i,1:t} = [\mathbf{Y}_{i1}, \dots, \mathbf{Y}_{it}]$ and $\mathbf{a}_{i,1:t} = [\mathbf{a}_{i1}, \dots, \mathbf{a}_{it}]$ denote the vectors of the corresponding sequences up to wave t. Furthermore let $f_{1:t}(\mathbf{a}_{1:t})$ be the joint p.d.f. of $\mathbf{a}_{i,1:t}$. With this notation, the likelihood contributions can be written as the joint conditional outcome probabilities

$$\ell_i = \Pr(\mathbf{Y}_{i,1:T} = \mathbf{y}_{i,1:T}|\mathbf{x}_i, \boldsymbol{\theta}). \tag{4.4}$$

Because of the presence of the latent process $\mathbf{a}_{i,1:T}$ in the conditional outcome probabilities of the measurement model in assumption 1, this expression can in general not be evaluated directly. Instead, it can be approximated numerically as will be discussed in the remainder of this section.

4.3.1 Joint Integration

Proposition 1 *For the class of models described in section 4.2, the likelihood contribution in equation 4.4 can be rewritten as*

$$\ell_i = \int \cdots \int \left(\prod_{t=1}^{T} P_{it}(\mathbf{y}_{it}|\mathbf{a}_t) \right) f_{1:T}(\mathbf{a}_{1:T})\, d\mathbf{a}_1 \cdots d\mathbf{a}_T. \tag{4.5}$$

For a proof, see section 4.6. Note that if the state space in each wave $\mathbf{a}_{it}$ is D-dimensional, the total dimension of the integral is DT. I discuss several approaches to evaluate this multidimensional integral numerically. The goal is to achieve a high accuracy of the numerical approximation with as little computational cost as possible.

The usual approach in econometrics to approximate multidimensional integrals is simulation. By assumption 2, the integrand in equation 4.5 can be evaluated for all $\mathbf{a}_{1:T}$. If draws from $f_{1:T}(\mathbf{a}_{1:T})$ can be obtained, Monte Carlo integration of the full integral is feasible. In the example presented above with an AR(1) state process, draws from this joint distribution can easily be generated. For each $i = 1, ..., N$ and $r = 1, ..., R$, first draw a random number a_{i1}^r from $f(a)$. Then sequentially for $t = 2, ..., T$, draw a random number a_{it}^r from $f_c(a|a_{t-1}^r)$. The resulting vector $a_{i,1:T}^r = [a_{i1}^r, ..., a_{iT}^r]$ is a draw from $f_{1:T}([a_1, ..., a_T])$. This allows the "brute force" simulation of the DT-dimensional integral in equation 4.5.

Algorithm 1: Simulation of the joint probability (JMC)

```
1. Start with $i = 1$
2. Draw a number of sequences $\mathbf{a}_{i,1:T}^r$ with $r = 1, ..., R$ from the joint
   distribution $f_{1:T}(\mathbf{a}_{1:T})$
3. For each $r = 1, ..., R$ and $t = 1, ..., T$, calculate $P_{it}^r = P_{it}(\mathbf{y}_{it}|\mathbf{a}_{it}^r)$
4. For each $r = 1, ..., R$, calculate $P_{i,1:T}^r = \prod_{t=1}^{T} P_{it}^r$.
5. Calculate the simulated likelihood contribution as $\tilde{\ell}_i = 1/R \sum_{r=1}^{R} P_{i,1:T}^r$.
6. Repeat steps 2 through 5 for all $i = 2, ..., N$
```

Under weak regularity conditions, $\tilde{\ell}_i$ is $\sqrt{R}$-consistent by a law of large numbers. The joint simulation approach can also be replaced by deterministic multivariate integration. By a change of variables, the integral in equation 4.5 would have to be reformulated by a change of variables to conform to a standard setting for which deterministic integration rules are available. Usually, the dimension of the integral is too high for standard product rule integration, but as shown in chapter 5 and Heiss and Winschel (2005), Smolyak cubature can be a very powerful alternative to simulation.

4.3.2 Sequential Integration: General Approach

Each numerical integration method suffers from high dimensionality. Smolyak cubature dramatically reduces the "curse of dimensionality" of product rule quadrature methods. For a given degree of approximation, the computational costs do not rise exponentially with the dimension of the integral. But they still rise significantly, see chapter 5. The same holds for simulation methods. While by a law of large numbers the convergence rate is independent of the dimension, the approximation error for given finite number of random draws can increase substantially with the number of dimensions since the sampling space may be poorly covered. Doucet and de Freitas (2001) discusses this for nonlinear time series models and Lee (1997) provides Monte Carlo evidence on this problem for the GHK simulator for panel probit models.

With a D-dimensional state space in each point in time, the integral in equation 4.5 has DT dimensions which can quickly become very high for numerical integration and rises with T. In the following, I discuss methods to separate the DT-dimensional integral into T separate D-dimensional integrals which are much easier to approximate. These approaches can be interpreted as a generalization of the Kalman filter to nonlinear models with possibly nonnormal disturbances.

By the rules of conditioning, the likelihood contribution can be decomposed into T factors. For $s \leq t$, define $P_{it|1:s}(\mathbf{y}_{it})$ to be the outcome probability of $\mathbf{y}_{it}$ conditional on the past sequence of observed outcomes up to wave s:

$$P_{it|1:s}(\mathbf{y}_{it}) = \begin{cases} \Pr\left(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{x}_i, \boldsymbol{\theta}\right) = P_{it}(\mathbf{y}_{it}) & \text{if } t = 1 \\ \Pr\left(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{Y}_{i,1:s} = \mathbf{y}_{i,1:s}, \mathbf{x}_i, \boldsymbol{\theta}\right) & \text{if } 2 \leq t \leq T \end{cases}$$

Equivalently, $f_{it|1:s}(\mathbf{a}) = f(\mathbf{a})$ if $t = 1$ and denotes the p.d.f. of $\mathbf{a}_{it}$ conditional on $\mathbf{Y}_{i,1:s} = \mathbf{y}_{i,1:s}$.

With this notation, the likelihood contribution can be written by standard rules of conditioning as

$$\ell_i = \prod_{t=1}^{T} P_{it|1:t-1}(\mathbf{y}_{it}). \tag{4.6}$$

In the following, each of these factors will be approximated separately by different methods. Let $\tilde{P}_{it}$ denote an approximation of $P_{it|1:t-1}(\mathbf{y}_{it})$. Then the approximated likelihood contribution is

$$\tilde{\ell}_i = \prod_{t=1}^{T} \tilde{P}_{it}. \tag{4.7}$$

The problem with with expression is that these factors are not straightforward to evaluate. The presence of the unobserved state process makes the outcome probabilities conditional on the sequence of past outcomes an involved expression. The model structure discussed in section 4.2 allows to write the conditional outcome probabilities as D-dimensional integrals:

Proposition 2 *The conditional outcome probabilities can be written as*

$$P_{it|1:t-1}(\mathbf{y}_{it}) \quad = \quad \int P_{it}(\mathbf{y}_{it}|\mathbf{a}) f_{it|1:t-1}(\mathbf{a})\, d\mathbf{a} \tag{4.8}$$

For a proof, see section 4.6. In this formulation, the information on the past outcomes is captured by the conditional state distribution $f_{it|1:t-1}(\mathbf{a})$. Each past outcome contains information about the state at that time and thereby about the current state. The conditional outcome probabilities are the expectations of $P_{it}(\mathbf{y}_{it}|\mathbf{a})$ with respect to this conditional distribution.

The integrand $P_{it}(\mathbf{y}_{it}|\mathbf{a})$ is a known function by assumption 1. Unfortunately, the conditional state distribution characterized by the p.d.f. $f_{it|1:t-1}(\mathbf{a})$ is itself nontrivial. By assumption 3 the marginal p.d.f. $f(\mathbf{a})$ is known, say it is the normal p.d.f. But conditioning on the past outcomes changes the whole shape of the distribution.

There are various attempts to solve this problem. Most are developed in the statistics literature and engineering, where various problems have a similar structure.[1] These methods also generate increasing attention in the econometric time series literature, see Fernández-Villaverde and Rubio-Ramírez (2004). The methods differ in the accuracy given a number R of evaluations of the integrand $P_{it}(\mathbf{y}_{it}|\mathbf{a})$, additional computational costs, complexity of implementation, and other properties like the smoothness of the approximation with respect to the model parameters. Furthermore, their relative performance depends on the underlying problem with the time-series dimension of the data as a critical factor.

In the following, different approaches known in the time series literature are briefly discussed. For a more extensive overview, see Tanizaki (2003) and Haug (2005). One class of algorithms is based on resampling techniques as discussed in

[1] One of the major applications in engineering is the automatic target recognition for intelligent weapons, e.g. Salmond and Gordon (2001). This is rocket science.

section 4.3.3. Section 4.3.4 discusses other approaches that rely on importance sampling. Finally, section 4.3.5 discusses a third class of methods that use deterministic integration rules and introduces an algorithm based on Gaussian quadrature or Smolyak cubature.

4.3.3 Resampling Techniques: Particle Filters

An obvious numerical approach to the integration problem in equation 4.8 is to make draws from $f_{it|1:t-1}(\mathbf{a})$ and perform standard Monte Carlo simulation. The idea of the nonlinear particle filter to obtain such draws was suggested by Gordon, Salmond, and Smith (1993) and adopted to econometric time series models by Fernández-Villaverde and Rubio-Ramírez (2004). The main idea is sequential resampling in the bootstrap spirit.

The simulation of $\tilde{P}_{i1}$ for the initial wave is straightforward since the relevant state distribution is the marginal distribution $f(\mathbf{a})$. For each $r = 1, ..., R$ draw a value $\mathbf{a}_{i1}^r$ from this distribution and calculate $P_{i1}^r = P_{i1}(\mathbf{y}_{i1}|\mathbf{a}_{i1}^r)$. The simulated outcome probability is $\tilde{P}_{i1} = 1/R \sum_{r=1}^{R} P_{i1}^r$. For the second wave, draws from $f_{i2|1}(\mathbf{a})$ are required by equation 4.8. These can be obtained in two steps. Step 1 is to obtain draws $\mathbf{a}_{i1}^{*r}$ from $f_{i1|1}(\mathbf{a})$. Given those, step 2 is to obtain draws $\mathbf{a}_{i2}^r$ from $f_{i2|1}(\mathbf{a})$.

Step 1 can be implemented by importance resampling. We know that the nodes $\mathbf{a}_{i1}^r$ are samples from $f(\mathbf{a})$. From this set of nodes, R values $\mathbf{a}_{i1}^{*r}$ are drawn with replacement, where for each $r, s = 1, ..., R$ the node $\mathbf{a}_{i1}^{*r} = \mathbf{a}_{i1}^s$ with probability $q_{i1}^r = f_{i1|1}(\mathbf{a}_{i1}^s)/f(\mathbf{a}_{i1}^s)$. The result is a sample from $f_{i1|1}(\mathbf{a})$ (Gordon, Salmond, and Smith 1993). How the resampling weights q_{i1}^r can be calculated will be discussed below. The nodes $\mathbf{a}_{i1}^r$ which are most "compatible" with the observed $\mathbf{y}_{it}$ are resampled more frequently than unlikely nodes. Given these draws $\mathbf{a}_{i1}^{*r}$, for each $r = 1, ..., R$ obtain a draw $\mathbf{a}_{i2}^r$ from the transition density $f_c(\mathbf{a}|\mathbf{a}_{i1}^{*r})$. The result is a sample from $f_{i2|1}(\mathbf{a})$.

The remaining probabilities can be equivalently evaluated in a sequential fashion. Given draws $\mathbf{a}_{it}^r$ from $f_{it|1:t-1}(\mathbf{a})$, the probabilities $P_{it}^r = P_{it}(\mathbf{y}_{it}|\mathbf{a}_{it}^r)$ and the simulated outcome probabilities $\tilde{P}_{it} = 1/R \sum_{r=1}^{R} P_{it}^r$ can easily be evaluated. To prepare the calculations for the next wave, draws $\mathbf{a}_{it}^{*r}$ from $f_{it|1:t}(\mathbf{a})$ are obtained by resampling with the resampling probabilities $q_{it}^r = f_{it|1:t}(\mathbf{a}_{it}^r)/f_{it|1:t-1}(\mathbf{a}_{it}^r)$. Then draw $\mathbf{a}_{i,t+1}^r$ from the transition density $f_c(\mathbf{a}|\mathbf{a}_{it}^{*r})$ to obtain a sample from $f_{i,t+1|1:t}(\mathbf{a})$. With these, the sequential procedure is repeated for $t + 1$ and so on. Proposition 3 provides the key to the evaluation of the resampling probabilities.

Proposition 3 *The structure of the model discussed in section 4.2 allows to evaluate the appropriate resampling weights for the nonlinear particle filter as*

$$q_{it}^r := \frac{f_{it|1:t}(\mathbf{a}_{it}^r)}{f_{it|1:t-1}(\mathbf{a}_{it}^r)} = \frac{P_{it}(\mathbf{y}_{it}|\mathbf{a}_{it}^r)}{P_{it|1:t-1}(\mathbf{y}_{it})}. \tag{4.9}$$

A proof of this proposition can be found in section 4.6. It is based on Bayes rule. The numerator and an approximation of the denominator of this expression have been readily calculated in the sequential algorithm and can be reused. The particle filter algorithm can be summarized as follows:

Algorithm 2: Nonlinear Particle Filter (NPF)

1. Start with $i = 1$
2. Start with $t = 1$ and draw R random numbers $\mathbf{a}_{i1}^r$ from the density $f(\mathbf{a})$.
3. Calculate $P_{it}^r = P_{it}(\mathbf{y}_{it}|\mathbf{a}_{it}^r)$ for all $r = 1, ..., R$ and approximate $P_{it|1:t-1}(\mathbf{y}_{it})$ as $\tilde{P}_{it} = \sum_{r=1}^R P_{it}^r$.
4. Draw R values $\mathbf{a}_{i,t+1}^{*r}$ from the set $[\mathbf{a}_{it}^1, ..., \mathbf{a}_{it}^R]$ with replacement, where the r^{th} element is drawn with probability $P_{it}^r/\tilde{P}_{it}$. This yields draws from $f_{it|1:t}(\mathbf{a})$.
5. For each r, one random number $\mathbf{a}_{i,t+1}^r$ is drawn from the density $f_c(\mathbf{a}|\mathbf{a}_{it}^{*r})$. This yields draws from $f_{i,t+1|1:t}(\mathbf{a})$.
6. Repeat steps 3 through 5 for all $t = 2, ..., T$
7. The individual likelihood contribution is simulated as $\tilde{\ell}_i = \prod_{t=1}^T \tilde{P}_{it}$.
8. Repeat steps 2 through 7 for all $i = 2, ..., N$

This algorithm is attractive since it is intuitive and easily implemented. Since the new information contained in each observation is captured sequentially in the nodes, the method works very well also for long time series. On the other hand, the resampling creates problems. First, it is computationally burdensome and the computational costs increase quickly with the number of draws R. Second, it introduces additional noise compared to the deterministic reweighting of algorithm 3. And third the simulated likelihood contributions are not smooth in the parameters. While small parameter changes do not affect the resampling, the simulated likelihood contribution jumps as soon as the parameter change is large enough to change the resampling. This impedes gradient-based maximization algorithms for the likelihood function (Fernández-Villaverde and Rubio-Ramírez 2004).

These problems can be expected to become less relevant with longer time series or more random draws R, since the resampling noise averages out. With the typical microeconomic panel data with moderate time dimension, a large number of draws is needed to reduce resampling noise and capture the shape of the conditional distributions. For some applications, the distributions $f_{it|1:t}(\mathbf{a}_{it})$ and $f_{it|1:t-1}(\mathbf{a}_{it})$ differ very much. This creates numerical problems since some values of q_{it}^r are close to zero whereas others are very high. Recent research suggests approaches to overcome these problems. Interested readers are referred to Pitt and Shephard (1999) and van der Merwe, Doucet, de Freitas, and Wan (2001).

4.3.4 Importance Sampling techniques

Instead of drawing random numbers from the target distribution, importance sampling allows to draw from a different distribution, the proposal density, and to capture the differences by reweighting the integrand. The most straightforward choice of the proposal density is the marginal density $f(\mathbf{a})$. Note that equation 4.8 can be rewritten as

$$P_{it|1:t-1}(\mathbf{y}_{it}) = \int P_{it}(\mathbf{y}_{it}|\mathbf{a}) \underbrace{\frac{f_{it|1:t-1}(\mathbf{a})}{f(\mathbf{a})}}_{:=q_{it}(\mathbf{a})} f(\mathbf{a})\, d\mathbf{a}. \tag{4.10}$$

It has already been discussed that it is possible to make draws from the proposal distribution $f(\mathbf{a})$. As long as the factor $q_{it}(\mathbf{a})$ can be evaluated, the integral in equation 4.10 can easily be approximated by weighted simulation. Two algorithms to obtain appropriate weights $q_{it}(\mathbf{a})$ are discussed in the following.

The first importance sampling approach discussed is based on draws from the joint distribution $f_{1:T}(\mathbf{a}_{1:T})$. As will be shown below, it actually corresponds to algorithm 1. But it is instructive to present it separately to clarify the sequential approach and the differences to the subsequently discussed algorithms.

Proposition 4 *The importance sampling factor $q_{it}(\mathbf{a})$ can be written as*

$$q_{it}(\mathbf{a}) \;=\; \int \cdots \int \left(\prod_{s=1}^{t-1} \frac{P_{is}(\mathbf{y}_{is}|\mathbf{a}_{is})}{P_{is|1:s-1}(\mathbf{y}_{is})} \right) f_{1:t-1}(\mathbf{a}_{i,1:t-1}|\mathbf{a}_{it} = \mathbf{a})\, d\mathbf{a}_{i,1:t-1}$$

A proof of this proposition is presented in section 4.6. By combining this results with equation 4.10,

$$P_{it|1:t-1}(\mathbf{y}_{it}) = \int \cdots \int P_{it}(\mathbf{y}_{it}|\mathbf{a}_{it}) \left(\prod_{s=1}^{t-1} \frac{P_{is}(\mathbf{y}_{is}|\mathbf{a}_{is})}{P_{is|1:s-1}(\mathbf{y}_{is})} \right) f_{1:t}(\mathbf{a}_{i,1:t})\, d\mathbf{a}_{i,1:t}. \quad (4.11)$$

Algorithm 3 describes one way to sequentially generate appropriate weights based on this result:

Algorithm 3: Importance Sampling with Joint Draws (JSS)

```
1. Start with $i = 1$
2. Draw $R$ random vectors $\mathbf{a}^r_{i,1:T}$ from the density $f_{1:T}(\mathbf{a}_{1:T})$.
3. Start with $t = 1$ and initialize $q^r_{i1} = 1$ for all $r = 1, ..., R$.
4. Calculate $P^r_{it} = P_{it}(\mathbf{y}_{it}|\mathbf{a}^r_{it})$ for all $r = 1, ..., R$ and approximate
   $P_{it|1:t-1}(\mathbf{y}_{it})$ as $\tilde{P}_{it} = \sum_{r=1}^{R} P^r_{it} q^r_{it}$.
5. Update $q^r_{i,t+1} = q^r_t \frac{P^r_{it}}{\tilde{P}_{it}}$.
6. Repeat steps 4 and 5 for all $t = 2, ..., T$
7. The individual likelihood contribution is simulated as $\tilde{\ell}_i = \prod_{t=1}^{T} \tilde{P}_{it}$.
8. Repeat steps 2 through 7 for all $i = 2, ..., N$
```

Proposition 5 *Algorithm 3 is equivalent to Algorithm 1 in the sense that they deliver the same simulated likelihood contributions if the same random draws are used.*

This can be shown by first noting that both algorithms are actually based on draws from the joint distribution $f_{1:T}(\mathbf{a}_{1:T})$. As a result, the conditional probabilities P^r_{it} are the same if the same random draws are used since they evaluate the same function at the same arguments. The sequentially updated weights in algorithm 3 can be written as

$$q^r_{it} = q^r_{t-1} \frac{P^r_{i,t-1}}{\tilde{P}_{i,t-1}} = \prod_{s=1}^{t-1} \frac{P^r_{is}}{\tilde{P}_{is}}.$$

Therefore, the simulated outcome probability for each wave t can be written as

$$\tilde{P}_{it} = \frac{1}{R}\sum_{r=1}^{R} q_{it}^{r} P_{it}^{r} = \frac{1}{R}\frac{\sum_{r=1}^{R}\prod_{s=1}^{t} P_{is}^{r}}{\prod_{s=1}^{t-1}\tilde{P}_{is}}.$$

So $\prod_{s=1}^{t}\tilde{P}_{is} = \frac{1}{R}\sum_{r=1}^{R}\prod_{s=1}^{t} P_{is}^{r}$. The simulated likelihood contribution is

$$\tilde{\ell}_i = \prod_{t=1}^{T}\tilde{P}_{it} = \frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T} P_{it}^{r},$$

which is equal to the corresponding expression in algorithm 1.

A direct consequence of this proposition is that nothing is gained by this procedure relative to algorithm 1. The reason is that this algorithm is still based on draws from the joint distribution $f_{1:T}(\mathbf{a}_{1:T})$. If the dimension of $\mathbf{a}_{1:T}$ is high, a finite number of draws R covers the support of its joint distribution only coarsely.

For pure time-series models, Tanizaki and Mariano (1994) and Tanizaki (1999) suggest an algorithm that can be interpreted as a refinement of algorithm 3. The latter approach was based on draws from the joint distribution of $\mathbf{a}_{i,1:T}$. The Tanizaki (1999) algorithm adopted to panel data draws values from the marginal distribution of each $\mathbf{a}_{it}$ for all i and t using antithetic samples. The full set of $R \times R$ transition probabilities between the states in adjacent waves are considered. Just as algorithm 3, start from the importance sampling equation 4.10. The weighting factors $q_{it}(\mathbf{a})$ are by definition equal to 1 for $t = 1$. The next proposition offers a way to write them in a sequential fashion.

Proposition 6 *For $t \geq 2$, the importance sampling weights can be written recursively as*

$$q_{i,t+1}(\mathbf{a}_{i,t+1}) = \int q_{it}(\mathbf{a})\frac{P_{it}(\mathbf{y}_{it}|\mathbf{a})}{P_{it|1:t-1}(\mathbf{y}_{it})}\frac{f_c(\mathbf{a}_{i,t+1}|\mathbf{a})}{f(\mathbf{a}_{i,t+1})}f(\mathbf{a})\,d\mathbf{a}. \tag{4.12}$$

A prof is presented in section 4.6. The idea of this algorithm is to simulate the integrals in both equations 4.10 and 4.12. Given draws $[\mathbf{a}_{it}^{1}, ..., \mathbf{a}_{it}^{R}]$ and $[\mathbf{a}_{i,t+1}^{1}, ..., \mathbf{a}_{i,t+1}^{R}]$ from $f(\mathbf{a})$ and corresponding previous weights $[q_{it}^{1}, ..., q_{it}^{R}]$, the appropriate weights $[q_{i,t+1}^{1}, ..., q_{i,t+1}^{R}]$ can be approximated as

$$q_{i,t+1}^{r} = \frac{1}{R}\sum_{s=1}^{R} q_{it}^{s}\frac{P_{it}(\mathbf{y}_{it}|\mathbf{a}_{it}^{s})}{P_{it|1:t-1}(\mathbf{y}_{it})}\frac{f_c(\mathbf{a}_{i,t+1}^{r}|\mathbf{a}_{it}^{s})}{f(\mathbf{a}_{i,t+1}^{r})}. \tag{4.13}$$

Algorithm 4: Sequential importance sampling (SIS)

```
1. Start with $i = 1$
2. Start with $t = 1$ and initialize $q_{i1}^r = 1$ for all $r = 1, ..., R$
3. Draw random numbers $a_{it}^r$ from the density $f(\mathbf{a})$.
4. Outcome probability: Calculate $P_{it}^r = P_{it}(\mathbf{y}_{it}|\mathbf{a}_{it}^r)$ for all $r = 1, ..., R$ and approximate $P_{it|1:t-1}(\mathbf{y}_{it})$ as $\tilde{P}_{it} = 1/R \sum_{r=1}^R P_{it}^r q_{it}^r$.
5. For all $r, s = 1, ..., R$, calculate $c_{it}^*(r, s) = \frac{f_c(\mathbf{a}_{i,t+1}^r|\mathbf{a}_{it}^s)}{f(\mathbf{a}_{i,t+1}^r)}$.
6. For all $s = 1, ..., R$, normalize $c_{it}(r, s) = c_{it}^*(r, s) R \sum_{r=1}^R c_{it}^*(r, s)$
7. For all $r = 1, ..., R$, update the weights $q_{i,t+1}^r = \sum_{s=1}^R q_{it}^s \frac{P_{it}^s}{\tilde{P}_{it}} c_{it}(r, s)$.
8. Repeat steps 3 through 7 for all $t = 2, ..., T$
9. The individual likelihood contribution is simulated as $\tilde{\ell}_i = \prod_{t=1}^T \tilde{P}_{it}$.
10. Repeat steps 1 through 9 for all $i = 2, ..., N$
```

Step 6 corrects for the fact that the approximated densities do not necessarily integrate to 1 exactly which can lead to an accumulation of these errors (Tanizaki 1999). For a given number of nodes R, this algorithm is computationally more expensive than algorithm 3. The main source of additional computational costs is that in step 5, R^2 relative densities have to be computed for each i and t. There are approaches to ease the computational burden of this approach, for example Tanizaki (2001) suggests to draw random numbers from the joint density of $\mathbf{a}_{it}$ and $\mathbf{a}_{i,t+1}$ conditional on $\mathbf{y}_{it}$ instead of integrating over $\mathbf{a}_{it}$.

Importance sampling works best numerically if the proposal and the target densities are similar so that the importance weights are not too far away from unity (Geweke 1989). This problem is similar to the mentioned problems with resampling when the distributions $f_{it|1:t}(\mathbf{a})$ and $f_{it|1:t-1}(\mathbf{a})$ differ "too much". The corresponding problem with the sequential importance sampling approach with $f(\mathbf{a})$ as the proposal density is worse since the importance weights capture the differences between $f_{it|1:t}(\mathbf{a})$ and $f(\mathbf{a})$. Especially with long time series, the accumulated information of $\mathbf{y}_{i,1:t}$ can drive these densities far apart. This creates poor approximations of the importance sampling approach. There have been approaches to solve this problem by choosing a proposal density that is closer to $f_{it|1:t}(\mathbf{a})$ than $f(\mathbf{a})$, see for example Julier and Uhlmann (1997).

For the moderate time series dimension typical for microeconometric panel data models, this problem can be expected to be less severe than for long time series. Since the modification of the proposal density to account for previous observations involves significant extra computations and complicates the implementation considerably, I stick to the simple version of using $f(\mathbf{a})$ as the proposal density.

4.3.5 Sequential Quadrature

Instead of simulation, the integrals in equations 4.10 and 4.12 can be approximated using deterministic integration rules. Kitagawa (1987) suggests a linear spline approximation of the densities. This approach has been found to be difficult to implement and computationally intensive. Computational costs rise exponentially with the dimension of the state vector. This finding led to the predominant use of different simulation techniques in the literature.

I suggest to use quadrature based methods for the numerical integration. This has two advantages compared to simulation. First, the approximation can be expected to achieve a high accuracy with a very low number of nodes R and does not suffer from random noise. This is widely appreciated for one-dimensional Gaussian quadrature that is appropriate if $\mathbf{a}_{it}$ is scalar such as in the AR(1) example introduced in section 4.4 and applied in section 4.2. Chapter 5 and Heiss and Winschel (2005) show that Smolyak cubature provides a promising approach to carry over these advantages to higher dimensions.

The second advantage of deterministic integration is that the nodes $\mathbf{a}_{it}$ are actually identical since the integral is defined over the marginal distributions and they are assumed to be identical. This allows to do the computationally expensive step of calculating the R^2 conditional distributions $f_c(\mathbf{a}^r_{i,t+1}|\mathbf{a}^s_{it})$ for all r and s only once instead of $N(T-1)$ times for each $i = 1, ..., N$ and $t = 2, ..., T$. This advantage is especially relevant for panel data with a large cross-sectional dimension.

The ideas of algorithm 4 can be used equivalently to approximate the integrals in equations 4.10 and 4.12. Instead of drawing R random numbers $\mathbf{a}^r_{it}$ from $f(\mathbf{a})$ for all $i = 1, ..., N$ and $t = 1, ..., T$, appropriate nodes $\mathbf{a}^r$ and weights w^r for a Gaussian quadrature or Smolyak cubature rule are obtained. This is efficiently implemented in many statistical packages for the univariate case so that it is both straightforward to implement and quick to calculate in this case.[2] The relative

[2] The implementation used in section 4.4 is based on Matlab code for Gaussian quadrature provided with the textbook of Miranda and Fackler (2002). It is available at `http://www4.ncsu.edu/~pfackler/compecon`.

densities $c(r,s) = \frac{f_c(\mathbf{a}^r|\mathbf{a}^s)}{f(\mathbf{a}^r)}$ can be calculated once for each $r,s = 1,...,R$. Given those and the initialization $q^r_{i1} = 1$, the problem can be solved sequentially.

Given q^r_{it}, calculate $P^r_{it} = P_{it}(\mathbf{y}_{it}|\mathbf{a}^r)$ for all $r = 1,...,R$ and $\tilde{P}_{it} = 1/R\sum_{r=1}^R P^r_{it}q^r_{it}w^r$. To update the weights for the next period, calculate $q^r_{i,t+1} = \sum_{s=1}^R q^s_t \frac{P^s_{it}}{\tilde{P}_{it}} c(r,s)w^r$ for all $r = 1,...,R$.

Algorithm 5: Sequential Quadrature (SGQ)

1. Obtain R appropriate nodes $\mathbf{a}^r$ and weights w^r for the deterministic integration rule corresponding to $f(\mathbf{a})$.
2. For all $r,s = 1,...,R$, calculate $c^*(r,s) = \frac{f_c(\mathbf{a}^r|\mathbf{a}^s)}{f(\mathbf{a}^r)}$
3. For all $s = 1,...,R$, normalize $c(r,s) = c^*(r,s)R\sum_{r=1}^R c^*(r,s)$
4. Start with $i = 1$
5. Start with $t = 1$ and initialize $q^r_{i1} = 1$ for all $r = 1,...,R$
6. Calculate $P^r_{it} = P_{it}(\mathbf{y}_{it}|\mathbf{a}^r)$ for all $r = 1,...,R$ and approximate $P_{it|1:t-1}(\mathbf{y}_{it})$ as $\tilde{P}_{it} = 1/R\sum_{r=1}^R P^r_{it}q^r_{it}w^r$.
7. For all $r = 1,...,R$, update the weights $q^r_{i,t+1} = \sum_{s=1}^R q^s_t \frac{P^s_{it}}{\tilde{P}_{it}} c(r,s)w^r$.
8. Repeat steps 6 and 7 for all $t = 2,...,T$
9. The individual likelihood contribution is simulated as $\tilde{\ell}_i = \prod_{t=1}^T \tilde{P}_{it}$.
10. Repeat steps 5 through 9 for all $i = 2,...,N$

Again, the normalization in step 3 corrects for potential approximation errors that accumulate over time as in Tanizaki (1999) and related methods. It is based on the fact that $\int f_c(\mathbf{a}_{i,t+1}|\mathbf{a}_{it})\, d\mathbf{a}_{i,t+1} = 1$ for all $\mathbf{a}_{it}$.

4.3.6 Summary of the Algorithms

The algorithms discussed in this chapter differ in the accuracy of approximation, the computational costs, and the simplicity of implementation. Obviously, JMC (algorithm 1) is the easiest to implement since the other algorithms all need additional calculations such as the evaluation of weights. On the other hand, most of these calculations are easily implemented as they mainly consist of evaluating fractions of readily available numbers and replacing the calculation of averages with weighted averages. NPF (algorithm 2) is an exception for which a resampling procedure has to be implemented. Computationally efficient resampling is conceptionally straightforward but a research area in its own right, see for example Bolić, Djurić, and Hong (2003).

Clearly, JMC is faster than SIS with a given number of function evaluations. For both, NTR random numbers have to be generated. SIS requires the numerical approximation of NT integrals for the likelihood contributions and $N(T-1)R$ integrals for the update of the importance weights. For the latter, R relative densities have to be computed for each integral. SGQ requires the solution of the same number of integrals, but the relative densities have to be computed only once so that the numerical integration basically boils down to a matrix multiplication.

SGQ has the additional advantage that only DR nodes and weights have to be determined, whereas JMC and SIS require $DNTR$ draws from $f(\mathbf{a})$. For large models, this can be computationally burdensome. If all those draws are to be stored in memory to save on recalculations, this can also quickly cause memory problems. In the application in section 4.4, $DTN \approx 100,000$. With $R = 1000$ and each random number requiring 8 bytes, the matrix of random draws would occupy roughly 800 megabytes of memory, whereas SGQ only needs to determine and store $2DR = 2000$ numbers for nodes and weights which would occupy only about 16 kilobytes.

The resampling used by NPF is computational burdensome even if sophisticated algorithms are used.[3] Another problem with NPF is that the approximated likelihood contributions are not smooth in the parameters. For maximum likelihood estimation, this rules out gradient-based numerical maximization of the likelihood function. While derivative-free likelihood maximization is possible, these algorithms require considerably more likelihood evaluations and thereby create substantial additional costs.

The main question for the comparison of the computational costs is how many function evaluations R are needed to achieve a given accuracy of approximation. As argued above, JMC and JSS integrate over DT dimensions whereas the integral dimension of the other approaches is only D. This suggests that JMC needs a higher R since the support of a higher-dimensional joint distribution has to be covered. For the other approaches based on importance sampling or resampling, numerical problems can arise if the proposal and the target densities are too different. For long time series, this problem is less severe for NPF than for SIS or SGQ, but for short and moderately long time series as typical for microeconometric analyses, this difference can be expected not to be severe. As

[3] In preliminary tests, more than 90% of the total run-time for NPF-based estimation was spend for resampling.

discussed above, these problems can be avoided if the proposal densities are chosen more carefully. But this would require substantial additional programming and computing costs.

SGQ is the only one of the discussed algorithms which is based on deterministic integration. For one-dimensional states $\mathbf{a}_{it}$ which are typical in microeconometric models, it is based on univariate Gaussian quadrature. With R function evaluations, integrals over polynomials of order $2R-1$ are evaluated exactly by these methods. Since the typical integrands are smooth functions, they can be expected to be well approximated by a relatively low-order polynomial. Note that with $R=10$, the order of polynomial exactness is $2\times 10-1=19$. For Monte Carlo integration, 10 random draws is an extremely small number to hope for sensible results. Since all these arguments are only qualitative and since no general quantitative statements are possible, section 4.4 provides empirical evidence for a typical microeconometric application.

4.4 Application: Ordered Logit Model of Health with an AR(1) Error Term

In this section, the performance of different algorithms is compared for a typical microeconometric panel data model. It combines limited dependent variables with a stationary AR(1) process in the error terms. The application was already discussed in chapter 3 together with alternative model specifications.

I use panel data from the Health and Retirement Study (HRS) to study the evolution of individual health over time. This survey is sponsored by the National Institute of Aging (NIA) and conducted by the University of Michigan. For our analyses I used the RAND HRS Data File (Version D). It was developed by the RAND Center for the Study of Aging with funding from the National Institute on Aging (NIA) and the Social Security Administration (SSA). 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. This applies to 25,499 respondents. After excluding respondents with missing information on essential variables, all analyses are based on a sample of 25,451 respondents with up to 6 observations over time each. A total of 103,250 observations are available.

In this paper, I concentrate on a frequently studied measure, the self-reported health (SRH). The wording of this question in the HRS is "Would you say your health is excellent, very good, good, fair, or poor?". This 5-scale variable is modeled as an ordered logit model with a stationary AR(1) error term. As discussed in chapter 3, this specification captures the data much better than random effects or markov chain models. I concentrate on SRH and ignore mortality in this analysis. The model structure was already presented in section 4.2 in a more general setting. The latent variable y^*_{it} represents the health status of respondent i at wave t and is modeled as

$$y^*_{it} \quad = \quad \mathbf{x}_{it}\boldsymbol{\beta} + a_{it} + e_{it}.$$

The observed dependent variable takes 5 different outcomes between 1 ("poor") and 5 ("excellent"). It is assumed to be generated as

$$y_{it} = y \Leftrightarrow \alpha_y \leq y^*_{it} < \alpha_{y+1} \quad \text{with } 1 \leq y \leq 5.$$

The stochastic specification is

$$\begin{aligned} a_{i1} &\sim \text{ i.i.d. } \mathcal{N}\left(0, \sigma^2\right), \quad a_{it} = \rho a_{i,t-1} + u_{it} \\ e_{it} &\sim \text{ i.i.d. logistic,} \quad u_{it} \sim \text{i.i.d. } \mathcal{N}\left(0, (1-\rho^2)\sigma^2\right) \end{aligned}$$

In the general notation, this implies

$$\begin{aligned} P_{it}(y_{it}|a_{it}) &= \Lambda\left(\alpha_{y_{it}+1} - \mathbf{x}_{it}\boldsymbol{\beta} - a_{it}\right) - \Lambda\left(\alpha_{y_{it}-\mathbf{x}_{it}\boldsymbol{\beta}-a_{it}}\right) \\ f(a) &= \frac{1}{\sigma}\phi\left(\frac{a}{\sigma}\right) \quad \text{and} \quad f_c(a|a') = \frac{1}{\sqrt{1-\rho^2}\sigma}\phi\left(\frac{a-\rho a'}{\sqrt{1-\rho^2}\sigma}\right), \end{aligned}$$

where Λ is the logistic c.d.f. and ϕ is the standard normal p.d.f. The vector of model parameters in this application is $[\boldsymbol{\beta}, \sigma, \rho, \alpha_2, ..., \alpha_5]$. For the exogenous variables, only age is considered specified as a linear spline with changing slopes at ages 50, 60, 70, 80, and 90. Obviously, the effect of covariates that are not included in the model and are correlated over time is captured by a_{it}.

I estimated this model with different algorithms and different number of nodes R. The algorithms are:

- JMC: Simulation of the joint outcome probability (algorithm 1) using the antithetic draws generated by the modified latin hypercube sampler (Hess, Train, and Polak 2005)[4]

- SIS: Sequential Monte Carlo integration (algorithm 4) using the antithetic draws generated by the modified latin hypercube sampler (Hess, Train, and Polak 2005). Different draws are used for each $i = 1, ..., N$, but are held constant for each $t = 1, ..., T$.

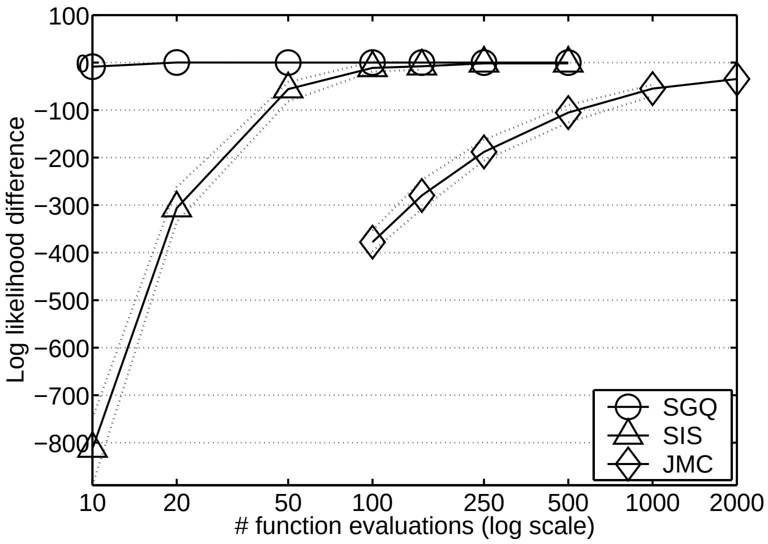
[4] Preliminary experiments indicate that the antithetic draws perform considerably better than draws from a standard random number generator.

- SGQ: Sequential Gauss quadrature (algorithm 5) based on a transformation of Gauss-Legendre quadrature of the conditional outcome probabilities and densities.

The main criterion of comparison between the algorithms is the accuracy of approximation given a number of function evaluations R or, equivalently, the number R needed for a given level of accuracy. As argued above, the difference between SIS and JMC is due to the different dimensionality of the integral. As long as the difference between the marginal and the conditional state distribution is not too large, SIS can be expected to perform better, since the dimension of integration is smaller. Otherwise, the importance sampling strategy of SIS can lead to numerical problems. The computational efficiency of Gaussian quadrature relative to Monte Carlo integration drives the difference between SGQ and SIS.

Figure 4.1 compares the results for the three approaches with different numbers of nodes R. It shows the differences between the approximated log likelihood at the respective estimated parameters to the value it converges to for all methods as $R \to \infty$. Since JMC and SIS are based on random draws, the results depend on the chosen random seed. For these algorithms, 10 estimates were obtained using different random seeds. The solid lines indicate the mean over these results and the dotted lines represent the minimum and maximum to provide a rough idea of the simulation noise.

Figure 4.1: Results: Log likelihood difference to limiting value



It is well known that the simulated log likelihood is biased downward with a finite R. This is due to the log transformation of the unbiased simulated outcome probabilities. The results show that this downward bias can be substantial for the simulation-based algorithms. All methods appear to converge to the same value with growing R. The proposed SGQ algorithm reaches this limiting value with only 10 or 20 evaluations, while the SIS algorithm needs about 150 evaluations to get close to it. The JMC approach typically used in empirical research converges very slowly. For R less than 100, the maximization algorithm did not converge at all. Even with 2000 evaluations, notable differences to the limiting value remain. Remember that Gaussian quadrature with $R = 20$ approximates the smooth conditional distributions as a 39$^{\text{th}}$ order polynomial, so it it no surprise that this strategy works very well.

Figures 4.2 and 4.3 show the estimates of the two most interesting parameters that drive the intertemporal correlation. The qualitative picture is the same as for the likelihood values. While with at most 20 evaluations, SGQ has reached its limiting value, SIS needs at least 150 and JMC has still not converged after R=2000. The simulation-based estimates of the correlation parameter ρ are biased upward toward 1. This might be due to the fact that the downward bias of the simulated log likelihood is stronger the higher the simulation noise. With $\rho = 1$, the model becomes a random effects model and the integration is in fact one-dimensional. The simulation of this one-dimensional integral is more accurate with a given R so the downward bias decreases. A more thorough investigation of this possible source of systematic bias is left for future research.

As discussed above, the computational costs for a given number R differs between the algorithms. To provide a complete comparison, Table 4.1 shows the time in seconds that the implemented algorithms need for each likelihood evaluation for the example application on a Pentium 4 PC. JMC is more expensive than the other algorithms with a low number R. The reason is the large number NTR of random numbers to be generated. Since for SIS the same values for the random numbers are used for each $t = 1, ..., T$, only NR numbers have to be generated. SGQ requires only R nodes. Another difference between SIS and SGQ is that the former needs to evaluate the R^2 relative densities for each $i = 1, ..., N$, whereas SGQ needs to do this only once.

As discussed above, the computational costs of JMC rise linearly with the number of draws R. This is due to the rising computational costs of each integral. For SIS and SGQ, the computational costs rise overproportionally, since in addition the number of integrals increase. As a result, for large R, JMC is the fastest algorithm. Noting that SGQ needs a very small number R for accurate results, the advantage of SGQ is even more pronounced. It does not only require much less function evaluations R, but it is also the fastest algorithm for small R.

Figure 4.2: Results: Estimated σ

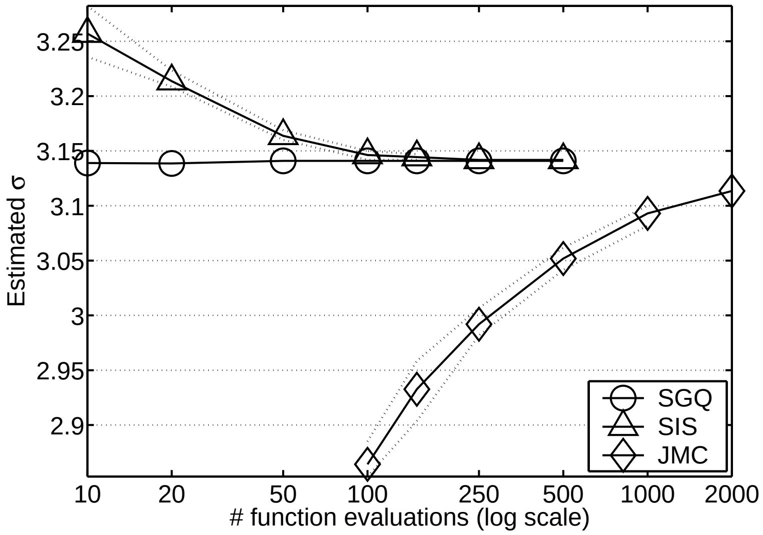


Figure 4.3: Results: Estimated ρ

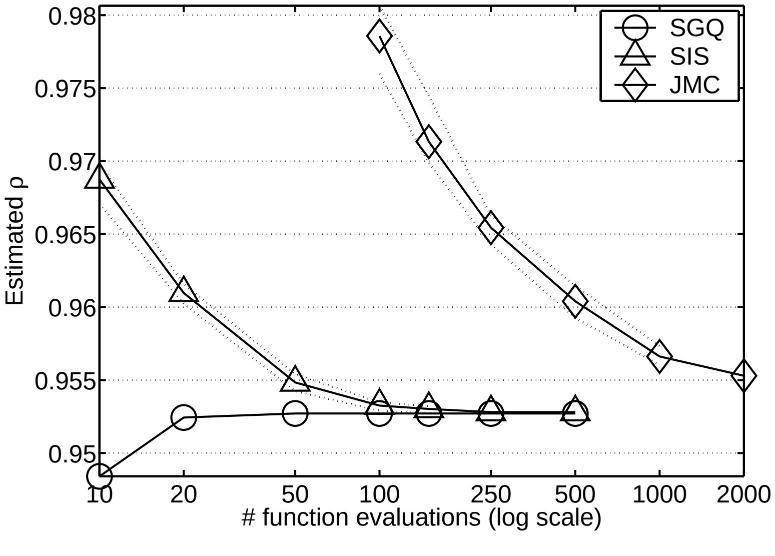


Table 4.1: Computational Costs
Seconds / Likelihood evaluation

R	JMC	SIS	SGQ
10	52.0	20.3	7.3
50	68.6	51.6	14.3
100	85.8	124.8	31.5
150	105.5	306.1	112.4
200	123.1	678.9	340.7
250	140.8	1171.0	658.5
2000	881.1		

4.5 Conclusions

This chapter discusses the numerical approximation of the likelihood for a certain class of nonlinear panel data models. Limited dependent variable models with AR(1) error terms are an important example of appropriate models. The numerical difficulties arise because the likelihood function involves multiple integrals. 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.

This chapter discusses how these models allow to split the multiple integrals into several integrals with lower dimensions. In the univariate AR(1) example, the integrals become one-dimensional. 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.

There are several approaches to actually implement the sequential evaluation of the likelihood function. In engineering where most of these methods were developed and in the econometric time series literature where they receive increased attention, the number of time periods is high compared to the typical microeconometric panel data. This affects the relative advantages and disadvantages of the algorithms. I suggest an approach that is plausibly very powerful for moderate time series dimensions. It is based on Gaussian quadrature for one-dimensional problems and can be extended by Smolyak cubature for multidimensional problems. This allows very precise approximations with little computational effort. For high longitudinal dimensions, this approach is likely to perform less well since the integrands become less well-behaved with more time-series observations. It will certainly help to investigate the relative performance of different algorithms in Monte Carlo study in which the relevant dimensions of the data and model can be varied. This will be left for future research.

In an application to an ordered logit model with an AR(1) error term for panel data with $T = 6$ observations over time, I show that the proposed method clearly outperforms the typically used approach of joint simulation and also a sequential simulation approach. While sequential Gaussian quadrature needs only 10 to 20 function evaluations R for an accurate parameter estimation, the joint simulation still suffers from bias with $R = 2000$. The method is also easily implemented and needs moderate additional computation for a given R.

4.6 Appendix: Proofs

Proof of proposition 1

The likelihood contribution in equation 4.4 is

$$\begin{aligned} \ell_i &= \Pr(\mathbf{Y}_{i,1:T} = \mathbf{y}_{i,1:T} | \mathbf{x}_i, \boldsymbol{\theta}) \\ &= \int \cdots \int \Pr(\mathbf{Y}_{i,1:T} = \mathbf{y}_{i,1:T} | \mathbf{a}_{1:T}, \mathbf{x}_i, \boldsymbol{\theta}) f_{1:T}(\mathbf{a}_{1:T})\, d\mathbf{a}_1 \cdots d\mathbf{a}_T. \end{aligned}$$

In general,

$$\begin{aligned} &\Pr(\mathbf{Y}_{i,1:T} = \mathbf{y}_{i,1:T} | \mathbf{a}_{1:T}, \mathbf{x}_i, \boldsymbol{\theta}) = \\ &\qquad \Pr(\mathbf{Y}_{i1} = \mathbf{y}_{i1} | \mathbf{a}_{1:T}, \mathbf{x}_i, \boldsymbol{\theta}) \Pr(\mathbf{Y}_{i2} = \mathbf{y}_{i2} | \mathbf{Y}_{i1} = \mathbf{y}_{i1}, \mathbf{a}_{1:T}, \mathbf{x}_i, \boldsymbol{\theta}) \cdots \\ &\qquad \Pr(\mathbf{Y}_{iT} = \mathbf{y}_{iT} | \mathbf{Y}_{i,1:T-1} = \mathbf{y}_{i,1:T-1}, \mathbf{a}_{1:T}, \mathbf{x}_i, \boldsymbol{\theta}) \end{aligned}$$

By assumption 2,

$$\Pr(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1}, \mathbf{a}_{1:T}, \mathbf{x}_i, \boldsymbol{\theta}) = P_{it}(\mathbf{y}_{it} | \mathbf{a}_t) \quad \forall t = 1, ..., T.$$

So proposition 1 follows directly:

$$\ell_i = \int \cdots \int \left(\prod_{t=1}^{T} P_{it}(\mathbf{y}_{it} | \mathbf{a}_t) \right) f_{1:T}(\mathbf{a}_{1:T})\, d\mathbf{a}_1 \cdots d\mathbf{a}_T.$$

Proof of proposition 2

By definition,

$$P_{it|1:t-1}(\mathbf{y}_{it}) = \Pr\left(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1}, \mathbf{x}_i, \boldsymbol{\theta}\right)$$

for the non-trivial case of $t \geq 2$. Obviously,

$$\Pr\left(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1}, \mathbf{x}_i, \boldsymbol{\theta}\right) = \\ \int \Pr\left(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{a}_{it} = \mathbf{a}, \mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1}, \mathbf{x}_i, \boldsymbol{\theta}\right) f_{it|1:t-1}(\mathbf{a})\, d\mathbf{a}.$$

By assumption 2,

$$\Pr\left(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{a}_{it} = \mathbf{a}, \mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1}, \mathbf{x}_i, \boldsymbol{\theta}\right) = P_{it}(\mathbf{y}_{it} | \mathbf{a}),$$

so proposition 2 follows directly.

Proof of proposition 3

Bayes rule implies

$$f_{it|1:t}(\mathbf{a}_{it}) = f_{it|1:t-1}(\mathbf{a}_{it}) \frac{\Pr(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{a}_{it}, \mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1}, \mathbf{x}_i, \boldsymbol{\theta})}{\Pr(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1}, \mathbf{x}_i, \boldsymbol{\theta})}$$

The denominator is defined as $P_{it|1:t-1}(\mathbf{y}_{it})$. By assumption 2 the numerator $\Pr(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{a}_{it}, \mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1}, \mathbf{x}_i, \boldsymbol{\theta}) = \Pr(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{a}_{it}, \mathbf{x}_i, \boldsymbol{\theta})$ which is defined as $P_{it}(\mathbf{y}_{it} | \mathbf{a}_{it})$. So proposition 3 follows directly:

$$\frac{f_{it|1:t}(\mathbf{a}_{it})}{f_{it|1:t-1}(\mathbf{a}_{it})} = \frac{P_{it}(\mathbf{y}_{it} | \mathbf{a}_{it})}{P_{it|1:t-1}(\mathbf{y}_{it})}.$$

Proof of proposition 4

By definition,

$$q_{it}(\mathbf{a}) \quad = \quad \frac{f_{it|1:t-1}(\mathbf{a})}{f(\mathbf{a})}.$$

Bayes' rule implies

$$\frac{f_{it|1:t-1}(\mathbf{a})}{f(\mathbf{a})} = \frac{\Pr\left(\mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1} | \mathbf{a}_{it} = \mathbf{a}, \mathbf{x}_i, \boldsymbol{\theta}\right)}{\Pr\left(\mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1} | \mathbf{x}_i, \boldsymbol{\theta}\right)}. \\ = \int \cdots \int \frac{\Pr\left(\mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1} | \mathbf{a}_{i,1:t-1} = \mathbf{a}_{1:t-1}, \mathbf{a}_{it} = \mathbf{a}, \mathbf{x}_i, \boldsymbol{\theta}\right)}{\Pr\left(\mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1} | \mathbf{x}_i, \boldsymbol{\theta}\right)} f_{1:t-1}(\mathbf{a}_{1:t-1} | \mathbf{a}_{it} = \mathbf{a})\, d\mathbf{a}_{1:t-1}.$$

Analogous to the argument in the proof of proposition 1, the numerator in this expression is equal to

$$\Pr\left(\mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1} | \mathbf{a}_{i,1:t-1} = \mathbf{a}_{1:t-1}, \mathbf{a}_{it} = \mathbf{a}, \mathbf{x}_i, \boldsymbol{\theta}\right) = \prod_{s=1}^{t-1} P_{is}(\mathbf{y}_{is} | \mathbf{a}_{is}).$$

The denominator can be written by standard rules of conditioning as

$$\Pr\left(\mathbf{Y}_{i,1:t-1} = \mathbf{y}_{i,1:t-1} | \mathbf{x}_i, \boldsymbol{\theta}\right) = \prod_{s=1}^{t-1} P_{is|1:s-1}(\mathbf{y}_{is}).$$

Proposition 4 follows directly.

Proof of proposition 6

By definition,

$$q_{i,t+1}(\mathbf{a}_{i,t+1}) = \frac{f_{i,t+1|1:t}(\mathbf{a}_{i,t+1})}{f(\mathbf{a}_{i,t+1})}.$$

Conditional on $\mathbf{a}_{it}$, $\mathbf{Y}_{i,1:t}$ and $\mathbf{a}_{i,t+1}$ are independent. Therefore,

$$f_{i,t+1|1:t}(\mathbf{a}_{i,t+1}) = \int f_c(\mathbf{a}_{i,t+1}|\mathbf{a}) f_{it|1:t}(\mathbf{a})\, d\mathbf{a}.$$

Bayes' rule in combination with assumption 2 implies

$$f_{it|1:t}(\mathbf{a}) = f_{it|1:t-1}(\mathbf{a}) \frac{P_{it}(\mathbf{y}_{it}|\mathbf{a})}{P_{it|1:t-1}(\mathbf{y}_{it})}.$$

Combining these results with the definition $q_{it}(\mathbf{a}) = \frac{f_{i,t|1:t-1}(\mathbf{a})}{f(\mathbf{a})}$ yields the expression of proposition 6:

$$q_{i,t+1}(\mathbf{a}_{i,t+1}) = \int q_{it}(\mathbf{a}) \frac{P_{it}(\mathbf{y}_{it}|\mathbf{a})}{P_{it|1:t-1}(\mathbf{y}_{it})} \frac{f_c(\mathbf{a}_{i,t+1}|\mathbf{a})}{f(\mathbf{a}_{i,t+1})} f(\mathbf{a})\, d\mathbf{a}$$

5 Multidimensional Integration in Estimation Problems

5.1 Introduction

Many econometric models imply likelihood and moment functions that involve multidimensional integrals without analytically tractable solutions. This problem arises frequently in microeconometric models in which all or some of the endogenous variables are only partially observed. Other sources include unobserved heterogeneity in nonlinear models and expectations of agents.

There are different approaches for numerical integration. It is well known that Gaussian quadrature performs very well in the case of one-dimensional integrals of smooth functions (Butler and Moffit 1982). Quadrature can be extended to multiple dimensions and is then also called cubature. The most direct extension is a tensor product of one-dimensional quadrature rules. However, computing costs rise exponentially with the number of dimensions and become prohibitive for more than four or five dimensions. This phenomenon is also known as the curse of dimensionality.

This led to the advancement and predominant use of simulation techniques for the numerical approximation of multidimensional integrals in the econometric literature, see for example McFadden (1989) or Börsch-Supan and Hajivassiliou (1993). Hajivassiliou and Ruud (1994) provide an overview over the general approaches of simulation and Train (2003) provides a textbook treatment with a focus on discrete choice models, one of the major classes of models for which these methods were developed and frequently used.

This chapter is based on joint work with Viktor Winschel and proposes and investigates the performance of a different approach that can be traced back to Smolyak (1963). It has been advanced in recent research in numerical mathematics, see for example Novak and Ritter (1999). It is based on one-dimensional Gaussian quadrature but extends it to higher dimensions in a more careful way than the tensor product rule. This dramatically decreases computational costs in higher dimensions.

Just like Gaussian cubature and simulation, Smolyak cubature evaluates the integrand at certain points and calculates a weighted average of these function values. The difference is how these points and weights are derived. The Smolyak approach is a general method for multivariate extensions of univariate operators and not only applicable to integration. Another example is function approximation which has been used for the solution of a overlapping generations model by Krüger and Kübler (2004).

After introducing the Smolyak approach, the results of Monte Carlo experiments are presented. They directly address the question of interest for estimation: Which method delivers the best estimates with a given amount of computing costs? The experiments are based on a panel data random parameters logit models which are widely used in applied discrete choice analysis. They vary the panel data dimensions, the number of alternatives, the dimension of unobserved taste components and the parameterization of the data generating process. The results show that the Smolyak-based cubature methods clearly outperform simulations based on both random number generators and the modified latin hypercube sampling proposed by Hess, Train, and Polak (2005).

The chapter is structured as follows: Section 5.2 briefly discusses the circumstances in which multiple integrals evolve in estimation problems. It then introduces an example, the random parameters logit model, in somewhat more detail since it will be used in the Monte Carlo experiments. Section 5.3 discusses the general approaches to numerical integration and introduces Smolyak-based cubature. Section 5.4 presents the Monte Carlo design and results. Section 5.5 concludes.

5.2 Econometric Models Requiring Numerical Integration

The log-likelihood function of microeconometric models can typically be written as a sum over a number N of independent log-likelihood contributions $\log(\ell_i(\boldsymbol{\theta}; \text{data})$. Maximum likelihood defines the estimated parameter vector $\hat{\boldsymbol{\theta}}$ as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log(\ell_i(\boldsymbol{\theta})). \tag{5.1}$$

The discussion is focused on maximum likelihood estimation, but the same problems and approaches are applicable for other methods like GMM or Bayesian analyses. In many models the likelihood contributions $\ell_i(\boldsymbol{\theta})$ involve multiple integrals which cannot be expressed in closed form and must be evaluated numerically. The approximation algorithm used is essential for the estimation task. Numerical maximization involves repeated evaluations of the likelihood function. Each evaluation in turn involves solving N multiple integrals, where N in practice can be several thousand. While the ongoing increase in computational power makes widespread use of such models feasible, the computational costs are still high and sometimes prohibitive. Instead of compromising on model specification or approximation quality, it is therefore important to choose an efficient method of numerical integration for a given model and accuracy.

There are various reasons why microeconometric models imply multiple integrals in $\ell_i(\boldsymbol{\theta})$. Typically, they represent the expectation of a function over several random variables. A major reason for their presence is that many models are specified in terms of latent random variables for which the observed endogenous variables provide only a partial indication. Another source in nonlinear models is an error term with a mixture distribution such as random effects or error components models which can be estimated by calculation of integrated likelihood functions. Finally, dynamic optimization models naturally involve multiple integrals, see for example Eckstein and Wolpin (1999). For a more general presentation see for example Hajivassiliou and Ruud (1994) or Gouriéroux and Monfort (1996), both from a perspective of simulation.

The random parameters logit (RPL) model or mixed logit model is widely used for studying choices between a finite set of alternatives. See McFadden and Train (2000) for an introduction to this model and a discussion of its estimation by simulation methods. Suppose discrete choices of N individuals are observed. The data has a panel structure, so that each of the subjects makes T choices. In each of these choice situations, the individual is confronted with a set of J alternatives and chooses one of them. These alternatives are described by K exogenous attributes. The $(K \times 1)$ vectors $\mathbf{x}_{itj}$ collect these attributes of alternative $j = 1, ..., J$ in choice situation $t = 1, ..., T$ of individual $i = 1, ..., N$.

Random utility maximization (RUM) models of discrete choices assume that the individuals make their choices by evaluating the utility that each of the alternatives yields and then picking the one with the highest value. The researcher obviously does not observe these utility levels. They are modeled as latent variables for which the observed choices provide an indication. Let the utility that individual i attaches to alternative j in choice situation t be represented by the random coefficients specification

$$U_{itj} \quad = \quad \mathbf{x}'_{itj}\boldsymbol{\beta}_i + e_{itj}. \tag{5.2}$$

It is given by a linear combination of the attributes of the alternative, weighted with individual-specific taste levels $\boldsymbol{\beta}_i$. These individual taste levels are distributed across the population according to a parametric joint p.d.f. $f(\boldsymbol{\beta}_i; \boldsymbol{\theta})$ with support $\Psi \subseteq \mathbb{R}^K$. The i.i.d. random variables e_{itj} capture unobserved utility components. They are assumed to follow an Extreme Value Type I (or Gumbel) distribution. Note that this model can be generalized, for example, the distribution $f(\boldsymbol{\beta}_i;\boldsymbol{\theta})$ can be specified as a function of observed individual characteristics.

Our goal is to estimate the parameters $\boldsymbol{\theta}$. Let y_{itj} denote an indicator variable that has the value 1 if individual i chooses alternative j in choice situation t and 0 otherwise. Denote the vector of observed individual outcomes as $\mathbf{y}_i = [y_{itj}; t = 1, ..., T, j = 1..., J]$ and the matrix of all exogenous variables as $\mathbf{x}_i = [\mathbf{x}_{itj}; t = 1, ..., T, j = 1..., J]$. Then, the probability that the underlying random variable $\mathbf{Y}_i$ equals the observed realization $\mathbf{y}_i$ conditional on $\mathbf{x}_i$ and the taste levels $\boldsymbol{\beta}_i$ can be expressed as

$$P_i^*(\boldsymbol{\beta}_i) = \Pr(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}_i) \;=\; \prod_{t=1}^{T} \frac{\prod_{j=1}^{J} \exp(y_{itj}\mathbf{x}'_{itj}\boldsymbol{\beta}_i)}{\sum_{j=1}^{J} \exp(\mathbf{x}'_{itj}\boldsymbol{\beta}_i)}. \tag{5.3}$$

Suppose the regularity conditions given by McFadden and Train (2000) hold. The likelihood contribution of individual i as a function of $\boldsymbol{\theta}$ can be written as

$$\ell_i(\boldsymbol{\theta};\mathbf{Y}_i) = \Pr(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) \;=\; \int_{\Psi} P_i^*(\boldsymbol{\beta}_i) f(\boldsymbol{\beta}_i;\boldsymbol{\theta})\, d\boldsymbol{\beta}_i. \tag{5.4}$$

A solution for this K-dimensional integral does not exist in closed form and has to be approximated numerically.

5.3 Numerical Integration in Multiple Dimensions

There are several methods to numerically approximate an integral of a function g over a D-dimensional vector $\mathbf{z}$.

$$I^D[g] = \int_{\Omega} g(\mathbf{z})\, w(\mathbf{z})\, d\mathbf{z}, \tag{5.5}$$

where $w(\mathbf{z})$ is some weighting function. As discussed above, in estimation problems, the integral often represents an expected value of g so that $w(\mathbf{z})$ is a p.d.f. and Ω its support. A computationally feasible approach that is common to all methods discussed in this chapter is to approximate the integral as a weighted sum of a number R of the integrand evaluated at certain points, referred to as nodes:

$$I^D[g] \approx \sum_{r=1}^{R} g(\mathbf{z}_r) w_r, \tag{5.6}$$

where w_r is the weights of node $\mathbf{z}_r$. The methods differ in the way they derive the nodes $\mathbf{z}_r$ and weights w_r. For Monte Carlo integration, the $\mathbf{z}_r$ are equally weighted draws with $w_r = R^{-1} \quad \forall r = 1, ..., R$ from the density $w(\mathbf{z})$. These draws can be generated by different strategies. Classically, a random number generator is used. Since draws generated by a computer can never be truly random, these draws are often labeled pseudo-random draws. These nodes are often clustered in certain areas. Antithetic sampling algorithms distribute the nodes more evenly but preserve properties of random numbers. In the Monte Carlo experiments, pseudo-random Monte Carlo simulation (PRMC) and antithetic draws from a modified latin hypercube sampling (MLHS) algorithm are used. It was shown to work well for the estimation of RPL models by Hess, Train, and Polak (2005).

Polynomial cubature methods are multidimensional extensions of Gaussian quadrature. They take a different strategy and determine nodes and weights such that $g(\mathbf{z})$ is approximated by a polynomial of a given order for which the integral is straightforward to solve. In the econometrics literature, one-dimensional Gaussian quadrature methods are known to work well if $g(\mathbf{z})$ is a smooth function and can therefore be well approximated by a (low-order) polynomial (Butler and Moffit 1982). There are different strategies for the generalization of this approach to multiple dimensions. The most straightforward method known as the tensor product rule suffers from the fact that the computational costs rise exponentially with the dimensionality of the problem. It is therefore of little use for more than four or five dimensions. Other methods are not as straightforward to implement and are therefore often considered impractical for econometric analyses (Bhat 2001, Geweke 1996). The complication lies in the calculation of the nodes $\mathbf{z}_r$ and weights w_r. Given those, they only have to be plugged into equation 5.6. Since the nodes and weights only depend on the dimension of the problem and the desired approximation level, it is also possible to use precalculated values.

Multidimensional cubature methods are derived from one-dimensional Gaussian quadrature formulas. The following discussion is based on the case of fully symmetric weight functions in the sense that the D-variate weighting function w can be multiplicatively decomposed into D univariate functions $\tilde{w}$ as $w(\mathbf{z}) = \tilde{w}(z_1) \cdots \tilde{w}(z_D)$ that are all symmetric such that $\tilde{w}(z_d) = \tilde{w}(-z_d)$ for all $d = 1, \ldots, D$. Most problems in econometrics can be expressed in such a way by a change of variables. See Novak and Ritter (1999) for a discussion of this and more general cases.

In the case of a one-dimensional variable z, the integral in equation 5.5 can approximated efficiently by Gaussian quadrature methods. Let

$$V_i[g] = \sum_{r=1}^{R(i)} g(z_r^i) w_r^i. \tag{5.7}$$

denote a one-dimensional quadrature rule. The parameter $i \in \mathbb{N}$ drives the precision of this rule. It requires the evaluation of g at a number $R(i)$ of nodes which depends on the intended precision. The nodes $z_1^i, ..., z_{R(i)}^i$ and weights $w_1^i, ..., w_{R(i)}^i$ are given by the quadrature rule. They are constructed such that $V_i[g]$ evaluates the integral exactly with minimum number of function evaluations if g is a polynomial of a certain degree. It is well known that with $R(i) = i$, Gaussian quadrature rules $V_i[g]$ are able to give exact solutions for all polynomials with an order of at most $2i-1$. The nodes and weights depend on the weight function w and the support Ω. For many standard cases, Gaussian quadrature rules are well known and implemented in many statistical software packages.

In the case of multidimensional $\mathbf{z}$, Gaussian quadrature can be extended by the product rule as discussed by Tauchen and Hussey (1991). Let $\mathbf{i} = [i_1, ..., i_D]$ denote the vector of precision indecees for each dimension. The product rule can be written as

$$\begin{aligned} T_{D,\mathbf{i}}[g] &= (V_{i_1} \otimes \cdots \otimes V_{i_D})[g] && (5.8) \\ &= \sum_{r_1=1}^{R(i_1)} \cdots \sum_{r_D=1}^{R(i_D)} g(z_{r_1}^{i_1}, ..., z_{r_D}^{i_D}) w_{r_1}^{i_1} \cdots w_{r_D}^{i_D}, && (5.9) \end{aligned}$$

where the nodes and weights are those implied by the underlying one-dimensional quadrature rules $V_{i_1}, ..., V_{i_D}$. Usually the precision is chosen equally in all dimensions, so the integral is approximated by $T_{D,[i,i,...,i]}[g]$. The curse of dimensionality lies in the fact that the evaluation of this rule requires $R(i)^D$ function evaluations which rises exponentially with D and is prohibitive for high D.

The Smolyak method proposed in this chapter extends Gaussian quadrature rules to multiple dimensions with substantially less function evaluations. This is achieved by combining the univariate rules in a “more clever” way than the product rule. The approach goes back to Smolyak (1963) and is a general method for multivariate extensions of univariate operators. Integration was already discussed in the original paper and is an active research area in numerical mathematics. Instead of taking the sophisticated one-dimensional rule and naïvely extending it to multiple dimensions by a full product grid, the Smolyak approach is specifically designed for multidimensional problems.

Given a approximation level k, the Smolyak rule linearly combines product rules with different combinations of precision indecees $\mathbf{i}$. It can be written as

$$A_{D,k}[g] = \sum_{\mathbf{i} \in S_k^D} (-1)^{D+k-|\mathbf{i}|} \binom{D-1}{D+k-|\mathbf{i}|} T_{D,\mathbf{i}}[g] \qquad (5.10)$$

where $S_k^D = \{\mathbf{i} \in \mathbb{N}^D : k+1 \leq |\mathbf{i}| \leq k+D\}$ and $|\mathbf{i}| = i_1 + ... + i_D$. The sum is over all D-dimensional vectors of natural numbers $\mathbf{i}$ which have a norm within certain bounds that are governed by D and k. These vectors translate into the number of nodes for each dimension in a tensor product cubature rule. The bound on the norm has the effect that the tensor product rules with a relatively fine sequence of nodes in one dimension are relatively coarse in the other dimensions.

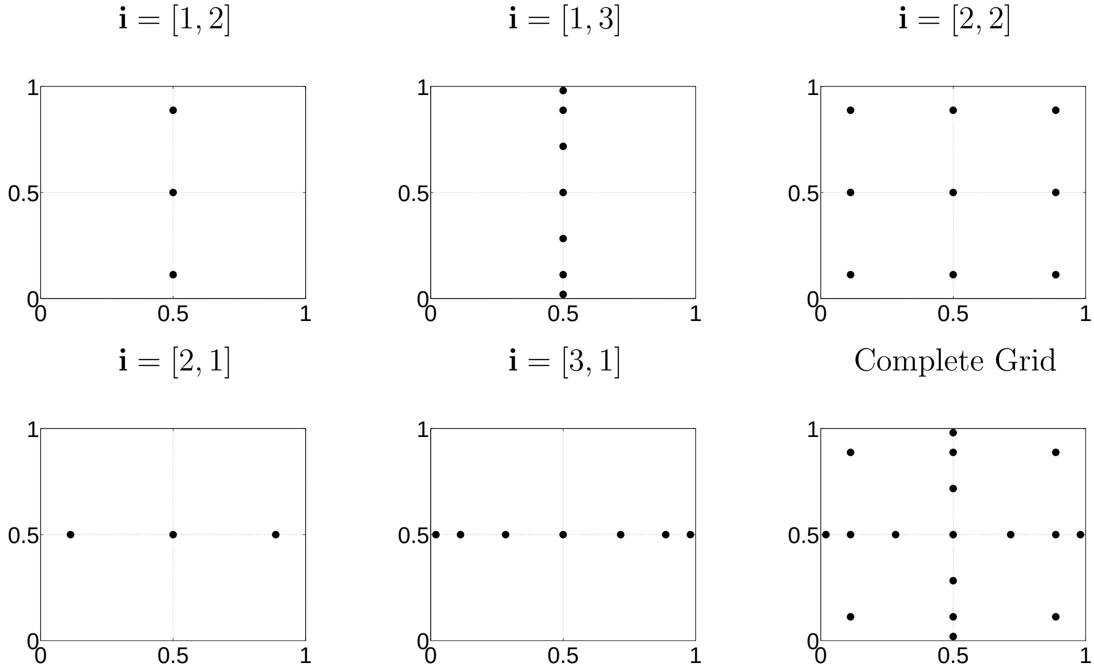
Equation 5.10 is based on a linear combination of product rules and those in turn are based on one-dimensional Gaussian quadrature rules. Any univariate quadrature rule can serve as a basis for this multivariate extension. A careful choice can further reduce the computational burden. This is true if for some $\mathbf{i} \in S_k^D$, the product rule $T_{D,\mathbf{i}}$ evaluates g at the same nodes as for other vectors in this set. Instead of repeatedly evaluating the function at the same nodes, this can be done once and only the corresponding weights have to be aggregated. This is possible if the nodes of the one-dimensional basis rules with a low precision index i are also used in those with a high precision index j so that $\{z_1^i, ..., z_{R(i)}^i\} \subset \{z_1^j, ..., z_{R(j)}^j\}$ if $i < j$. For a discussion of this issue also see (Novak and Ritter 1996).

Different rules for generating sets of nodes with this property for Gaussian quadrature are discussed by Petras (2003). In the Monte Carlo experiments for the RPL model shown below, a Smolyak cubature rule based on delayed Kronrod-Patterson sequences as suggested by Petras is used. It is defined for $w(\mathbf{z}) = 1$ and $\Omega = [0,1]^D$. This is adequate for the evaluation of expectations over uniformly distributed random variables and is the extension of a Gauss-Legendre quadrature rule. A problem that frequently occurs is the expectation over standard normal random variables. For this case, Genz and Keister (1996) apply the Smolyak extension to a Gauss-Hermite quadrature rule, also based on nested sets of nodes, also see Novak and Ritter (1999).

A simple example may help to clarify the approach. Let $D = k = 2$. This implies $S_k^D = \{\mathbf{i} \in \mathbb{N}^2 : 3 \leq |\mathbf{i}| \leq 4\} = \{[1,2],[2,1],[1,3],[3,1],[2,2]\}$. The strategy of Petras (2003) is based on nested sets of nodes. This is achieved with delayed Kronrod-Patterson sequences with $R(1) = 1$, $R(2) = 3$, and $R(3) = 7$. The sets of nodes are $[0.5]$ for $i = 1$, $[0.11, 0.5, 0.89]$ for $i = 2$, and $[0.02, 0.11, 0.28, 0.5, 0.72, 0.89, 0.98]$ for $i = 3$. The set of nodes for lower i are subsets of those with higher i.

For all $\mathbf{i} \in S_k^D$, Figure 5.1 shows the nodes used by the corresponding product rule. Obviously, the nodes for $\mathbf{i} = [1,2]$ and $\mathbf{i} = [2,1]$ are a subset of the nodes for $\mathbf{i} = [1,3]$ and $\mathbf{i} = [3,1]$, respectively. The grid for $\mathbf{i} = [2,2]$ adds four distinct nodes. The complete set of all nodes used by the Smolyak rule only consists of the 17 nodes depicted in the lower right panel of the graph. The full product rule with the corresponding degree of exactness would require the evaluation of the function at the full grid of $7^2 = 49$ nodes. In higher dimensions, this difference becomes more dramatic. With $D = 10$ and $k = 2$, The Smolyak rule needs $1,201$ and the product rule $7^{10} = 282,475,249$ function evaluations.

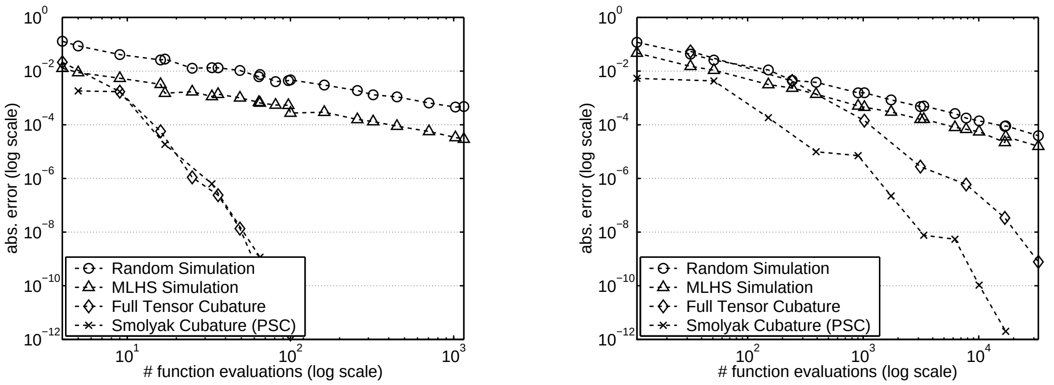
Figure 5.1: Construction of the Smolyak grid



Monte Carlo integration is very general in the sense that under mild regularity conditions, the approximated integral is $\sqrt{R}$-consistent by a law of large numbers. This convergence rate is independent of the number of dimensions D. However, the error given a finite number R does increase with D. Gauss quadrature and Smolyak cubature rely on the approximation of g by a polynomial. For smooth functions this allows a faster convergence than simulation (Gerstner and Griebel 2003). For a given number of evaluations, the performance depends on how well g can be approximated by a polynomial of the corresponding order.

Figure 5.2 shows the performance of different methods in terms of absolute errors for a simple example for which a closed-form expression exists: $\int_{[0,3]^D} \lambda(\mathbf{z}) d\mathbf{z}$, where λ is the joint p.d.f. of D i.i.d. logistic random variables. The figure shows results for $D = 2$ and $D = 5$. Since λ is very smooth, the cubature methods exhibit a much faster convergence rate than simulation. The difference between the tensor product and the Smolyak rule is apparent from a comparison between $D = 2$ and $D = 5$: The higher the dimension, the more efficient is the Smolyak relative to the tensor rule.

Figure 5.2: Approximation of a Logistic c.d.f. in 5 dimensions



(a) $D = 2$ Dimensions

(b) $D = 5$ Dimensions

5.4 Monte Carlo Experiments

A number of Monte Carlo experiments help to evaluate the relative performance of the numerical integration algorithms. Different random parameters logit (RPL) models as discussed in section 5.2 are implemented. The models are specified for N individuals with T choices between J alternatives each, where each alternative is characterized by K properties. In most model specifications, the K individual taste parameters are normally distributed across the population with mean μ, variance σ^2, and zero covariance. Below also results for uniformly distributed taste levels are reported in order to test the sensitivity of the results with respect to the specification of the distribution.

As a starting point, a reference model is specified with $N = 1000$, $T = 5$, $J = 5$, $K = 10$, $\mu = 1$, and $\sigma = 0.5$. Then each of these numbers is varied separately to demonstrate their impact on the approximation errors of the different methods. For each of these settings, estimates were obtained for 100 artificial data sets. The properties of the alternatives $\mathbf{x}_{itj}$ were drawn from a standard uniform distribution. The model parameters μ and σ were estimated for each data set.

In order to approximate the integral in equation 5.4 with normally distributed $\boldsymbol{\beta}_i$ by Gauss-Hermite quadrature, it has to be expressed in terms of standard normal random variables. This can be easily done by a change of variables:

$$\int_{\mathbb{R}^K} P_i^*(\boldsymbol{\beta}_i) f_\beta(\boldsymbol{\beta}_i; \boldsymbol{\mu}, \boldsymbol{\sigma})\, d\boldsymbol{\beta}_i = \int_{\mathbb{R}^K} P_i^* \left(\boldsymbol{\mu} + L(\boldsymbol{\sigma})\mathbf{e}_i)\right) \phi^K(\mathbf{e}_i)\, d\mathbf{e}_i, \qquad (5.11)$$

where L is the Cholesky factorization of the covariance matrix of β_i and ϕ^K denotes the joint p.d.f. of K i.i.d. standard normal random variables.

Two simulation-based and two Smolyak-based estimators are implemented and compared in terms of performance. In applications of the RPL models, the predominant method for estimation is maximum simulated likelihood. The pseudo-random Monte Carlo simulation (PRMC) method uses random numbers and the modified latin hypercube sampling simulation (MLHS) method uses the antithetic quasi random numbers suggested by Hess, Train, and Polak (2005) for the RPL model. In addition, two versions of Smolyak-based cubature for the approximation of the likelihood contributions discussed in section 5.3 are implemented. In the first part, results from the Genz and Keister (1996) Smolyak cubature (GKSC) rules for the integration over Gaussian distributions are reported, since equation 5.11 has the form required for this method. In addition, results obtained by the Petras (2003) Smolyak cubature (PSC) rules are reported to test for the sensitivity of the results with respect to the choice of the cubature rule. Both were discussed in section 5.3.

The maximization algorithm for the likelihood function does not affect the relative performance of the estimators based on different numerical integration rules. The standard Newton method with numerically approximated gradients and a BHHH approximation of the Hessian works fine in this application and was used for all estimates.

Reference Model

Figure 5.3 shows the results for the reference model. The performance measure on the ordinate is a relative root mean squared error. Given a certain number of nodes at which the functions are evaluated, the parameters were estimated for 100 simulated data sets using all three methods. For both parameters μ and σ, the mean squared errors were calculated over the 100 replications. They were then normalized by the respective variance of the best performing method with the maximal number of function evaluations. This makes the MSEs of both parameters comparable between each other and across different model specifications.

The number of 100 replications with different data sets is sufficient for conclusive comparisons. In Figure 5.3, the remaining randomness is visualized with error bars indicating the 95% confidence intervals for respective result. They were generated by resampling from the estimated parameters with replacement and recalculating the performance measure for each of the samples.

Figure 5.3: Monte Carlo Results: Reference Model

$N = 1000$, $T = 5$, $J = 5$, $K = 10$, $\mu = 1$, $\sigma = 0.5$

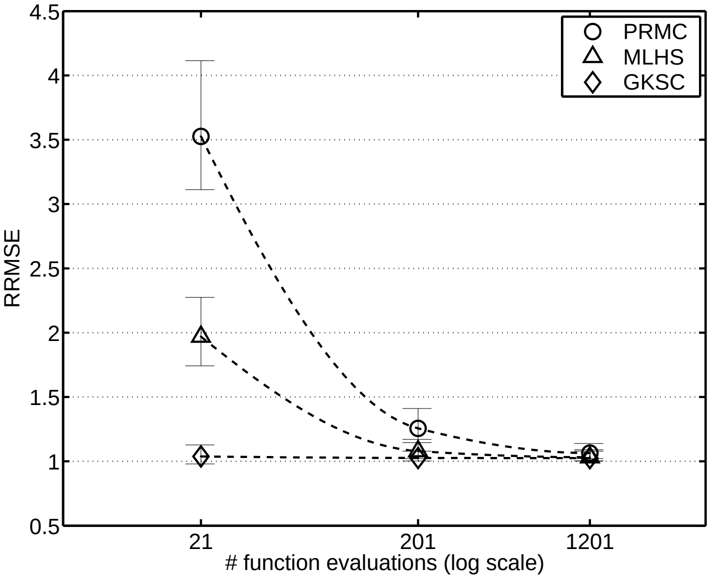


Table 5.1 shows the same data, but the results are relative to GKSC with the same number of function evaluations. The confidence intervals take correlations of the results into account. The results are striking. For a large number of function evaluations $R = 1201$, all methods perform equally well. But for moderate and small numbers of evaluations, there are enormous differences. While the Smolyak cubature method only needs 21 evaluation to achieve a negligible approximation error, the error of MLHS and PRMC is higher by a factor of 1.9 and 3.4, respectively. Put differently: The error of GKSC with 21 function evaluations is lower than the error of MLHS with 201 and the error of PRMC with 1201 evaluations.

Table 5.1: Monte Carlo Results: Reference model

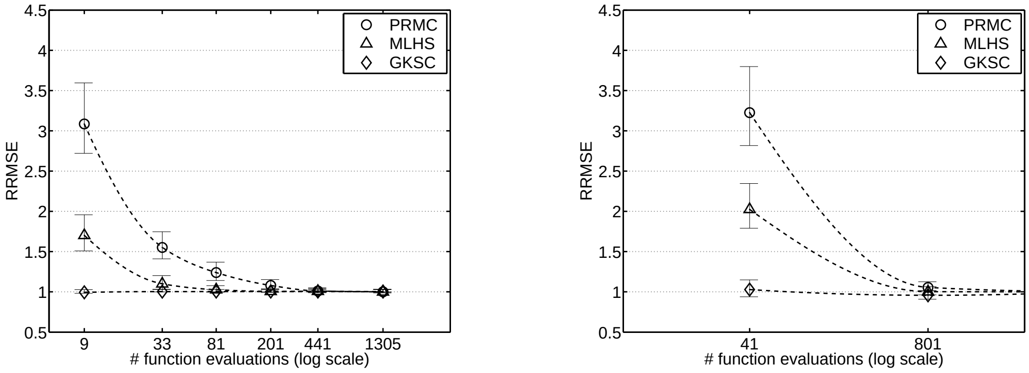
	$\text{RRMSE}_{\text{GKSC}}$		$\frac{\text{RRMSE}_{\text{PRMC}}}{\text{RRMSE}_{\text{RRMSE}}}$		$\frac{\text{RRMSE}_{\text{MLHS}}}{\text{RRMSE}_{\text{RRMSE}}}$	
R	est.	95% CI	est.	95% CI	est.	95% CI
21	1.04	0.98–1.13	3.40	3.04–3.87	1.90	1.73–2.11
201	1.03	1.00–1.08	1.22	1.13–1.33	1.05	1.01–1.10
1201	1.03	1.00–1.08	1.04	1.01–1.06	1.01	0.99–1.02

Varying the Number of Integral Dimensions K

In the following, the results are discussed for similar models in which the parameters of the data generating process are varied. One of the most important parameters is the number K of characteristics defining each of the alternatives, since it equals the dimension of the integral. Figure 5.4 shows the results for the dimensions $K = 4$ and 20. The Smolyak-based GKSC method performs well with a very small number of replications, whereas the simulation-based methods require significantly more computations. A closer look shows that this effect increases somewhat in higher dimensions: While 201 evaluations suffice for MLHS to catch the GKSC performance in 4 dimensions, it does not in ten and the relative difference of PRMC is even higher. The results for other dimensions are qualitatively the same. They are presented together with tables equivalent to table 5.1 in the appendix.

Figure 5.4: Monte Carlo Results: Different Dimensions K

All results for: $N = 1000, T = 5, J = 5, \mu = 1, \sigma = 0.5$



(a) $K = 4$ Dimensions

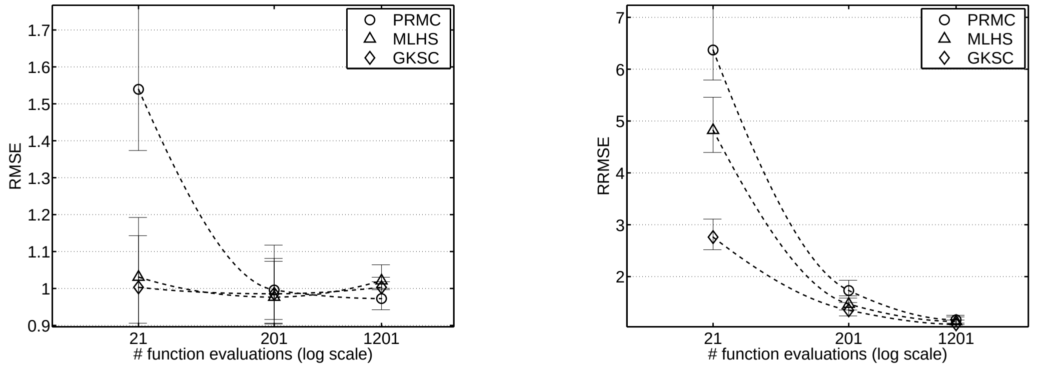
(b) $K = 20$ Dimensions

Varying the Variance σ^2

Another important parameter of the data-generating process is σ. The higher its value, the more the integrand varies with the unobserved individual tastes over which the integration is performed. Figure 5.5 shows the results of models with $\sigma = 0.25$ and 1. As expected, all methods perform worse with a higher σ. The relative performances remain similar. With very high σ, all methods fail to give reasonable results.

Figure 5.5: Monte Carlo Results, Differences by σ

All results for: $N = 1000, T = 5, J = 5, K = 10, \mu = 1$



(a) $\sigma = 0.25$

(b) $\sigma = 1$

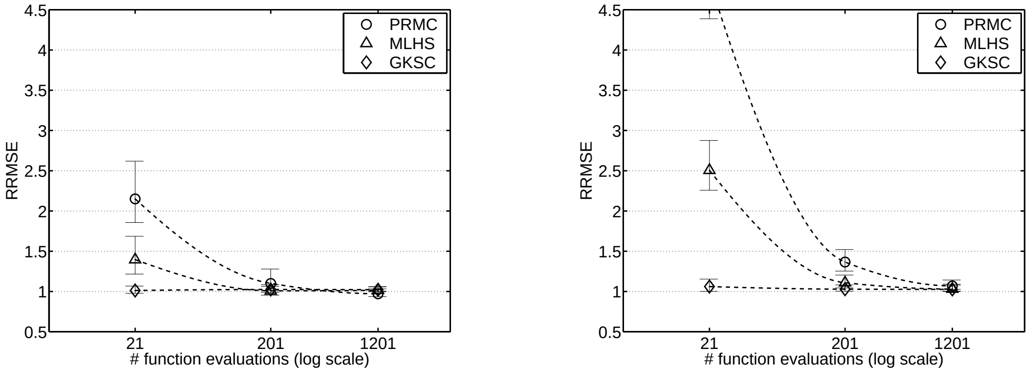
Other Variations: N, T, J, and μ

The number N of independent observations corresponds to the number of integrals to be solved for each likelihood evaluation. With a higher N, random fluctuations of the approximation error “average out”, but a systematic bias remains. This explains the results shown in Figure 5.6 which might be counter-intuitive at first sight: the higher N, the worse do the simulation-based approaches perform relative to Smolyak cubature. With rising N, the variance of the simulation-based estimators decreases, also relative to the sampling variance. But the bias is largely unaffected by N and since the results are expressed in terms of the sampling variance, this drives the MSE up.

Variations of the number of alternatives J or the number of individual choice situations T give similar results. The more data, the higher is the advantage of Smolyak-based cubature over simulation. The parameter μ does not have any impact on the performance of the approximation methods. Loosely speaking, it merely shifts the function to be integrated. These and other results are shown in the appendix.

Figure 5.6: Monte Carlo Results, Differences by N

All results for: $T = 5, J = 5, K = 10, \mu = 1, \sigma = 0.5$



(a) $N = 500$

(b) $N = 2000$

Taste Distributions and the Cubature Rules

The original formulation of the model with normally distributed taste levels suggests the Genz/Keister cubature method, since it is appropriate for expectations over normally distributed random variables. But the problem can also be restated by another change of variables to become an expectation over standard uniform random variables. The integral in equation 5.11 can be rewritten as

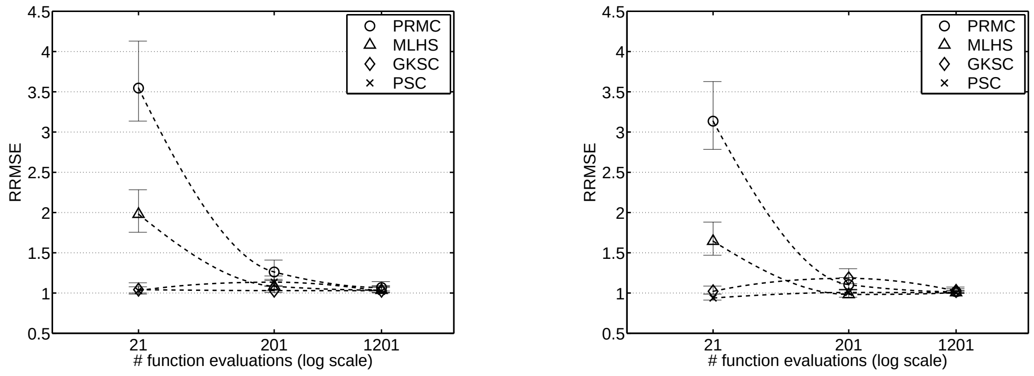
$$\int_{\mathbb{R}^K} P_i^* \left(\boldsymbol{\mu} + L(\boldsymbol{\sigma})\mathbf{e}_i\right)\right) \phi^K(\mathbf{e}_i)\, d\mathbf{e}_i = \int_{[0,1]^K} P_i^* \left(\boldsymbol{\mu} + L(\boldsymbol{\sigma})\Phi(\mathbf{u}_i)\right)\right)\, d\mathbf{u}_i, \qquad (5.12)$$

where Φ denotes the element-wise standard normal c.d.f. This is the form of integral that the Smolyak cubature method of Petras (2003) requires. The same idea can be applied vice versa: if the model specifies $\boldsymbol{\beta}_i$ to be uniformly distributed, the Petras method can be applied directly and the Genz/Keister method requires a change of variables.

Figure 5.7 compares the performance of both methods with the simulation estimators for these different model specifications. The Genz/Keister rule performs slightly better than Petras in the normal case and slightly worse in the uniform case. But the differences are insignificant and both methods clearly outperform the simulation-based estimators. This evidence suggests that the choice of the Smolyak cubature method is of minor importance for our application.

Figure 5.7: Monte Carlo Results: Petras (2003) vs. Genz/Keister (1996)

All results for: $N = 1000, T = 5, J = 5, K = 10, \mu = 1, \sigma = 0.5$



(a) Normal $\boldsymbol{\beta}_i$ (b) Uniform $\boldsymbol{\beta}_i$

5.5 Conclusions

Multidimensional integrals are prevalent in econometric estimation problems. Only for special cases, closed-form solutions exist. With a flexible model specification, the researcher frequently has to resort to numerical integration techniques. For one-dimensional integrals, Gaussian quadrature is known to be a powerful tool. Its most straightforward extension to multiple dimensions involves full tensor products of the one-dimensional rules. This implies computational costs that rise exponentially with the number of dimensions, making it infeasible for more than four or five dimensions.

The development of simulation techniques made numerical integration available in general settings. This inspired further development of models for which estimation was previously infeasible. One important example is the mixed or random parameters logit (RPL) model which became one of the most widely used discrete choice models in applied work in recent years. While simulation techniques provide a powerful and flexible approach, they often still require a lot of evaluations of the integrand until the approximation error becomes negligible. Thereby they often impose substantial computational costs.

An intuitive explanation of the advantage of quadrature over simulation in low dimensions is that it efficiently uses the smoothness of the integrand to recover its shape over the whole domain. This chapter proposes a strategy to extend this approach and its advantages to multiple dimensions with dramatically less computational costs than the full tensor rule. It is based on the general method to extend univariate operators to multivariate settings by Smolyak (1963) which is also appropriate for problems like function approximation.

Extensive Monte Carlo evidence is presented for the RPL model. The results show that the computational costs to achieve a negligible approximation error are much lower with the Smolyak-based approaches than with simulation estimators. Since they only depend on the dimension of the integral and the desired approximation level, the nodes at which to evaluate the integrand and weights can be calculated once or obtained from external sources. Given these, the Smolyak-based methods are straightforward to implement since they only involve calculating weighted averages of integrand values.

Recent research in numerical mathematics suggests possible refinements of the Smolyak approach. First, instead of predefining an approximation level in terms of the number of nodes, a critical value of the approximation error which is easily measured can be specified and the required number of function evaluations can be determined automatically. Second, the approximation does not have to be refined in each dimension symmetrically. It is also possible to invest more effort in the most relevant dimensions. These dimensions can also be determined automatically in an adaptive fashion (Gerstner and Griebel 2003). Third, quadrature-based methods can handle functions that are not well-behaved for example due to singularities by piecewise integration. These areas can also be determined in an automated fashion. The exploration of the usefulness of these extensions is left for further research.

Well-behaved integrands are typical in econometric analyses. For these, Smolyak-based cubature provides an efficient and easily applicable alternative to simulation. The efficiency gains can be invested in a reduction of computer time, an improvement of the estimates, and/or a more flexible model specification.

5.6 Appendix: Further results

This appendix presents further results mentioned but not shown in the main text.

Table 5.2: Monte Carlo Results: Differences by K: 2–10

	$\text{RRMSE}_{\text{GKSC}}$		$\frac{\text{RRMSE}_{\text{PRMC}}}{\text{RRMSE}_{\text{RRMSE}}}$		$\frac{\text{RRMSE}_{\text{MLHS}}}{\text{RRMSE}_{\text{RRMSE}}}$	
R	est.	95% CI	est.	95% CI	est.	95% CI
Dimension $K = 2$						
5	1.01	0.99–1.05	2.67	2.38–3.05	1.85	1.63–2.12
9	1.01	1.00–1.05	2.31	2.03–2.67	1.22	1.12–1.35
17	1.01	1.00–1.05	1.66	1.49–1.89	1.07	1.01–1.14
45	1.01	1.00–1.05	1.25	1.14–1.40	1.03	1.01–1.07
401	1.01	1.00–1.05	1.02	0.99–1.06	1.00	0.99–1.01
961	1.01	1.00–1.05	1.00	0.98–1.02	1.00	0.99–1.00
Dimension $K = 4$						
9	0.99	0.98–1.03	3.10	2.72–3.60	1.71	1.52–1.95
33	1.00	1.00–1.03	1.55	1.40–1.72	1.10	1.03–1.18
81	1.00	1.00–1.03	1.23	1.13–1.35	1.02	0.99–1.06
201	1.00	1.00–1.03	1.08	1.02–1.13	1.00	0.99–1.02
441	1.00	1.00–1.03	1.01	0.98–1.03	1.00	0.99–1.02
1305	1.00	1.00–1.03	1.00	0.98–1.01	1.00	0.99–1.00
Dimension $K = 6$						
13	0.98	0.96–1.02	3.03	2.67–3.53	1.52	1.39–1.70
73	1.00	1.00–1.03	1.38	1.26–1.52	1.03	0.98–1.09
257	1.00	1.00–1.03	1.07	1.01–1.14	1.01	0.99–1.02
749	1.00	1.00–1.03	1.02	0.99–1.05	1.00	0.99–1.01
2021	1.00	1.00–1.03	1.01	0.99–1.02	1.00	0.99–1.00
Dimension $K = 8$						
17	1.00	0.97–1.05	3.64	3.24–4.14	1.87	1.69–2.10
129	1.00	1.00–1.03	1.35	1.25–1.47	1.00	0.95–1.06
609	1.00	1.00–1.03	1.04	1.01–1.08	1.00	0.98–1.01
2193	1.00	1.00–1.03	1.01	0.99–1.02	1.00	0.99–1.01
Dimension $K = 10$						
21	1.04	0.98–1.12	3.40	3.04–3.89	1.90	1.72–2.12
201	1.03	1.00–1.08	1.22	1.14–1.33	1.05	1.01–1.10
1201	1.03	1.00–1.08	1.04	1.01–1.06	1.01	0.99–1.02

Table 5.3: Monte Carlo Results: Differences by K: 12–20

	$\text{RRMSE}_{\text{GKSC}}$		$\frac{\text{RRMSE}_{\text{PRMC}}}{\text{RRMSE}_{\text{RRMSE}}}$		$\frac{\text{RRMSE}_{\text{MLHS}}}{\text{RRMSE}_{\text{RRMSE}}}$	
R	est.	95% CI	est.	95% CI	est.	95% CI
Dimension $K = 12$						
25	0.96	0.90–1.03	3.36	3.07–3.74	1.94	1.78–2.14
289	1.02	1.00–1.05	1.10	1.00–1.19	0.97	0.94–1.02
2097	1.00	1.00–1.03	1.00	0.98–1.02	1.00	0.99–1.02
Dimension $K = 14$						
29	0.96	0.91–1.03	3.28	2.95–3.73	1.86	1.67–2.10
393	1.01	0.99–1.04	1.03	0.95–1.10	0.98	0.92–1.03
3361	1.00	1.00–1.02	1.00	0.98–1.02	1.01	1.00–1.03
Dimension $K = 16$						
33	0.96	0.89–1.07	3.29	2.93–3.74	1.97	1.79–2.20
513	1.00	1.00–1.03	1.05	0.96–1.14	1.00	0.95–1.06
Dimension $K = 18$						
37	1.04	0.93–1.19	3.66	3.29–4.16	2.14	1.95–2.41
649	1.00	1.00–1.03	1.03	0.97–1.10	0.98	0.92–1.04
Dimension $K = 20$						
41	1.03	0.94–1.15	3.14	2.82–3.52	1.97	1.82–2.15
801	0.96	0.91–1.01	1.10	1.05–1.15	1.05	0.99–1.10
10001	1.00	1.00–1.02	1.00	0.98–1.03	1.00	0.97–1.03

Table 5.4: Monte Carlo Results: Differences by σ

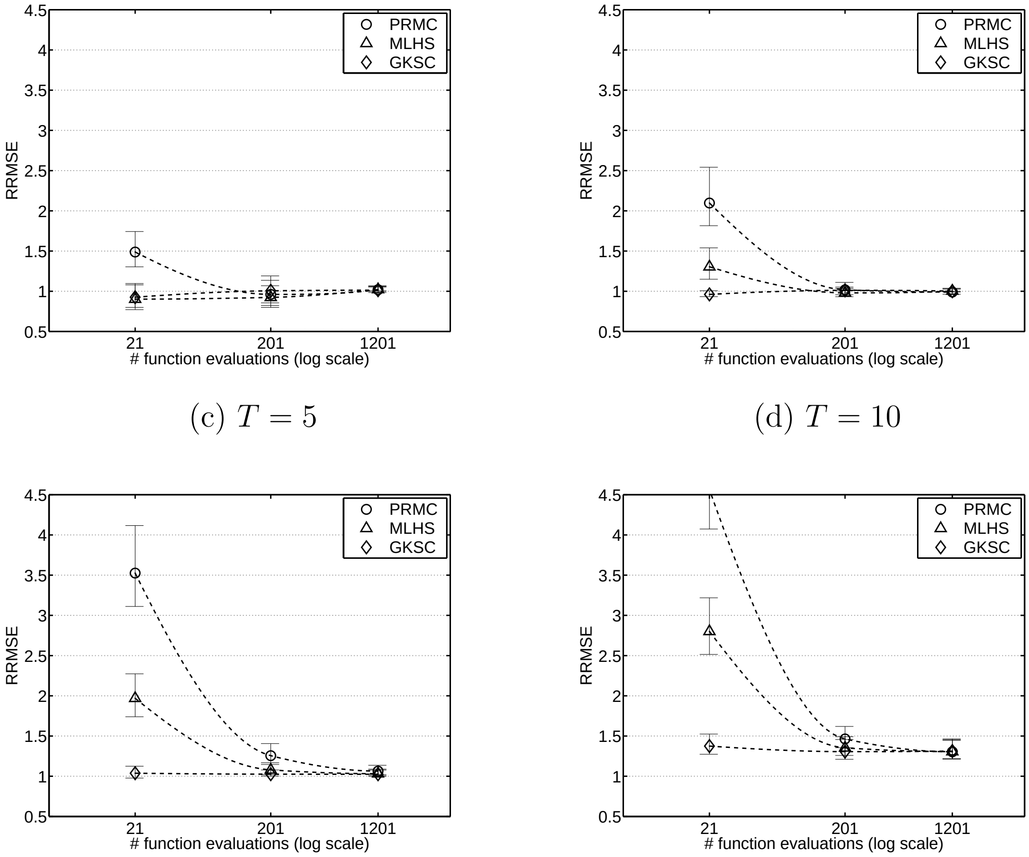
	$\text{RRMSE}_{\text{GKSC}}$		$\frac{\text{RRMSE}_{\text{PRMC}}}{\text{RRMSE}_{\text{RRMSE}}}$		$\frac{\text{RRMSE}_{\text{MLHS}}}{\text{RRMSE}_{\text{RRMSE}}}$	
R	est.	95% CI	est.	95% CI	est.	95% CI
$\sigma = 0.25$						
21	1.00	0.89–1.14	1.53	1.38–1.72	1.03	0.94–1.14
201	0.99	0.92–1.08	1.01	0.94–1.09	0.99	0.96–1.02
1201	1.00	1.00–1.03	0.97	0.93–1.01	1.02	0.99–1.04
$\sigma = 0.5$						
21	1.04	0.98–1.12	3.40	3.04–3.89	1.90	1.72–2.12
201	1.03	1.00–1.08	1.22	1.14–1.33	1.05	1.01–1.10
1201	1.03	1.00–1.08	1.04	1.01–1.06	1.01	0.99–1.02
$\sigma = 1$						
21	2.76	2.52–3.11	2.31	2.23–2.39	1.75	1.69–1.80
201	1.35	1.24–1.50	1.28	1.22–1.35	1.09	1.03–1.14
1201	1.08	1.03–1.16	1.08	1.04–1.12	1.05	1.01–1.09

Table 5.5: Monte Carlo Results: Differences by N

	$\text{RRMSE}_{\text{GKSC}}$		$\frac{\text{RRMSE}_{\text{PRMC}}}{\text{RRMSE}_{\text{RRMSE}}}$		$\frac{\text{RRMSE}_{\text{MLHS}}}{\text{RRMSE}_{\text{RRMSE}}}$	
R	est.	95% CI	est.	95% CI	est.	95% CI
$N = 200$						
21	0.95	0.84–1.09	1.44	1.28–1.64	1.10	1.01–1.21
201	0.99	0.95–1.05	0.98	0.92–1.05	0.98	0.93–1.04
1201	1.00	1.00–1.03	0.99	0.95–1.02	0.97	0.96–0.99
$N = 500$						
21	1.01	0.98–1.07	2.12	1.83–2.57	1.38	1.20–1.64
201	1.02	1.01–1.06	1.07	0.96–1.23	0.98	0.93–1.04
1201	1.02	1.01–1.06	0.95	0.91–1.00	0.99	0.98–1.01
$N = 1000$						
21	1.04	0.98–1.12	3.40	3.04–3.89	1.90	1.72–2.12
201	1.03	1.00–1.08	1.22	1.14–1.33	1.05	1.01–1.10
1201	1.03	1.00–1.08	1.04	1.01–1.06	1.01	0.99–1.02
$N = 2000$						
21	1.06	1.00–1.15	4.58	4.16–5.13	2.36	2.17–2.63
201	1.03	1.00–1.08	1.33	1.24–1.42	1.08	1.03–1.13
1201	1.03	1.01–1.08	1.04	1.02–1.07	1.00	0.99–1.02

Figure 5.8: Monte Carlo Results, Differences by T

All results for: $N = 1000, J = 5, K = 10, \mu = 1, \sigma = 0.5$



(a) $T = 2$

(b) $T = 3$

(c) $T = 5$

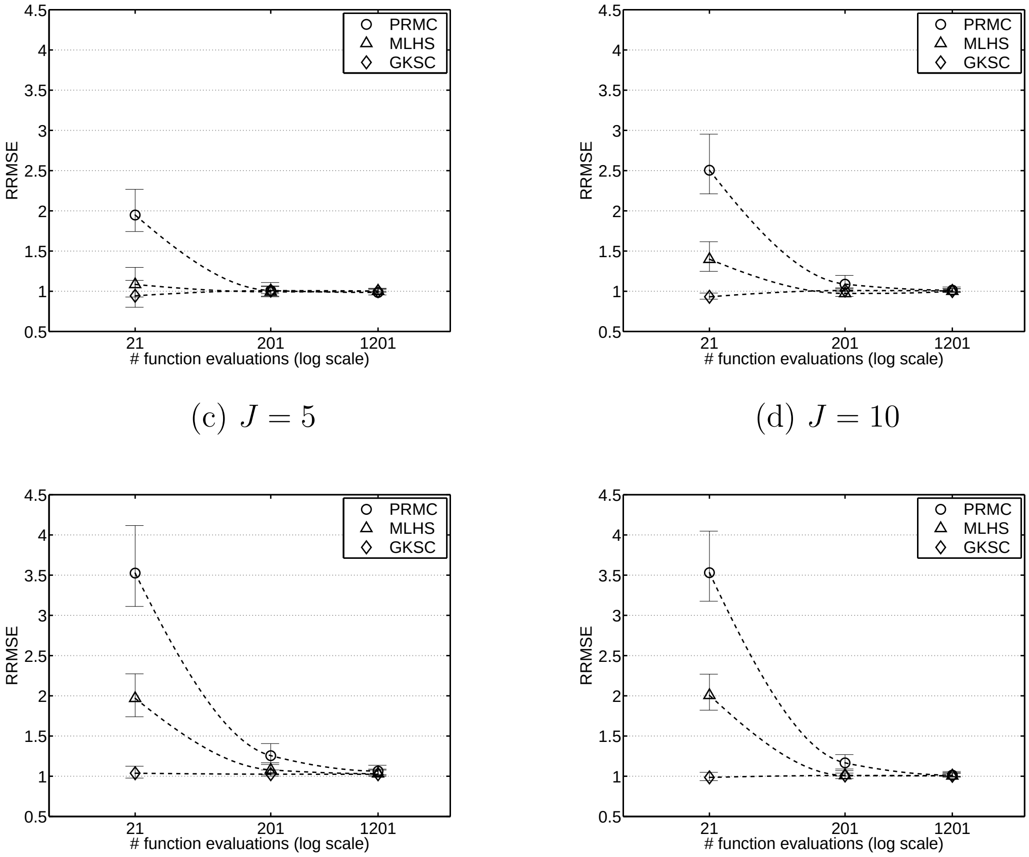
(d) $T = 10$

Table 5.6: Monte Carlo Results: Differences by T

	$\text{RRMSE}_{\text{GKSC}}$		$\frac{\text{RRMSE}_{\text{PRMC}}}{\text{RRMSE}_{\text{RRMSE}}}$		$\frac{\text{RRMSE}_{\text{MLHS}}}{\text{RRMSE}_{\text{RRMSE}}}$	
R	est.	95% CI	est.	95% CI	est.	95% CI
$T = 2$						
21	0.93	0.80–1.10	1.61	1.40–1.85	0.97	0.87–1.10
201	1.01	0.86–1.19	0.96	0.89–1.04	0.92	0.87–0.98
1201	1.01	1.00–1.05	1.01	0.97–1.05	1.01	0.99–1.03
$T = 3$						
21	0.96	0.93–1.01	2.18	1.88–2.64	1.36	1.20–1.59
201	1.01	1.00–1.04	1.01	0.94–1.09	0.97	0.92–1.03
1201	1.00	1.00–1.03	0.99	0.96–1.03	1.00	0.99–1.02
$T = 5$						
21	1.04	0.98–1.12	3.40	3.04–3.89	1.90	1.72–2.12
201	1.03	1.00–1.08	1.22	1.14–1.33	1.05	1.01–1.10
1201	1.03	1.00–1.08	1.04	1.01–1.06	1.01	0.99–1.02
$T = 10$						
21	1.37	1.27–1.52	3.32	2.94–3.72	2.04	1.81–2.27
201	1.31	1.21–1.46	1.12	1.04–1.21	1.04	0.99–1.09
1201	1.31	1.22–1.46	0.99	0.98–1.01	0.99	0.98–1.01

Figure 5.9: Monte Carlo Results, Differences by J

All results for: $N = 1000, T = 5, K = 10, \mu = 1, \sigma = 0.5$



(a) $J = 2$

(b) $J = 3$

(c) $J = 5$

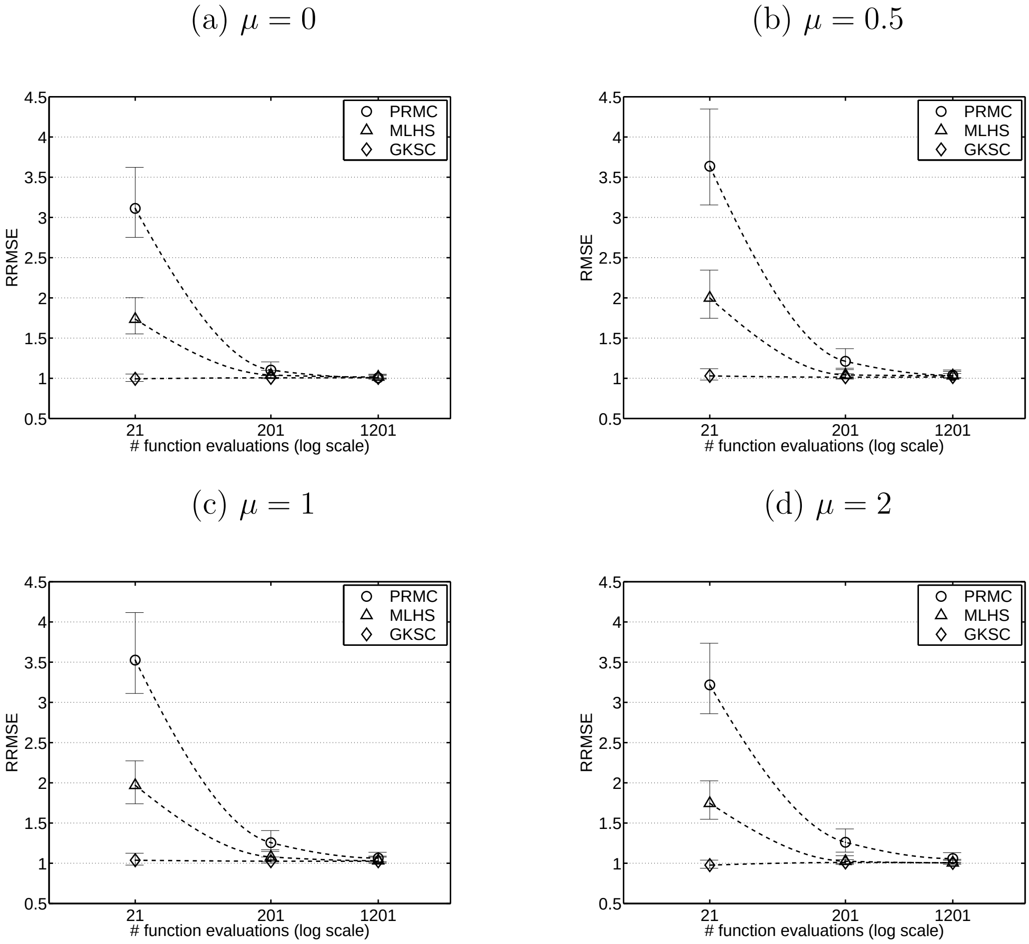
(d) $J = 10$

Table 5.7: Monte Carlo Results: Differences by J

	$\text{RRMSE}_{\text{GKSC}}$		$\frac{\text{RRMSE}_{\text{PRMC}}}{\text{RRMSE}_{\text{RRMSE}}}$		$\frac{\text{RRMSE}_{\text{MLHS}}}{\text{RRMSE}_{\text{RRMSE}}}$	
R	est.	95% CI	est.	95% CI	est.	95% CI
$J = 2$						
21	0.94	0.80–1.14	2.06	1.82–2.37	1.15	1.03–1.29
201	1.01	0.97–1.06	1.00	0.92–1.09	0.99	0.92–1.05
1201	1.00	1.00–1.03	0.98	0.95–1.02	1.00	0.98–1.02
$J = 3$						
21	0.93	0.90–0.98	2.69	2.36–3.17	1.50	1.34–1.72
201	1.01	1.00–1.04	1.08	0.99–1.18	0.96	0.91–1.02
1201	1.00	1.00–1.03	1.01	0.98–1.04	1.00	0.98–1.02
$J = 5$						
21	1.04	0.98–1.12	3.40	3.04–3.89	1.90	1.72–2.12
201	1.03	1.00–1.08	1.22	1.14–1.33	1.05	1.01–1.10
1201	1.03	1.00–1.08	1.04	1.01–1.06	1.01	0.99–1.02
$J = 10$						
21	0.99	0.95–1.05	3.58	3.22–4.05	2.04	1.87–2.25
201	1.01	1.00–1.04	1.16	1.08–1.24	1.00	0.96–1.04
1201	1.01	1.00–1.04	1.01	0.99–1.03	1.00	0.99–1.01

Figure 5.10: Monte Carlo Results, Differences by μ

All results for: $N = 1000, T = 5, J = 5, K = 10, \sigma = 0.5$



(a) $\mu = 0$

(b) $\mu = 0.5$

(c) $\mu = 1$

(d) $\mu = 2$

Table 5.8: Monte Carlo Results: Differences by μ

	$\mathrm{RRMSE}_{\mathrm{GKSC}}$		$\frac{\mathrm{RRMSE}_{\mathrm{PRMC}}}{\mathrm{RRMSE}_{\mathrm{RRMSE}}}$		$\frac{\mathrm{RRMSE}_{\mathrm{MLHS}}}{\mathrm{RRMSE}_{\mathrm{RRMSE}}}$	
R	est.	95% CI	est.	95% CI	est.	95% CI
$\mu = 0$						
21	0.99	0.96–1.05	3.13	2.79–3.57	1.75	1.58–1.96
201	1.01	1.00–1.04	1.10	1.03–1.17	1.03	1.00–1.07
1201	1.01	1.00–1.03	1.00	0.99–1.02	1.01	1.00–1.02
$\mu = 0.5$						
21	1.03	0.98–1.12	3.53	3.16–4.00	1.94	1.77–2.15
201	1.01	1.00–1.06	1.19	1.10–1.30	1.03	0.99–1.07
1201	1.02	1.00–1.06	1.02	1.00–1.05	1.02	1.00–1.03
$\mu = 1$						
21	1.04	0.98–1.12	3.40	3.04–3.89	1.90	1.72–2.12
201	1.03	1.00–1.08	1.22	1.14–1.33	1.05	1.01–1.10
1201	1.03	1.00–1.08	1.04	1.01–1.06	1.01	0.99–1.02
$\mu = 2$						
21	0.98	0.94–1.04	3.29	2.96–3.71	1.79	1.61–2.00
201	1.01	1.00–1.04	1.25	1.13–1.39	1.02	0.97–1.06
1201	1.00	1.00–1.04	1.05	1.01–1.10	1.00	0.98–1.02

Bibliography

BHAT, C. (2001): "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model," *Transportation Research B*, 35, 677–693.

BOLIĆ, M., P. M. DJURIĆ, AND S. HONG (2003): "New resampling algorithms for particle filters," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*

BÖRSCH-SUPAN, A. (1990): "On the Compatibility of Nested Logit Models with Utility Maximization," *Journal of Econometrics*, 43, 373–388.

BÖRSCH-SUPAN, A., AND V. HAJIVASSILIOU (1993): "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models," *Journal of Econometrics*, 58, 347–368.

BUTLER, J. S., AND R. MOFFIT (1982): "A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model," *Econometrica*, 50(3), 761–764.

CAMERON, S. V., AND J. J. HECKMAN (1998): "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy*, 106, 262–333.

CASE, A., AND C. PAXSON (2004): "Sex Differences in Morbidity and Mortality," Discussion paper, NBER Working Paper 10653.

CHAMBERLAIN, G. (1984): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. II, pp. 1247–1318. Elsevier, Amsterdam, New-York.

CONTOYANNIS, P., A. M. JONES, AND N. RICE (2003): "The dynamics of health in the British Household Panel Survey," Mimeo, University of York.

CROSSLEY, T. F., AND S. KENNEDY (2002): "The Reliability of Self-Assessed Health Status," *Journal of Health Economics*, 21(4), 643–658.

DOUCET, A., N. DE FREITAS, AND N. GORDON (eds.) (2001): *Sequential Monte Carlo Methods in Practice*. Springer Verlag, New York.

DOUCET, A., AND N. . N. G. DE FREITAS (2001): "An Introduction to Sequential Monte Carlo Methods," in *Sequential Monte Carlo in Practice*, ed. by A. Doucet, and N. . N. G. de Freitas, chap. 1. Springer-Verlag, New York.

ECKSTEIN, Z., AND K. WOLPIN (1999): "Why Youths Drop out of High School: The Impact of Preferences, Opportunities, and Abilities," *Econometrica*, 67, 1295–1339.

FERNÁNDEZ-VILLAVERDE, J., AND J. F. RUBIO-RAMÍREZ (2004): "Estimating Nonlinear Dynamic Equilibrium Economies: A Likelihood Approach," PIER Working Paper 04-001, Penn Institute for Economic Research.

GENZ, A., AND B. D. KEISTER (1996): "Fully Symmetric Interpolatory Rules for Multiple Integrals Over Infinite Regions with Gaussian Weight," *Journal of Computational and Applied Mathematics*, 71, 299–309.

GERSTNER, T., AND M. GRIEBEL (2003): "Dimension-Adaptive Tensor-Product Quadrature," *Computing*, 71, 65–87.

GEWEKE, J. (1989): "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57(6), 1317–1339.

——— (1996): "Monte Carlo Simulation and Numerical Integration," in *Handbook of Computational Economics Vol. 1*, ed. by H. M. Amman, D. A. Kendrick, and J. Rust, pp. 731–800. Elsevier Science, Amsterdam.

GORDON, N. J., D. J. SALMOND, AND A. F. M. SMITH (1993): "Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation," *IEE Proceedings F*, 140(2), 107–113.

GOURIÉROUX, C., AND A. MONFORT (1996): *Simulation-Based Econometric Methods*. Oxford University Press.

GREENE, W. H. (2000): *Econometric Analysis*. Prentice Hall, London, 4 edn.

HAJIVASSILIOU, V. A., AND D. L. MCFADDEN (1998): "The Method of Simulated Scores for the Estimation of LDV Models," *Econometrica*, 66(4), 863–896.

HAJIVASSILIOU, V. A., AND P. A. RUUD (1994): "Classical Estimation Methods for LDV Models Using Simulation," in *Handbook of Econometrics Vol. IV*, ed. by R. F. Engle, and D. L. McFadden, pp. 2383–2441. Elsevier, New-York.

Hamilton, J. D. (1994): “State-Space Models,” in *Handbook of Econometrics Volume 4*, ed. by R. Engle, and D. McFadden, chap. 50. North-Holland.

Haug, A. J. (2005): “A Tutorial on Bayesian Estimation and Tracking Techniques Applicable to Nonlinear and Non-Gaussian Processes,” MITRE Technical Report 05-0211.

Heckman, J. J. (1981a): “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. by C. F. Manski, and D. McFadden, pp. 179–195. MIT Press, Cambridge, Mass.

——— (1981b): “Statistical Models for Discrete Panel Data,” in *Structural Analysis of Discrete Data and Econometric Applications*, ed. by C. F. Manski, and D. McFadden, pp. 114–178. MIT Press, Cambridge, Mass.

Heiss, F. (2002): “Structural Choice Analysis with Nested Logit Models,” *Stata Journal*, 2(3), 227–252.

Heiss, F., A. Börsch-Supan, M. Hurd, and D. Wise (2005): “Predicting Disability Trajectories,” Paper prepared for the NBER Conference on Disability in Jackson, WY, October 2004.

Heiss, F., M. Hurd, and A. Börsch-Supan (2005): “Healthy, Wealthy, and Knowing Where to Live: Trajectories of Health, Wealth and Living Arrangements Among the Oldest Old,” in *Analyses in the Economics of Aging*, ed. by D. Wise, pp. 241–275. University of Chicago Press.

Heiss, F., and V. Winschel (2005): “Smolyak Cubature for Multiple Integration in Estimation Problems,” mimeo, University of Mannheim.

Hensher, D. A., and W. H. Greene (2000): “Specification and Estimation of the Nested Logit Model: Alternative Normalizations,” mimeo, New York University.

Hernandez-Quevedo, C., A. M. Jones, and N. Rice (2004): “Reporting Bias and Heterogeneity in Self-Assessed Health. Evidence from the British Household Panel Survey,” Discussion paper, University of York.

Hess, S., K. Train, and J. Polak (2005): “On the Use of a Modified Latin Hypercube Sampling (MLHS) Method in the Estimation of a Mixed Logit Model for Vehicle Choice,” *Transportation Research B*, forthcoming.

HUNT, G. L. (2000): “Alternative Nested Logit Model Structures and the Special Case of Partial Degeneracy,” *Journal of Regional Science*, 40, 89–113.

JULIER, S. J., AND J. K. UHLMANN (1997): “A New Extension of the Kalman Filter to Nonlinear Systems,” in *11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls*.

KITAGAWA, G. (1987): “Non-Gaussian State-Space Modeling of Nonstationary Time Series,” *Journal of the American Statistical Association*, 82, 1032–1041.

KOPPELMAN, F. S., AND C.-H. WEN (1998): “Alternative Nested Logit Models: Structure, Properties and Estimation,” *Transportation Research-B*, 32, 289–298.

KRÜGER, D., AND F. KÜBLER (2004): “Computing equilibrium in OLG models with stochastic production,” *Journal of Economic Dynamics and Control*, 28, 1411–1436.

LEE, L.-F. (1997): “Simulated Maximum Likelihood Estimation of Dynamic Discrete Choice Statistical Models: Some Monte Carlo Results,” *Journal of Econometrics*, 82, 1–35.

LOUVIERE, J. J., D. A. HENSHER, AND J. D. SWAIT (2000): *Stated Choice Methods*. Cambridge University Press, Cambridge.

MADDALA, G. S. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge, MA.

MCFADDEN, D. (1974): “Conditional Logit Analysis of Qualitative Choice Behavior,” in *Frontiers in Econometrics*, ed. by P. Zarembka, pp. 105–142. Academic Press, New York.

——— (1981): “Econometric Models of Probabilistic Choice,” in *Structural Analysis of Discrete Data and Econometric Applications*, ed. by C. F. Manski, and D. L. McFadden, pp. 198–272. MIT Press, Cambridge, MA.

——— (1989): “A Method of Simulated Moments for Estimation of Discrete Choice Models Without Numerical Integration,” *Econometrica*, 57, 995–1026.

MCFADDEN, D., AND K. TRAIN (2000): “Mixed MNL Models for Discrete Response,” *Journal of Applied Econometrics*, 15, 447–470, Unveröffentlichtes Manuskript, University of California, Berkeley.

MIRANDA, M. J., AND P. L. FACKLER (2002): *Applied Computational Economics and Finance*. MIT Press, Cambridge MA.

NOVAK, E., AND K. RITTER (1996): "High Dimensional Integration of Smooth Functions over Cubes," *Numerische Mathematatik*, 75, 79–97.

——— (1999): "Simple cubature formulas with high polynomial exactness," *Constructive Approximation*, 15, 499–522.

PETRAS, K. (2003): "Smolyak Cubature of Given Polynomial Degree with Few Nodes for Increasing Dimension," *Numerische Mathematik*, 93, 729–753.

PITT, M. K., AND N. SHEPHARD (1999): "Filtering via Simulation: Auxiliary Particle Filters," *Journal of the American Statistical Association*, 94, 590–599.

SALMOND, D., AND N. GORDON (2001): "Particles and mixtures for tracking and guidance," in *Sequential Monte Carlo Methods in Practice*, ed. by A. Doucet, and N. de Freitas, pp. 517–532. Springer-Verlag.

SMOLYAK, S. A. (1963): "Quadrature and Interpolation Formulas for Tensor Products of Certain Classes of Functions," *Soviet Mathematics Doklady*, 4, 240–243.

TANIZAKI, H. (1999): "Nonlinear and Nonnormal Filter Using Importance Sampling: Antithetic Monte-Carlo Integration," *Communications in Statistics, Simulation and Computation*, 28, 463–486.

——— (2001): "Nonlinear and Non-Gaussian State-Space Modeling Using Sampling Techniques," *Ann. Inst. Statist. Math.*, 53, 63–81.

——— (2003): "Nonlinear and Non-Gaussian State-Space Modeling with Monte Carlo Techniques: A Survey and Comparative Study," in *Handbook of Statistics*, ed. by D. Shanbhag, and C. Rao, vol. 21, pp. 871–929. Elsevier.

TANIZAKI, H., AND R. MARIANO (1994): "Prediction, Filtering and Smoothing in Non-Linear and Non-Normal Cases Using Monte Carlo Integration," *Journal of Applied Econometrics*, 9(2), 163–179.

TAUCHEN, G., AND R. HUSSEY (1991): "Quadrature-Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models," *Econometrica*, 59(2), 371–396.

TRAIN, K. (2003): *Discrete Choice Methods with Simulation.* Cambridge University Press.

VAN DER MERWE, R., A. DOUCET, N. DE FREITAS, AND E. WAN (2001): "The Unscented Particle Filter," in *Advances in Neural Information Processing Systems 13*, vol. 13, pp. 584–590.

YASHIN, A. I., AND K. G. MANTON (1997): “Effects of Unobserved and Partially Observed Covariate Processes on System Failure: A Review of Models and Estimation Strategies,” *Statistical Science*, 12(1), 20–34.

ZHANG, W., AND L.-F. LEE (2004): “Simulation Estimation of Dynamic Discrete Choice Panel Models with Accelerated Importance Samplers,” *Econometrics Journal*, 7(1), 120–142.

Ehrenwörtliche Erklärung

Hiermit erkläre ich ehrenwörtlich, dass ich diese Dissertationsschrift selbständig angefertigt habe und mich anderer als der in ihr angegebenen Hilfsmittel nicht bedient habe. Entlehnungen aus anderen Schriften sind ausdrücklich als solche gekennzeichnet und mit Quellenangaben versehen.

Mannheim, 08. April 2005 Florian Heiß

Lebenslauf

Persönliche Daten

Florian Heiß
Mannheim Research Institute for the Economics of Aging (MEA)
Universität Mannheim
L 13, 17
68131 Mannheim

Tel.: (06 21) 181 1858
E-Mail: mail@florian-heiss.de

Geb. am 30. 07. 1973 in Bremen

Ausbildung

seit April 2000	Promotion Universität Mannheim, Fakulät für Volkswirtschaftslehre Betreuer: Prof. Axel Börsch-Supan, Ph.D.
April 2000	Abschluß Diplom-Volkswirt, Universität Mannheim
1998 – 1999	University of California, Berkeley. Graduiertenprogramm
1994 – 2000	Universität Mannheim: Studium der Volkswirtschaftslehre
1980 – 1993	Schule, Bremen – Abschluß: Abitur

Aktuelle Tätigkeit

seit 15.04.2000	Wissenschaftlicher Mitarbeiter Universität Mannheim, Fakulät für Volkswirtschaftslehre Mannheim Research Institute for the Economics of Aging

Vorherige Tätigkeiten

1999 – 2000	Wissenschaftliche Hilfskraft. Universität Mannheim, Lehrstuhl Prof. Axel Börsch-Supan, Ph.D.
1998	Tutor für Mikroökonomik. Universität Mannheim, Lehrstuhl Prof. Konrad Stahl, Ph.D.
1996 - 1998	Wissenschaftliche Hilfskraft. Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim