Structural Estimation by Homotopy Continuation with an Application to Discount Factor Estimation*

Philipp Müller
Dept. of Business Adminstration
University of Zurich
philipp.mueller@business.uzh.ch

Gregor Reich
Dept. of Strategy and Management
Norwegian School of Economics
gregor.reich@nhh.no

This version: March 2021
First version: December 2019

Abstract

We develop a method to robustly estimate parameters of structural economic models with potential identification issues. Using homotopy path continuation applied to the MPEC formulation of the estimation problem (Su and Judd, 2012), we trace the parameter estimates and their confidence intervals as a function of a controlled parameter. As the discount factor is commonly assumed to be poorly identified in DDCMs, we trace the parameter estimates of the bus engine replacement model by Rust (1987) as a function of the discount factor β. Applying methods developed for undiscounted dynamic programming, we find that β is well identified and statistically significantly larger than 1. We establish an economically reasonable qualitative link between the decision-maker's discounting and the real interest rates: in an extended model with an unanticipated structural break in β, the decrease in β qualitatively agrees with the macroeconomic regime change in the real interest rates during the great inflation. These rates were low or even negative, and increased after Paul Volcker took office as chairman of the Fed. In this period of negative

*We are heavily indebted to Kenneth Judd and Karl Schmedders. We further thank Michel Bierlaire, Øystein Daljord, Robert Erbe, Walter Farkas, Bo Honoré, Diethard Klatte, Felix Kübler, Jasmin Maag, János Mayer, John Rust, Mark Watson, Johannes Willkomm, the participants of the 3rd Conference on Structural Dynamic Models in Chicago and seminar participants at the University of Zurich for helpful comments and discussions. We gratefully acknowledge the support of Kenneth Judd, Senior Fellow at the Hoover Institution. Reich gratefully acknowledges the support of the Fonds zur Förderung des akademischen Nachwuchses (FAN), Zurich. Finally, we thank Dave Brooks for editorial comments.

real interest rates, a time value of money argument cannot reject the estimate $\beta > 1$.

Keywords: Structural estimation, parametric optimization, mathematical programming with equilibrium constraints, homotopy path continuation, identification, multiplicity of equilibria.

1 Introduction

The estimation of the parameters of structural models through maximum likelihood estimation is known to be very efficient, and thus its application is particularly desirable in situations where, for example, the amount of data is limited. However, the underlying optimization problem spans a wide range of methodological complexity: from likelihood-maximizing estimators that allow for closed-form solutions, through relatively easy to solve smooth convex, unconstrained problems, to non-convex constrained problems, potentially with multiple local solutions. The numerical difficulty within such a class of optimization problems varies greatly and depends heavily on the problem itself (and potentially the data): for example, a lack of identification of a subset of the estimated parameters turns the Hessian of the optimization problem singular; similarly, multiplicity in the solutions to the model can induce—depending on the estimation method in use—discontinuities in the likelihood function, or failures of important constraint qualifications, such as the linear independence constraint qualification.

In this paper, we develop a *parametric optimization* approach that mitigates a variety of these problems: Suppose we fix a particular parameter of the model referred to as the *controlled parameter*—ideally a "troublemaker"—and estimate the remaining free parameters. Standard results from the optimization literature, in particular the implicit function theorem and the envelope theorem, state rigorous conditions under which the estimates of the free parameters are continuous (often even smooth) functions of the controlled parameter in a neighborhood of their initial estimates. If we restrict ourselves to equality constrained problems, we can apply *homotopy continuation* to efficiently *trace* the estimates as a function of the controlled parameter on a compact interval, and compute their corresponding likelihood function values. These likelihood values—which are optimal w.r.t. the free parameters only—can be seen as the *profile likelihood* in the controlled parameter. Moreover, we demonstrate the

estimation of confidence interval functions of the controlled and the free parameters within our approach.

A prototypical example of structural models with identification issues are dynamic discrete choice models (DDCM), whose discount factor is considered to be notoriously hard to identify.[1] Indeed, the discount factor in these models is *non-parametrically non-identified* (Rust, 1987, 1994; Magnac and Thesmar, 2002), and even under strong assumptions on the utility function, it is often only "poorly" identified (Aguirregabiria and Mira, 2010). We refer to a parameter as *poorly* identified if the profile likelihood function of this parameter is flat, that is, a change in the parameter produces only a negligible change in its likelihood, although the parameter is identified in theory. The estimation of such parameters can be cumbersome, and thus they are often fixed to some a-priori "reasonable" value. In this paper, we will apply our method to estimate the discount factor—and analyze its identification—in the seminal bus engine replacement model of Rust (1987); specifically, the author states (Rust, 1987, p. 1023):

> I was not able to precisely estimate the discount factor β. Changing β to .98 or .999999 produced negligible changes in the likelihood function and parameter estimates.

The observation by the author motivates the application of our method to the bus engine replacement model of Rust (1987). Before proceeding to the estimation results, we briefly present the model and the required numerical tools to solve a dynamic programming problem with infinite horizon for the discount factor $\beta \geq 1$: In this model, the decision-maker, Harold Zurcher, inspects the buses monthly and decides whether to carry out regular maintenance work or to replace the engine which resets the mileage to 0. While the maintenance costs increase in the mileage driven by the buses, the engine replacement costs are fixed. The decision-maker acts dynamically optimally by maximizing the expected discounted utility over an infinite horizon. The per-period utility equals the negative costs plus an unobserved utility shock. Crucially, the monthly mileage transitions of the buses are modeled by a stationary law of motion, and the discount factor is assumed to be constant. While Rust (1987) could not estimate the cost parameters jointly with the discount factor, he noted

[1] There has been substantial interest in the estimation of the discount factor in the literature on two-step estimators (going back to Hotz and Miller, 1993); see, e.g., Abbring and Daljord (2017); Komarova et al. (2018); Daljord et al. (2018).

"a systematic tendency for the estimated value of β to be driven to 1" (Rust, 1987, p. 1023), which is problematic in infinite horizon problems as the value function diverges unless the cost per state is bounded and there exists a cost-free absorbing state. The literature refers to the case of $\beta = 1$ as average cost per stage problem and solves for the value function relative to some value—the relative value—instead of the absolute values by *relative value iteration* (see, e.g., Bertsekas, 2012). This procedure also applies for $\beta < 1$ (see, e.g., Puterman, 2014) and numerically, the extension to $\beta \in \mathbb{R}^+$ is straightforward. Economically, "discount" factors greater than or equal to 1 might seem unappealing, because the expected sum of discounted payoffs might diverge; the policy function, however, still can be well-defined—especially in dynamic logit models where the policy is calculated by value differences.

We solve Rust (1987) by tracing the estimates as a function of the discount factor and find that the profile likelihood in the discount factor is well-behaved and features a locally unique maximum, which implies that the discount factor is actually well identified. The estimate $\hat{\beta}$ is strictly larger than 1 for all subsamples considered in the original paper, and even statistically significantly greater than 1 for bus groups 1–4.[2] This estimate strikes to be odd at first, as the discount factor is usually restricted to the half-open interval [0,1) using a time value of money argument.

The surprising results of $\beta > 1$ prompts the question if this might be an artifact of misspecification. The literature widely ignores the *context* in which the decision-maker in Rust (1987) acts: first, the individual bus group's embedding in the expanding bus fleet, and second, the historical macroeconomic context—the *great inflation*. The buses are grouped by their purchase date and bus type into eight bus groups which are part of the bus fleet managed by the decision-maker. Four of the eight bus groups were purchased by the company during the sample period, which evidently affected the transition probabilities of the already existing bus groups. Historically, the sample period started during the *great inflation* in which the US faced rising inflation rates leading to low or even negative *real interest rates*. After Paul Volcker became chairman of the Federal Reserve and introduced new monetary policies targeting the money supply in 1979, the nominal interest rates increased while the inflation decreased and in turn, which resulted in increasing real interest rates.

[2]We recently became aware of Blom Västberg and Karlström (2017, unpublished; available on request from the authors), which obtains a similar empirical result through a grid search for bus group 1–4 of Rust's original data set.

We find empirical evidence that both contexts affect the decision-maker by estimating an extended version of Rust (1987) with an unanticipated structural break at an unknown time (i) in the individual bus groups' transition probabilities and (ii) in the discount factor. For each possible time of the break, we apply the proposed estimation method to trace the estimates as a function of the discount factor. We reject the restricted models without a structural break for (i) and (ii) with a p-value much lower than 0.0001 using the entire dataset including bus groups 1–8. The results from both extensions reveal insights into the estimate $\hat{\beta} > 1$: In extension (i), $\hat{\beta}$ falls below 1 with a confidence interval that barely includes 1 suggesting that β is sensitive to misspecifications in the law of motion. In extension (ii), the optimal time of the structural break agrees closely with the time Paul Volcker took office as chairman of the Fed. The estimated discount factor falls from 1.03 before to 1.00 after the structural break, which, qualitatively, agrees with the rise in real interest rates. Moreover, in light of the negative real interest rates before the structural break, the estimate of $\beta > 1$ can not be rejected based on a time value of money argument.

The application of homotopy path continuation to solve parametrized systems of equations is fairly established in economics and operations research: Eaves and Schmedders (1999) as well as Besanko et al. (2010) use it to trace model solutions in dependence on a parameter value, in particular, to detect the presence of multiple solutions of the model for a given parameter value; Judd et al. (2012) use polynomial homotopy continuation to find (proveably) all solutions to systems of equations arising from equilibrium equations; for more details on the application of homotopy path continuation methods in economics, see Judd (1998) and Borkovsky et al. (2010).[3][4]

There are several examples of $\beta > 1$ in the literature as, for example, in Erdem and Keane (1996) where the authors derive the discount factor estimate $\hat{\beta} > 1$ in their model of consumer behavior, and in the New Keynesian literature to allow the zero lower bound on the nominal interest rates to bind (see e.g., Christiano et al., 2011). Regardless of the relation to the nominal or real interest rate environment, Erdem and Keane (1996) argue that there is no inherent behavioral reason to re-

[3]In a technical report, DiCiccio and Tibshirani (1991) develop a first-order approximation of the curve of an implicit many-to-one transformation of parameters to obtain its confidence intervals.

[4]Homotopy path continuation has also been applied to general parametric constrained optimization problems, as, e.g., by Tiahrt and Poore (1990) and Poore (1990). Since we restrict ourselves to equality constrained problems, the necessary mathematical results on parametric optimizations can also be found in textbooks, like Fiacco and McCormick (1990); Simon and Blume (1996); Klatte and Kummer (2002); Nocedal and Wright (2006).

strict the discount factor to $[0, 1)$. For state-dependent discounting, Stachurski and Zhang (2021) develop the theory to solve dynamic programming with state-dependent discount factors while allowing for a positive probability of $\beta \geq 1$.

The remainder of this paper is organized as follows: Section 2 defines the original estimation problem, formalizes the concept of the profile likelihood problem and its representation by first-order conditions, briefly outlines homotopy path continuation, and synthesizes the concepts. Moreover, an extension to likelihood ratio confidence intervals of the free parameters is given. Section 3 introduces the model of Rust (1987), introduces relative value iteration for discount factors $\beta > 1$, traces the estimates of the model parameters as a function of β, and establishes the qualitative link between β and the real interest rates. Section 4 concludes.

2 Estimation by Homotopy Path Continuation

2.1 The Economic Model

We closely follow the notation of Su and Judd (2012) due to its generality and consider structural economic models featuring the following three types of variables and parameters:

Structural parameters A p-dimensional vector of structural parameters, $\theta \in \Theta \subset \mathbb{R}^p$ and a scalar, real-valued parameter within a compact interval, $\beta \in [a, b]$. These include parameters of statistical distributions, preference parameters, cost functions, policy related parameters, and discount factors.

Endogenous variables An m-dimensional vector of endogenous variables, $\sigma \in \Sigma \subseteq \mathbb{R}^n$.[5]

State variables An m-dimensional vector of states, $x \in \mathcal{X} \subseteq \mathbb{R}^m$; note that the state space does not have to be discrete in order to obtain a finite dimensional σ.

An economic model further requires specific relations between these variables and parameters. We subsume these relations into a system of (nonlinear) equations denoted

[5]Since the aim of this paper is computational, we restrict ourselves to real, finite vectors σ, and therefore assume that all objects have been discretized and truncated if necessary.

as $h : \Sigma \times \Theta \times [a, b] \to \mathbb{R}^k$ with $h \in \mathcal{C}^2$, i.e., twice continusly differentiable, and require

$$h(\sigma; \theta, \beta) = 0. \tag{1}$$

The state vector x might enter h, e.g., in that (1) has to hold for all admissible values of the state variable or in (conditional) expectation, etc. Like Su and Judd (2012), we denote an equilibrium outcome of the endogenous variables as $\dot{\sigma}(\theta, \beta)$. Note that $\dot{\sigma}(\cdot)$ is *implicitly* defined by the structural parameters and the model equations. We denote the set of *all* such equilibrium outcomes for a particular parameter value as

$$\dot{\Sigma}(\theta, \beta) \equiv \{\sigma \in \Sigma : h(\sigma; \theta, \beta) = 0\}. \tag{2}$$

At this point, we do not impose any structure on $\dot{\Sigma}$ and thus allow it to be a singleton, a finite set of outcomes—commonly referred to as "multiple equilibria"—or even a non-trivial connected set if (1) is under-identified.

2.2 The Parameter Estimation Problem

To estimate the structural parameters θ and β, the econometrician obtains data on the variables predicted by the model. These are usually functions of the state variables and the endogenous variables.[6] We denote the data variables by $\tilde{x} \equiv \tilde{x}(x, \sigma)$ and a full sample of s concrete observations by $\tilde{\boldsymbol{x}}_{1:s} \equiv \{\tilde{x}_i\}_{i=1}^{s}$. We utilize the *likelihood function* as measure assessing the degree to which the parameters "explain" the observations and assume it to be twice continuously differentiable. Two main approaches to estimation based on the likelihood function exist: The nested fixed point (NFXP) approach of Rust (1987) maximizes the likelihood over θ and β, while replacing σ by an implicit function representation,

$$(\hat{\theta}, \hat{\beta}, \hat{\sigma})^{ML_i} = \arg \max_{\theta \in \Theta,\, \beta \in [a,b]} L(\theta, \beta, \dot{\sigma}(\theta, \beta); \tilde{\boldsymbol{x}}_{1:s}); \tag{3}$$

$\hat{\sigma}$ can be obtained as $\dot{\sigma}(\hat{\theta}, \hat{\beta})$.

Mathematical programming with equilibrium constraints (MPEC) by Su and Judd (2012) maximizes the *augmented likelihood* over θ, β, and σ all together, and at the

[6] Usually only a subset of the variables that the model predicts can be observed; the details of the treatment of unobserved variables generally depend on the estimation method, but often forms some kind of expectation and thus an integral; see Reich (2018).

same time imposes the model equations as constraints:

$$(\hat{\theta}, \hat{\beta}, \hat{\sigma})^{ML_c} = \arg \max_{\theta \in \Theta,\, \beta \in [a,b],\, \sigma \in \Sigma} L(\theta, \beta, \sigma; \tilde{\boldsymbol{x}}_{1:s}) \quad \text{s.t. } h(\sigma; \theta, \beta) = 0. \tag{4}$$

To develop our concepts in this paper, we rely on the constrained optimization formulation (4).

2.3 The Profile Likelihood Function

Suppose we are not able or willing to point estimate a function of structural parameters $g(\theta_j, \beta)$ on the parameter subset $\theta_j \subset \theta$; possible reasons can be its poor identification and/or multiplicity in the model solutions. Rather, we aim to obtain a *parametric solution* to the estimation problem (4) in β on $[a, b]$, as $\hat{\theta}(\beta; \tilde{\boldsymbol{x}}_{1:s})$ and $\dot{\sigma}(\hat{\theta}(\beta; \tilde{\boldsymbol{x}}_{1:s}), \beta)$.

Suppose we are not able or willing to point estimate the structural parameter β; possible reasons can be its poor identification and/or multiplicity in the model solutions. Rather, we aim to obtain a *parametric solution* to the estimation problem (4) in β on $[a, b]$, as $\hat{\theta}(\beta; \tilde{\boldsymbol{x}}_{1:s})$ and $\dot{\sigma}(\hat{\theta}(\beta; \tilde{\boldsymbol{x}}_{1:s}), \beta)$. We refer to β as the *controlled parameter*. The statistical concept to achieve this is the *profile likelihood*:

$$L_p(\beta; \tilde{\boldsymbol{x}}_{1:s}) \equiv \max_{\theta \in \Theta,\, \sigma \in \Sigma} L(\theta, \beta, \sigma; \tilde{\boldsymbol{x}}_{1:s}) \quad \text{s.t. } h(\sigma; \theta, \beta) = 0, \tag{5}$$

where $\hat{\theta}(\beta)$ and $\dot{\sigma}(\hat{\theta}, \beta)$ are the maximizers of (5) as a function of β; we skip the dependency on the sample for notational brevity below.

In principle, each evaluation of the profile likelihood for a particular β requires solving the nonlinear constrained optimization problem (5). Solving the parametrized system of nonlinear first-order necessary conditions—and checking the second-order sufficient conditions—is, however, mathematically equivalent.

Therefore, consider the Lagrangian function $\mathcal{L}$ of (5),

$$\mathcal{L}(\theta, \sigma, \mu; \beta) \equiv L(\theta, \beta, \sigma) - \mu^T h(\sigma; \theta, \beta), \tag{6}$$

with Lagrange multipliers $\mu \in \mathbb{R}^k$. The gradient of the Lagrangian (6) reads

$$\nabla_{\mu,\theta,\sigma}\mathcal{L}(\theta,\sigma,\mu;\beta) \equiv \begin{pmatrix} -h(\sigma;\theta,\beta) \\ \nabla_{\theta,\sigma}L(\theta,\beta,\sigma) - (\mu^T D_{\theta,\sigma}h(\sigma;\theta,\beta))^T \end{pmatrix}. \tag{7}$$

If the gradients of the constraints are linearly independent—that is, if the Jacobian $D_{\theta,\sigma}h$ has full rank[7]—then the Karush-Kuhn-Tucker first-order necessary conditions hold: Suppose $(\hat{\theta},\hat{\sigma};\beta)$ is a local optimum of the constrained optimization problem (5) and $\hat{\mu}$ are the corresponding Lagrange multipliers; then:

$$\nabla_{\mu,\theta,\sigma}\mathcal{L}(\hat{\theta},\hat{\sigma},\hat{\mu};\beta) = 0. \tag{8}$$

Note that the first-order necessary conditions only imply stationary points; see Section A.2 for a discussion on appropriate second-order criteria.

Suppose $(\hat{\theta},\hat{\sigma},\beta_0)$ and the corresponding $\hat{\mu}$ satisfy (8) as well as the second-order sufficient conditions—implying full rank of the Jacobian of (8)—and recall that we assumed the likelihood function and the constraints to be twice continuously differentiable. Then, there exists the maximizer function $(\hat{\theta}(\beta),\hat{\sigma}(\beta),\hat{\mu}(\beta))$ in some neighborhood $\mathcal{N} = [\beta_0 - \varepsilon, \beta_0 + \varepsilon]$ such that the optimality conditions hold and that this function is at least once continuously differentiable w.r.t. β for any $\beta \in \mathcal{N}$; see, e.g., Fiacco (1976, Thm. 2.1). Furthermore, the envelope theorem yields that the profile likelihood function exists and is smooth for $\beta \in \mathcal{N}$.

2.4 Homotopy Parameter Continuation

We use homotopy parameter continuation to efficiently solve the parameterized system of first-order necessary conditions (8) for $\beta \in [a,b]$. Homotopy parameter continuation solves the exemplary system of (nonlinear) equations $F(x,p) = 0 \in \mathbb{R}^N$ with a parameter $p \in [0,1]$ by defining a continuous map—referred to as the homotopy map—$\rho \in C^2 : \mathbb{R}^N \times [0,1] \to \mathbb{R}^N$ as

$$\rho(x,\lambda) = F(x,\lambda). \tag{9}$$

[7]This condition is referred to as the linear independence constraint qualification (LICQ) and constitutes a constraint qualification to the system of KKT conditions.

Note that the homotopy parameter $\lambda \in [0,1]$ parameterizes $F(x,\cdot)$ [8]. By construction, the desired solution set to $F(x,p)=0$ equals the zero set $\rho^{-1}(0) \equiv \{(x,\lambda) : \rho(x,\lambda) = 0\}$. In the following, we consider the zero set ξ emanating from the initial solution $(x_0, 0)$ — a subset of $\rho^{-1}(0)$.

To approximate the solution set $\{(x,p)\,|F(x,p)=0\}$, homotopy parameter continuation starts at the initial solution $(x_0,0)$ and traces ξ. Numerically, the curve $\rho(x,\lambda)$ is reparameterized in terms of the arc length and solved for by ordinary differential equation algorithms. For details, see, e.g., Borkovsky et al. (2010).

2.5 Estimation by Parametric Mathematical Programming with Equilibrium Constraints

Estimation by *parametric mathematical programming with equilibrium constraints* applies homotopy continuation to the parameterized FOCs that represent the profile likelihood (5) by defining the homotopy map as

$$\rho((\theta,\sigma,\mu),\lambda) \equiv \nabla_{\mu,\theta,\sigma}\mathcal{L}(\theta,\sigma,\mu;c(\lambda)), \tag{10}$$

where (θ,σ,μ) denotes the model parameters, the endogenous variables, and the corresponding Lagrange multipliers, respectively. The controlled parameter β is replaced by the linearly transformed homotopy parameter $c(\lambda) \equiv (1-\lambda)a+\lambda b$ allowing for a controlled parameter $\beta \in [a,b]$. By construction, all points in $\rho^{-1}(0)$ satisfy the first-order necessary conditions and thus, are stationary points denoted as $(\hat{\theta}(c(\lambda)), \hat{\sigma}(c(\lambda)), \hat{\mu}(c(\lambda)))$ for $\lambda \in [0,1]$.

We denote the subset of $\rho^{-1}(0)$ that emanates from the initial estimate $(\hat{\theta}(c(0)), \hat{\sigma}(c(0)), \hat{\mu}(c(0)))$ by ξ. We obtain the initial estimate by standard constrained optimization algorithms. Numerically tracing ξ solves efficiently for a discrete and finite subset of the stationary points $\hat{\xi} \subseteq \xi$. If the Jacobian $D_x\rho(x,\lambda)$ is regular for all points on ξ, it is sufficient to ensure the optimum type at $\lambda = 0$ (see, e.g., Poore, 1990). If there exist singularities along the path as, e.g., turning points, the second-order sufficient conditions must be checked on the interior of ξ.

The method explicitly does not require identification w.r.t. the controlled parameter β, but only w.r.t.the parameter vector θ given β. Furthermore, we allow for

[8] (Non-)linear transformations of λ allow for $p \in I \subseteq \mathbb{R}$.

multiplicity in the solutions of the model as long as the likelihood discriminates the model solutions properly;[9] see Appendix A.1 for a more detailed discussion on identification and multiplicity.

As a numerical implementation for tracing the path ξ, we propose the homotopy solution methods included in Hompack 90 (Watson et al., 1997) combined with the automatic differentiation (AD) tool CasADi (Andersson et al., 2018). We interface to the Fortran 90 library Hompack 90 through our interface M-Hompack such that the underlying model can be implemented entirely in Matlab.

2.6 Confidence Interval Functions

Estimating dimension-wise *likelihood ratio confidence intervals* (LRCI) typically involves finding their boundaries by solving for the roots of two level set problems on the profile likelihood function. This naturally integrates into our tracing approach such that we can efficiently trace dimension-wise LRCI of $\hat{\theta}$ as a function of β.

The $\gamma \cdot 100\%$ LRCI of the parameter θ_j in dependence of β reads

$$CI_\gamma(\hat{\theta}_j;\beta) \equiv \left\{ \theta_j : \max_{\theta_{-j}} L(\theta;\beta) - (L(\hat{\theta}(\beta);\beta) - 0.5\chi_1^2(\gamma)) \geq 0 \right\}, \tag{11}$$

where $\theta \equiv (\theta_j, \theta_{-j})$; $\chi_1^2(\gamma)$ is the γ quantile of the χ^2 distribution with one degree of freedom; $\hat{\theta}(\beta)$ denotes the maximum likelihood estimate in dependence of β.

To estimate the boundary of $CI_\gamma(\hat{\theta}_j;\beta)$, we consider the following system of equations:

$$\begin{pmatrix} L(\theta;\beta) - (L(\hat{\theta}(\beta);\beta) - 0.5\chi_1^2(\gamma)) \\ \nabla_{\mu,\theta_{-j},\sigma}\mathcal{L}(\mu,\theta,\sigma;\beta) \end{pmatrix} = 0. \tag{12}$$

The first equation of (12) ensures that the level of the likelihood at (θ_j, θ_{-j}) is equal to the critical value of the likelihood ratio test statistic.[10] While we vary θ_j to obtain the critical value of the likelihood, the remaining dimensions must optimize the likelihood; this is ensured by the first-order conditions of the Lagrangian w.r.t.to θ_{-j} (and σ).

[9] However, for the profile likelihood to be a unique function, a necessary condition for the Hessian of the Lagrangian (28) is to be nonsingular. This does not hold if the gradients of the constraints are linearly dependent which can happen at points where the solution is indeed unique but "splits up" into several solutions corresponding to *turning points* or *bifurcations*.

[10] In order to efficiently evaluate $L(\hat{\theta}(\beta);\beta)$, we found it useful to interpolate the discrete set $\hat{\xi}$ of previously obtained tracing results, e.g. by cubic spline interpolation.

3 Application: Bus Engine Replacement

This section applies the structural estimation by homotopy continuation approach to the bus engine replacement model by Rust (1987). We present the model in Section 3.1, the concept of relative value functions in Section 3.2, and estimate the discount factor β and the other parameters of the original model in Section 3.3. In Section 3.4, we examine our discount factor estimate of $\hat{\beta}$ larger than 1 and find empirical evidence that the context of the decision-maker acts matters: first, the embedding of bus groups in the expanding bus fleet in Section 3.4.2, and second, the historical macroeconomic context—the great inflation—in Section 3.4.3.

3.1 The Bus Engine Replacement Model of Rust (1987)

In the bus engine replacement model of Rust (1987), a manager of a fleet of public transportation buses regularly (monthly) inspects his buses. During this inspection, he assesses their roadworthiness and quantifies the need for regular maintenance work and its costs. Finally, for each bus in each period, he decides whether to carry out the work or completely overhaul (or replace) the most critical part of the bus, its engine, which would reset its odometer to 0. It is assumed that the cost of regular maintenance work increases with the bus's age (measured by its odometer), whereas engine replacement comes at a cost that is independent of a bus's age. This will expose the manager to a dynamic trade-off—namely whether to spend a (usually) larger amount of money for full replacement but reducing expected future maintenance costs or to spend less in the current period but incurring higher regular costs with an aging bus.

To account for cost parameter and transition probability heterogeneity, we partition the bus groups into partitions p and denote the set of partitions as $\mathcal{P}$. Typically, a partition comprises bus groups of the same or similar bus model and engine type. The per-period cost function for one individual bus of partition p reads

$$u(x, i; \theta_{11}^p, RC^p) + \epsilon(i) \equiv \begin{cases} -RC^p + \epsilon(1) & i = 1 \\ -\theta_{11}^p \cdot x + \epsilon(0) & i = 0 \end{cases} \tag{13}$$

where i denotes the decision (1: replacement, 0: no replacement), RC^p and θ_{11}^p are scalar, positive parameters to be estimated, and $\epsilon \equiv (\epsilon(0), \epsilon(1))$ are choice specific,

random utility shocks, which are—as it is common to assume in the discrete choice literature—modeled as two i.i.d. extreme value type I (Gumbel) random variables; note that both components of ϵ are observed by the manager prior to making his decision. The mileage of a bus is discretized to bins of $5,000$ miles with a maximum of 450,000 miles—with the state variable $x \in \{1, \dots, 90\}$ denoting the index of the bin—and assumed to follow a Markov process with conditional transition probabilities $\theta_3^p \equiv (\theta_{30}^p, \theta_{31}^p, \theta_{32}^p)$:

$$\theta_{3\Delta}^p \equiv Pr(x_{t+1} = (1 - i_t)x_t + \Delta \mid x_t, i_t; \theta_3^p),\ \Delta \in \{0, 1, 2\}, \tag{14}$$

for $x_{t+1} \in \{1, \dots, 90\}$, and zero otherwise. The structural vector $\theta^p = (RC^p, \theta_{11}^p, \theta_3^p)$ is to be estimated.

The manager is assumed to act dynamically optimally, i.e., maximizing the sum of his expected discounted future costs over an infinite time horizon,

$$V_{\theta^p}(x_t, \epsilon_t) = \sup_{f_{\theta^p}(\cdot,\cdot)} \mathbb{E}\left[\sum_{j=t}^{\infty} \beta^{j-t} \left(u(x_j, f_{\theta^p}(x_j, \epsilon_j); \theta_{11}^p, RC^p) + \epsilon(f_{\theta^p}(x_j, \epsilon_j))\right) \middle| x_t, \epsilon_t; \theta_3^p\right], \tag{15}$$

where β denotes the discount factor, and where the *decision rule* $f_{\theta^p} : x, \epsilon \mapsto i$ maps states to decisions. If $\beta \in [0, 1)$, the Bellman equation forms a sufficient optimality condition for (15):

$$V_{\theta^p}(x, \epsilon) = \max_{i \in \{0,1\}} \{u(x, i; \theta_{11}^p, RC^p) + \epsilon(i) + \beta \mathbb{E}\left[V_{\theta^p}(x', \epsilon') | x, i; \theta_3^p\right]\} \tag{16}$$

where x' and ϵ' denote next period's values of the states. When estimating the discount factor β—which was fixed in the original specification by Rust (1987)—we have to be aware of the following: The discount factor is a property of the manager; thus, when estimating β, we can not treat the bus groups separately, but use all bus groups to estimate the discount factor, while allowing for cost parameter and transition probability heterogeneity for arbitrary partitions $p \in \mathcal{P}$.

In Rust (1987), the author derives from the distributional assumptions on ϵ a

(partial) closed-form solution for the expectation over the next period's value as

$$EV_{\theta^p}(x,i) \equiv \mathbb{E}\left[V_{\theta^p}(x',\epsilon')|x,i;\theta_3^p\right] \tag{17}$$

$$= \sum_{\Delta\in\{0,1,2\}} \log\Bigg(\sum_{j\in\{0,1\}} \exp(u((1-i)x+\Delta,j;\theta_{11}^p,RC^p) \tag{18}$$

$$+ \beta EV_{\theta^p}((1-i)x+\Delta,j))\Bigg)\theta_{3\Delta}^p$$

$$\equiv T(EV_{\theta^p})(x,i). \tag{19}$$

Note that equation (17) defines an operator equation on the function $EV_\theta(\cdot,\cdot)$, and has—for $\beta\in[0,1)$—a unique solution (Rust, 1988).

Since the mileage state x is discretized, the function EV_{θ^p} is discrete, too. By assuming $EV_{\theta^p}(x,1) = EV_\theta^p(1,0)$ for all x, we can denote the finite vector of values characterizing the function EV_{θ^p} by $\overline{EV}^p \in \mathbb{R}^{90}$. A single element of $\overline{EV}^p$ corresponds to state x by $\overline{EV}_x^p$. The functional equation reduces to the system

$$\overline{EV}_x^p = T_{\theta^p}(\overline{EV}^p)_x \quad \forall x\in\{1,\ldots,90\}. \tag{20}$$

Note that due to the sparsity of the transition matrix implied by the mileage transition probabilities (14), the Jacobian matrix of (20) is very sparse.

Using data on (partial) states and decisions, the structural parameters of the model, $\theta^p \equiv (RC^p,\theta_{11}^p,\theta_3^p)$ can be estimated using maximum likelihood. Rust (1987) shows that the decision probabilities equal the multinomial logit formula

$$Pr(i_t=1|x_t;\theta^p,EV_{\theta^p}) = \Bigg(1+\exp\Big(u(x_t,0;\theta_{11}^p,RC^p)+\beta EV_{\theta^p}(x_t,0) \tag{21}$$

$$- u(x_t,1;\theta_{11}^p,RC^p) - \beta EV_{\theta^p}(x_t,1)\Big)\Bigg)^{-1}$$

due to the fact that the difference of two extreme value type 1 random variables is logistically distributed. As the likelihood for partitions p is mutually independent,

the likelihood for T periods and partition set $\mathcal{P}$ can be written as

$$\begin{aligned} L(\theta; \{x_t, i_t\}_{t=1}^T, EV_\theta) &= \prod_{p \in \mathcal{P}} L(\theta^p; \{x_t, i_t\}_{t=1}^T, EV_{\theta^p}) \\ &= \prod_{p \in \mathcal{P}} \prod_{t=1}^{T} Pr(i_t | x_t; \theta^p) Pr(x_t | x_{t-1}, i_{t-1}; \theta_3^p, EV_{\theta^p}). \end{aligned} \tag{22}$$

After taking the logarithm of the likelihood function (22), we can maximize it over θ using the mathematical programming with equilibrium constraints approach (MPEC) by Su and Judd (2012) or the nested fixed-point algorithm by Rust (1987). The MPEC optimization problem over the partition set $\mathcal{P}$ and fixed β reads

$$\begin{aligned} &\max_{(\theta^p, \overline{EV}^p)_{p \in \mathcal{P}}} \sum_{p \in \mathcal{P}} \log L(\theta^p, \overline{EV}^p; \{x_t, i_t\}_{t=1}^T) \\ &\text{s.t. } \overline{EV}_x^p = T_{\theta^p}(\overline{EV}^p)_x \quad \forall x \in \{1, \dots, 90\}, \quad \forall p \in \mathcal{P}. \end{aligned} \tag{23}$$

3.2 Relative Value Function

To allow for $\beta \geq 1$, we introduce a slightly different formulation of the value function. Rust (1987) formulates the dynamic programming problem (15) as a discounted Markov decision problem (MDP) with an infinite-horizon and the discounted utility optimality criterion. In his formulation, the discount factor β is restricted to the half-open interval $[0, 1)$ as the value function V—as well as the expected value EV—diverges otherwise; restricting the discount factor to $\beta \in [0, 1)$ leads to a discounting of future costs and in turn—under mild assumptions—to a finite V.

By global recentering, we can solve for the *relative (expected) values* $\overline{ev}^p$ by

$$\overline{ev}_x^p = \overline{EV}_x^p - \overline{EV}_k^p, \quad \forall x \in \{1, \dots, 90\}, \tag{24}$$

with some fixed $k \in \{1, \dots, 90\}$, which we fix to $k = 1$. This idea of solving for the relative values is proposed in White (1963) to solve MDPs with $\beta = 1$ where the absolute (expected) values diverge. Numerically, the extension to $\beta > 1$ is reasonable and recently Blom Västberg and Karlström (2017) (unpublished) have shown independently the applicability for the model of Rust (1987) for $\beta > 1$.[11]

[11] Furthermore, Puterman (2014) has shown the applicability of relative values to the standard problem $\beta < 1$.

Analogously to the (expected) value iteration, the relative (expected) value iteration solves the fixed-point equation

$$\overline{ev}_x^p = T_{\theta^p}(\overline{ev}^p)_x - T_{\theta^p}(\overline{ev}^p)_1 \quad \forall x \in \{1, \ldots, 90\}. \tag{25}$$

By reformulating the choice probabilities (21) as

$$Pr(i_t = 1|x_t; \theta^p) = \left(1 + \exp\left(u(x_t, 0; \theta_{11}^p, RC^p) - u(x_t, 1; \theta_{11}^p, RC^p) + \beta(\underbrace{\overline{EV}_{x_t}^p - \overline{EV}_1^p}_{\overline{ev}_{x_t}^p}) \right) \right)^{-1} \tag{26}$$

it becomes apparent that the relative expected value vector $\overline{ev}^p$ is sufficient to evaluate the choice probabilities in the likelihood function (22).

3.3 Estimation Results

This section applies the estimation by parametric mathematical programming with equilibrium constraints to the bus engine replacement model of Rust (1987). The tracing of the estimates as a function of the discount factor is motivated by the common belief that discount factors of dynamic models are often poorly identified (Aguirregabiria and Mira, 2010). For the model at hand, the author states (Rust, 1987, p. 1023):

> I was not able to precisely estimate the discount factor β. Changing β to .98 or .999999 produced negligible changes in the likelihood function and parameter estimates of (RC, θ_{11}). The reason for this insensitivity is that β is highly collinear with the replacement cost parameter RC ... Thus, if I treated β as a free parameter, the estimated information matrix was nearly singular, causing difficulties for the maximization algorithm.

At the same time, he notes that (ibid.):

> I did note a systematic tendency for the estimated value of β to be driven to 1. This curious behavior may be an artifact of computer round-off errors, or it could indicate a deeper result. ... if Harold Zurcher is actually minimizing long-run average costs, an estimation algorithm based on discounted costs would use Abel's theorem and attempt to drive β to 1.

In fact, we will confirm both statements by showing that indeed (i) the conditioning of the Hessian matrix of the estimation problem explodes as $\beta \to 1$ for absolute expected values, making a direct estimation of any β close to 1 using the absolute expected value formulation hard; and (ii) the observed "tendency for the estimated value of β to be driven to 1" is very real, as the estimate using the relative expected value formulation is even larger than 1.

To assess Rust's hypothesis, we trace the profile likelihood for the original data set as a function of the discount factor β. We consider the bus groups 1–4 in two distinct settings: (i) the "restricted" model in which the partition set equals the singleton $\mathcal{P} = \{\{1, 2, 3, 4\}\}$, i.e., bus groups 1–4 share the same cost parameters and transition probabilities, and (ii) the "unrestricted" model in which the partition set equals $\mathcal{P} = \{\{1, 2, 3\}, \{4\}\}$ which allows for cost parameter and transition probability heterogeneity across the partitions $\{1, 2, 3\}$ and $\{4\}$.

Figure 1 depicts the main estimation results for the restricted model (top) and the unrestricted model (bottom). On the left, the value of the profile likelihood is plotted as a function of β as well as the original point estimate of Rust (1987) at $\beta = .9999$ and our point maximum at the peak of the profile likelihood. The other plots depict the estimates $(\hat{\theta}_{11}^p, \widehat{RC}^p)$ as functions of β including their 75% and 95% confidence interval boundary functions.

We interpret the estimation results as follows: First, from the shape of the profile likelihood, we conclude that β is well identified. Its maximizer—and thus the maximizer of the full likelihood function—is well above 1; indeed, the likelihood ratio confidence interval reveals that β is *significantly* larger than 1 in the full sample. Assessing the original estimates of Rust (1987), we find that both the value of the likelihood function as well as the estimates for $\beta = 0.9999$ match. At the same time, the cost parameters estimates $(\hat{\theta}_{11}^p(\beta), \widehat{RC}^p(\beta))$ vanish and diverge, respectively, as β becomes larger than 1. In particular, given the degree of replacement in the data, limiting $\beta < 1$ is compensated in the estimation by making replacement too cheap relative to regular maintenance.

Table 1 reports the quantitative estimation results, including their parameter-wise 90% likelihood ratio confidence intervals for the restricted and unrestricted model. For both models, we estimate β to be greater than 1. For the restricted model, the likelihood ratio test rejects $H_0 : \beta = 1$ with a p-value of less than 1%, while for the unrestricted model, the likelihood ratio test with a p-value of about 15% cannot reject

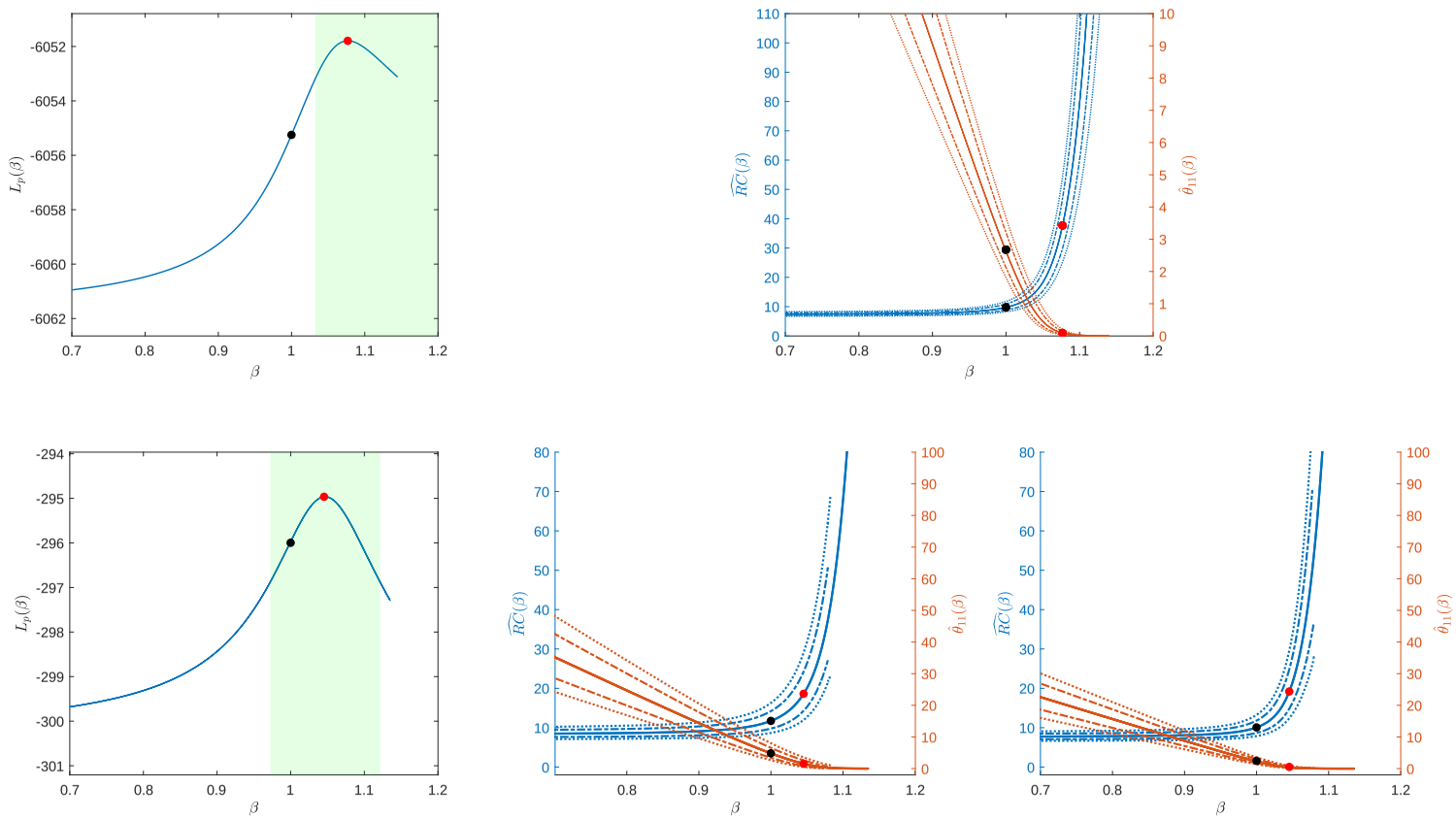


Figure 1: Top: Results for restricted model ($\mathcal{P} = \{\{1,2,3,4\}\}$). Bottom: Results for unrestricted model ($\mathcal{P} = \{\{1,2,3\},\{4\}\}$). Left: Profile likelihood $L_p(\beta)$ including 95% likelihood ratio confidence interval for β (light green). Middle/Right: Parameter estimates for (θ_{11}^p, RC^p) as functions of β (solid lines, red and blue, respectively, for $p \in \mathcal{P}$), $(\hat{\theta}_{11}^p(\beta), \widehat{RC}^p(\beta))$, and the corresponding 75% and 95% likelihood ratio confidence interval boundaries as functions of β (dash-dotted and dotted lines, respectively). Original estimates at $\beta = .9999$ (black dot) and full estimates at their maximizers (red dot).

	Restricted model	Unrestricted model	
p	$\{1,2,3,4\}$	$\{1,2,3\}$	$\{4\}$
β	1.0768	1.0467	
	$[1.0245, \infty)$	$[0.9897, 1.1042]$	
RC	37.7109	18.9955	19.8543
	$[12.999, 354.896]$	$[10.2491, 397.8716]$	$[9.0370, 426.3007]$
θ_{11}	0.0905	1.4711	0.3903
	$[0.001, 1.029]$	$[0.0366, 6.1420]$	$[0.0068, 2.6053]$
LL	-6,051.79	-6,011.51	
p-value $(H_0 : \beta = 1)$	0.0086	0.1552	
p-value $(H_0 : \theta^{\{1,2,3\}} = \theta^{\{4\}})$		0.0000	

Table 1: Joint estimation of the main structural parameters, (θ_{11}, RC, β), for the restricted model ($\mathcal{P} = \{\{1,2,3,4\}\}$) and the unrestricted model ($\mathcal{P} = \{\{1,2,3\},\{4\}\}$); 90% likelihood ratio confidence intervals are reported in brackets (if available; half-closed interval containing the confidence interval otherwise); p-value is reported for the likelihood ratio test of $H_0 : \beta = 1$.

H_0 at conventional significance levels. While $\beta > 1$ is not statistically significant for the unrestricted model, it is economically significant.

To verify the results' numerical accuracy, we report the violation of the first-order conditions in terms of the L^∞ norm as a function of β in Figure 2 (left); we conclude that the approximation error is well within accepted numerical tolerances. We further assess the numerical error pattern by investigating the conditioning of the problem. The right side of Figure 2 depicts the condition number of the augmented Jacobian once for absolute expected values (red) and once for relative expected values (blue).[12] The condition of the augmented Jacobian deteriorates in both cases. Although the condition number also diverges for relative expected values $\overline{ev}$, it only does so for some $\beta > 1$.[13] This is very much in line with the near-singular information matrix

[12]The augmented Jacobian equals the Jacobian of the first-order conditions, with the derivative w.r.t. the tracing parameter added as an additional column

[13]Also the convergence issues reported by Blom Västberg and Karlström (2017) for $\beta > 1.05$ can be explained by the conditioning of the problem. We argue that NFXP might fail for $\beta > 1$ if the starting points are not chosen very locally to the solution, as it is done using homotopy path continuation with predictor steps.

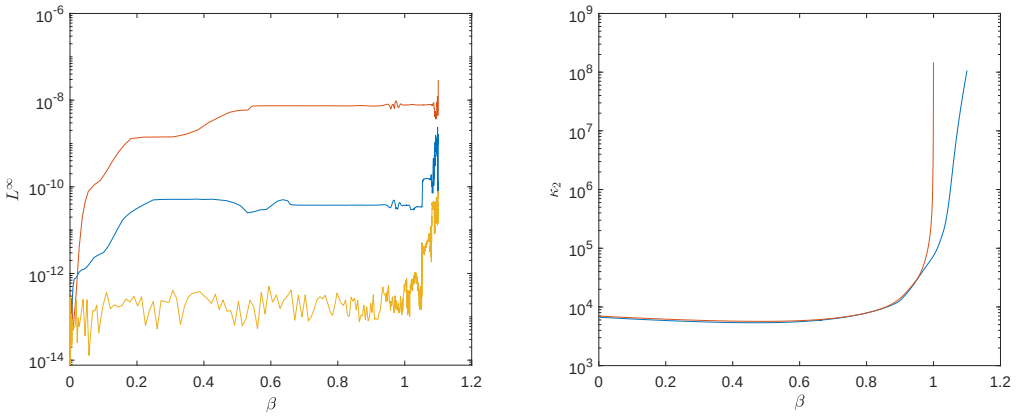


Figure 2: Left: Violation of the first-order conditions as a function of β for three error tolerance configurations of Hompack90 (1e-08, 1e-10, 1e-12). Right: Condition number of the augmented Jacobian, i.e., the Jacobian of the first-order conditions with an additional column containing the partial derivative w.r.t. the tracing parameter, as a function of β, using absolute expected values (red) and relative expected values (blue).

reported by Rust (1987) for β approaching 1 when using absolute expected values.

3.4 The Context of Rust (1987) and its Effect on Discounting

The literature widely ignores the *context* in which the decision-maker in Rust (1987) acts: first, the individual bus group's embedding in the expanding bus fleet, and second, the historical macroeconomic context—the *great inflation*. By adding a structural break (i) to the transition probabilities and (ii) to the discount factor, we find statistical evidence that both affect the decision-maker's replacement behavior. Extension (i) suggests that the parameter β is sensitive to misspecifications in the transition probabilities; extension (ii) establishes a qualitative link between the change in the discount factor estimates around the structural break and the change in the *real interest rates* during the great inflation. We argue that due to the partly prevailing negative real interest rates during this period, the estimate of $\beta > 1$ cannot be rejected by a time value of money argument.

3.4.1 Rust (1987) with Bus Groups 1–8

This section presents the estimation results for the entire data set, including bus groups 1–8. Rust (1987) uses only bus groups 1–4 for the estimation, which covers a subset of the sample period. We use bus groups 1–8 to have data for the whole sample period. Analogously to Section 3, we account for the buses' heterogeneity by allowing for cost heterogeneity. The buses group naturally into three partitions matching the bus and engine types: (i) Bus groups 4, 5, and 7 of bus model 5308A and engine type 8V71, (ii) bus groups 6 and 8 of bus model 4523A and engine type 6V71, and (iii) bus groups 1, 2, and 3, which are of heterogeneous type, but not divided further.[14] Thus, the partition set comprising the three partitions equals $\mathcal{P} = \{\{1,2,3\},\{4,5,7\},\{6,8\}\}$. Moreover, we allow for individual transition probabilities for each bus group. This extends the system of constraints of the standard MPEC formulation to one Bellman equation for each bus group which we solve simultaneously.

Figure 3 shows the estimation results: On the left, it depicts the profile likelihood as a function of β and on the right, the cost parameter estimates for each partition in $\mathcal{P}$.[15] The findings are qualitatively consistent with the findings in Section 3.3 for $\mathcal{P} = \{\{1,2,3\},\{4\}\}$: The profile likelihood indicates that the maximum likelihood estimate of β is larger than 1, and not even the confidence interval includes 1. The shape of the profile likelihood implies that the maximum likelihood estimates are well-identified. The replacement cost parameters diverge with increasing β whereas the maintenance cost parameters vanish.

3.4.2 Structural Break in the Transitions Probabilities

Rust (1987) uses Harold Zurcher's maintenance records from December 1974 to May 1984, comprising monthly observations on mileage and maintenance decisions. Figure 4 depicts the number of buses per bus group for each month. Evidently, bus groups 5–8 were purchased before the dataset begins, while bus groups 1–4 were purchased afterward, which almost tripled the bus fleet from 58 to 162 buses. We have no

[14] In grouping bus groups 1, 2, and 3 we follow Rust (1987). A further subdivision of these actually heterogeneous bus types is not identified as there exist no observations of engine replacements in bus groups 1 and 2.

[15] For expositional purposes, we drop the confidence interval functions that were included in Section 3.3.

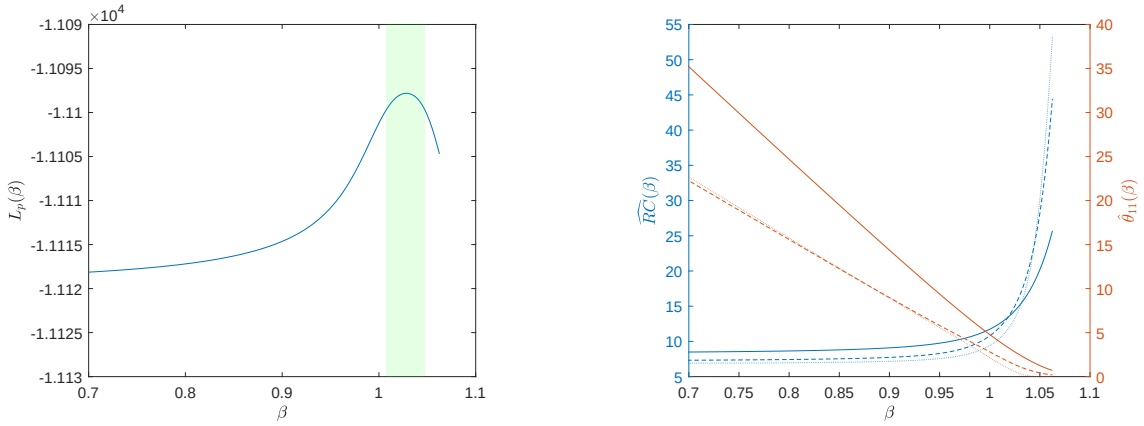


Figure 3: Results for the unrestricted model ($\mathcal{P} = \{\{1,2,3\},\{4,5,7\},\{6,8\}\}$). Left: Profile likelihood $L_p(\beta)$ including 95% likelihood ratio confidence interval for β (light green). Right: Parameter estimates for (θ_{11}^p, RC^p) as functions of β (red and blue, respectively) for partitions $\{1,2,3\},\{4,5,7\}$, and $\{6,8\}$ (solid, dashed, and dotted).

information on whether these buses were purchased to extend the bus service with new routes, increase the frequency on existing routes, or relieve existing buses. However, the purchase of the buses evidently impacts the already existing buses as depicted in Figure 5. This Figure plots a 12-month centered moving average of the aggregated mileage transitions for bus groups 3–6 and the entire fleet. After the addition of bus group 3, the aggregated mileage transitions for bus groups 1, 2, and 4 drop; bus group 4 is particularly affected and its aggregated mileage transitions drop by almost 40%. A reasonable explanation for this could, e.g., be a less frequent scheduling of the "old" bus groups or a change in the company's route schedule.

The original model assumes the transition probabilities of the bus groups, $\theta_{3,g}$, to be stationary. We relax this stationarity assumption by introducing a single unanticipated structural break in the transition probabilities for bus groups 4–8.[16] We denote the transition probabilities before the structural break for bus group g by $\theta_{3,g}^1$, and after by $\theta_{3,g}^2$. Formally, the decision-maker considers $\theta_{3,g}^1$ to be constant from today to infinity in any month before the structural break at time $t_{\theta_{3,g}}$. Starting from $t_{\theta_{3,g}}$, he considers $\theta_{3,g}^2$ to be constant from today to infinity. This extends the system of constraints to two Bellman equations per bus group for bus groups 4–8, which we solve

[16] We assume that bus groups 1–3 have no structural break as the company purchased these late, and thus, we only have few observations.

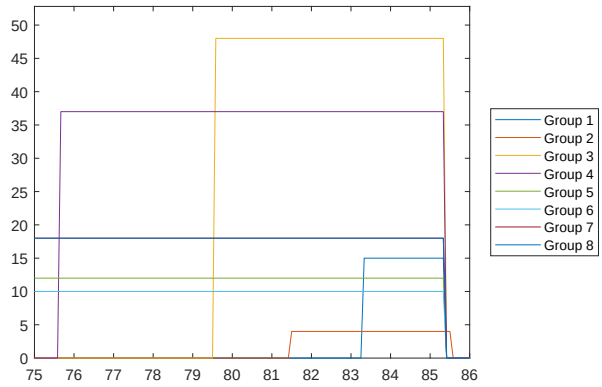


Figure 4: Number of buses per bus group in each month.

simultaneously, once with $\theta^1_{3,g}$ and once with $\theta^2_{3,g}$. We follow the literature (Casini and Perron, 2018) in choosing the estimates $\hat{t}_{\theta_{3,g}}$ that maximize the likelihood.

Figure 6 depicts the estimation results: On the left, it shows the profile likelihood as a function of β with a 95% confidence interval around $\hat{\beta}$ and on the right, the corresponding cost parameter estimates. The shape of the profile likelihood indicates a well-identified β with a maximum likelihood estimate below one, and the confidence interval barely includes 1. The other maximizers show qualitatively the same behavior as before. While the specification of the structural break—a single unanticipated structural break—is ad hoc, we can reject the restricted model with a p-value of below 10^{-16}.

The discount factor estimate drops from larger than 1 in the original model specification to less than 1 after relaxing the stationarity assumption of the transition probabilities. This suggests that β is sensitive to misspecifications in the law of motion, which is reasonable as both "discount" future costs. The decision-maker forms his expectation about future costs using the law of motion. A misspecification which causes too high probabilities for low transitions yields too low expected costs. The discount factor discounts these expected costs and thus, a high biased discount factor estimate can compensate, to some extent, the misspecification the law of motion.

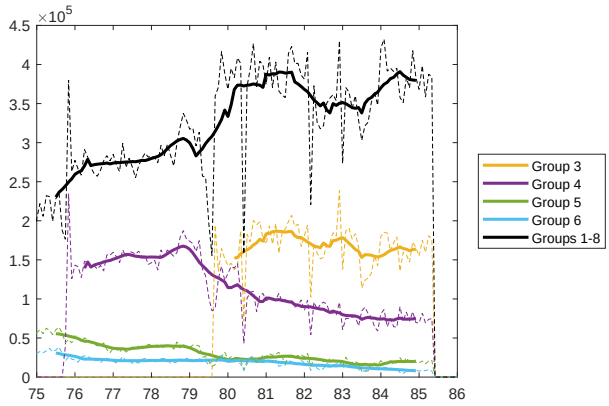


Figure 5: Centered 12-month moving average aggregated mileage (solid) and monthly mileage transitions (dashed) for bus groups 3–6 and the entire fleet.

3.4.3 Structural Break in the Discount Factor β

A period of considerable economic turmoil lasted from 1965 to the mid-1980s and is often referred to as *great inflation*. In the US, inflation rates rose from below 2% to above 15% in 1979, leading to an economic environment with striking price uncertainty. Even though the US faced periods of high inflation before, the great inflation was the only instance of a long period of inflation during peace times. During its peak, it made "every business decision a speculation on monetary policy" (**?**). Due to the low nominal interest rates compared to the inflation, the *real interest rates* were widely negative starting from 1974, as depicted in Figure 7.

This lasted until President Jimmy Carter nominated Paul Volcker as chairman of the FED in 1979. Already in his confirmation hearing in July 1979, he pledged to make fighting inflation his top priority. While not specific about any planned policies, he made clear that money supply had been "rising at a pretty good clip" even though there was no evidence the nation was "suffering grievously from a shortage of money" (**?**). After an unscheduled Federal Open Market Committee Meeting on October 6, 1979, Paul Volcker announced new monetary policies which targeted the growth rate of money stock in the economy instead of stabilizing the federal funds rate as has been the practice before. This led to a rise in the federal funds rate to 19%, and falling inflation rates. Hence, the real interest rates turned positive shortly after the beginning of the monetary policies.

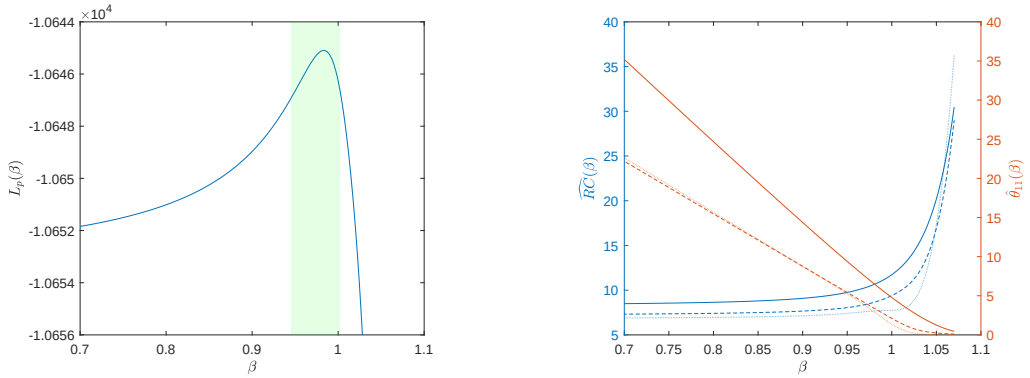


Figure 6: Results for the unrestricted model ($\mathcal{P} = \{\{1,2,3\},\{4,5,7\},\{6,8\}\}$) with a structural shock to the transition probabilities for bus groups 4-8. Left: Profile likelihood $L_p(\beta)$ including 95% likelihood ratio confidence interval for β (light green). Right: Parameter estimates for (θ_{11}^p, RC^p) as functions of β (red and blue, respectively) for $p = \{1,2,3\},\{4,5,7\},\{6,8\}$ (solid, dashed, dotted).

The decision-maker in Rust (1987) maximizes the sum of discounted utilities which equal the negative expected costs plus utility shock. We argue that the nominal value of these costs increased over time due to the prevailing high inflation. It is reasonable to assume that the decision-maker expects this trend to continue after observing several years of high inflation. In the model, however, the cost parameters are constant over time; consequently, we argue that they are stated in real terms. This implies that the decision-maker's discount factor relates to the real interest rate instead of the nominal interest rate to account for inflation.

We argue that this historical context constitutes a natural experiment. The decision-maker acts in a period of a largely unanticipated macroeconomic regime change in the real interest rates: from low or even negative to economically significantly higher rates. We hypothesize that this regime change affects the decision-maker's discounting. To test this hypothesis, we extend the model from the previous section with an unanticipated structural break to the discounting. Before the structural break at t_β, Zurcher discounts by a constant discount factor β_1 and after the shock he discounts by β_2. Formally, this implies that he considers β_1 to be constant from today to infinity in any month before the structural break at time t_{θ_β}. Starting from t_β, he assumes β_2 to be constant from today to infinity. This again extends the system of constraints to a Bellman equation for each parameter combination we solve

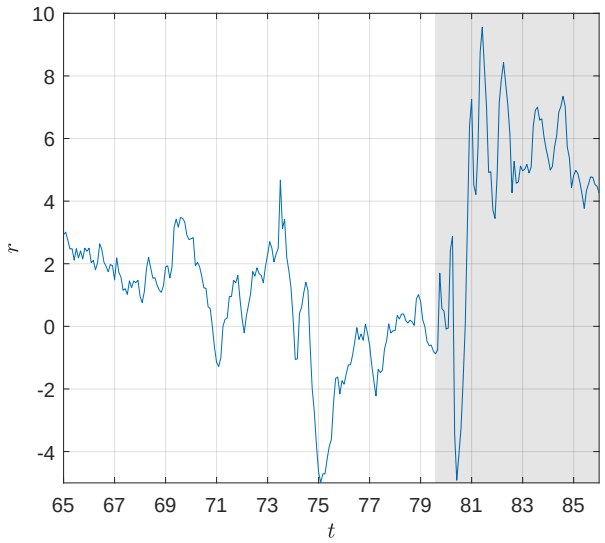


Figure 7: Ex-post real interest rate calculated as the difference of CPI and the federal fund rate. Grey shaded area depicts period of Paul Volcker as chairman of the Fed.

simultaneously.

We employ our proposed homotopy continuation estimation approach by tracing the parameter estsimates as a function of the difference of the discount factor before and after the break, $\Delta\beta = \beta_1 - \beta_2$, as the controlled parameter. This is a natural choice as it enables us to examine the degree of identification of the potentially poorly identified $\Delta\beta$. Starting from the restricted model, we fix t_β and trace the parameter estimates for all $t_\beta \in \{\text{Jan } 1976, \ldots, \text{Dec } 1983\}$.

Figure 8 depicts the profile likelihood $L(\Delta\beta, t_\beta)$ which is two-dimensional—continuous in $\Delta\beta$ and discrete in t_β. The profile likelihood in $\Delta\beta$—for fixed t_β—allows for a similar analysis as before. The black points denote the restricted model for which $\Delta\beta = 0$. The profile likelihood increases in $\Delta\beta$ up to its peak for all $t_\beta \in \{\text{Jan } 1976, \ldots, \text{Dec } 1983\}$. The shape and location of the peaks indicate that $\Delta\beta$ is well-identified and its estimate larger than zero, i.e., $\hat{\beta}_1 > \hat{\beta}_2$.

Next, we further profile $L(\Delta\beta, t_\beta)$ w.r.t. $\Delta\beta$ by taking the respective maximum $L(t_\beta) = \max_{\Delta\beta} L(\Delta\beta, t_\beta)$. This reduces the two-dimensional profile likelihood to the one-dimensional profile likelihood depicted in Figure 9. The Figure shows $L(t_\beta)$, $\beta_1(t_\beta)$, and $\beta_2(t_\beta)$ as functions of t_β. The green shaded area depicts the area for which

the likelihood is not worse than the typical 95% likelihood ratio confidence interval around the maximum likelihood estimate of the time of the structural break, $\hat{t}_\beta$. The vertical black line denotes the time Paul Volcker took office. The log-likelihood takes its maximum for $\hat{t}_\beta$ equal to September 1979. The estimates for the discount factors equal $\hat{\beta}_1 = 1.03$ and $\hat{\beta}_2 = 1.00$. The t_β within the confidence interval include February 1979 to November 1979 and October 1978 to December 1978. The orange dashed horizontal line denotes the log-likelihood value for the restricted model with $\beta_1 = \beta_2$, which we reject with a p-value of $2.17 \cdot 10^{-09}$.

We focus in the following on the maximum likelihood estimate $\hat{t}_\beta$ equal to September 1979. September 1979 corresponds to the month after Paul Volcker took office as chairman of the Fed, who introduced monetary policies which led to economically significantly higher real interest rates. The change in the discount factor estimates from $\hat{\beta}_1 = 1.03$ to $\hat{\beta}_2 = 1.00$ follows the expected economically sensible qualitative link to the real interest rates: a rise in the real interest rates leads to a fall in the discount factor. The estimate of the discount factor before September 1979, $\hat{\beta}_1 = 1.03$, falls in a period of largely negative real interest rates. In this economic environment, the time value of money is inversed, and thus, the discount factor estimate larger than 1 cannot be rejected. In fact, it agrees with the popular notion of discout factor $= 1/(1 + \text{interest rate})$. The estimate of the discount factor starting from September 1979, $\hat{\beta}_2 = 1.00$, corresponds to an agent that maximizes his long-run average utility.

3.4.4 Implied Demand for $\beta > 1$

Econometricians are, in general, not only interested in the estimates but also study policy questions using the estimated models. In Rust (1987), econometricians could, e.g., study the impact of increasing part costs for the replacement engine on the expected annual engine replacements (the implied demand). While the policy question itself is of no interest to us, we use it to study the implications of a discount factor estimate of $\hat{\beta} > 1$.

Figure 10 traces the expected annual engine replacement as a function of the replacement cost and the discount factor for a single bus with the characteristics of bus group 4. For $\beta = 0$, the expected annual engine replacement is sensitive to changes in the replacement bus engine cost. In the region with low part costs, the demand is over-predicted, while in the region with high part costs, the demand is

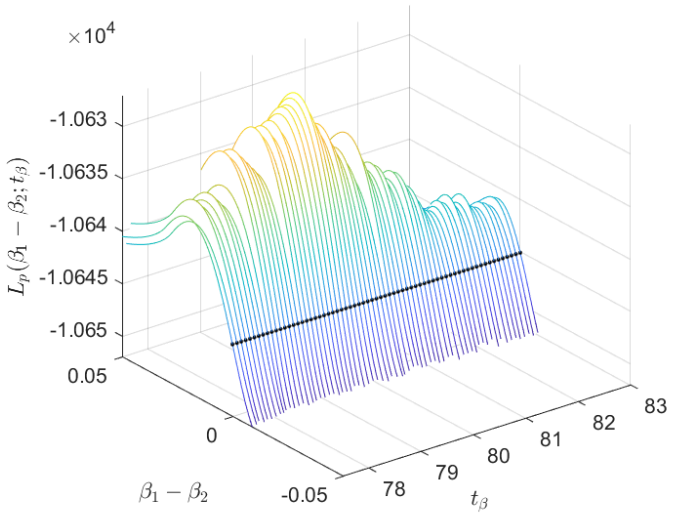


Figure 8: Profile likelihood $L(\Delta\beta, t_\beta)$ as function of $\Delta\beta$ and the time of the structural break t_β. Restricted model with $\Delta\beta = 0$ denoted by black dots.

under-predicted. The same holds for the implied demand for $\beta = 0.9999$. The curve of the expected annual engine replacements flattens with increasing β. This appears to be a continuous deformation of the demand curve in β without the structural change one might expect at $\beta = 1$. The prediction for $\beta > 1$ is a reasonable extension of the model predictions for $\beta < 1$: the price elasticity of demand continues to decrease with increasing β.

4 Conclusion

In this paper, we presented the necessary mathematical, statistical, and numerical tools to trace the maximum likelihood estimates of the structural parameters of a model and its confidence intervals, in dependence on a controlled parameter, based on the profile likelihood and using homotopy path continuation. Applying the method to the bus engine replacement model of Rust (1987), we find that—in contrary to a common belief—the discount factor is well identified. The application of relative value iteration allows us to solve the model for values of the discount factor equal to or beyond 1, and we can actually show the corresponding estimate to lie above 1 with

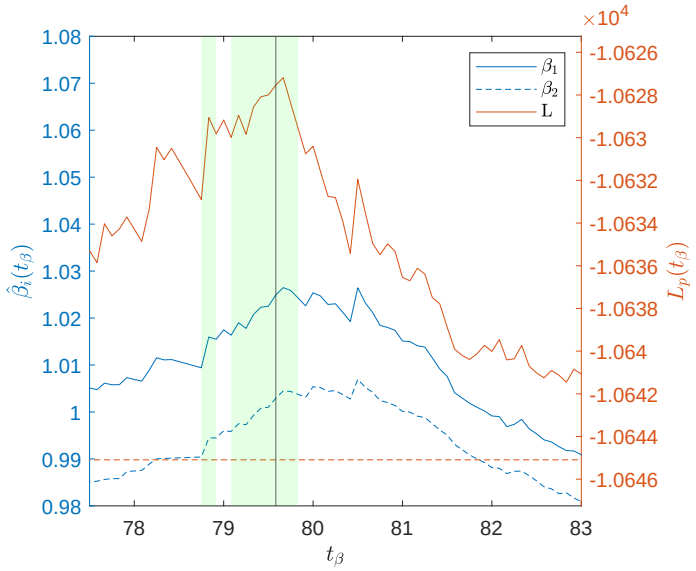


Figure 9: Profile likelihood $L(t_\beta)$ (orange), and $\beta_1(t_\beta)$ (blue solid) and $\beta_2(t_\beta)$ (blue dashed) maximum likelihood estimates for $\mathcal{P} = \{\{1, 2, 3\}, \{4, 5, 7\}, \{6, 8\}\}$ as a function of the structural break at time t_β. The t_β for which the log-likelihood is not worse than the typical 95% confidence interval w.r.t. t_β (green shaded). Log-likelihood for the restricted model with $\beta_1 = \beta_2$ (orange dashed). Paul Volcker takes office as chairman of the Federal Reserve (black solid vertical line).

statistical significance. We present further insight by examining the context in which the decision-maker acts: First, a misspecification in the transition probabilities biases the discount factor to an estimate larger than 1, and second, a historical macroeconomic regime switch qualitatively relates the real interest rate to the discount factor.

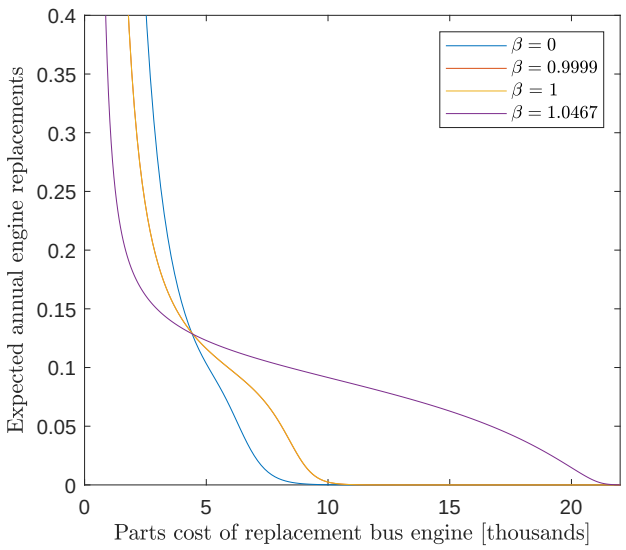


Figure 10: Expected annual engine replacement in group 4 as a function of the part cost of replacement bus engine in 1985 dollars for varying discount factors.

References

Abbring, J. H. and Daljord, Ø. (2017). Identifying the Discount Factor in Dynamic Discrete Choice Models. *Working paper*.

Aguirregabiria, V. and Mira, P. (2010). Dynamic Discrete Choice Structural Models: A Survey. *Journal of Econometrics*, 156(1):38–67.

Andersson, J. A. E., Gillis, J., Horn, G., Rawlings, J. B., and Diehl, M. (2018). CasADi – A software framework for nonlinear optimization and optimal control. *forthcoming in: Mathematical Programming Computation*.

Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control* , volume 2 of *Approximate dynamic programming*. Athena Scientific, Belmont, MA, 4th edition.

Besanko, D., Doraszelski, U., Kryukov, Y., and Satterthwaite, M. (2010). Learning-by-Doing, Organizational Forgetting, and Industry Dynamics. *Econometrica: Journal of the Econometric Society*, 78(2):453–508.

Blom Västberg, O. and Karlström, A. (2017). Discount factors greater than or equal to one in infinite horizon dynamic discrete choice models. *unpublished; available on request from the authors*.

Borkovsky, R. N., Doraszelski, U., and Kryukov, Y. (2010). A User's Guide to Solving Dynamic Stochastic Games Using the Homotopy Method. *Operations Research*, 58(4):1116–1132.

Casini, A. and Perron, P. (2018). Structural Breaks in Time Series. *arXiv*.

Christiano, L., Eichenbaum, M., and Rebelo, S. (2011). When Is the Government Spending Multiplier Large? *Journal of Political Economy*, 119(1):78–121.

Daljord, Ø., Nekipelov, D., and Park, M. (2018). A Simple and Robust Estimator for Discount Factors in Optimal Stopping Dynamic Discrete Choice Models. *Working paper*.

DiCiccio, T. J. and Tibshirani, R. (1991). On the implementation of profile likelihood methods. Technical report.

Eaves, B. C. and Schmedders, K. (1999). General equilibrium models and homotopy methods. *Journal of Economic Dynamics and Control*, 23(9-10):1249–1279.

Erdem, T. and Keane, M. (1996). Decision-Making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets. *Marketing Science*, 15(1):1–20.

Fiacco, A. V. (1976). Sensitivity analysis for nonlinear programming using penalty methods. *Mathematical Programming*, 10(1):287–311.

Fiacco, A. V. and McCormick, G. P. (1990). *Nonlinear Programming*. Sequential Unconstrained Minimization Techniques. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Hotz, V. J. and Miller, R. A. (1993). Conditional Choice Probabilities and the Estimation of Dynamic Models. *The Review of Economic Studies*, 60(3):497–529.

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

Judd, K. L., Renner, P., and Schmedders, K. (2012). Finding all pure-strategy equilibria in games with continuous strategies. *Quantitative Economics*, 3(2):289–331.

Klatte, D. and Kummer, B. (2002). *Nonsmooth Equations in Optimization*. Regularity, Calculus, Methods and Applications. Springer, Berlin, Heidelberg.

Komarova, T., Sanches, F., Silva Junior, D., and Srisuma, S. (2018). Joint analysis of the discount factor and payoff parameters in dynamic discrete choice models. *Quantitative Economics*, 9(3):1153–1194.

Magnac, T. and Thesmar, D. (2002). Identifying Dynamic Discrete Decision Processes. *Econometrica: Journal of the Econometric Society*, 70(2):801–816.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, Berlin, Heidelberg.

Poore, A. B. (1990). Bifurcations in parametric nonlinear programming. *Annals of Operations Research*, 27(1):343–369.

Puterman, M. L. (2014). *Markov Decision Processes*. Discrete Stochastic Dynamic Programming. John Wiley & Sons, Hoboken, NJ.

Reich, G. (2018). Divide and Conquer: Recursive Likelihood Function Integration for Hidden Markov Models with Continuous Latent Variables. *Operations Research*, 66(6):1457–1470.

Rust, J. (1987). Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica: Journal of the Econometric Society*, 55(5):999–1033.

Rust, J. (1988). Maximum Likelihood Estimation of Discrete Control Processes. *SIAM Journal on Control and Optimization*, 26(5):1006–1024.

Rust, J. (1994). Structural estimation of markov decision processes. In McFadden, D. and Engle, R. F., editors, *Handbook of Econometrics*, pages 3081–3143. Elsevier, New York, NY.

Simon, C. P. and Blume, L. (1996). *Mathematics for Economists*. WW Norton & Company, New York, NY.

Stachurski, J. and Zhang, J. (2021). Dynamic programming with state-dependent discounting. *Journal of Economic Theory*, 192:105190.

Su, C.-L. and Judd, K. L. (2012). Constrained Optimization Approaches to Estimation of Structural Models. *Econometrica: Journal of the Econometric Society*, 80(5):2213–2230.

Tiahrt, C. A. and Poore, A. B. (1990). A bifurcation analysis of the nonlinear parametric programming problem. *Mathematical Programming*, 47(1-3):117–141.

Watson, L. T., Sosonkina, M., Melville, R. C., Morgan, A. P., and Walker, H. F. (1997). Algorithm 777: HOMPACK90: a suite of Fortran 90 codes for globally convergent homotopy algorithms. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):514–549.

White, D. J. (1963). Dynamic programming, Markov chains, and the method of successive approximations. *Journal of Mathematical Analysis and Applications*, 6(3):373–376.

A Appendix

A.1 Multiplicity and Identification

We introduced the profile likelihood function, and with it, due to its special nature, the conditions under which it exists in the strict sense of a function. In this subsection, we briefly examine how these conditions are related to two important aspects of the application domain—that is to say structural estimation: identification and multiplicity in the model solution.

As commonly done in the literature, we restrict our attention to models where the parameter vector θ is identified—or, more precisely, the distribution of an observation implied by the model is different for different values of θ. However, this is no guarantee that, for finite samples, the likelihood function has a (locally) unique maximum; in fact, since the likelihood function (i) aggregates information by multiplying over the probability or density of all data points from a random sample, and (ii), for continuous data, in fact, compares distributions on sets of measure zero, a locally unique maximum is sufficient, but not necessary for identification.[17] We explicitly add that no identification requirements are made with respect to the controlled parameter β.

Suppose we have a local maximum $(\hat{\theta}, \hat{\sigma}, \beta_0)$. As we have argued above, a maximum—even if locally unique—does not imply the second-order sufficient conditions for optimality. In particular, even if the gradients of the constraints are linearly independent, the Hessian is only guaranteed to be negative semi-definite, with some eigenvalues potentially equal to zero and thus singular. If, however, we find the Hessian to be regular, we can conclude that the profile likelihood function exists and is smooth within some neighborhood of β_0. Obviously this argument can be applied recursively to some $\beta_1 \gtrless \beta_0$ in that neighborhood, effectively creating some kind of "tube" around the profile likelihood function within which it is unique. However, this neighborhood will shrink and tend to zero if it approaches a singularity; we give an example of how this can easily arise from multiplicity below.

As we have noted, no identification requirements are stated for β. In fact, a key motivation of this paper is to show how poorly or non-identified parameters can be

[17] This argument can be taken even further by arguing that for continuous likelihood function, even in the very vicinity of a maximum there exist infinitely many parameter values with equal likelihood.

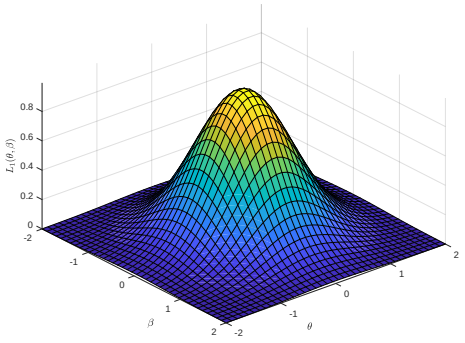
traced out using the profile likelihood function. Consider the examples in Figure 11: The top left panel shows a function for which the maximum is unique; consequently, a projection as implemented through the profile likelihood (5) has a unique maximum; see the bottom left panel. At the same time, the top right panel shows a function where, for each value of β, the maximum w.r.t. θ is the same; consequently, if it were a likelihood function, β could not be identified jointly with θ, and the corresponding profile likelihood would be flat; see the bottom right panel; note in particular the non-uniqueness of the optimal θs. The mathematical explanation of the independence of our approach from any identification assumptions about β is that—at this stage—the numerical properties of the problem are mostly delimited by the regularity and definiteness of the Hessian (28), which, however, does not contain any derivatives w.r.t. β.

Closely related to identification is multiplicity in the solution of the model. In the description of an abstract model above, we made no restrictions on the set of solutions to the model for given structural parameters, $\dot{\Sigma}(\theta, \beta)$. In fact, multiplicity does not contradict identification per se, as long as the likelihood itself discriminates the model solutions properly—that is, as long as the maximum of the profile likelihood w.r.t.ãll structural parameters,

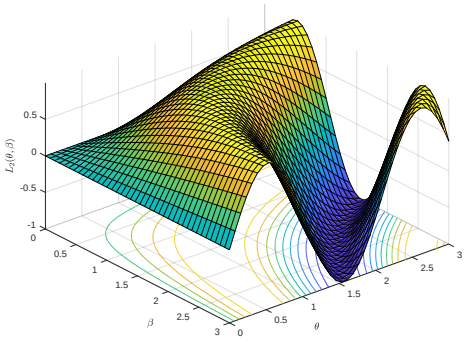
$$L_p(\hat{\theta}, \beta) \equiv \max_{\sigma \in \dot{\Sigma}(\hat{\theta}, \beta)} L(\hat{\theta}, \beta, \sigma), \tag{27}$$

at all local solutions of $\hat{\theta}$ is locally unique. (Note that local uniqueness of (27) is necessary but not sufficient for local uniqueness of $L_p(\beta)$.)

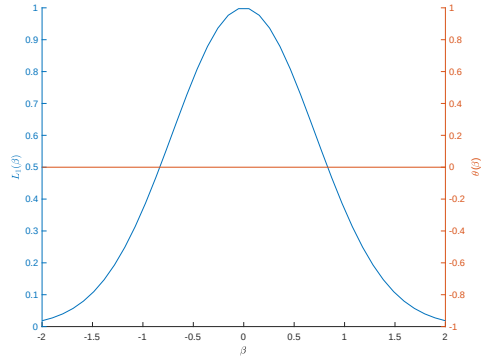
However, we have argued several times that it is essential for the profile likelihood to be a unique function, a necessary condition for the Hessian of the Lagrangian (28) is to be nonsingular; however, this cannot be true if the gradients of the constraints are linearly dependent—that is, if the Jacobian $D_{\theta,\sigma}h$ drops in rank. This can, e.g., happen at points where the solution is indeed unique but "splits up" into several solutions, e.g., at *turning points* or *bifurcations*. The following simple algebraic example might give an intuition for this: Consider the function x^2, which has a *double zero* at 0. While the solution to $x^2 = 0$ is unique, the Jacobian is zero and thus singular. If one further generalizes this example to the solution set of $x^2 - y = 0$, we observe that as we increase y from 0 to a positive value, the corresponding solutions x solving the equation are unique only for $y = 0$, but ambiguous for $y > 0$; the Jacobian at strictly positive y is, however, nonsingular. We provide more extensive examples in



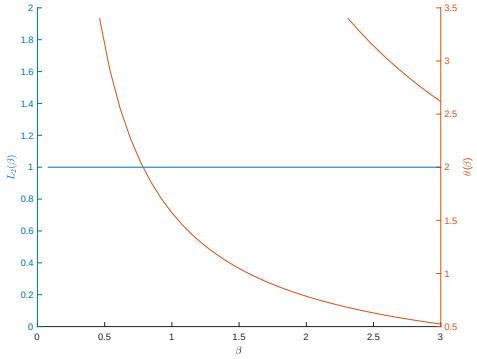
(a) $L_1 : \exp(-\theta^2 - \beta^2)$



(b) $L_2 : \sin(\theta \cdot \beta)$



(c) Optimal value and maximizer functions of L_1.



(d) Optimal value and maximizer functions of L_2

Figure 11: Two numerical examples of 2-dimensional functions (top panels), and their optimal value (blue, left axis) and maximizer functions (red, right axis) (bottom panels).

the following sections.

A.2 Second-Order Sufficient Conditions

The first-order necessary conditions only give us stationary points. A sufficient condition can be formulated based on a second-order argument. Consider the Hessian of the Lagrangian (6):

$$\nabla^2_{\mu,\theta,\sigma}\mathcal{L}(\theta,\sigma,\mu;\beta) \equiv \begin{pmatrix} 0 & -D_\theta h(\sigma;\theta,\beta) & -D_\sigma h(\sigma;\theta,\beta) \\ -D_\theta h(\sigma;\theta,\beta)^T & \multicolumn{2}{c}{\multirow{2}{*}{$\nabla^2_{\theta,\sigma}\mathcal{L}(\theta,\sigma,\mu;\beta)$}} \\ -D_\sigma h(\sigma;\theta,\beta)^T & & \end{pmatrix}, \quad (28)$$

where $\nabla^2_{\theta,\sigma}\mathcal{L}(\theta,\sigma,\mu;\beta) \equiv \nabla^2_{\theta,\sigma}L(\theta,\beta,\sigma) - \nabla^2_{\theta,\sigma}\mu^T h(\sigma;\theta,\beta)$. Obviously, the well-known second-order sufficient condition from unconstrained optimization requiring the full Hessian $\nabla^2_{\theta,\sigma,\mu}\mathcal{L}$ to be negative-definite cannot hold for any point because of the block of zeros in the northwest corner of the Hessian. Rather, it is sufficient to require that the Hessian of the Lagrangian w.r.t. σ and θ is negative-definite on a linearization of the constraint set (1):

$$v^T\nabla^2_{\theta,\sigma}\mathcal{L}(\hat{\theta},\hat{\sigma},\hat{\mu};\beta)v < 0 \quad \forall v \neq 0 : D_{\theta,\sigma}h(\hat{\sigma};\hat{\theta},\beta)v = 0. \quad (29)$$

In summary, if the point $(\hat{\theta},\hat{\sigma},\beta)$ together with $\hat{\mu}$ satisfies (8) and (29), it is a strict (i.e., locally unique) local maximum of the parametric optimization problem (5). On the other hand, the converse is not necessarily true; in fact, second-order necessary optimality conditions only imply semi-definiteness of the Hessian at the optimum, and, additionally, explicitly require that the gradients of the constraints are linearly independent. If, however, $(\hat{\theta},\hat{\sigma},\beta)$ together with $\hat{\mu}$ is a solution to (5), and if, moreover, the Hessian $\nabla^2_{\mu,\theta,\sigma}\mathcal{L}$ is nonsingular at $(\hat{\theta},\hat{\sigma},\hat{\mu},\beta)$, then the second-order sufficient conditions are satisfied (Fiacco and McCormick, 1990, Cor. 7).

A.3 pMPEC

In this section, we provide details on the application of pMPEC to the optimal replacement of bus engines model. As we have argued above, the dynamic programming problem in the model can be reformulated using relative expected values, still yielding valid decision probabilities and thus a valid likelihood function. Therefore, the profile

likelihood function of β w.r.t. θ_1 is established by the following constrained optimization formulation (recall $\theta_1 \equiv (\theta_{11}, RC)$), and note that we separately estimate θ_3 at this point, but present simultaneous estimations below):

$$\begin{aligned} L_p(\beta) = & \max_{\theta_1, \overline{ev}} L(\theta_1, \overline{ev}; \beta, \theta_3) \\ & \text{s.t. } \overline{ev} - T_\theta(\overline{ev}) + T_\theta(\overline{ev})_1 = 0, \end{aligned} \tag{30}$$

which yields first-order necessary conditions

$$\nabla_{\mu,\theta_1,\overline{ev}} \mathcal{L}(\theta_1, \overline{ev}, \mu; \beta, \theta_3) \equiv \begin{pmatrix} dT_\theta(\overline{ev}) \\ \nabla_{\theta_1,\overline{ev}} L(\theta_1, \overline{ev}; \beta, \theta_3) - \nabla_{\theta_1,\overline{ev}} \mu^T dT_\theta(\overline{ev}) \end{pmatrix} = 0, \tag{31}$$

where we write $dT_\theta(\overline{ev}) \equiv \overline{ev} - T_\theta(\overline{ev}) + T_\theta(\overline{ev}_1)$ for notational brevity. The Jacobian matrix of the system (31) (or the Hessian matrix of problem 30) reads

$$\nabla^2_{\mu,\theta_1,\overline{ev}} \mathcal{L}(\theta_1, \overline{ev}, \mu; \beta, \theta_3) \equiv \begin{pmatrix} 0 & -D_{\theta_1} dT_\theta(\overline{ev}) & -D_{\overline{ev}} dT_\theta(\overline{ev}) \\ -D_{\theta_1} dT_\theta(\overline{ev}) & \nabla^2_{\theta_1,\overline{ev}} \mathcal{L}(\theta_1, \overline{ev}, \mu; \beta, \theta_3) & \\ -D_{\overline{ev}} dT_\theta(\overline{ev}) & & \end{pmatrix}. \tag{32}$$

Note that the full Jacobian $\nabla^2_{\mu,\theta_1,\overline{ev}} \mathcal{L}$ inherits the sparsity of the Hessian of the likelihood function w.r.t. the structural parameters, $\nabla^2_{\theta_1,\overline{ev}} L$, in combination with the Jacobian of the constraints, $D_{\overline{ev}} dT_\theta$.

Following Section 2.5, we define the homotopy map $\rho(\mu, \theta_1, \overline{ev}, c(\lambda))$ as the gradient of the Lagrangian w.r.t.ãll free parameters, i.e., explicitly written as

$$\rho(\mu, \theta_1, \overline{ev}, c(\lambda)) \equiv \nabla_{\mu,\theta_1,\overline{ev}} \mathcal{L}(\theta_1, \overline{ev}, \mu; c(\lambda), \theta_3), \tag{33}$$

with the linear transformation $c(\lambda) = (1 - \lambda)a + \lambda b = \beta$. The augmented Jacobian reads

$$\rho(\mu, \theta_1, \overline{ev}, c(\lambda)) = \left(\nabla^2_{\mu,\theta_1,\overline{ev}} \mathcal{L}(\theta_1, \overline{ev}, \mu; c(\lambda), \theta_3), \quad \frac{\partial}{\partial \lambda} \nabla_{\mu,\theta_1,\overline{ev}} \mathcal{L}(\theta_1, \overline{ev}, \mu; c(\lambda), \theta_3) \right), \tag{34}$$

which equals the full Jacobian $\nabla^2_{\mu,\theta_1,\overline{ev}} \mathcal{L}$ with an additional column containing the derivative of the first-order conditions w.r.t. λ. Note that since the Bellman operator T in (20) constitutes a contraction mapping for $\beta \in (0, 1)$, the solution to the sys-

	absolute expected value ev_θ		Absolute expected value EV_θ	
	θ_3 partial	θ_3 joint	θ_3 partial	θ_3 joint
RC	9.7557	9.7558	9.7557	9.7558
θ_{11}	2.6276	2.6275	2.6276	2.6275
θ_{30}	(0.3488)	0.3489	(0.3488)	0.3489
θ_{31}	(0.6394)	0.6394	(0.6394)	0.6394
LL	-6,055.2504	-6,055.2504	-6,055.2504	-6,055.2504

Table 2: Replication of the original configuration of Rust (1987) with $\beta = .9999$ fixed, 90 mileage bins, and bus groups $\{1, 2, 3, 4\}$, using relative and absolute expected values, and estimating transition probabilities θ_3 beforehand using partial likelihood (values reported in parentheses) and jointly.

tem of constraints is unique, and its Jacobian has full rank, which is an important necessary condition for the interpretation of our results as solutions to the profile likelihood problem (30); see Sections 2.3 and A.1. As we will show, in the absolute expected value formulation this property indeed breaks down as $\beta \to 1$; moreover, we will demonstrate that also in the absolute expected value formulation, the profile likelihood will diverge starting for some βs larger than 1.

A.4 Results

Figure 12 depicts the absolute expected value function ev_θ obtained from the absolute expected value fixed point equation for various values of β. As an overlay, we also plot the EV_θ function of the original specification at $\beta = .9999$ obtained using absolute expected values. We observed that the absolute expected value formulation indeed nests the absolute expected value formulation for $\beta \in [0, 1)$. For the estimation, this is confirmed numerically in Table 2, where we present estimations of the original specification using both absolute and absolute expected values.

We estimate the model with and without the transition probability vector θ_3 (in the latter case taking the estimates from the partial likelihood of state transitions only), and observe that, as pointed out by Rust (1987), this sequential estimation procedure is very efficient; see the first two columns in Table 3.

We also briefly verify that the result is not an artifact of the mileage state discretization as maybe too coarse a transition probability vector—which is essentially a non-parametric density estimation—might affect the estimates. Table 3 presents our

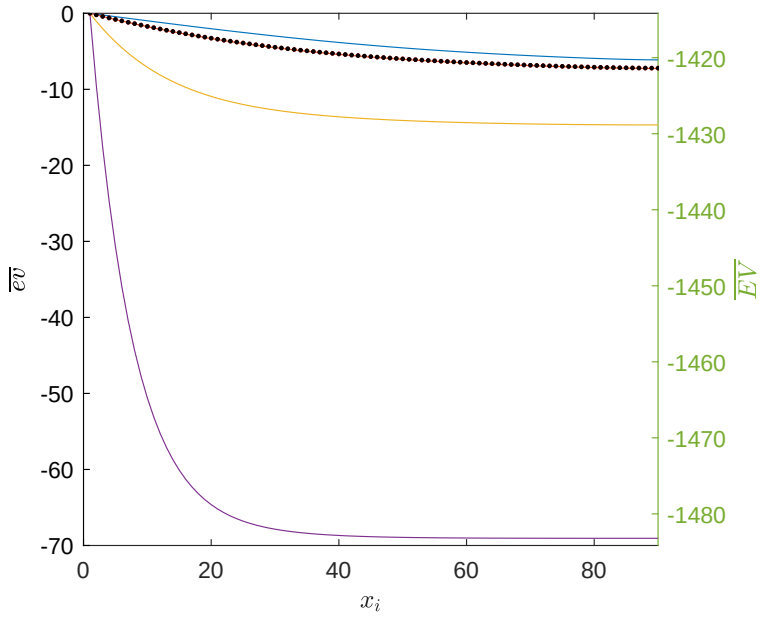
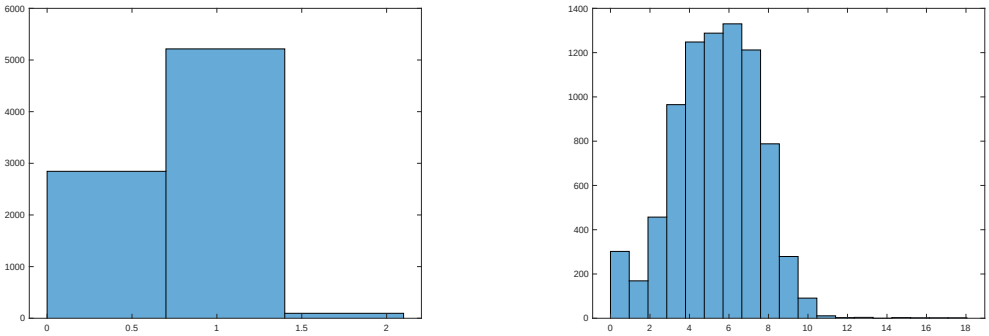


Figure 12: Expected value function ev_θ using absolute expected values for various values of $\beta \in \{.95, .9999, 1.05, 1.1\}$ (top to bottom; solid lines, left axis); expected value function EV_θ using absolute expected values for $\beta = .9999$ (dotted line, right axis); mileage discretization: 90 bins.



(a) Density for 90 bins.

(b) Density for 700 bins.

Figure 13: Non-parametric density of mileage transition from partial likelihood estimation, using 90 bins (left) and 700 bins (right), for bus groups $\{1, 2, 3, 4\}$.

findings, implying that the coarseness of the discretization has no significant effect on the estimates $\hat{\theta}_1$ and $\hat{\beta}$.[18] Figure 13 depicts the non-parametric density implied by $\hat{\theta}_3$ for 90 (original paper) and 700 mileage bins.

As we illustrate in Figure 14, we note that the steps taken by Hompack 90 to trace the path are becoming smaller in terms of β, as it approaches the region of poor conditioning. However, note that all plots—be it the profile likelihood or the estimates as a function of β—are *projections* of the full-dimensional path, and so are the steps. Therefore, assessing the step length requires us to check all dimensions, and indeed, the step lengths in the $\widehat{RC}$ dimension are not vanishing for $\beta > 1$.

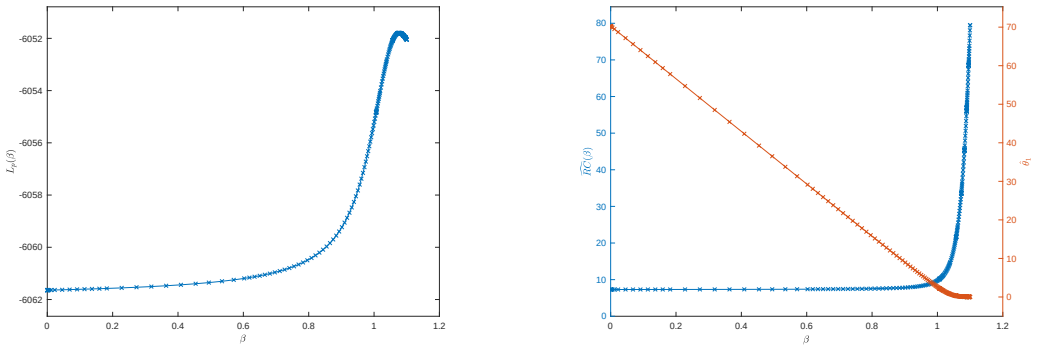
Finally, we demonstrate the feasibility of our approach by reporting running times for various configurations in terms of L^∞ tolerances and number of states in Table 4. We compare them to the running times of the corresponding full point estimation (including β). All experiments are carried out in MATLAB using our tool M-HOMPACK to interface to Hompack 90 (Watson et al., 1997) for the ODE homotopy solution algorithm, CasADi (Andersson et al., 2018) for exact derivatives, and KNITRO for constrained nonlinear optimization. All computations are performed on a Lenovo Yoga 520-14IKB Laptop with Intel Core i7-8550U @ 1.8 GHz and 8GB RAM.

Most importantly, the running times prove the feasibility of the approach: e.g., tracing the parameter estimates of the configuration with 90 mileage bins for β from 0 to 1.1 with a targeted precision of 1e-8 takes less than 10 seconds; increasing the

[18] The coarseness of the discretization has no significant effect on the estimates $\hat{\theta}_1$ up to the fact that since θ_{11} regresses on the *index* of the state rather than on its mileage value, it is expected to roughly half if the number of mileage states is doubled, which is indeed what we observe.

	90 bins (θ_3 partial)	90 bins	175 bins	350 bins	700 bins
β	1.0768	1.0768	1.0760	1.0764	1.0764
RC	37.7107	37.7109	36.2772	37.0743	37.3942
θ_{11}	0.0905	0.0905	0.0490	0.0240	0.0119
θ_{30}	(0.3488)	0.3488	0.1070	0.0463	0.0370
θ_{31}	(0.6394)	0.6394	0.5152	0.1263	0.0207
θ_{32}	(0.0118)	0.0118	0.3622	0.2897	0.0560
θ_{33}	—	—	0.0143	0.3183	0.1183
θ_{34}	—	—	0.0009	0.1896	0.1530
θ_{35}	—	—	0.0004	0.0272	0.1579
θ_{36}	—	—	—	0.0012	0.1631
θ_{37}	—	—	—	0.0004	0.1486
θ_{38}	—	—	—	0.0006	0.0966
θ_{39}	—	—	—	0.0004	0.0342
θ_{310}	—	—	—	—	0.0112
θ_{311}	—	—	—	—	0.0013
θ_{312}	—	—	—	—	0.0004
θ_{313}	—	—	—	—	0.0005
θ_{314}	—	—	—	—	0.0000
θ_{315}	—	—	—	—	0.0004
θ_{316}	—	—	—	—	0.0002
θ_{317}	—	—	—	—	0.0002
θ_{318}	—	—	—	—	0.0002
LL	-6,051.7915	-6,051.7915	-8,604.4930	-13,011.2883	-18,159.8869

Table 3: Joint estimation of all structural parameters, $(\theta_{11}, RC, \beta, \theta_3)$ (transition probabilities θ_3 jointly, except first column), for bus groups $\{1, 2, 3, 4\}$, for various mileage bin configurations.



(a) Steps of the path continuation.

(b) Steps of the path continuation.

Figure 14: Steps of a path continuation laid over the profile likelihood $L_p(\beta)$ (left), and the estimates in dependence of β (right). Estimation is based on bus groups $\{1, 2, 3, 4\}$; mileage discretization: 90 bins.

problem size to 700 bins—resulting in a system with more than 1,400 equations—the tracing still takes less than 10 minutes. As one would expect, doing a point estimation only is faster in terms of running time, but the tracing algorithm clearly outperforms the optimizer in terms of time per solution (recall that each step taken by the continuation algorithm equals a full solution to the profile likelihood problem for fixed β). Hereby, the targeted tolerances for the homotopy solution method are fulfilled, and only when RC diverges, the maximum absolute error increases (c.f. Figure **??**). Note that the time to trace a path increases more than linearly if the number of mileage bins is increased, even though we fully exploit the sparsity of the problem. This is due to the superlinear growth of the number of nonzero elements in the Jacobian caused by the increasing number of transition probabilities.Also note that the number of steps taken to trace a path only depends on the targeted accuracy, and not on the problem size. Conversely, the time per step only depends on the size of the problem.

We note, though, that the transition probabilities when traced as a function of β are not constant, in particular for large values of the discount factor, but their variation is not significant in order of magnitude; see Figure 15.[19]

[19]Note that in order to trace the complete estimation problem including θ_3, we have to add another constraint to (30), stating that the probabilities sum up to 1: $\sum_{i\in\{0,1,2\}} \theta_{3i} = 1$; however, we do not explicitly impose $0 \leq \theta_{3i} \leq 1$ (but rather verify ex post) as the incorporation of (binding) inequality constraints is the subject of further research.

		90 bins	175 bins	350 bins	700 bins
1e−8	time [s]	8.69e+00	2.40e+01	1.15e+02	5.24e+02
	L^∞	2.89e−08	9.22e−09	2.67e−08	2.21e−08
	steps	3.58e+02	3.67e+02	3.54e+02	3.64e+02
	time [s]/steps	2.43e−02	6.54e−02	3.26e−01	1.44e+00
1e−10	time [s]	1.40e+01	4.37e+01	2.00e+02	8.78e+02
	L^∞	2.36e−09	3.96e−09	1.19e−09	3.59e−09
	steps	5.92e+02	5.86e+02	6.16e+02	5.91e+02
	time [s]/steps	2.36e−02	7.46e−02	3.25e−01	1.49e+00
1e−12	time [s]	5.70e+01	1.37e+02	7.47e+02	3.04e+03
	L^∞	2.05e−10	1.24e−10	3.73e−11	1.95e−10
	steps	2.40e+03	2.41e+03	1.97e+03	2.01e+03
	time [s]/steps	2.37e−02	6.69e−02	3.80e−02	1.52e−02
	nnz	2.24e+03	6.41e+03	1.84e+04	5.89e+04
	time [s] full estimation	4.79e+00	1.26e+00	1.84e+00	4.39e+00

Table 4: Running times in seconds, realized L^∞ tolerance, number of steps taken, and time per step in seconds for each combination of mileage discretization and targeted tolerance for the homotopy continuation algorithm (ODE). All quantities are averages over 5 runs. The last two rows report the number of nonzeros (nnz) of the Jacobian for each mileage discretization, and the time for the full estimation.

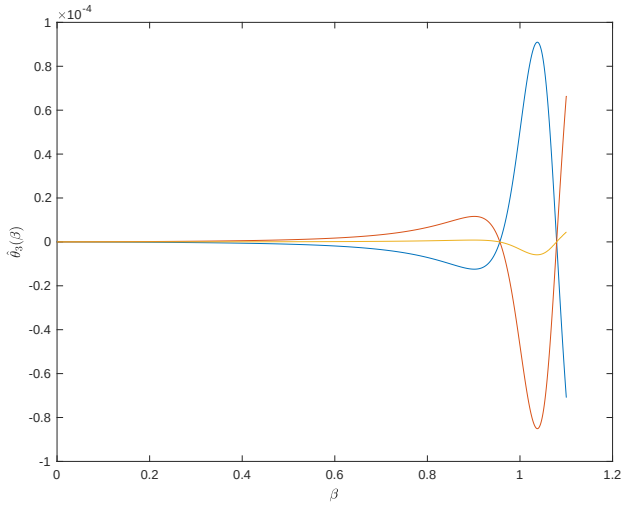


Figure 15: Estimated normalized transition probabilities in dependence of the value of the discount factor, $\hat{\theta}_3(\beta)$, normalized by $\hat{\theta}_3(0)$ (red: $\hat{\theta}_{30}$; blue: $\hat{\theta}_{31}$; orange: $\hat{\theta}_{32}$). Estimation is based on bus groups $\{1, 2, 3, 4\}$; mileage discretization: 90 bins.

A.5 Robustness

We test the robustness of our results under various model specifications. While exploring functional forms for the utility functions and allowing for heteroscedacity of risk, we apply our method and trace the controlled parameter to the maximum likelihood estimates. Our result of $\beta > 1$ proved to be robust across all tested model specifications.

A crucial choice in the construction of dynamic discrete choice models is the specification of the utility function as Magnac and Thesmar (2002) have shown that the discount factor is not identified without strong assumptions on the utility function. In constructing the utility function, Rust (1987) relies on the resulting maximum likelihood in addition to non-quantitative information which suggest a linear and square root functional form as "best fit". Without any specific non-quantitative information, we test the linear, square root, cubic polynomial, exponential and log functional form. Figure 16 plots the profile likelihood for all tested functional forms as

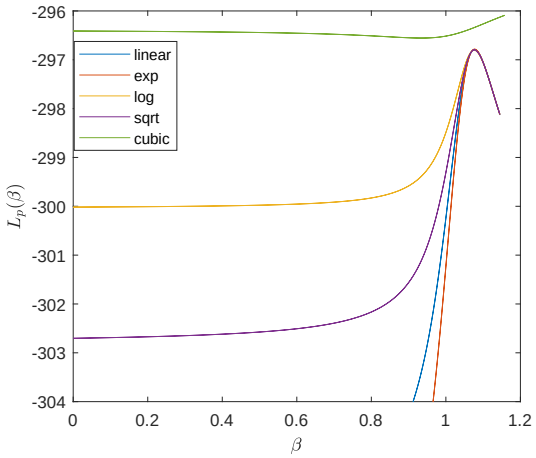


Figure 16: Profile Likelihood as a function of the controlled parameter β for the linear (blue), exponential (red), logarithmic (yellow), square root (violet) and cubic polynomial (green) functional form.

function of β. The cost functions allowing for one parameter yield the same MLE for β with $\beta = 1.0768$. The linear, exponential and square root functional form sustain the statistical significance of $\beta > 1$. The tracing for the cubic polynomial stops before the MLE is attained as RC grows in the same magnitude as before leading to a bad

condition of the FOC's Jacobian. The plot of the profile likelihood function suggests a β distinctly greater than 1.0768.

The model assumes the utility shock ϵ to be extreme value type I distributed for all mileage states. This simplifying assumption is counter-intuitive; it seems reasonable to assume that buses with higher mileages have a higher probability of breaking down and in turn a higher variance in the unobserved utility shocks. Thus, the model ignores any heteroscedasticity w.r.t.t̃he increasing mileage. We model the heteroscedasticity by adding a second shock η to the utility function $u+\epsilon$. We assume the second shock to be distributed as $\eta \sim N(0, x^2)$ and introduce the heteroscedasticity parameter θ_h. The resulting utility reads

$$u(x; \theta, d) + \epsilon(d) + \theta_h \eta. \tag{35}$$

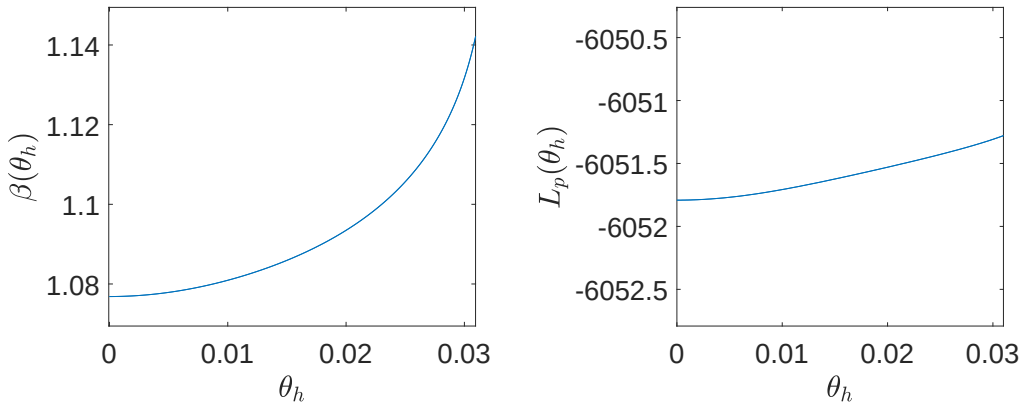


Figure 17: Maximizer $\beta(\theta_h)$ and profile likelihood $L_p(\theta_h)$ as a function of θ_h traced from $(\theta_h = 0)$ to $(\theta_h = 0.031)$. The left plot depicts the maximum likelihood estimate $\beta(\theta_h)$ and the right plot depicts $\mathcal{L}_p(\theta_h)$.

Figure 17 illustrates on the left the maximum likelihood estimate $\beta(\theta_h)$ and on the right the corresponding profile likelihood value as functions of the heteroscedacticity parameter θ_h. The likelihood is monotonically increasing in θ_h, but the maximum of the likelihood cannot be attained due to the same numerical reasons as before. Note that the original homoscedastic model is nested for $\theta_h = 0$. Opposed to our expectation, $\beta(\theta_h)$ is monotonically increasing in θ_h. Thus, even with heteroscedasticity included, the maximum likelihood estimate of β is greater than 1.

This section shows that our main findings are robust with respect to varying model specification. The maximum likelihood estimator for the discount factor β—using the data for bus groups 1-4—is greater than 1 for all tested model specification.

References

Andersson, J. A. E., Gillis, J., Horn, G., Rawlings, J. B., and Diehl, M. (2018). CasADi – A software framework for nonlinear optimization and optimal control. *forthcoming in: Mathematical Programming Computation*.

Fiacco, A. V. and McCormick, G. P. (1990). *Nonlinear Programming*. Sequential Unconstrained Minimization Techniques. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Rust, J. (1987). Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica: Journal of the Econometric Society*, 55(5):999–1033.

Watson, L. T., Sosonkina, M., Melville, R. C., Morgan, A. P., and Walker, H. F. (1997). Algorithm 777: HOMPACK90: a suite of Fortran 90 codes for globally convergent homotopy algorithms. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):514–549.