Harold Zurcher and Paul Volcker—The Protocol of a Meeting that Never Took Place*

Philipp Müller[†] Gregor Reich[‡]

This version: September 2021
First version: December 2018

Abstract

We develop a method to robustly estimate dynamic models even if their discount factor is "poorly identified", or not identified at all: We use homotopy continuation to trace the solution to the first-order conditions of the MPEC formulation of the estimation problem (Su and Judd, 2012)—and thus the estimates of the model parameters—as a function of the discount factor. We apply this method to the bus engine replacement model of Rust (1987), a.k.a. the "Harold Zurcher model", whose discount factor is believed to be poorly identified. Instead, we find that it is well identified with estimates that are greater than one in various model specifications. We demonstrate that this does not contradict economic theory, given the prolonged periods of negative real interest rates in the US during the sampling period, 1974 to 1985. Moreover, we implement a natural experiment: In August 1979, Paul Volcker became chairman of the US Federal Reserve, immediately taking drastic steps to curb inflation, which led to a significant rise in nominal and real interest rates. We introduce an unanticipated structural break in discounting, finding that the estimate of Harold Zurcher's discount factor decreases significantly one month after Volcker took office.

Keywords: Structural estimation, homotopy continuation, discount factor, dynamic programming, structural break, great inflation.

---

*We are heavily indebted to Kenneth Judd, Felix Kübler, and Karl Schmedders. We further thank Aleksander Berentsen, Michel Bierlaire, Øystein Daljord, Robert Erbe, Walter Farkas, Bo Honoré, Diethard Klatte, 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 financial support of Kenneth Judd, Senior Fellow at the Hoover Institution. Reich gratefully acknowledges the financial support of the Fonds zur Förderung des akademischen Nachwuchses (FAN), Zurich. Finally, we thank Dave Brooks for his editorial comments.

[†]Department of Business Administration, University of Zurich, Switzerland. Email: philipp.mueller@business.uzh.ch

[‡]Independent researcher. Email: gregor.reich@me.com

1 Introduction

The estimation of the discount factor in dynamic discrete choice models (DDCMs) is considered to be difficult: First, it is *non-parametrically non-identified* (Rust, 1994). That is, without strong assumptions on the primitives of the model, there exist a function mapping state and actions to (immediate) utility for all values of the discount factor—which is commonly restricted to the range of the half-open interval $[0, 1)$—explaining the observed state and action combinations equally well. Furthermore, as shown by Magnac and Thesmar (2002), the discount factor is—together with the distribution of the error terms—key to identification. Therefore, substantial effort has been put into finding conditions and restrictions under which the former is indeed identified, among them Komarova et al. (2018); Daljord et al. (2018); Abbring and Daljord (2020). Second, even if the model complies with the structure and the assumptions needed for its discount factor to be identified theoretically, a widespread conception is that it is still often difficult to estimate "because it is poorly identified" (Aguirregabiria and Mira, 2010, Footnote 7).

In this paper, we develop a *homotopy continuation approach to estimation* for models with poorly or even non-identified discount factors in a maximum likelihood setting. Our approach traces the estimates of the model primitives, such as the parameters of the utility function and those of the state transition distribution, together with the solution of the model as a function (or, more precisely, a correspondence) of the non-identified parameter. This approach is inspired by the insight of Su and Judd (2012) that the estimation problem can be solved as a constrained optimization problem, maximizing the likelihood function subject to the constraints induced by the model equations, an approach referred to as mathematical programming with equilibrium constraints (MPEC). We go one step beyond, taking the first-order conditions of the MPEC formulation *parametrized* by non-identified parameter, and trace the resulting system of equations using homotopy continuation. It is well-known in the literature that this approach is numerically well-behaved even if the optimization problem *including* the non-identified parameter has a singular Hessian (i.e., the discount factor is non-identified); for an introduction to homotopy continuation, see, e.g., Zangwill and Garcia (1981); Allgower and Georg (2012). Although we consider discount factors in this paper, our approach is not limited to that; in fact, it is applicable in many situations where a parameter constitutes a "trouble-maker" in terms of identification—even if defined implicitly only; we demonstrate this in our application below.

We apply this homotopy estimation method to the bus engine replacement model and the original data set of Rust (1987). The model features a manager of the maintenance service department for a fleet of public transportation buses located in Madison (Wisconsin), a person named Harold Zurcher (1926–2020): On a monthly basis, he has to decide for each of his buses whether completely overhaul them (most importantly, replacing their engine), or whether to do small maintenance work only; in general, the complete overhaul is much more expensive compared to the regular maintenance work. However, Rust (1987) quotes evidence that maintenance costs increase with "age", measured in terms of mileage traveled since the last complete overhaul. Therefore, this decision constitutes a dynamic decision problem, as the decision to overhaul the bus impacts not only immediate but also future maintenance costs. Rust (1987) models the behavior of Harold Zurcher as a dynamic programming problem and, using a data set of his decisions between 1974 to 1985, estimates the parameters of the cost function underlying Zurcher's decisions.

The "bus engine replacement model", or the "Harold Zurcher model" as some call it, is considered to be pioneering the work on the estimation of dynamic discrete choice problems, and Rust (1987), together with Rust (1988), also provides a complete, general numerical approach to estimate such models: the nested fixed-point algorithm (NFXP). At the same time, it might also have contributed to the discount factor's nimbus of not being estimable—from the very early days of this literature on. In fact, the discount factor in the paper is not estimated, but fixed at a value of 0.9999. The author explains this as follows (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.

However, 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.

We investigate this conundrum using our homotopy estimation method, which appears to be perfectly suited in the presence of a near-singular Hessian, and indeed find a deeper root behind it.

Before we can address the econometric problem of estimating a supposedly poorly identified parameter, we have to mitigate an issue that comes with the second statement from above. Rust notes that his algorithm tends to push the estimate of the discount factor towards one, and thus we have to be prepared for the potential implications of this situation in the context of the underlying dynamic programming problem. Generally speaking, discount factors equal to—or greater than—one are not a problem, as long as the dynamic problem which attempts to maximize the (expected) sum of discounted future utilities has a finite horizon. In fact, it is a standard procedure in the field of operations research to solve total reward problems through finite horizon dynamic programming without a discount factor, i.e., a discount factor identical to one; see, for example, Barz (2007). Of those who attempt to estimate it, Erdem and Keane (1996), among others, find it to be even larger than one; importantly, they do not preclude this result, arguing (Erdem and Keane, 1996, Footnote 11):

> The intertemporal factor is usually assumed to be between 0 and 1 because it is assumed to be 1 / (1 + interest rate) although behaviorally this does not have to hold.

The problem with discount factors equal to or larger than one becomes more apparent if the horizon of the dynamic programming problem is infinite. In such a setup, the sum of (expected) future discounted utilities diverges unless there exists an absorbing state with zero utility. Therefore, it is a legitimate question if such configuration makes economic sense. Technically speaking, an infinite horizon problem is simply the limit of a finite horizon problem with T periods, as $T \to \infty$. Indeed, quite early in the literature on dynamic programming, Morton and Wecker (1977) provide conditions under which the infinite horizon problem constitutes a valid *approximation* to the finite horizon problem even for discount factors equal to or larger than one, and they find that these conditions are fulfilled for many important classes of applied problems (the one studied in this paper included). Consequently, if one is willing to accept discount factors greater than one for finite horizon problems, accepting the same for the infinite case is almost inevitable, given that no human being can seriously be assumed to look ahead billions of years or more when making their decisions (and particularly not when deciding about bus engines. . .).

Still, a problem that remains with a discount factor equal to or greater than one in infinite horizon dynamic programming is the question of how to actually solve for a value function that is equal to infinity. In fact, the operations research literature has developed concepts for this case quite early on: In particular, White (1963) developed what is known as "relative value iteration", for example in order to solve the frequently occurring "minimum average cost problem". Here, one does not solve for the values themselves but instead for the values *relative* to some reference state. It can be shown that such procedure converges for many classes of problems (again, the one we address included), and the policies implied by relative values are the same as for conventional values, whenever the latter can be computed; see, e.g., Puterman (2014) for more details. Interestingly, relative values have a very natural economic intuition: Even if an agent has to choose between two alternatives that will give him or her infinite value down the road, he or she might still prefer one over the other if the aggregated per-period utility is *growing* faster.[1]

Finally, the most blatant economic interpretation of a discount factor is simply the inverse of the gross interest rate. Obviously, this cannot be taken literally in models and data sets that span a significant period of time, because discount factors typically are constant whereas interest rates are not. If one accepts that the link between the discount factor and interest rates is imperfect and more like a link to the inverse of some average gross interest rate, this definition does, however, imply that discount factors greater than one go with negative net interest rates, and vice versa.

Equipped with our homotopy estimation approach as well as with relative value iteration, we investigate the question of whether the discount factor in the model and the data set of Rust (1987) is identified or not, and to which extent it is "driven to 1", or even beyond that. Indeed, in a series of estimation experiments using different sub-samples of the original data set and different degrees of observed heterogeneity among the types of buses in the sample, we find all estimates for the discount factors to be larger than one, some of them even with statistical significance[2] (cf. Rust, 1987, Table IX): The estimate for the simplest model with no cost parameter heterogeneity (i.e., bus-type specific cost functions) based on the sub-sample used in the original paper is as high as 1.08 and significant, but drops to 1.05—and becomes insignificant—once we

[1]A similar argument outlining the possibility of a macroeconomic model with well-defined outcomes despite a discount factor greater than one was introduced and debated following Kocherlakota (1990).

[2]In the following, we call a discount factor estimate that is larger than one "significant" if the null hypothesis of it being equal to one can be rejected.

respect the difference in bus types. If we use the full data set containing some buses that Rust did not take into account in his reported experiments, but which are relevant to us as their engine replacements happen early in the sampling period, the estimated discount factor further drops to 1.03, but is, again, significantly larger than one. Moreover, in all experiments, the likelihood is nicely shaped and has a clearly distinguished, unique local maximum. We conclude, at this point, that the reported identification issues were indeed artifacts from numerical instabilities arising when the likelihood maximization process attempted to push the discount factor towards one, causing the conventional value functions used in the paper to diverge. We give detailed evidence based on the condition numbers of the corresponding Hessian matrices in Section 3.3 to back this claim.

The more interesting question is, however, how realistic an estimated discount factor greater than one really is from an economic perspective. In fact, looking at the federal funds rate which peaked at an unprecedented level of 20% in 1980, and was well in the two-digits even in the years before, it is hard to argue for negative interest rates at first sight. However, this is only one side of the medal, as the US was in a situation of extraordinarily high inflation in the 1970s and early 1980s—a period now rereferred to as the "great inflation". That same year, the annual increase in consumer prices peaked at almost 15%. And for the arguments to follow, note that all these dates lie in the very interior of the sampling period of Rust (1987).

The embedding of the decision maker, Harold Zurcher, in this macroeconomic environment immediately raises two questions relevant for the model and data set at hand: First, what is the appropriate notion of interest rate when it comes to the link to discounting—nominal or real—and second, is the aforementioned "average interest rate" a good measure for discounting in our context. The former question can be answered rather unambiguously: Given that the accumulated inflation over the sampling period is beyond 100%, but, at the same time, the model primitives which quantify the costs of the alternatives have no time dependence and therefore do not account for increasing prices in any way, interpreting the model in nominal terms would not be very credible.[3] To address the second question, a little more historical context is needed.

After a period of already high inflation, high interest rates, significant unemployment, and a substantial devaluation of the Dollar, there was a widespread concern that the 1979 oil price shock would act as the final straw that pushed the US into a serious recession. Nevertheless, there also was a consensus that inflation had gotten out of control, but that the measures taken against it by then Federal Reserve Chairman William Miller, who was appointed in 1978 as by President Jimmy Carter, did not show sufficient effect. Therefore, after only a little more than a year, President Carter appointed a new Chairman of the Fed: Paul Volcker (1927–2019).

While the procedure that led to Volcker's appointment was unprecedentedly short, his statements in the course of it—most notably the congressional hearings on July 30, 1979—were the beginning of an enormously disruptive shock to the US economy. In particular, he made very clear that even while at the dawn of a recession, he was putting a higher priority on fighting inflation, even if the rise in interest rates that would accompany the monetary discipline needed to achieve this goal would put extra harm on the economy. Being asked by a Senator whether he wouldn't "see anything abnormal about going into an international recession with interest

[3] Although Rust (1987) does not explicitly comment on this question, all reported dollar values in the paper are assigned a year, predominantly 1985.

rates at 10, 12, to 15 percent?" he answered (US House, 1979, p. 10):

> I think our economic situation is abnormal. [...] You say going into a recession; I think there is a clear danger there. [...] But what's abnormal is that we're going into it with an inflation rate of the sort we have. That's what's unprecedented, and I think that the interest rates are a byproduct.

Volcker then continued to express his view on how the macroeconomic environment impacted every day (micro)economic decision making in America in a negative way, and how important a change of people's expectations was (US House, 1979, p. 29):

> the American people have, I suspect, become convinced as never before that inflation is here to stay and that it may rise. That affects activity; it affects the way they invest; it affects what they buy; it affects what they do. It makes our job more difficult. The first priority is to demonstrate that that's not the case, that we don't face a situation where inflation inevitably has to rise. And I hope—and perhaps at this point I should put it no more strongly than that—that we can get that psychology turned around through persistence and disciplined policies, and as we do we can find it possible to move more rapidly to restore stability.

The measures that Volcker took to break the spiral of increasing interest rates and inflation created some short-term turmoil in all measures of interest—including inflation as well as nominal and real interest rates—but eventually achieved their intention: Both the level and the volatility of all those quantities stabilized, and real interest rates increased from low or even negative values to higher levels, without becoming excessive. Due to the enormous commitment that the Fed showed by allowing the nominal rates to reach historically high levels, the real interest rate adjusted already in 1980, while it took a little longer for inflation to reach the levels that then characterized the environment of the "great moderation".

In the context of the data set of Rust (1987), this unique historical context allows for a revealing *natural experiment*: If Volcker's view about the American people factoring in the macroeconomic environment in their decisions is correct, and if discounting is really inversely linked to gross real interest rates in this model, we should observe decisions that indicate a drop in discounting somewhen after Paul Volcker announced and implemented his policies. Moreover, this drop should be mostly unanticipated, as the appointment of Volcker was itself not anticipated given the short time into the term of his predecessor, William Miller. Using our model, implementing such an experiment can be done in a straightforward manner through a *structural break*: We assume that Harold Zurcher discounts at some factor, say β_1, before the break, and at some other factor, β_2, thereafter; the change in discounting is unanticipated, as we just "switch" the underlying dynamic programming problems that yield the corresponding decision probabilities for the likelihood function. We can obtain estimates for both discount factors and, moreover, the timing of the break, and finally run a statistical test on the hypothesis $H_0 : \beta_1 = \beta_2$.

Indeed, we find strong empirical evidence for this economically reasonable, qualitative link between the discount factor and the real interest rates: Estimating the model with two discount factors, we find that the estimate of the discount factor before the break is larger than the estimate of the discount factor thereafter: $\hat{\beta}_1 = 1.04$ and $\hat{\beta}_2 = 1.02$, respectively. In fact, the likelihood ratio test for equality rejects the null hypothesis with a p-value in the order of 10^{-4}. This is in line with the significant rise in real interest rates due to Paul Volcker's monetary policy shortly after he took office—just about half way into the sample period. Moreover, we also estimate the most likely timing of the structural break in discounting. Our estimate is September 1979, which happens to be just two months after the congressional hearings, and one month after Paul Volcker actually became Chair in August 1979. Finally, we assess the simultaneous identification of the two discount factors by tracing the difference between them, and hence the implicitly defined parameter $\Delta\beta = \beta_1 - \beta_2$, using our homotopy approach. Again, we find that everything is well-behaved and fully identified.

In a second extension, we demonstrate that the overall level of the discount factor estimates decreases if—in addition to the structural break in discounting—we allow for a more flexible state transition process. This extension is motivated by our observation that mileage transitions are indeed non-stationary in the original sample. We estimate $\hat{\beta}_1 = 1.03$ and $\hat{\beta}_2 = 1.00$, respectively. This suggests that miss-specification of state transition probabilities can bias the discount factor estimator, because it potentially corrects for too high (or too low) an expected value function. We conclude that the estimates from both experiments for the discount factor before the structural break, $\hat{\beta}_1$—although being larger than 1—do not contradict the time value of money since the real interest rates have been largely negative in that period. Moreover, the discount factor estimate after the structural break, $\hat{\beta}_2$, from the second extension is consistent with an agent maximizing his or her long-run average utility, or, in Harold Zurcher's case, minimizing long-run average cost.

We conclude the introduction with a couple of bibliographic remarks: Our estimation method loosely relates to some approaches from the statistics and computational economics literature. First, DiCiccio and Tibshirani (1991), a technical report, develops an approximation of the profile likelihood function of unconstrained likelihood maximization problems in which all parameters are well identified. Second, Eaves and Schmedders (1999), Besanko et al. (2010), and Aguirregabiria (2012) apply homotopy continuation to trace general equilibrium model solutions and dynamic game equilibria as a function of a parameter value, in particular, to detect and handle the presence of multiple solutions of the model for a given parameter value.[4] Third, Tiahrt and Poore (1990) and Poore (1990) apply homotopy continuation to general parametric constrained optimization problems.

The remainder of this paper is organized as follows: Section 2 defines the original estimation problem, formalizes the concept of the augmented profile likelihood problem and its representation by first-order conditions, briefly outlines homotopy 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 empirically establishes the qualitative link between β and the real interest rates. Section 4 concludes.

[4] For more details on the application of homotopy continuation methods in economics, see Judd (1998) and Borkovsky et al. (2010).

2 Structural Estimation by Homotopy Continuation

2.1 The Structural Model

We closely follow Su and Judd (2012) in its notation of a structural model due to its generality:

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, e.g., parameters of statistical distributions, preference parameters, and the discount factor.

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

Endogenous variables A n-dimensional vector of endogenous variables, $\sigma \in \Sigma \subseteq \mathbb{R}^n$. For computational feasibility, we restrict σ to a finite-dimensional vector, and assume that all objects have been discretized and truncated if necessary. Important examples of endogenous variables are value or policy functions of the state variables after they have been discretized by projection onto a finite set of basis functions.

Model equations A structural model further specifies relations between these variables and parameters. These can, for example, be first-order optimality conditions or Bellman equations approximated through projection or collocation conditions. We subsume these relations into a system of (nonlinear) equations $h : \Sigma \times \Theta \times [a, b] \to \mathbb{R}^k$ and require

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

Note that equation (1) is symbolic and the state vector x may enter h, e.g., in that (1) has to hold for all $x \in \mathcal{D} \subseteq \mathcal{S}$. The domain $\mathcal{D}$ depends on the chosen discretization method; for projection methods, we typically choose the number of elements in $\mathcal{D}$ to be n.

Equilibrium outcomes We denote the set of *all* equilibrium outcomes of σ as a function of (θ, β) by

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

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

2.2 The Parameter Estimation Problem

To estimate the structural parameters, (θ, β), the econometrician obtains data on the variables predicted by the model. These variables are usually a subset of the state variables and the endogenous variables, or functions thereof.[5] We denote the full sample of observations by $\mathcal{X}$. In the case of DDCMs, the sample usually comprises a subset of the state variables and the decision outcomes; the latter depend on the endogenous variables through the decision probabilities.

[5]Often, we can only observe a subset of the variables that the model predicts; the details of the treatment of unobserved and occasionally observed variables depend on the estimation method, but often forms some kind of expectation and thus an integral; see Reich (2018) and Lanz et al. (2021).

In this paper, we utilize the *likelihood function* implied by the model (1) as a measure to which extent particular values of the parameters "explain" these observations. We assume the likelihood function and the system of equations h to be twice continuously differentiable w.r.t. the structural parameters and the endogenous variables. There exist two main likelihood-based approaches to the estimation of structural models that fit into our definition:

The nested fixed point (NFXP) approach of Rust (1987) maximizes the likelihood over (θ, β). For each candidate vector (θ, β), we are required to find all equilibrium outcomes of σ—that is, the set $\Lambda(\theta, \beta)$—and choose the equilibrium outcome that returns the highest likelihood value:

$$(\hat{\theta}, \hat{\beta}) = \operatorname*{arg\,max}_{\theta \in \Theta,\, \beta \in [a,b]} \left\{ \max_{\sigma \in \Lambda(\theta, \beta)} L(\theta, \beta, \sigma; \mathcal{X}) \right\}. \tag{3}$$

We denote the maximum likelihood estimates of the structural parameters θ and β by $(\hat{\theta}, \hat{\beta})$, respectively.

The mathematical programming with equilibrium constraints (MPEC) approach by Su and Judd (2012) maximizes the *augmented likelihood* over θ, β, and σ altogether, while imposing the model equations, h, as constraints:

$$\begin{aligned} (\hat{\theta}, \hat{\beta}, \hat{\sigma}) = \operatorname*{arg\,max}_{\theta \in \Theta,\, \beta \in [a,b],\, \sigma \in \Sigma} \quad & L(\theta, \beta, \sigma; \mathcal{X}) \\ \text{s.t.} \quad & h(\sigma; \theta, \beta) = 0. \end{aligned} \tag{4}$$

We additionally denote the maximum likelihood estimate of σ by $\hat{\sigma}$. Su (2014) and Egesdal et al. (2015) argue that this approach avoids searching for all equilibrium solutions for each candidate vector (θ, β). To develop our concepts in this paper, we rely on the constrained optimization formulation (4).

2.3 First-Order Conditions

We consider estimation problems in which the point or interval estimation of the structural parameter β is insufficient or infeasible: Suppose the structural parameter β is "poorly identified". In this case, solving the estimation problem for β fixed to values close to its maximum likelihood estimate, $\hat{\beta}$, yields a very similar likelihood. Thus these estimates explain the observed data (almost) equally well, even though they may be drastically different from their maximum likelihood estimates. We argue that only β together with *parametric estimates* of the remaining parameters and variables as a *function* of β, $(\hat{\theta}(\beta), \hat{\sigma}(\beta))$, convey a complete picture.

We derive this parametric estimation in β by extending the well known notion of a profile likelihood function from unconstrained to constrained maximum likelihood estimation, and refer to the new notion as the *augmented profile likelihood*:

$$\begin{aligned} L_p(\beta; \mathcal{X}) \equiv \max_{\theta \in \Theta,\, \sigma \in \Sigma} \quad & L(\theta, \beta, \sigma; \mathcal{X}) \\ \text{s.t.} \quad & h(\sigma; \theta, \beta) = 0. \end{aligned} \tag{5}$$

In the following, we refer to β as the *controlled parameter*, and skip the dependency on the sample for notational brevity.

In principle, each evaluation of the augmented profile likelihood for a particular β requires solving the nonlinear constrained optimization problem (5). Solving the system of first-order necessary conditions (FOCs) of (5) for this particular β—and checking the second-order sufficient conditions—is, however, mathematically equivalent. We use these FOCs to develop our estimation approach further.

Consider the Lagrangian $\mathcal{L}$ of the augmented profile likelihood (5),

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

with $\mu_h \in \mathbb{R}^k$ denoting the Lagrange multipliers. The gradient of the Lagrangian (6) with respect to μ_h, θ, and σ reads

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

where $D_{\theta,\sigma}h(\sigma; \theta, \beta)$ denotes the Jacobian of $h(\sigma; \theta, \beta)$ w.r.t. θ and σ. If the gradients of the constraints are linearly independent,[6] then the following Karush-Kuhn-Tucker FOCs hold: suppose $(\hat{\theta}, \hat{\sigma}; \beta)$ is a local optimum of the constrained optimization problem (5) for fixed β, and $\hat{\mu}$ are the corresponding Lagrange multipliers then

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

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

2.4 Homotopy Parameter Continuation

After deriving the FOCs of the augmented profile likelihood (8), we face a system of nonlinear equations—parameterized by β—instead of the continuum of constrained optimization problems (8). We apply homotopy continuation algorithms to solve this system of equations robustly for "all" β in the compact interval $[a, b]$.

Let us define the homotopy map, $\rho : \Theta \times \sigma \times \mathbb{R}^k \times [a, b] \to \mathbb{R}^{p+n+k}$, as the parameterized system of FOCs:

$$\rho((\theta, \sigma, \mu_h), \beta) \equiv \nabla_{\mu_h,\theta,\sigma}\mathcal{L}(\theta, \sigma, \mu_h; \beta). \tag{9}$$

The zero set of this homotopy map, $\rho^{-1}(0) \equiv \{(\theta, \sigma, \mu_h, \beta) : \rho((\theta, \sigma, \mu_h), \beta) = 0\}$, equals the set of stationary points of the FOCs (8). Homotopy continuation algorithms approximate this zero set by tracing the curve $\xi \subseteq \rho^{-1}(0)$ that emanates from an initial solution; the initial solution equals the estimates for fixed $\beta = a$, which we obtain by standard constrained optimization.

There exist two popular types of homotopy continuation algorithms: predictor-corrector and piecewise-linear; see Allgower and Georg (2012) for details. We employ predictor-corrector homotopy algorithms for our estimation approach. The idea behind them is to numerically

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

approximate the curve ξ through a sequence of points $\hat{\xi}_i,\ i=1,2,\dots$ along ξ, by alternating predictor and corrector steps: Starting from a point $\hat{\xi}_i$, the prediction step produces a guess for the next point, say $\tilde{\xi}_{i+1}$, along the curve ξ; popular choices include ordinary differential equation solvers and extrapolation schemes. The correction step moves this guess "closer" to the curve, producing $\hat{\xi}_{i+1}$; due to the contractive properties of ξ, a common choice are Newton-like methods. By alternating both steps, we can produce a sequence of points $\{\hat{\xi}_i\}_{i=1}^N$ which is arbitrary close to ξ. Interpolating this sequence of points yields a continuous approximation of ξ.

By construction, all points on ξ satisfy the first-order *necessary* conditions (8)—that is, all points on ξ are stationary points which may indicate either minima, saddle points, or maxima. We are, however, only interested in maxima. One way to select the maximum points is to check the second-order sufficient conditions of all points on ξ. However, if the Jacobian $D_{\theta,\sigma,\mu}\rho((\theta,\sigma,\mu),\beta)$ is regular for all points on ξ, it is sufficient to ensure that the initial solution at $\beta=a$ is a maximum point (see, e.g., Poore, 1990). Singularities along the curve ξ, such as turning points and bifurcations, indicate a non-regular Jacobian; in this case, we must check the second-order sufficient conditions on ξ.

For our application in Section 3, we employ the predictor-corrector algorithms implemented in the Hompack 90 library (Watson et al., 1997) and use the automatic differentiation (AD) tool CasADi (Andersson et al., 2019) for all derivatives. We interface to the Fortran 90 library Hompack 90 through our interface M-Hompack to implement the underlying model entirely in Matlab.

2.5 Extensions

2.5.1 Tracing Functions of Parameters

In the previous subsections, we focus on a structural parameter which is part of the model. Yet, our approach also extends to other quantities of interest, say δ, which are implied by a *function* of the structural parameters and endogenous variables, which we denote by $g(\theta,\beta,\sigma,\delta)$. The function g can be as simple as the difference of two structural parameters, but also more complex, such as the implied demand.

To obtain a parametric solution to the estimation problem in the quantity of interest δ, we extend the augmented profile likelihood (5) to

$$\begin{aligned} L_p(\delta) \equiv & \max_{\theta\in\Theta,\,\sigma\in\Sigma,\beta\in[a,b]} L(\theta,\beta,\sigma) \\ & \text{s.t. } h(\sigma;\theta,\beta)=0 \\ & \qquad g(\theta,\beta,\sigma,\delta)=0. \end{aligned} \tag{10}$$

Note that the structural parameter β now becomes part of the maximization problem, and the augmented profile likelihood function is a function of δ. This extends Lagrangian function $\mathcal{L}$ to

$$\mathcal{L}(\theta,\sigma,\mu_h,\mu_g;\delta) = L(\theta,\beta,\sigma) - \mu_h^T h(\sigma;\theta,\beta) - \mu_g g(\theta,\beta,\sigma,\delta), \tag{11}$$

with an additional Lagrange multiplier $\mu_g \in \mathbb{R}$. The homotopy map (9) changes to

$$\rho((\theta, \sigma, \mu_h, \mu_g), \delta) \equiv \nabla_{\mu_h, \mu_g, \theta, \sigma, \beta} \mathcal{L}(\theta, \sigma, \mu_h, \mu_g; \delta). \tag{12}$$

We can directly apply the homotopy continuation method discussed in Section 2.4 to the homotopy map (12).

2.5.2 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

$$CI_\gamma(\hat{\theta}_j; \beta) \equiv \left\{ \theta_j : \max_{\theta_{-j}} L(\theta; \beta) - (L(\hat{\theta}(\beta); \beta) - 0.5\chi_1^2(\gamma)) \geq 0 \right\}, \tag{13}$$

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:

$$\begin{pmatrix} L(\theta; \beta) - (L(\hat{\theta}(\beta); \beta) - 0.5\chi_1^2(\gamma)) \\ \nabla_{\mu, \theta_{-j}, \sigma} \mathcal{L}(\mu, \theta, \sigma; \beta) \end{pmatrix} = 0. \tag{14}$$

The first equation of (14) ensures that the level of the likelihood at (θ_j, θ_{-j}) is equal to the critical value of the likelihood ratio test statistic.[7] While we vary θ_j to obtain the critical value of the likelihood, we optimize the likelihood w.r.t. θ_{-j}; this is ensured by the first-order conditions of the Lagrangian w.r.t. to θ_{-j} (and σ).

3 Application: Identifying the Discount Factor in the Bus Engine Replacement Model

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 which is larger than 1. We find empirical evidence that the context in which the decision-maker acts matters: first, the historical macroeconomic context—the great inflation—in Section 3.4.2 and second, the embedding of bus groups in the expanding bus fleet in Section 3.4.3.

[7] 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.

3.1 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. 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 exposes 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

$$u(x, i; \theta_{11}^p, RC^p) + \epsilon(i) \equiv \begin{cases} -RC^p + \epsilon(1) & i = 1 \\ -\theta_{11}^p \cdot x + \epsilon(0) & i = 0 \end{cases} \tag{15}$$

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 $x \in \{1, \ldots, 90\}$ denoting the index of the bin— and assumed to follow a Markov process with conditional transition probabilities $\theta_3^p \equiv (\theta_{30}^p, \theta_{31}^p, \theta_{32}^p)$:

$$\theta_{3\Delta}^p \equiv Pr(x_{t+1} = (1 - i_t)x_t + \Delta \mid x_t, i_t; \theta_3^p),\ \Delta \in \{0, 1, 2\}, \tag{16}$$

for $x_{t+1} \in \{1, \ldots, 90\}$, 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,

$$V_{\theta^p}(x_t, \epsilon_t) = \sup_{f_{\theta^p}(\cdot,\cdot)} \mathbb{E}\left[\sum_{j=t}^{\infty} \beta^{j-t} \left(u(x_j, f_{\theta^p}(x_j, \epsilon_j); \theta_{11}^p, RC^p) + \epsilon(f_{\theta^p}(x_j, \epsilon_j)) \right) \middle| x_t, \epsilon_t; \theta_3^p \right], \tag{17}$$

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 (17):

$$V_{\theta^p}(x, \epsilon) = \max_{i \in \{0,1\}} \left\{ u(x, i; \theta_{11}^p, RC^p) + \epsilon(i) + \beta \mathbb{E}\left[V_{\theta^p}(x', \epsilon') | x, i; \theta_3^p \right] \right\} \tag{18}$$

where x' and ϵ' denote next period's values of the states. When estimating the discount factor β—which is fixed in the original specification in 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 a (partial) closed-form solution for the expectation over the next period's value from the distributional assumptions on ϵ as

$$\begin{aligned} EV_{\theta^p}(x,i) &\equiv \mathbb{E}\left[V_{\theta^p}(x',\epsilon')|x,i;\theta_3^p\right] && (19)\\ &= \sum_{\Delta\in\{0,1,2\}} \log\Bigg(\sum_{j\in\{0,1\}} \exp(u((1-i)x+\Delta,j;\theta_{11}^p,RC^p) && (20)\\ &\qquad + \beta EV_{\theta^p}((1-i)x+\Delta,j))\Bigg)\theta_{3\Delta}^p \\ &\equiv T(EV_{\theta^p})(x,i). && (21) \end{aligned}$$

Note that equation (19) 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 construction $EV_{\theta^p}(x,1) = EV_\theta^p(1,0)$ for all x, we can denote the finite vector of values characterizing the function EV_{θ^p} by $\overline{EV}^p \in \mathbb{R}^{90}$.The value of $\overline{EV}^p$ at state x is denoted by $\overline{EV}_x^p$. The functional equation reduces to the system of equations

$$\overline{EV}_x^p = T_{\theta^p}(\overline{EV}^p)_x \quad \forall x \in \{1,\dots,90\}. \tag{22}$$

Note that due to the sparsity of the transition matrix implied by the mileage transition probabilities (16), the Jacobian matrix of (22) is 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

$$\begin{aligned} Pr(i_t = 1|x_t;\theta^p, EV_{\theta^p}) = \Bigg(1 + \exp\Bigg(& u(x_t,0;\theta_{11}^p,RC^p) + \beta EV_{\theta^p}(x_t,0) && (23)\\ & - u(x_t,1;\theta_{11}^p,RC^p) - \beta EV_{\theta^p}(x_t,1)\Bigg)\Bigg)^{-1} \end{aligned}$$

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

$$
\begin{aligned}
L(\theta;\{x_t,i_t\}_{t=1}^T,EV_\theta) &= \prod_{p\in\mathcal{P}} L(\theta^p;\{x_t,i_t\}_{t=1}^T,EV_{\theta^p}) \\
&= \prod_{p\in\mathcal{P}}\prod_{t=1}^T Pr(i_t|x_t;\theta^p)Pr(x_t|x_{t-1},i_{t-1};\theta_3^p,EV_{\theta^p}).
\end{aligned} \tag{24}
$$

After taking the logarithm of the likelihood function (24), 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

$$
\begin{aligned}
&\max_{(\theta^p,\overline{EV}^p)_{p\in\mathcal{P}}} \sum_{p\in\mathcal{P}} \log L(\theta^p,\overline{EV}^p;\{x_t,i_t\}_{t=1}^T) \\
&\text{s.t. } \overline{EV}_x^p = T_{\theta^p}(\overline{EV}^p)_x \quad \forall x\in\{1,\dots,90\}, \quad \forall p\in\mathcal{P}.
\end{aligned} \tag{25}
$$

3.2 Dynamic Programming with a Discount Factor $\beta \geq 1$

The value function reflects the best attainable expected present value of a utility stream. In infinite horizon problems, this present value diverges for a discount factor $\beta \geq 1$ unless there exists an absorbing state with a utility of 0. However, even if the value function diverges, the *difference* between the value function and its value at some reference state k can still be bounded (Bertsekas, 2012). The literature refers to this difference as the *relative value function*, $ev(x)$. Intuitively, $ev(x)$ corresponds to the difference in the present value of the utility stream between starting from state x as opposed to from the reference state, k. For the long-run average cost problem in which $\beta = 1$, White (1963) proposes the optimality equation

$$
ev(x) = T(ev)(x) - T(ev)(k). \tag{26}
$$

This optimality equation extends to the standard case, $\beta < 1$, (Puterman, 2014). Morton and Wecker (1977) shows its extension to $\beta > 1$ by viewing infinite horizon problems as a approximation of finite horizon problems. Thus, they show the existence and uniqueness for the limiting value and policy functions in finite horizon problems. If the model satisfies their assumptions—which is indeed the case for model studied in this paper—and if the relative value function converges, then the policy is also an optimal policy for the infinite horizon problem.[8]

It remains to show that we can estimate the bus engine replacement model using the *relative* value function instead of the value function. In its standard formulation, the discretized value function for partition p, $\overline{EV}^p \in \mathbb{R}^{90}$, is used to evaluate the choice probabilities. A reformulation of the choice probabilities reveals that the discretized relative values for partition p, $\overline{ev}^p \in \mathbb{R}^{90}$,

[8] Blom Västberg and Karlström (2017) (only available on request) provides an alternative proof for dynamic discrete choice models without cost-free terminal state.

are sufficient to evaluate the choice probabilities:

$$Pr(i_t = 1|x_t; \theta^p) = \left(1 + \exp\left(u(x_t, 0; \theta_{11}^p, RC^p) - u(x_t, 1; \theta_{11}^p, RC^p) + \beta(\underbrace{\overline{EV}_{x_t}^p - \overline{EV}_1^p}_{\equiv \overline{ev}_{x_t}^p})\right)\right)^{-1}. \tag{27}$$

As the likelihood only depends on the choice and the transition probabilities, we can estimate the model using the (expected) relative values, $\overline{ev}^p$.

3.3 Estimation Results

Tracing the estimates as a function of the discount factor in the bus engine replacement model of Rust (1987) 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.

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 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.

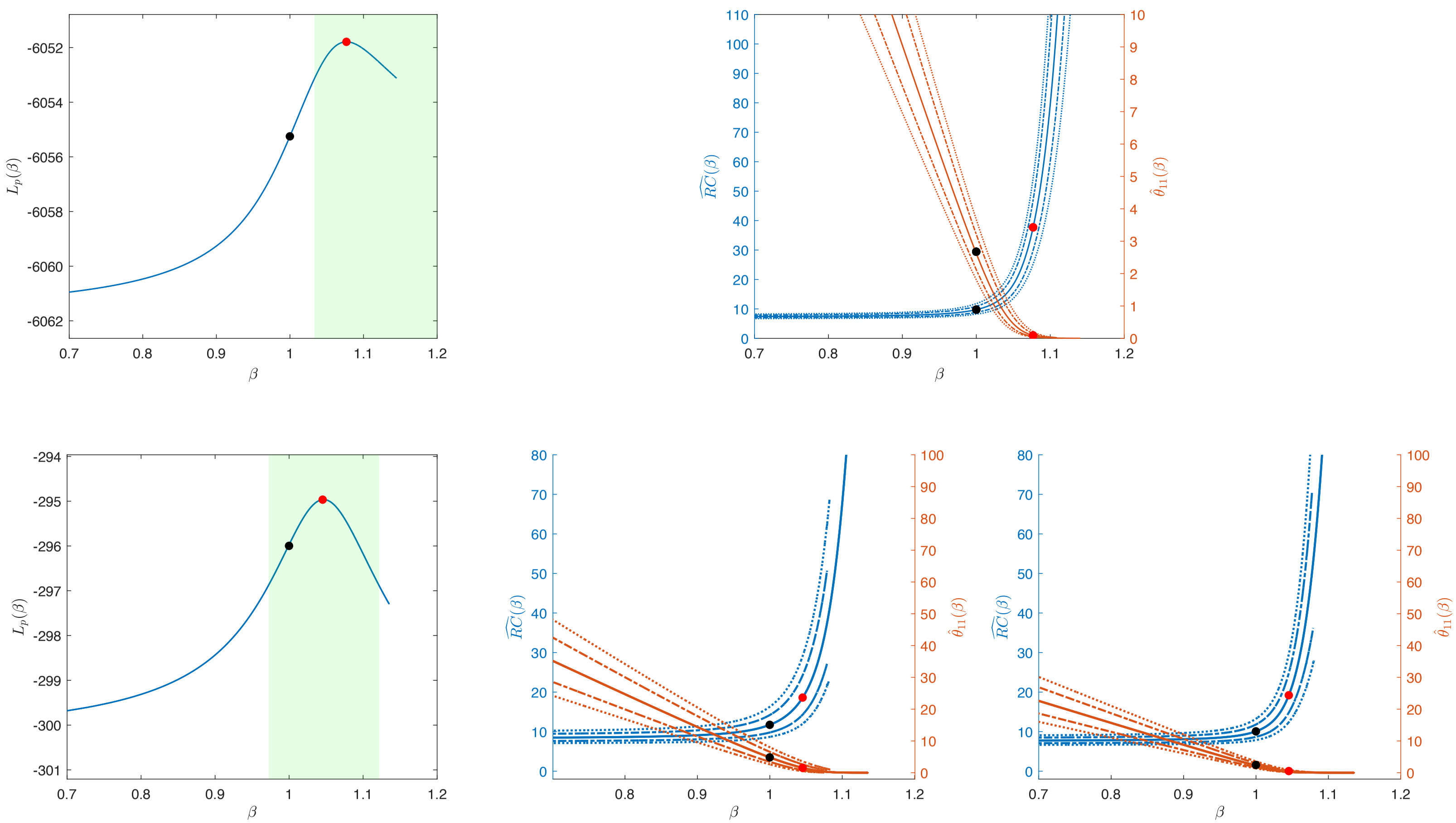


Figure 1: Top: Results for restricted model ($\mathcal{P} = \{\{1,2,3,4\}\}$). Bottom: Results for unrestricted model ($\mathcal{P} = \{\{1,2,3\},\{4\}\}$). Left: Profile likelihood $L_p(\beta)$ including 95% likelihood ratio confidence interval for β (light green). Middle/Right: Parameter estimates 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).

p	Restricted model $\{1,2,3,4\}$	Unrestricted model $\{1,2,3\}$	$\{4\}$
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 $H_0 : \beta = 1$.

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 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).[9] 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

[9]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.

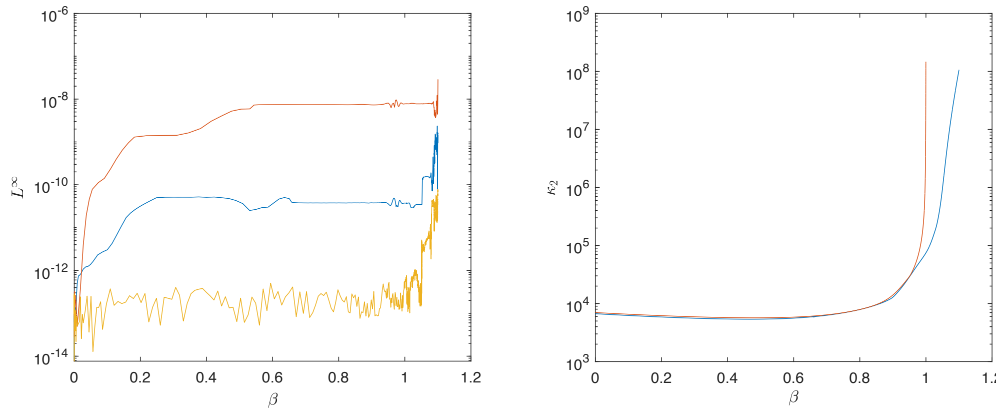


Figure 2: Left: Violation of the first-order conditions as a function of β for three error tolerance configurations of Hompack90 (1e-08, 1e-10, 1e-12). Right: Condition number of the augmented Jacobian, i.e., the Jacobian of the first-order conditions with an additional column containing the partial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

some $\beta > 1$.This is very much in line with the near-singular information matrix 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

During the sample period of the data from 1974 to 1984 the US experienced a major shock to *real interest rates*, a fact widely ignored by the empirical IO literature. By adding a structural break to the discount factor, we find statistical evidence that this shock affected the decision-maker's replacement behavior. We establish a qualitative link between the change in the discount factor estimates around the structural break and the change in the real interest rate. Further, the monthly mileage transitions of the buses changed after four additional bus groups were purchased by the bus company during the sample period. In an extended model with a structural break in the transition probabilities, we find that the discount factor estimate is sensitive to misspecifications in the transition probabilities. Due to the partly prevailing negative real interest rates during this period, the estimate of $\beta > 1$ can not 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. Although Rust (1987) uses only bus groups 1–4 for the estimation, 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.[10] 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 $\mathcal{P}$.[11] The

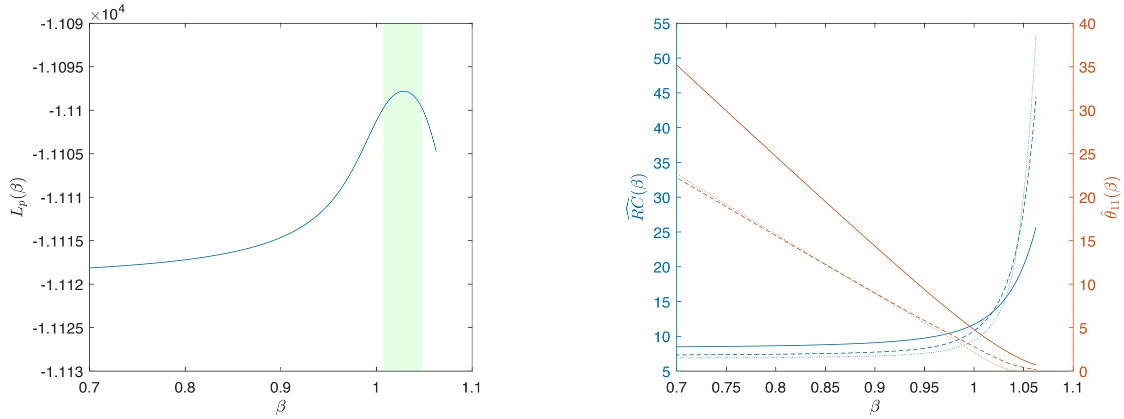


Figure 3: Results for the unrestricted model ($\mathcal{P} = \{\{1,2,3\},\{4,5,7\},\{6,8\}\}$). Left: Profile likelihood $L_p(\beta)$ including 95% likelihood ratio confidence interval for β (light green). Right: Parameter estimates for (θ^p_{11}, RC^p) as functions of β (red and blue, respectively)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

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 95% 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. Table 2 reports the estimates of the main structural parameters and their 95% confidence intervals; Table 5 reports the estimates of the transition probabilities.

3.4.2 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 high inflation during peace times. During its peak, it made “every business decision a speculation on monetary policy” (Long, 1996). 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 4.

This lasted until President Jimmy Carter nominated Paul Volcker as chairman of the FED in

[10] 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.

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

p	$\{1,2,3\}$	$\{4,5,7\}$	$\{6,8\}$
β		1.0283	
		[1.0073, 1.0476]	
RC	14.8725	15.9887	14.2516
	[9.8236, 24.7372]	$[10.6700, \infty)$	$[8.5583, \infty)$
θ_{11}	2.5809	1.2083	0.4154
	[0.7881, 5.6452]	[0.3931, 2.6385]	[0.0114, 1.9924]
LL		-11,097.8159	
p-value ($H_0 : \beta = 1$)		0.0092	

Table 2: Estimates of the main structural parameters, (θ_{11}, RC, β) for the partition set ($\mathcal{P} = \{\{1,2,3\},\{4,5,7\},\{6,8\}\}$); 95% likelihood ratio confidence intervals are reported in brackets (if available; half-closed interval containing the confidence interval otherwise).

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" (U.S. House. Committee On Banking, Housing, And Urban Affairs United States Senate, 1979, p. 4). 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. For more details, see Medley (2013).

Rust (1987) does not specify whether the model is stated in nominal or real terms. The nominal value of goods and services increased in the sample period due to the prevailing high inflation. In the model, however, all quantities are constant over time. In order for the model to fit in its historical context, we argue that all quantities are stated in real terms. Consequently, the decision-maker's discount factor relates to the real interest rate.

Furthermore, we argue that this historical context allows for a natural experiment: the decision-maker acts in a period of a largely unanticipated macroeconomic regime change in the nominal and real interest rates. We test if this regime change affected the decision-maker's discounting by extending the model with an unanticipated structural break in the discount factor: Before the structural break at time t_β, Zurcher discounts by a constant discount factor β_1, whereas he discounts by β_2 after the shock. As the shock is unanticipated for Zurcher, he considers β_1 to be constant from today to infinity in any month before the structural break at time t_{θ_β}. Analogously, he assumes β_2 to be constant from today to infinity after t_{θ_β}. This extends the system of constraints to two Bellman equations per bus group, which we solve simultaneously, once with β_1 and once with β_2.

We employ our proposed homotopy continuation estimation approach by tracing the parameter estimates as a function of the difference of the discount factor before and after the break,

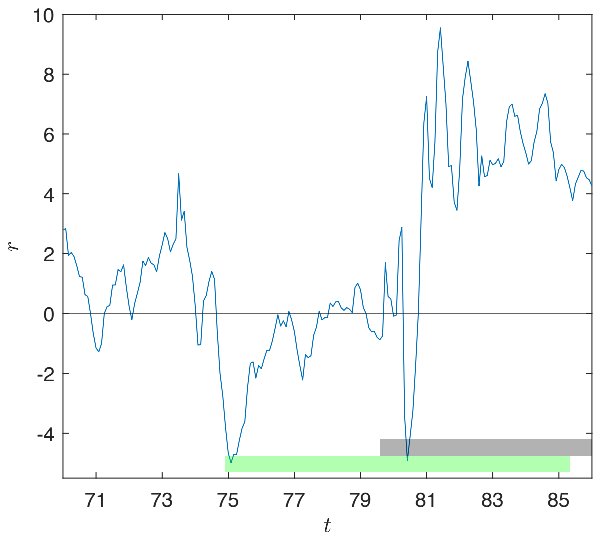


Figure 4: Ex-post real interest rate calculated as the difference of consumer price index and the federal fund rate for July 1970 to August 1985. Grey bar denotes the period of Paul Volcker as chairman of the Fed and the green bar the sample period of Rust (1987).

$\Delta\beta = \beta_1 - \beta_2$, as the controlled parameter. For each $t_\beta \in \{\text{Jan 1976}, \ldots, \text{Dec 1983}\}$, we trace the parameter estimates starting from the restricted model in which $\Delta\beta = 0$. This yields the two-dimensional profile likelihood $L(\Delta\beta, t_\beta)$—continuous in $\Delta\beta$ and discrete in t_β.

Figure 5 depicts this profile likelihood, which increases monotonically 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., $\hat{\beta}_1 > \hat{\beta}_2$.

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 6. The figure shows $L(t_\beta)$, $\beta_1(t_\beta)$, and $\beta_2(t_\beta)$ as functions of t_β. Table 3 reports the maximum likelihood estimates of the main structural parameters and their 95% confidence interval; Table 5 reports the estimates of the transition probabilities. The likelihood takes its maximum for t_β equal to September 1979. The 95% confidence interval around $\hat{t}_\beta$ includes October 1978 to December 1979. The estimates for the discount factors equal $\hat{\beta}_1 = 1.04$ and $\hat{\beta}_2 = 1.02$. We can reject the restricted model in which $\beta_1 = \beta_2$—denoted by the orange dashed line—with a p-value of $1.09 \cdot 10^{-4}$.

We argue that these empirical findings make economic sense: The maximum likelihood estimate of the time of the structural break, $\hat{t}_\beta$, equals September 1979, just one month after Paul Volcker took office as chairman of the Fed. He introduced monetary policies which led to economically significantly higher real interest rates. The drop in the discount factor estimates from $\hat{\beta}_1 = 1.04$ to $\hat{\beta}_2 = 1.02$ follows the expected economically reasonable, qualitative link to the real interest rates: the rise in the real interest rates leads to a fall in the discount factor. The discount factor estimate before September 1979, $\hat{\beta}_1 = 1.04$, falls into a period of largely negative real interest rates. In this period, time value of money is inversed, and thus, the $\beta_1 > 1$

cannot be rejected.

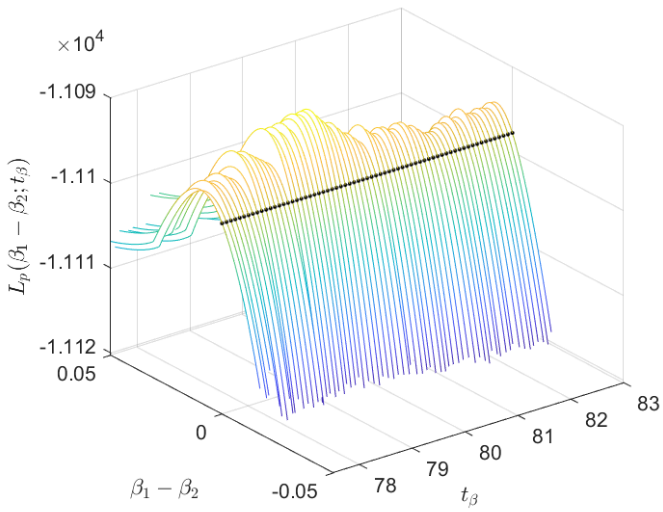


Figure 5: Profile likelihood $L(\Delta\beta, t_\beta)$ as function of $\Delta\beta$ and the time of the structural break t_β. Restricted model with $\Delta\beta = 0$ denoted by black dots.

3.4.3 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 7 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 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 8. 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 be, e.g., a less frequent scheduling of the "old" bus groups or a change in the company's route schedule.

The model assumes the transition probabilities of the bus groups, $\theta_{3,g}$, to be stationary. We extend the model from Section 3.4.2 by relaxing this stationarity assumption with an unanticipated structural break in the transition probabilities for bus groups 4–8.[12] We denote the transition probabilities before the structural break for bus group g by $\theta^1_{3,g}$ and after by $\theta^2_{3,g}$. As the shock is unanticipated for Zurcher, he considers $\theta^1_{3,g}$ to be constant from today to infinity in any month before the structural break at time $t_{\theta_{3,g}}$. Analogously, he assumes $\theta^2_{3,g}$ to be constant from today to infinity after $t_{\theta_{3,g}}$. This extends the system of constraints to a Bellman equation for each parameter combination which we solve simultaneously.

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

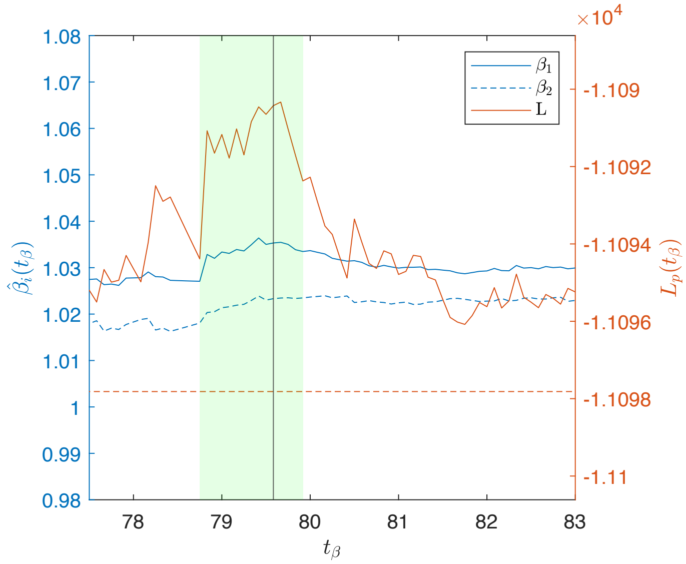


Figure 6: Profile likelihood $L(t_\beta)$ (orange), and $\beta_1(t_\beta)$ (blue solid) and $\beta_2(t_\beta)$ (blue dashed) maximum likelihood estimates for $\mathcal{P} = \{\{1,2,3\},\{4,5,7\},\{6,8\}\}$ as a function of the structural break at time t_β. The t_β for which the log-likelihood is not 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).

Figure 5 depicts the profile likelihood $L(\Delta\beta, t_\beta)$ for tracing the parameter estimates as a function of $\Delta\beta$ for each $t_\beta \in \{\text{Jan } 1976, \ldots, \text{Dec } 1983\}$. The black points denote the restricted model with $\Delta\beta = 0$, i.e., it only has a structural break in the transition probabilities. Note that we can reject this restricted model with a p-value of $2.17 \cdot 10^{-9}$. The shape and location of the peaks indicate that $\Delta\beta$ is well identified and its estimate larger than zero, i.e., again $\beta_1 > \beta_2$.

Further profiling the profile likelihood $L(\Delta\beta, t_\beta)$ in $\Delta\beta$ yields the profile likelihood $L(t_\beta)$. Figure 6 depicts $L(t_\beta)$ as well as the discount factor estimate before the structural break, $\hat{\beta}_1(t_\beta)$, and after the structural break, $\hat{\beta}_2(t_\beta)$. Table 4 reports the maximum likelihood estimates of the main structural parameters and the p-values for two hypothesis test: the restricted model from Section 3.4.1, $(H_0 : \beta_1 = \beta_2, \theta_3^1 = \theta_3^2)$, and the restricted model from Section 3.4.2, $(H_0 : \theta_3^1 = \theta_3^2)$. We can reject both with a p-value of below 10^{-16}. Table 6 reports the estimates of the transition probabilities and of the time of their structural break. The maximum likelihood estimate of t_β equals September 1979. The 95% confidence interval around $\hat{t}_\beta$ includes October 1978 to December 1978 and February 1979 to November 1979.

As in the previous model, $\hat{t}_\beta$ equals September 1979, one month after Paul Volcker took office. Compared to the previous model, the discount factor estimates drop from $\hat{\beta}_1 = 1.04$ and $\hat{\beta}_2 = 1.02$ to $\hat{\beta}_1 = 1.03$ and $\hat{\beta}_2 = 1.00$, respectively. 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

p	$\{1,2,3\}$	$\{4,5,7\}$	$\{6,8\}$
t_β		Sept 1979	
		[Oct 1978, Dec 1979]	
β_1		1.0355	
		$[1.0121, 1.0592]$	
β_2		1.0235	
		$[1.0008, 1.0448]$	
RC	14.1252	16.8884	17.88962
	$[9.5101, 23.2771]$	$[10.8426, \infty)$	$[9.8090, \infty)$
θ_{11}	2.9240	1.4358	0.6183
	$[0.9328, 6.1994]$	$[0.5077, 2.9938]$	$[0.0162, 2.5486]$
LL		-11,090.3327	
p-value		$1.09 \cdot 10^{-4}$	
$(H_0 : \beta_1 = \beta_2)$			
p-value		$3.13 \cdot 10^{-6}$	
$(H_0 : \beta_1 = \beta_2 = 1)$			

Table 3: Estimates of the main structural parameters, (θ_{11}, RC, β), and the time of the structural break in β, t_β, for the partition set $\mathcal{P} = \{\{1,2,3\},\{4,5,7\},\{6,8\}\}$ in the extended model with a structural break in β; 95% likelihood ratio confidence intervals are reported in brackets (if available; half-closed interval containing the confidence interval otherwise).

can compensate the misspecification the law of motion.

The discount factor estimates' drop from $\hat{\beta}_1 = 1.03$ to $\hat{\beta}_2 = 1.00$ follows the expected qualitative link to the real interest rates: the rise in the real interest rates leads to a fall in the discount factor. The estimate $\hat{\beta}_1$ again falls into the period of largely negative real interest rates and thus, cannot be rejected by a time value of money argument. The estimate $\hat{\beta}_2 = 1.00$ corresponds to an agent who maximizes his long-run average utility.

3.4.4 Counterfactual Analyses for $\beta > 1$

In the bus engine replacement model and its extensions, we estimate the discount factor to be larger than or equal to 1. In the prevailing periods of negative real interest rates, a time value of money argument cannot reject these estimates.[13] Regardless of the interest rate environment, some authors have argued that the discount factor can be seen as behavioural parameter. For example, Erdem and Keane (1996) argues that there exist no behavioral reason to preclude $\beta > 1$. This is supported by empirical studies as, e.g., Frederick et al. (2002), showing that the discount factor estimates highly depends on the choice contexts and populations. Thus, we examine the policy implications of $\beta > 1$ by exemplary studying the impact of increasing part costs for the replacement engine on the expected annual engine replacements—the implied demand.

Figure 11 traces the implied demand as a function of the replacement costs and the discount

[13]Even for positive real interest rates, Kocherlakota (1990) shows that competitive equilibria may exist in infinite horizon growth economies for $\beta > 1$.

p	$\{1,2,3\}$	$\{4,5,7\}$	$\{6,8\}$
t_β		Sept 1979	
		[Oct 1978, Dec 1978] $\bigcup$ [Feb 1979, Nov 1979]	
β_1		1.0269	
		$[1.0045, 1.0465]$	
β_2		1.0047	
		$[0.9852, 1.0197]$	
RC	12.0845	12.3563	11.2340
	$[8.7581, 17.5448]$	$[9.3412, 17.8037]$	$[8.0013, 18.0643]$
θ_{11}	4.3878	2.0506	1.1751
	$[2.0846, 7.9242]$	$[0.9921, 3.6144]$	$[0.2321, 3.3417]$
LL		-10,627.1853	
p-value $(H_0 : \theta_3^1 = \theta_3^2)$		$< 10^{-16}$	
p-value $(H_0 : \beta_1 = \beta_2, \theta_3^1 = \theta_3^2)$		$< 10^{-16}$	
p-value $(H_0 : \beta_1 = \beta_2 = 1)$		$6.55 \cdot 10^{-10}$	

Table 4: Estimates of the main structural parameters, (θ_{11}, RC, β), and the time of the structural break in β, t_β, for the partition set $\mathcal{P} = \{\{1,2,3\}, \{4,5,7\}, \{6,8\}\}$ in the extended model with a structural break in the discount factor and in the transition probabilities; 95% likelihood ratio confidence intervals are reported in brackets.

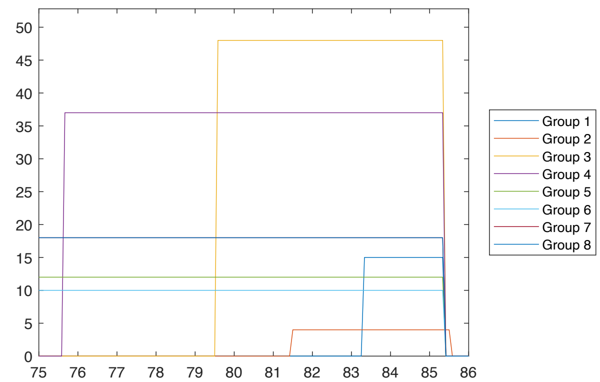


Figure 7: Number of buses per bus group in each month.

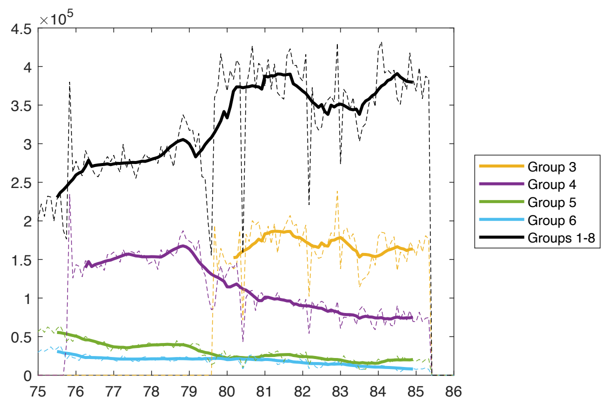


Figure 8: Centered 12-month moving average aggregated mileage (solid) and monthly mileage transitions (dashed) for bus groups 3–6 and the entire fleet.

factor for a single bus with the characteristics of bus group 4. We first focus on the left panel of Figure 11a. For $\beta = 0$, the implied demand is highly sensitive to changes in the replacement bus engine costs. The implied demand seen as a function of β appears to flatten continuously in β without a (visible) singularity one might expect at $\beta = 1$. The solution manifold in the right panel of Figure 11b supports this observation. The predictions for $\beta > 1$ are a reasonable extension of the model predictions for $\beta < 1$: the price elasticity of demand decreases with increasing β. This exemplary counterfactual analysis shows that policy implications derived from a model with $\beta > 1$ can be reasonable.

4 Conclusion

In this paper, we present the necessary mathematical, statistical, and numerical tools to trace the estimates of the structural parameters of a model and their confidence intervals, in dependence on a controlled parameter, based on the augmented profile likelihood using homotopy parameter

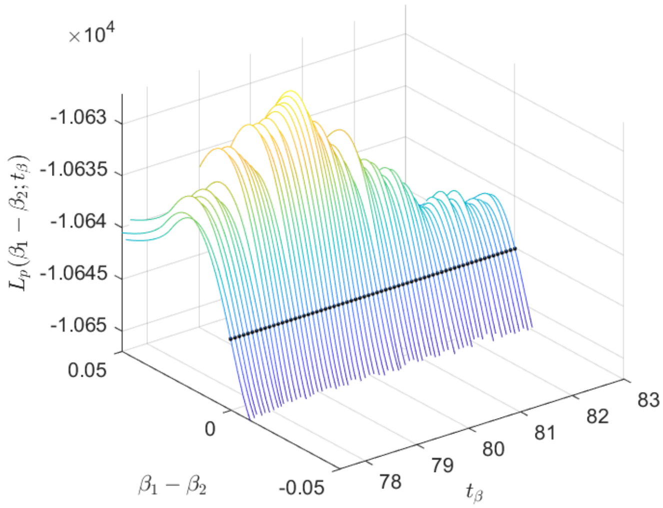


Figure 9: Profile likelihood $L(\Delta\beta, t_\beta)$ as function of $\Delta\beta$ and the time of the structural break t_β. Restricted model with $\Delta\beta = 0$ denoted by black dots.

continuation.

Applying the method to the bus engine replacement model of Rust (1987), we find that—contrary to common belief—the discount factor is well identified. Its estimate is significantly larger than 1 for the sample considered in the original paper. In two extensions of the original model with an unanticipated structural break in β, we find empirical evidence that the historical monetary policy shock after Paul Volcker took office affected the decision-maker's discounting. In particular, we empirically establish a economically reasonable, qualitative link: The discount factor estimates drop considerably from before to after the breaks, whose optimal times we estimate to be just one month after Paul Volcker took office. This drop in the discount factor estimates matches the economic intuition. The discount factor estimates before the structural break, $\beta > 1$, fall into a period of negative interest rates; thus, they do not contradict the time value of money.

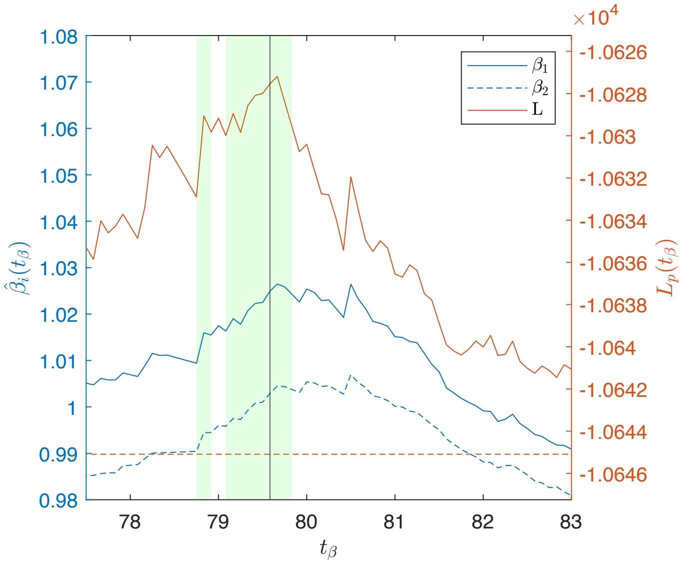
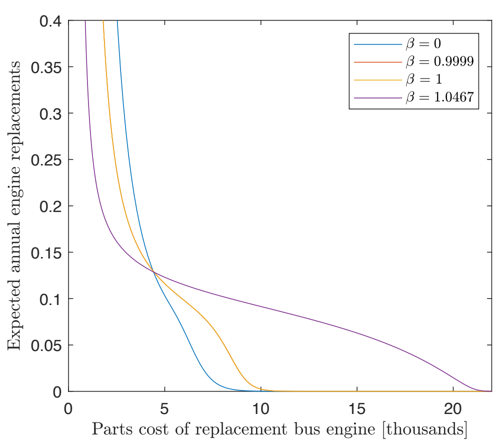
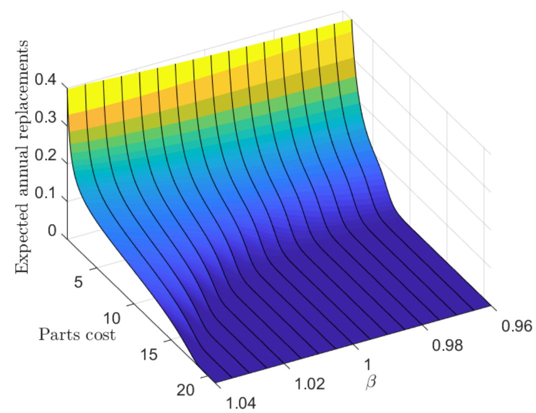


Figure 10: Profile likelihood $L(t_\beta)$ (orange), and $\beta_1(t_\beta)$ (blue solid) and $\beta_2(t_\beta)$ (blue dashed) maximum likelihood estimates for $\mathcal{P} = \{\{1,2,3\},\{4,5,7\},\{6,8\}\}$ as a function of the structural break at time t_β. The t_β for which the log-likelihood is not 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).



(a) Implied demand as a function of RC for $\beta \in [0, 0.9999, 1, 1.0467]$



(b) Implied demand as a function of RC and of β

Figure 11: Expected annual engine replacement—the implied demand—in group 4 in 1985 dollars for varying discount factors, β.

References

Abbring, J. H. and Daljord, Ø. (2020). Identifying the Discount Factor in Dynamic Discrete Choice Models. *Quantitative Economics*, 11(2):471–501.

Aguirregabiria, V. (2012). A method for implementing counterfactual experiments in models with multiple equilibria. *Economics Letters*, 114(2):190–194.

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

Allgower, E. L. and Georg, K. (2012). *Introduction to Numerical Continuation Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Andersson, J. A. E., Gillis, J., Horn, G., Rawlings, J. B., and Diehl, M. (2019). CasADi: a software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 11(1):1–36.

Barz, C. (2007). *Risk-Averse Capacity Control in Revenue Management*, volume 597 of *Lecture Notes in Economics and Mathematical Systems*. Springer Science & Business Media, Berlin, Heidelberg.

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.

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.

Egesdal, M., Lai, Z., and Su, C. (2015). Estimating dynamic discrete-choice games of incomplete information. *Quantitative Economics*, 6(3):567–597.

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

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*.

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

Kocherlakota, N. R. (1990). On the 'discount' factor in growth economies. *Journal of Monetary Economics*, 25(1):43–47.

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.

Lanz, A., Müller, P., Reich, G., and Wilms, O. (2021). 'Small Data': Efficient Inference with Occasionally Observed States. *SSRN Electronic Journal*.

Long, J. B. D. (1996). America's only peacetime inflation: the 1970s. *NBER Working Paper Series*.

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

Medley, B. (2013). Volcker's Announcement of Anti-Inflation Measures.

Morton, T. E. and Wecker, W. E. (1977). Discounting, Ergodicity and Convergence for Markov Decision Processes. *Management Science*, 23(8):890–900.

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.

Su, C.-L. (2014). Estimating discrete-choice games of incomplete information: Simple static examples. *Quantitative marketing and economics*, 12(2):167–207.

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.

U.S. House. Committee On Banking, Housing, And Urban Affairs United States Senate (1979). *Hearing Before The Committee ... On The Nomination Of Paul A. Volcker To Be Chairman, Board Of Governors, Federal Reserve System.* Accessed via FRASER—Federal Reserve Bank of St. Louis.

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.

Zangwill, W. I. and Garcia, C. B. (1981). *Pathways to solutions, fixed points, and equilibria.* Prentice Hall.

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.

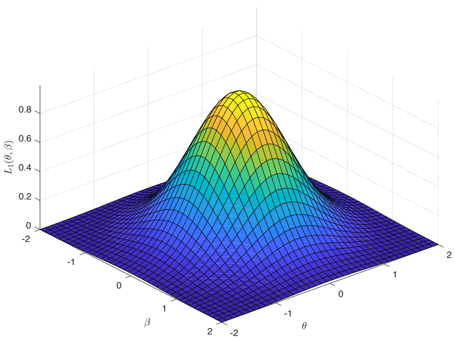
As commonly done in the literature, we restrict our attention to models where the parameter vector θ is identified—or, more precisely, the distribution of an observation implied by the model is different for different values of θ. However, this is no guarantee that, for finite samples, the likelihood 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.[14] 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.

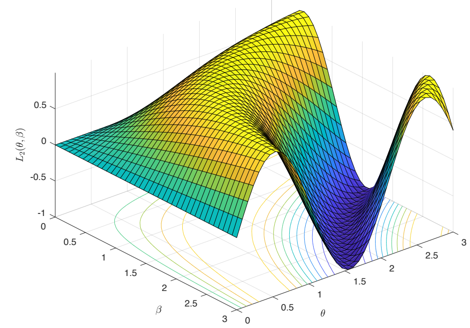
As we have noted, no identification requirements are stated for β. In fact, a key motivation of this paper is to show how poorly or non-identified parameters can be traced out using the profile likelihood function. Consider the examples in Figure 12: The top left panel shows a function for which the 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 (29), 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, $\mathbf{\Sigma}(\theta, \beta)$. In fact, multiplicity does not contradict identification per se, as long as the likelihood itself discriminates the model solutions properly—that is, as long

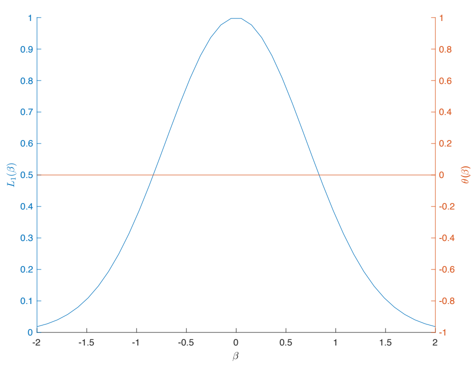
[14]This argument can be taken even further by arguing that for continuous likelihood function, even in the very vicinity of a maximum there exist infinitely many parameter values with equal likelihood.



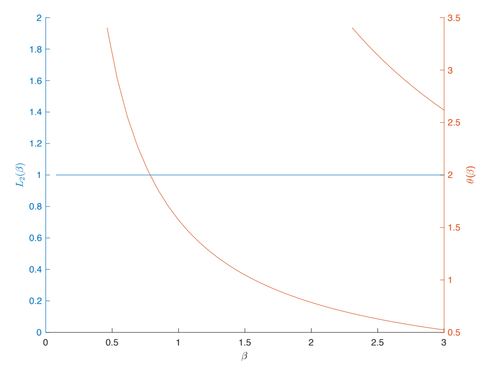
(a) $L_1 : \exp(-\theta^2 - \beta^2)$



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

as the maximum of the profile likelihood w.r.t.ãll structural parameters,

$$L_p(\hat{\theta}, \beta) \equiv \max_{\sigma \in \mathbf{\Sigma}(\hat{\theta},\beta)} L(\hat{\theta}, \beta, \sigma), \tag{28}$$

at all local solutions of $\hat{\theta}$ is locally unique. (Note that local uniqueness of (28) 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 (29) 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 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):

$$\nabla^2_{\mu,\theta,\sigma}\mathcal{L}(\theta, \sigma, \mu; \beta) \equiv \begin{pmatrix} 0 & -D_\theta h(\sigma; \theta, \beta) & -D_\sigma h(\sigma; \theta, \beta) \\ -D_\theta h(\sigma; \theta, \beta)^T & \multicolumn{2}{c}{\multirow{2}{*}{$\nabla^2_{\theta,\sigma}\mathcal{L}(\theta, \sigma, \mu; \beta)$}} \\ -D_\sigma h(\sigma; \theta, \beta)^T & & \end{pmatrix}, \tag{29}$$

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):

$$v^T \nabla^2_{\theta,\sigma}\mathcal{L}(\hat{\theta}, \hat{\sigma}, \hat{\mu}; \beta) v < 0 \quad \forall v \neq 0 : D_{\theta,\sigma}h(\hat{\sigma}; \hat{\theta}, \beta) v = 0. \tag{30}$$

In summary, if the point $(\hat{\theta}, \hat{\sigma}, \beta)$ together with $\hat{\mu}$ satisfies (8) and (30), 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 (**?**, Cor. 7).

g	1	2	3	4	5	6	7	8
$\theta_{30,g}$	0.1972	0.3906	0.3071	0.3919	0.4887	0.6184	0.6000	0.7218
$\theta_{31,g}$	0.7889	0.5990	0.6827	0.5953	0.5067	0.3816	0.3973	0.2782
$\theta_{32,g}$	0.139	0.104	0.0103	0.0128	0.0047	0.0000	0.0027	0.0000

Table 5: Estimates of the transition probabilities θ_3^g.

g	1	2	3	4	5	6	7	8
$\theta^1_{30,g}$	0.1972	0.3906	0.3071	2305	0.2778	0.5447	0.4818	0.6566
$\theta^1_{31,g}$	0.7889	0.5990	0.6827	0.7430	0.7108	0.4553	0.5121	0.3434
$\theta^1_{32,g}$	0.139	0.104	0.0103	0.0265	0.0114	0.0000	0.0061	0.0000
$\theta^2_{30,g}$	0.1972	0.3906	0.3071	0.5374	0.6340	0.7750	0.6929	0.8264
$\theta^2_{31,g}$	0.7889	0.5990	0.6827	0.4621	0.3660	0.2250	0.3071	0.1736
$\theta^2_{32,g}$	0.139	0.104	0.0103	0.0004	0.0000	0.0000	0.0000	0.0000
$t_{\theta_{3,g}}$	-	-	-	May ’80	Apr ’79	Feb ’82	Aug ’79	Jun ’81

Table 6: Estimates of the transition probabilities, $\theta_{3,g}$, and of the time of their structural break, $t_{\theta_{3,g}}$.

A.3 Transition Probability Estimates