Submitted to *Operations Research*
manuscript 1

Dynamic Discrete Choice Estimation with Partially Observable States and Hidden Dynamics

Yanling Chang
Department of Engineering Technology and Industrial Distribution and Department of Industrial and Systems Engineering, Texas A&M University, TX 77843, yanling.chang@tamu.edu

Alfredo A. Garcia
Department of Industrial and Systems Engineering, Texas A&M University, TX 77843, alfredo.garcia@tamu.edu

Zhide Wang
Department of Industrial and Systems Engineering, Texas A&M University, TX 77843, liang93429@tamu.edu

Dynamic discrete choice models are used to estimate the intertemporal preferences of an agent as described by a reward function based upon observable histories of states and implemented actions. However, in many applications, such as reliability and healthcare, the system state is only partially observable or hidden (e.g., the level of deterioration of an engine, the condition of a disease), and the decision maker only has access to information imperfectly correlated with the true value of the hidden state. In this paper, we consider the estimation of a dynamic discrete choice model with state variables and system dynamics hidden to *both* the agent and the modeler, thus generalizing the model in Rust (1987) to partially observable cases. We examine the structural properties of the model and prove that this model is still identifiable if the cardinality of the state space, the discount factor, the distribution of random shocks, and the rewards for a given (reference) action are given. We analyze both theoretically and numerically the potential mis-specification errors that may be incurred when the Rust's model is improperly used in partially observable settings. We further apply the model to a subset of dataset in Rust (1987) for bus engine mileage and replacement decisions. The results show that our model can improve model fit as measured by the log-likelihood function by 17.7% and the log-likelihood ratio test shows that our model statistically outperforms the Rust's model. Interestingly, our hidden state model also reveals an economically meaningful route assignment behavior in the dataset which was hitherto ignored, i.e. routes with lower mileage are assigned to buses believed to be in worse condition.

1. Introduction

We consider the task of training a model of dynamic decisions by a single human agent based upon the history of implemented actions with hidden (or partially observable) states and system dynamics. Under the assumption of complete state observability, this problem has been widely studied in two strands in the literature where it is referred to as *inverse reinforcement learning* (IRL) or alternatively as *structural estimation*.

The bus engine replacement model developed in Rust's seminal paper (Rust 1987) has served as a point of reference for this literature over decades. The model assumes that at each time period and for every bus in the fleet, Mr. Zurcher (a superintendent of maintenance of city of Madison's Metropolitan Bus Company) had to choose between replacing the engine or continuing to operate at a cost which included maintenance and loss of ridership in the case of a breakdown. The model assumes the accumulated mileage is a sufficient statistic for determining the on-going cost of operation for a given bus at any time period. To capture other unobservables likely affecting Mr. Zurcher's decisions, the model features random cost perturbations which are privately observed by Mr. Zurcher. Hence, Mr. Zurcher's decisions are modeled as resulting from a Markov Decision Process (MDP) subject to independent and identically distributed (i.i.d.) cost perturbations.

Assuming cumulative mileage as a sufficient statistic for making replacement decisions goes counter to much of the reliability literature on models of cumulative damage (see e.g., Sanchez and Klutke 2016). Cumulative mileage and other specialized tests only provide informative signals of the true underlying engine state. In fact, identifying a set of performance indicators closely related to the engine degradation state is a critical issue in predictive maintenance literature (Simoes et al. 2017). In addition, while some extensions of the Rust's model have considered serial correlation in cost perturbations (see e.g., Reich 2018), serial correlation may in fact be an *endogenous* feature of the dynamic decision making process: for example, in order to minimize maintenance costs, Mr. Zurcher could have conceivably assigned higher mileage routes to buses with engines he *believed* were in a better condition.

Rather than modeling Mr. Zurcher's decisions as resulting from a MDP with (possibly serially correlated) cost perturbations, it seems more appropriate to model such decisions as consistent with an evolving *belief* on the level of *unobservable* damage or deterioration of the engine. In this paper, we develop a dynamic discrete choice model of an agent's decisions as resulting from a partially observable Markov decision process (POMDP). At each stage, the agent (e.g., Mr. Zurcher) collects observations (e.g., mileage) that are imperfectly correlated with the state and makes decisions based on the *entire* history of his/her observations and actions. As in the original Rust's model, we assume that not all these observations are available to the modeler. For a given set of publicly revealed observations and action sequences, our objective is to develop a new estimation method for determining maximum likelihood estimates of the primitive parameters of the underlying controlled stochastic process, including the agent's reward structure and the system *hidden* dynamics. We show that the belief of the unobserved system state is a sufficient statistic for the proposed hidden state dynamic discrete choice model. Numerical testings on both synthetic datasets and on the Rust's dataset demonstrate that our model significantly outperforms the (completely observable) Rust's model as it identifies features which the Rust's model was unable to detect.

Making dynamic decisions when the relevant state variable is only partially observable is a general trait in many application domains. For example, in healthcare settings, the true physiological state of a patient, especially for cancers and chronic diseases (Steimle and Denton 2017), is only imperfectly known even with the most advanced testing technologies. In these cases, it is not reasonable to use a model relying on the assumption of the state being observable as it may inevitably incur model mis-specification errors. It is well known that model mis-specification errors will not diminish as sample size increases, and more importantly, model estimation can lead to erroneous and/or misleading results.

Unfortunately, this fundamental issue has not been examined in literature. The econometrics research has been focused on (i) developing efficient algorithms to address computational challenges in estimating the structural parameters such as determining the value function used in the likelihood function estimation (Hotz and Miller 1993, Hotz et al. 1994, Aguirregabiria and Mira 2002, 2010, Su and Judd 2012, Kasahara and Shimotsu 2018); (ii) relaxing restrictive assumptions on the unobservables to the modeler such as considering permanent heterogeneity (Arcidiacono and Miller 2011, Arcidiacono and Jones 2003), serially correlated unobservables (Blevins 2016, Reich 2018), unobservables correlated across choices (Keane and Wolpin 1997, Diermeier et al. 2005), etc.; (iii) analyzing approximation errors, inference and validation (Hendel and Nevo 2006, Imai et al. 2009); (iv) examining nonparametric and semiparametric identification issues and estimation (Abbring and Daljord 2020, Magnac and Thesmar 2002). Dynamic discrete choice estimation methods have been widely used in applications including retirement behaviors (Rust and Phelan 1997), occupational choices and career decisions of young men (Eckstein and Wolpin 1999), incentives to get teachers to work (Duflo et al. 2012), adult women's mammography decisions (Fang and Wang 2015), trade and labor markets (Caliendo et al. 2019), car ownership (Cirillo et al. 2016), etc. We remark that the "unobservables" mentioned in the literature specifically refers to state components that are not observed by the modeler (but they are completely observable to the agent). On the contrary, the "hidden state" in this paper refers to the situations where the system state is not observable to both the agent and the modeler.

A maximum entropy method proposed in Ziebart et al. (2008) has been highly influential in the computer science literature. This method can be seen as an information theoretic derivation of the Rust's nested loop estimation (Rust 1987). Sample-based algorithms for implementing the maximum entropy method have scaled to scenarios with nonlinear reward functions (see e.g., Boularias et al. 2011, Finn et al. 2016). In Choi and Kim (2011), the authors extended the maximum entropy estimation method to a partially observable environment, assuming both the transition probabilities and observation probabilities are known with domain knowledge. These methods have also been used for apprenticeship learning where a robot learns from expert-based demonstrations (Ziebart et al. 2008). To our best knowledge, none of these existing works have developed a general methodology to jointly estimate the reward structure and system dynamics based on trajectories of a POMDP.

We now summarize the main contributions of this paper. First, we develop a new dynamic discrete choice model which generalizes the Rust's model (Rust 1987) by taking into consideration that both the system state and the system dynamics may be hidden. Secondly, although the system state and dynamics are not directly observable from a dataset like in the Rust's model, we prove that the model is still identifiable if the cardinality of the system state space, the distribution of random shocks, the discount factor, and the reward structure in a reference action are known. Thirdly, when the dimensions of the possible system states and observations are the same, we characterize the deviation between the estimated quantities from the Rust's model and the quantities that the modeler intends to measure. This deviation is induced by using a mis-specified model, namely, applying a model for completely observable cases to situations where the system state and dynamics are hidden. The potential discrepancies are further illustrated via numerical examples. Lastly, we apply the model to the widely studied Rust's engine replacement dataset. We show that our new estimation method can dramatically improve the data fit by 17.7% in terms of the log-likelihood. The log-likelihood ratio test also suggests that our approach performs significantly better than the Rust's model. Furthermore, the results from our model indicate that Mr. Zurcher was trying to optimize route assignments based on engine conditions. This is a feature of Mr. Zurcher's decisions that has been hitherto ignored.

The paper is organized as follows. Section 2 provides an overview of the classical dynamic discrete choice model with observable states, which we will compare with and generalize. Section 3 presents our dynamic discrete choice model for partially observable states and hidden dynamics with model formulation in Section 3.1 and structural results in Section 3.2. The identification results are presented in Section 4. In Section 5, we present the deviation between the quantities that the modeler intends to analyze from what he/she may get when the modeler applies the Rust's model to situations where the system state and dynamics are hidden. This deviation represents the so called "specification error" in statistical analysis when the selected model poorly represents the underlying data generation process. The numerical evaluation in Section 6 tests and validates our hidden state model from two perspectives. Section 6.1 applies our hidden state model to the bus engine replacement data in Rust (1987), and compares to the results from the Rust's model. In Section 6.2, we perform the model validation using synthetic datasets generated by predefined models. Section 7 summarizes research results and discusses future research directions.

2. Dynamic Discrete Choice Estimation with Observable States

We now briefly review the problem of constructing a model of dynamic decision making by a single human agent based upon the history of implemented actions and states with observable state in Rust (1994).

At time $t \geq 0$, the human agent implements an action a_t from the action space A and receives a reward $r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t)$, where s_t is the system state at time t from the state space S, $r_{\theta_1}(s_t, a_t)$ is the parametrized reward associated to the state-action pair (s_t, a_t), $\theta_1 \in \mathbb{R}^{p_1}$ for some $p_1 \in N_+$, and $\epsilon_t(a_t)$ is a random perturbation. The cardinality of A is finite, $|A| < \infty$.

Upon implementing the action $a_t \in A$, the system state evolves according to a Markov process with parametrized conditional probabilities $P_{\theta_2}(s_{t+1}, \epsilon_{t+1}|s_t, \epsilon_t, a_t)$ where $\theta_2 \in \mathbb{R}^{p_2}$ for some $p_2 \in N_+$. The conditional independence assumption in Rust (1994) assumes:

$$P(s_{t+1}, \epsilon_{t+1}|s_t, \epsilon_t, a_t) = P(\epsilon_{t+1}|s_{t+1}) P(s_{t+1}|s_t, a_t).$$

In addition, the reward perturbation vectors $\{\epsilon_t \in \mathbb{R}^{|A|}, t > 0\}$ are assumed to be independently and identically distributed (i.i.d.) over time with probability measure μ. Thus, the decision process of the agent can be modelled as

$$V_\theta(s_t, \epsilon_t) = \max_{a_t \in A} [r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t) + \beta \sum_{s_{t+1} \in S} \int V_\theta(s_{t+1}, \epsilon_{t+1}) P_{\theta_2}(s_{t+1}|s_t, a_t) d\mu(\epsilon_{t+1}|s_{t+1})], \tag{1}$$

where $\theta = (\theta_1, \theta_2)$, and $\beta \in [0, 1)$ is the discount factor. The above equation can be rewritten as

$$V_\theta(s_t, \epsilon_t) = \max_{a_t \in A} [r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t) + \beta \sum_{s_{t+1} \in S} \bar{V}_\theta(s_{t+1}) P_{\theta_2}(s_{t+1}|s_t, a_t)] \tag{2}$$

where $\bar{V}_\theta(s_{t+1}) = \int V_\theta(s_{t+1}, \epsilon_{t+1}) d\mu(\epsilon_{t+1}|s_{t+1})$.

A (Markovian) model of the human agent's decisions is a function $\pi_\theta(a|s)$ (belonging to a dimension $|A|-1$ simplex), which gives the probability that the human agent implements action a when the state is s. This function is also referred to as the conditional choice probability (CCP) function in the literature. When the distribution of the random perturbations $\{\epsilon_t(a), t \geq 0\}$ are i.i.d. standard Gumbel for all $a \in A$, the CCP's are of the form:

$$\pi_\theta(a|s) = \frac{\exp Q_\theta(s,a)}{\sum_{a' \in A} \exp Q_\theta(s,a')} \tag{3}$$

where

$$Q_\theta(s,a) = r_{\theta_1}(s,a) + \beta \sum_{s' \in S} \bar{V}_\theta(s') P_{\theta_2}(s'|s,a) \tag{4}$$

$$\bar{V}_\theta(s) = \gamma + \log \sum_{a \in A} \exp(Q_\theta(s,a)), \tag{5}$$

and $\gamma > 0$ is the Euler's constant. For data in the form of a collection of $N > 0$ independent finite sequences of state-action pairs $\{(s_{t,i}, a_{t,i}), 0 \leq t \leq T\}_{i=1}^N$, a likelihood function is defined as:

$$\begin{aligned}\log \ell(\theta) &\triangleq \log \prod_{i=0}^{N} \prod_{t=0}^{T} \pi_\theta(a_{t,i}|s_{t,i}) P_{\theta_2}(s_{t+1,i}|s_{t,i}, a_{t,i}) \\ &= \sum_{i=1}^{N} \left(\sum_{t=0}^{T} \log \pi_\theta(a_{t,i}|s_{t,i}) + \sum_{t=0}^{T-1} \log P_{\theta_2}(s_{t+1,i}|s_{t,i}, a_{t,i}) \right)\end{aligned} \tag{6}$$

A model of the agent can be obtained by finding the parameter $\theta^* \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ that maximizes the log likelihood in (6).

3. Dynamic Discrete Choice for Hidden States and Dynamics

This section generalizes the Rust's framework in Rust (1994) for the completely observable case to the partially observable case.

3.1. Model Formulation

At each decision epoch $t \geq 0$, the value of state s_t is not directly observable to the human agent nor to the external modeler. Both the human agent and the external modeler are able to receive the value of a *public* random variable $z_t \in Z$ correlated with the underlying state s_t, assuming the cardinality of the observation space Z is finite. As in the Rust's model, we assume the human agent observes a private signal (or reward perturbation) $\epsilon_t(a_t)$ when implementing action $a_t \in A$. If the hidden state is s_t, the reward accrued is $r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t)$ where $\theta_1 \in \mathbb{R}^{p_1}$ for some $p_1 \in N_+$. The system dynamics is described by probabilities $P_{\theta_2}(z_{t+1}, \epsilon_{t+1}, s_{t+1}|s_t, a_t)$ where $\theta_2 \in \mathbb{R}^{p_2}$ for some $p_2 \in \mathbb{N}_+$; see Figure 1 for a schematic representation.

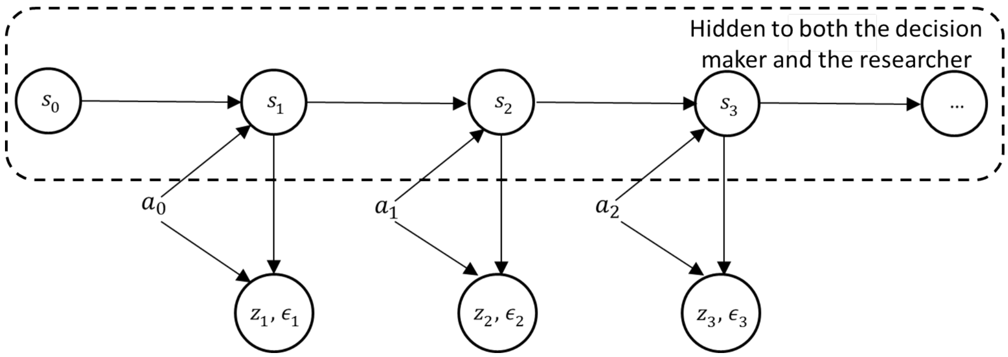


Figure 1 Graphical illustration of the proposed dynamic discrete choice model with partially observable states and hidden dynamics. At each stage, z_t is observed by both the human agent and the modeler, and ϵ_t is privately observed by the human agent.

Let $\zeta_t = \{z_t, ..., z_1, a_{t-1}, ..., a_0, x_0\}$ be the publicly received history of the dynamic decision process including all past and present revealed observations and all past actions at time $t > 0$, where $x_0 = \{P(s_0), s_0 \in S\}$ is the prior belief vector over S, assuming $|S| < \infty$. Similar to Rust (1994), we assume

Additively Separable (AS) Rewards:

$$\sum_{s_t} P(s_t|\zeta_t, \epsilon_t) r_{\theta_1}(s_t, a_t) = \sum_{s_t} P(s_t|\zeta_t) r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t); \tag{7}$$

Conditional Independence (CI) Assumption:

$$P_{\theta_2}(z_{t+1}, \epsilon_{t+1}|\zeta_t, \epsilon_t, a_t) = P(\epsilon_{t+1}|z_{t+1}) P_{\theta_2}(z_{t+1}|\zeta_t, a_t). \tag{8}$$

Note that different from the CI Assumption in Rust (1987), $\{\zeta_t, \epsilon_t\}$ is not Markovian. However, z_{t+1} is a sufficient statistic for ϵ_{t+1}, indicating ϵ_t and ϵ_{t+1} are independent given z_{t+1} and

$$P_{\theta_2}(z_{t+1}, \epsilon_{t+1}, s_{t+1}|s_t, a_t) = P(\epsilon_{t+1}|z_{t+1}) P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t). \tag{9}$$

In addition, the conditional probability $P(z_{t+1}|\zeta_t, a_t)$ does not depend on ϵ_t.

Under Assumptions AS and CI, the decision making process of the agent can be modelled as

$$\begin{aligned} V_{t,\theta}(\zeta_t, \epsilon_t) = \max_{a_t \in A} \Big\{ & \sum_{s_t} P(s_t|\zeta_t) r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t) \\ & + \beta \sum_{z_{t+1}} \int P_{\theta_2}(z_{t+1}|\zeta_t, a_t) V_{t+1,\theta}(\zeta_{t+1}, \epsilon_{t+1}) d\mu(\epsilon_{t+1}|z_{t+1}) \Big\}, \end{aligned} \tag{10}$$

which can be viewed as a POMDP with perturbations $\{\epsilon_t : t \geq 0\}$ and $\{P(s_t|\zeta_t) : t \geq 0\}$ is commonly called the belief over system state s_t, given the history at stage t. We thus explain the random shock as follows. If the modeler knew the values of (ζ_t, ϵ_t), then the modeler could also replicate the solution of the POMDP in (10). Because of this noise, the modeler does not know exactly the true value of belief $\{P(s_t|\zeta_t, \epsilon_t)\}$, resulting in a perturbation to the one-step reward.

The objective of the modeler is to identify θ_1 in the reward structure $r_{\theta_1}(s_t, a_t)$, and θ_2 in the dynamics $P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t)$ from the publicly received histories $\{\zeta_T^i\}_{i=1}^N$. Consequently, the underlying hidden state transition probabilities can be determined by

$$P_{\theta_2}(s_{t+1}|s_t, a_t) = \sum_{z_{t+1}} P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t), \tag{11}$$

and the observation probabilities $P_{\theta_2}(z_{t+1}|s_{t+1}, s_t, a_t)$ can also be obtained from

$$P_{\theta_2}(z_{t+1}|s_{t+1}, s_t, a_t) = \frac{P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t)}{\sum_{z_{t+1}} P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t)}, \tag{12}$$

assuming $\sum_{z_{t+1}} P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t) > 0$.

3.2. Structural Results

The main challenge of Eq. (10) is that the cardinality of ζ_t grows to infinity as t increases. We now present a computationally attractive form of ζ_t for Eq. (10) and correspondingly, a new estimation method for the partially observable case.

Denote $P(z, a) = \{P_{s,s'}(z, a)\}$, where (White 1991)

$$P_{s,s'}(z, a) = P(z_{t+1} = z, s_{t+1} = s'|s_t = s, a_t = a). \tag{13}$$

Let $x_t = \{x_t(s) : s \in S\}$, where $x_t(s) = P(s_t = s|\zeta_t)$ and $x_0 = \{P(s_0) : s_0 \in S\}$ is given. Then $\forall t$, $x_t \in X = \{x \in R^{|S|} : x(s) \geq 0, s \in S, \sum_{s\in S} x(s) = 1\}$, and since $|S| < \infty$, x_t is finite dimensional. By the Bayes' rule, White (1991) shows $\{x_t, t = 0, 1, ...\}$ is a controlled Markov process since there is a function $\lambda(z_{t+1}, x_t, a_t)$ such that

$$x_{t+1} = \lambda(z_{t+1}, x_t, a_t) = \frac{x_t P(z_{t+1}, a_t)}{\sigma(z_{t+1}, x_t, a_t)}, \quad (\sigma(z_{t+1}, x_t, a_t) \neq 0) \tag{14}$$

$$\sigma(z_{t+1}, x_t, a_t) = P(z_{t+1}|\zeta_t, a_t) = x_t P(z_{t+1}, a_t)1, \tag{15}$$

where

$$[x_t P(z_{t+1}, a_t)]_{s'} = \sum_s x_t(s) P_{s,s'}(z_{t+1}, a_t),$$

$$x_t P(z_{t+1}, a_t)1 = \sum_{s'} \sum_s x_t(s) P_{s,s'}(z_{t+1}, a_t).$$

Define $r(x_t, a_t) = \sum_{s_t} x_t(s_t) r(s_t, a_t)$. The next theorem leverages results in the POMDP literature and shows that the belief process $\{x_t, t \geq 0\}$ is a sufficient statistic for our hidden state dynamic discrete choice model.

Theorem 1 *$x_t(s_t) = P(s_t|\zeta_t)$ is a sufficient statistic to Eq.* (10)*, and thus*

$$\begin{aligned} V_{t,\theta}(\zeta_t, \epsilon_t) = V_{t,\theta}(x_t, \epsilon_t) = \max_{a_t \in A} \Big\{ & r_{\theta_1}(x_t, a_t) + \epsilon_t(a_t) \\ & + \beta \sum_{z_{t+1}} \int \sigma_{\theta_2}(z_{t+1}, x_t, a_t) V_{t+1,\theta}(\lambda(z_{t+1}, x_t, a_t), \epsilon_{t+1}) d\mu(\epsilon_{t+1}|z_{t+1}) \Big\}. \end{aligned} \tag{16}$$

Define

$$Q_t(x_t, a_t) = r(x_t, a_t) + \beta \sum_{z_{t+1}} \int \sigma(z_{t+1}, x_t, a_t) V_{t+1}(\lambda(z_{t+1}, x_t, a_t), \epsilon_{t+1}) d\mu(\epsilon_{t+1}|z_{t+1}). \tag{17}$$

Then,

$$V_t(x_t, \epsilon_t) = \max_{a_t \in A} \{Q_t(x_t, a_t) + \epsilon_t(a_t)\}. \tag{18}$$

We extend the concept of social surplus function in McFadden (1981) and Rust (1994) by

$$G[\{u(x_t, a_t), a_t \in A\}|x_t, z_t] = \int \max_{a_t \in A} [u(x_t, a_t) + \epsilon_t(a_t)] d\mu(\epsilon_t|z_t), \tag{19}$$

for a measurable function $u : X \times A \to R$. Then by the dominated convergence theorem and probability theory, we have the following result (analogous to Theorem 3.1 in Rust (1994)).

Theorem 2 *If $\mu(d\epsilon_t|z_t)$ has finite first moments, then the social surplus function* (19) *exists and it is*

$$G[\{u(x_t,a_t),a_t\in A\}|x_t,z_t]=\sum_{a_t\in A}\pi(a_t|x_t)(u(x_t,a_t)+E[\epsilon_t|Y(u,\epsilon)=a_t,z_t]), \tag{20}$$

where $Y(u,\epsilon)=\arg\max_{a_t\in A}(u(x_t,a_t)+\epsilon_t(a_t))$. Furthermore,

(i) G is a convex function of $\{u(x,a),a\in A\}$;

(ii) G satisfies the additivity property, i.e., for any $\alpha\in R$,

$$G[\{u(x_t,a_t)+\alpha,a_t\in A\}|x_t,z_t]=\alpha+G[\{u(x_t,a_t),a_t\in A\}|x_t,z_t]; \tag{21}$$

(iii) The partial derivative of G with respect to $u(x_t,a_t)$ is the conditional choice probability:

$$\frac{\partial G[\{u(x_t,a_t),a_t\in A\}|x_t,z_t]}{\partial u(x_t,a_t)}=\pi(a_t|x_t). \tag{22}$$

Let $\mathscr{B}$ be the Banach space of bounded, Borel measurable functions $H:X\times A\to R$ under the supremum norm. Define an operator $H:\mathscr{B}\to\mathscr{B}$ by

$$[Hv](x,a)=r(x,a)+\beta\sum_{z'}\sigma(z',x,a)G[\{v(\lambda(z',x,a),a'),a'\in A\}|\lambda(z',x,a),z']. \tag{23}$$

Assume the following regularity conditions (similar to Rust 1994 with state replaced by belief x):

(i) (Bounded Upper Semicontinuous) For each $a\in A$, $r(x,a)$ is upper semicontinuous in belief x with bounded expectation and

$$\begin{aligned}
&h(x):=\sum_{t=1}^{\infty}\beta^t h_t(x)<\infty,\\
&h_1(x)=\max_{a\in A}\sum_{z'\in Z}\sigma(z',x,a)\int\max_{a'\in A}|r(\lambda(z',x,a),a')+\epsilon'(a')|d\mu(\epsilon'|z'),\\
&h_t(x)=\max_{a\in A}\sum_{z'\in Z}\sigma(z',x,a)h_{t-1}(\lambda(z',x,a));
\end{aligned}$$

(ii) (Weakly Continuous) The stochastic kernel $\sigma(\cdot,x,a)=\{\sigma(z,x,a)\}_{z\in|Z|}$ is weakly continuous in $X\times A$;

(iii) (Bounded Expectation) The reward $r\in\mathscr{B}$ and for each $v\in\mathscr{B}$, $Ev\in\mathscr{B}$, where

$$[Ev](x,a)=\sum_{z'\in Z}\sigma(z',x,a)G[\{v(\lambda(z',x,a),a'),a'\in A\}|\lambda(z',x,a),z'].$$

Theorem 3 *Under AS, CI, and the regularity conditions, $H : \mathscr{B} \to \mathscr{B}$ is a contraction mapping with modulus β. Hence, H has a unique fixed point satisfying $Q^* = HQ^*$ and the optimal decision rule δ^* is*

$$\delta^*(x, \epsilon) = \arg\max_{a \in A} \Big\{ Q^*(x, a) + \epsilon(a) \Big\}. \tag{24}$$

Furthermore, the controlled process $\{z_{t+1}, x_t, a_t\}$ is Markovian with

$$\pi(a|x) = \frac{\partial G[\{Q^*(x, a), a \in A\}|x, z]}{\partial Q^*(x, a)}. \tag{25}$$

Theorem 4 *If the probability measure of ϵ is multivariate extreme-value, i.e.,*

$$\mu(d\epsilon|z) = \prod_{a \in A} \exp\{-\epsilon(a) + \gamma\} \exp[-\exp\{-\epsilon(a) + \gamma\}], \tag{26}$$

where γ is the Euler constant. Then, the agent will select its action a with probability

$$\pi(a|x) = \frac{\exp Q(x, a)}{\sum_{a' \in A} \exp Q(x, a')}, \tag{27}$$

where

$$Q(x, a) = r(x, a) + \beta \sum_{z'} \sigma(z', x, a) \log(\sum_{a'} \exp Q(\lambda(x, z', a), a')).$$

Let

$$V(x, \epsilon) = \max_{a \in A} \{Q(x, a) + \epsilon(a)\}. \tag{28}$$

Proposition 1 *Functions of $Q(x, a)$ and $V(x, \epsilon)$ are convex in x.*

For example, under the extreme-value assumption, $Q(x, a)$ in Theorem 4 is convex on X because $f(x) = \log \sum_i \exp(x_i)$ is a convex function (by Hölder's inequality).

Independence from Irrelevant Alternatives (IIA) Under the extreme-value assumption, it is clear to see per Theorem 4 that

$$\log\left(\frac{\pi(a|x)}{\pi(a = 0|x)}\right) = Q(x, a) - Q(x, a = 0). \tag{29}$$

Thus, the IIA property still holds with respect to belief (rather than the hidden state), i.e., the odds of choosing alternative a over the reference action 0 only depends on the attributes of the two choices.

For data in the form of a collection of $N > 0$ independent finite sequences of observables $\{(z_{t,i}, a_{t,i}), 0 \leq t \leq T\}_{i=1}^{N}$, a likelihood function can now be factored as

$$\begin{aligned}
\log \ell(\theta) &\triangleq \log \prod_{i=1}^{N} \prod_{t=0}^{T-1} P(z_{t+1,i}|\zeta_{t,i}, a_{t,i}) P(a_{t,i}|\zeta_{t,i}) \\
&= \log \prod_{i=1}^{N} \prod_{t=0}^{T-1} \sigma_{\theta_2}(z_{t+1,i}, x_{t,i}, a_{t,i}) \pi_\theta(a_{t,i}|x_{t,i}) \\
&= \sum_{i=1}^{N} \left(\sum_{t=0}^{T-1} \log \sigma_{\theta_2}(z_{t+1,i}, x_{t,i}, a_{t,i}) + \sum_{t=0}^{T-1} \log \pi_\theta(a_{t,i}|x_{t,i}) \right),
\end{aligned} \tag{30}$$

where the second equality is by Theorem 1 that x_t is a sufficient statistic for the problem. As a result, we still can obtain a model of the agent by finding the parameter that maximizes the log likelihood in (30).

Remark: It can be easily verified that Theorems 3-4 and Proposition 1 continue to hold for the case in which the agent is solving a finite horizon problem. Evidently, the results in this case require that δ, CCP, Q, V all depend on t.

4. Identification Results

A distinct feature of our model from the classical Rust's model and its derivatives is that the dynamics of the system under study cannot be directly observed as the system state is only partially observable. We now show that we could identify the hidden dynamics using two periods of data, assuming we know the cardinality of the state space of the system. To this end, we say the dynamic can be identified by the data if it is always possible to discriminate between two dynamics.

Theorem 5 *Assume $|S|$ is known. The hidden dynamic $\{P(z,a)\}$ (not rank-1) can be uniquely identified from the first two periods of data* [1].

Theorem 5 allows us to generalize the identification results in Hotz and Miller (1993) and Magnac and Thesmar (2002) for the completely observable case to the partially observable case, by realizing that our model is a belief version of the Rust's model once the hidden dynamics is determined.

Theorem 6 *Assuming $|S|$ is known. Fix the discount factor β, the distribution of random shock μ, and reward function in the reference action, there exists only one reward structure $\{r(s,a): s \in S, a \in A\}$ rationalizing the data.*

For example, under the Gumbel assumption, it is clear to see that

$$\bar{V}(x) = E_{\epsilon}[V(x,\epsilon)] = \log \sum_{a=0}^{|A|-1} \exp\left(Q(x,a) - Q(x,a=0)\right) + Q(x,a=0). \tag{31}$$

Note that

$$Q(x,a=0) = r(x,a=0) + \beta E_z[\bar{V}(\lambda(x,z,a))|x,a=0] \tag{32}$$

Combining Eq. (31) and Eq. (32),

$$\begin{aligned} Q(x,a=0) = r(x,a=0) + \beta E_z[\log \sum_{a'=0}^{|A|-1} \exp\left(Q(\lambda(x,z,a),a') - Q(\lambda(x,z,a),a'=0)\right)|x,a=0] \\ + \beta E_z[Q(\lambda(x,z,a),a'=0))|x,a=0], \end{aligned}$$

By Eq. (29) we have

$$Q(x,a=0) = r(x,a=0) + C + \beta E_z[Q(\lambda(x,z,a),a'=0))|x,a=0], \tag{33}$$

where

$$C = \beta E_z\Big[\log \sum_{a'=0}^{|A|-1} \frac{\pi(a'|\lambda(x,z,a))}{\pi(a'=0|\lambda(x,z,a))}|x,a=0\Big],$$

and it can be obtained from the dataset per Theorem 1 and Theorem 5. It is easy to show that Eq. (33) is a contraction mapping; hence, there is a unique solution for $Q(x,a=0)$, given $r(a=0)$. Thus, $Q(x,a)$ can be identified by Eq. (29) for all $a \in A$ and consequently $\bar{V}(x)$ as well. Next, since

$$r(x,a) = Q(x,a) - \beta E_z[\bar{V}(\lambda(x,z,a))|x,a]$$

we can get $r(x,a)$, and consequently, $r(s,a)$.

If $|S|$ is unknown, it may not be reasonable to expect the partially observable model is identifiable. In practice, the number of possible states can be obtained by domain knowledge for a particular application. For example, the possible stages of diseases or system degradation are likely obtainable.

Corollary 1 *For $T < \infty$, the hidden model is identifiable if μ, and both the reward structure and the terminal value function Q_T in the reference action, are known.*

5. Model Misspecification Errors

In this section, we examine how far the estimation may deviate from its true value if the (completely observable) Rust's model is used to fit data generated in a partially observable setting. We first analyze the mis-specification errors on the system dynamics, and then discuss the potential errors on the reward structure, assuming $|S| = |Z|$.

In a partially observable setting, we first observe that the dynamic $P(z'|z,a)$ estimated by the Rust's model is in fact a random variable, as its value depends on all past histories ζ_t ending with $z_t = z$. That is, given z', z, a, $P(z'|z,a)$ can be any value in $\{P(z_{t+1} = z'|z_t = z, a_t = a, \tilde{\zeta}_{t-1}, \tilde{a}_{t-1}) : \forall \tilde{\zeta}_{t-1}, \forall \tilde{a}_{t-1} \in A\} = \{\sigma(z', a, \lambda(z, \tilde{x}, \tilde{a})) : \tilde{x} \in X, \tilde{a} \in A\}$. For a sample path ζ_t, Proposition 2 quantifies the maximal difference between $P(z'|z,a)$ and its counterpart $\sigma(z', a, x_*)$ in the hidden state model, where $x_* = \lambda(z, \tilde{a}_*, \tilde{x}_*)$ is the current belief at ζ_t with $z_t = z$, and $\tilde{x}_*, \tilde{a}_*$ are the previous period belief x_{t-1} and action a_{t-1} along the history ζ_t. This difference is caused by (improperly) neglecting the information history leading to the current stage.

Proposition 2

$$P(z'|z,a) - \sigma(z', a, x_*) \geq \max_{j \in S} \left[\lambda(z, e_j, \tilde{a}_*) - \lambda(z, \tilde{x}_*, \tilde{a}_*)\right] P(z', a) 1 \geq 0, \tag{34}$$

where e_j is the unit vector in $R^{|S|}$ with the j^{th} element being 1 and all other elements being 0.

We remark that the selection of $j^* \in \arg\max \left[\lambda(z, e_j, \tilde{a}_*) - \lambda(z, \tilde{x}_*, \tilde{a}_*)\right] P(z', a) 1$ depends on z'. Hence, Proposition 2 also implies that

$$||P(\cdot|z,a) - \sigma(\cdot, a, x_*)|| \geq \max_{z' \in Z} \left\{ \max_{j \in S} \left[\lambda(z, e_j, \tilde{a}_*) - \lambda(z, \tilde{x}_*, \tilde{a}_*)\right] P(z', a) 1 \right\}, \tag{35}$$

where $||v|| = \max_{z \in Z} |v(z)|$ for any vector $v \in R^{|Z|}$, $P(\cdot|z,a) = \{P(z'|z,a)\}_{z' \in Z}$, and $\sigma(\cdot, a, x_*) = \{\sigma(z', a, x_*)\}_{z' \in Z}$.

Lemma 1

$$\min_z \{\frac{\sigma(z,a,x)}{\sigma(z,a,x')} : \sigma(z,a,x') > 0\} \geq \min_s \{\frac{x(s)}{x'(s)} : x'(s) > 0\} \tag{36}$$

$$\max_z \{\frac{\sigma(z,a,x)}{\sigma(z,a,x')} : \sigma(z,a,x') > 0\} \leq \max_s \{\frac{x(s)}{x'(s)} : x'(s) > 0\} \tag{37}$$

Define

$$\begin{aligned} d(x,x') &= 1-\min\left\{\frac{x(s)}{x'(s)} : x'(s)>0\right\}, \\ D_1(x,x') &= \max\{d(x,x'), d(x',x)\}, \\ D(z,a) &= \max\{D_1(\lambda(z,e_i,a),\lambda(z,e_j,a)) : i,j\in S\}, \\ D(z,a,a') &= \max\{D_1(\lambda(z,e_i,a),\lambda(z,e_j,a')) : i,j\in S\}. \end{aligned} \tag{38}$$

We remark that $D(z,a)$ is a contraction coefficient (coefficient of ergodicity) for $P(z,a)$ (White and Scherer 1994, Platzman 1980). The next proposition provides an upper bound on the potential difference between $P(z'|z,a)$ and $\sigma(z',x_*,a)$.

Proposition 3 *For any $\tilde{a}\in A$,*

$$D_1(\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),\sigma(\cdot,a,x_*)) \le D(z,\tilde{a},\tilde{a}_*), \forall \tilde{x}\in X;$$

thus, if $\tilde{a}=\tilde{a}_*$,

$$D_1(\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),\sigma(\cdot,a,x_*)) \le D(z,\tilde{a}_*), \forall \tilde{x}\in X.$$

Proposition 3 shows that the difference is bounded above by the ergodicity of $P(z,a)$ if $\tilde{a}=\tilde{a}_*$. If $\tilde{a}\neq\tilde{a}_*$, then the discrepancy caused by the two different actions $\tilde{a}$ and $\tilde{a}_*$ will also contribute to the upper bound.

Assume the observation probability $\mathscr{Q}(z'|s',s,a)$ is independent of s,a and

$$\mathscr{Q}(z|s) = \begin{cases} 1-\eta, & z=s, \\ \kappa_{s,z}\eta, & z\neq s \end{cases} \tag{39}$$

where $0\le\eta\le 1, \kappa_{s,z}\ge 0, \kappa_{s,s}=0, \sum_z \kappa_{s,z}=1, \forall z\in Z, s\in S$.

Proposition 4 *Given $x_t=x, a_t=a, s_t=s$,*

$$\begin{aligned} &\sigma(z'=i,x,a) - P(s'=i|s,a) \\ &= \sum_{\tilde{s}} x(\tilde{s})[P(i|\tilde{s},a)-P(i|s,a)] + \eta\sum_{\tilde{s}} x(\tilde{s})[\sum_{s''}\kappa_{s'',z=i}P(s''|\tilde{s},a)-P(i|\tilde{s},a)]. \end{aligned}$$

Moreover,

$$D_1[\sigma(\cdot,x,a),P(\cdot|s,a)]=D_1[P(\cdot|s,a),(1-\eta)\sum_{\tilde{s}}x(\tilde{s})P(.|\tilde{s},a)+\eta\sum_{\tilde{s}}x(\tilde{s})\sum_{s''}\kappa_{s'',z=.}P(s''|\tilde{s},a)],$$

and for $0<\eta<1$*,*

$$\begin{aligned}&D_1[\sigma(\cdot,x,a),P(\cdot|s,a)]\\&\leq\max\left\{D_1\left[P(\cdot|s,a),\sum_{\tilde{s}}x(\tilde{s})P(.|\tilde{s},a)\right],D_1\left[P(\cdot|s,a),\sum_{\tilde{s}}\sum_{s''}x(\tilde{s})\kappa_{s'',z=.}P(s''|\tilde{s},a)\right]\right\}.\end{aligned}$$

Thus, Proposition 4 indicates that even if $x=e_s$,

$$\sigma(z'=i,x,a)-P(s'=i|s,a)=\eta\sum_{s''}\kappa_{s'',z=i}[P(s''|s,a)-P(i|s,a)]\neq 0.$$

Theorem 7 *Assume a sample path* ζ_t *is given.*

$$\begin{aligned}&P(z'=i|z,a)-P(s'=i|s,a)\\&\geq\max_{j\in S}\left[\lambda(z,e_j,\tilde{a}_*)-\lambda(z,\tilde{x}_*,\tilde{a}_*)\right]P(z',a)1+\sum_{\tilde{s}}x_*(\tilde{s})[P(i|\tilde{s},a)-P(i|s,a)]\\&+\eta\sum_{\tilde{s}}\sum_{s''}x_*(\tilde{s})[\kappa_{s'',z=i}P(s''|\tilde{s},a)-P(i|\tilde{s},a)].\end{aligned}$$

Moreover,

$$\begin{aligned}&D_1(P(.|z,a),P(.|s,a))\leq\max_{\tilde{a}\in A}D(z,\tilde{a},\tilde{a}_*)\\&+\max\left\{D_1\left[P(\cdot|s,a),\sum_{\tilde{s}}x_*(\tilde{s})P(.|\tilde{s},a)\right],D_1\left[P(\cdot|s,a),\sum_{\tilde{s}}\sum_{s''}x_*(\tilde{s})\kappa_{s'',z=.}P(s''|\tilde{s},a)\right]\right\}.\end{aligned}$$

Corollary 2 *When* $\eta=0$*,* $P(z'|z,a)=P(s'|s,a)$*.*

For the reward structure, the Rust's model shows $\log\frac{\pi(a|z)}{\pi(a=0|z)}=Q^R(z,a)-Q^R(z,a=0)$, whereas in our hidden state model $\log\frac{\pi(a|x)}{\pi(a=0|x)}=Q(x,a)-Q(x,a=0)$. Thus, we can analyze how the CCP ratios measured by the two models affect the estimated reward. Namely, given

$$\rho_*(z,a)\leq|\log\frac{\pi(a|z)}{\pi(a=0|z)}-\log\frac{\pi(a|x_*)}{\pi(a=0|x_*)}|\leq\rho^*(z,a),\tag{40}$$

where the difference is again caused by ignoring history ζ_t leading to the current observation z, we examine how r^R may deviate from the true reward structure r. As only the relative difference between the choice-specific function Q matters, we let $Q^R(Z,a=0)=Q(X,a=0)=0$.

Define

$$\triangle r^R = \max_{z,z',a} |r^R(z,a) - r^R(z',a)|,$$

$$||\sigma - P^R||_1^* = \max_{z,a} \frac{1}{2} \sum_{z'} |[\sigma(z',a,x_*) - P(z'|z,a)]| = \max_{z,a,\tilde{x},\tilde{a}} \frac{1}{2} \sum_{z'} |[\sigma(z',a,x_*) - \sigma(z',a,\lambda(z,\tilde{x},\tilde{a}))]|.$$

Theorem 8 below shows that both the deviation in the CCPs and the system dynamic will contribute to the estimation error in the reward structure.

Theorem 8 *Given a sample path ζ_t,*

$$[||r - r^R|| + \triangle r^R] + 2\beta[||\sigma - P^R||_1^*(||Q^R|| + \log|A|) + \frac{1}{2}\log|A|] \geq (1-\beta)\max_{z,a} \rho_*(z,a).$$

Furthermore,

$$||r^R - r|| \leq (1+\beta)\max_{z,a} \rho^*(z,a) + 2\beta||\sigma - P^R||_1^*||Q^R|| + \beta \max_{z,a,x_*} |h(z,a,x_*)|,$$

where $h: X \times A \times Z \to R$ is a continuous function satisfying

(i) $h(z,a,x_*) \to 0$ *if the observation probability $\mathcal{Q} \to I$, I is the identity matrix,*

(ii) $\max_{z,a,x_*} |h(z,a,x_*)| \leq \log|A| + \max_{z,a,a'} |Q^R(z,a) - Q^R(z,a')|$.

6. Numerical Results

In this section, we first apply our hidden state model to the well-known engine replacement data set for Mr. Zurcher's decisions in Rust (1987), and then we test and validate our approach using synthetic data sets.

6.1. Rust's Model Revisited

To illustrate the application of the proposed methodology we revisit a subset of Rust's dataset. Specifically, Group 4 consisting of buses with 1975 GMC engines. As in Rust (1987), we discretize the state space in 175 bins of 2.5K miles. Evidence of positive serial correlation in mileage increments is quite strong as the Durbin-Watson statistic is less than 1.13 for all buses except one with a value of 1.32. The action $a_t = 1$ is associated with engine replacement at a cost $RC > 0$ whereas

$a_t = 0$ is the continued operation. The state space S is $S = \{0, 1\}$, where "0" is being the "good state" and "1" is being the "bad state". The model for hidden state dynamics is

$$P_{\theta_2}(s_{t+1}|s_t, a_t = 0) = \begin{pmatrix} \theta_{2,0} & 1-\theta_{2,0} \\ 1-\theta_{2,1} & \theta_{2,1} \end{pmatrix}$$

and $P_{\theta_2}(s_{t+1} = 0|s_t, a_t = 1) = 1$. Per mile maintenance costs are parametrized by $\theta_{1,0}$ (in good state) and $\theta_{1,1}$ (in bad state). With a belief $x_t \in [0, 1]$ of the engine being in good state and z_t cumulative mileage after t months, the expected (monthly) maintenance cost is of the form $(\theta_{1,0}z_t)x_t + (\theta_{1,1}z_t)(1 - x_t)$. Monthly mileage increments $\Delta \in \{0, 1, 2, 3\}$ correspond to values between $[0, 2.5K), [2.5K, 5K), [5K, 7.5K)$ and $[7.5K, 10K)$, respectively. The distribution is parametrized as follows

$$P_{\theta_3}(\, z_{t+1} - z_t = \Delta|\, s_t = 0, a_t = 0) = \theta_{3,0,\Delta} \qquad \Delta \in \{0, 1, 2\}$$

$$P_{\theta_3}(\, z_{t+1} - z_t = \Delta|\, s_t = 0, a_t = 0) = 1 - \theta_{3,0,0} - \theta_{3,0,1} - \theta_{3,0,2} \quad \Delta = 3$$

Similarly, we define $P_{\theta_3}(\, z_{t+1} - z_t = \Delta|\, s_t = 1, a_t = 0) = \theta_{3,1,\Delta}, \Delta \in \{0, 1, 2, 3\}$. The estimation results are described in Table 1 where the belief space (i.e. the unit interval) is discretized in a uniform grid of 100 intervals.

Parameter	$\theta_{3,0,0}$	$\theta_{3,0,1}$	$\theta_{3,0,2}$	$\theta_{3,1,0}$	$\theta_{3,1,1}$	$\theta_{3,1,2}$	$\theta_{2,0}$	$\theta_{2,1}$	$\theta_{1,0}$	$\theta_{1,1}$	RC
Good State	0.04	0.33	0.59	$*$	$*$	$*$	0.94	$*$	0.0011	$*$	10.11
Bad State	$*$	$*$	$*$	0.181	0.757	0.061	$*$	0.99	$*$	0.0011	10.11
Log-Likelihood	-3818										

Table 1 Parameter estimates and log-likelihood of our hidden state model

Compared to the original Rust's model displayed in Table 2, the hidden state model captures an optimal route assignment: the distribution of mileage increments for engines considered in bad state is dominated (in the first-order stochastic sense) by the distribution of mileage increments of engines in good state. Assigning routes with lower mileages to buses in bad state decreases the operational cost of buses in such state. We find that the marginal operation cost (per mile) for a bus in *bad* vs. *good* is approximately the same, namely $\theta_{1,0} \simeq \theta_{1,1}$. This is consistent with the optimal route assignment in the sense that if for example $\theta_{1,0} < \theta_{1,1}$ then buses in good condition would be *under*-utilized vis-á-vis those in bad condition for which replacement is not yet justified. This is an economically meaningful feature of Mr. Zurcher's behavior (ignored by Rust's model) which improves model fit as measured by log-likelihood by $\frac{4496-3818}{3818} = 17.7\%$.

Parameter	$\theta_{3,0}$	$\theta_{3,1}$	$\theta_{3,2}$	θ_1	RC
Rust's Model (Rust 1987, p. 1022)	0.119	0.576	0.287	0.0012	10.89
Log-Likelihood	-4496				

Table 2 Parameter estimates and log-likelihood Rust's model

The significant increment in the log-likelihood function is achieved at the cost of 6 more parameters in dynamics and rewards. Considering that the Rust's model is a special case of our model, we use the log-likelihood ratio test to examine whether our model statistically outperforms the Rust's model in this example. The underlying null hypothesis is H0: there is no significant difference between the Rust's model and our model. The log-likelihood ratio test result in Table 3 shows that the null hypothesis is rejected with a p value very close to zero, indicating that our model statistically outperforms the Rust's model on the Group 4 of the 1975 GMC bus engine dataset.

	Log-likelihood	Degree of Freedom	Statistic	p Value
Hidden state model	-3818	11	1356	$< 10^{-100}$
Rust's model	-4496	5		

Table 3 The log-Likelihood ratio test for both models on the 1975 GMC engine (Group 4) data set

6.2. Synthetic Data

We randomly generate synthetic data sets where both the system state and dynamics are hidden, in order to test if our approach can correctly identify the reward structure $r_{\theta_1}(s,a)$ and the hidden dynamics $P_{\theta_2}(z', s'|s, a)$, assume the discount factor $\beta(=0.95)$ is known and and $r(a=0)$ is fixed. The belief space is uniformly discretized to 100 intervals over $[0, 1]$. Table 4 lists the parameter values (values without parentheses) of an example in the size of $|S| = 2, |Z| = 2$, and $|A| = 3$, where the values fitted by our model are included in the parentheses ($r(a=0) = [10, 3]$ is fixed and given during the estimation). The result shows that our model can identify the model parameters fairly well, with maximal element-wise deviation of 0.026 in dynamics and 0.08 in reward, using 800 sample trajectories.

$a=0$	(z', s')			
	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
$s=0$	0.72 (0.724)	0.08 (0.083)	0.02 (0.013)	0.18 (0.180)
$s=1$	0.00 (0.000)	0.00 (0.000)	0.10 (0.120)	0.90 (0.880)

$a=1$	(z', s')			
	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
$s=0$	0.81 (0.784)	0.09 (0.100)	0.01 (0.000)	0.09 (0.116)
$s=1$	0.00 (0.002)	0.00 (0.000)	0.10 (0.118)	0.90 (0.880)

$a=2$	(z', s')			
	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
$s=0$	0.90 (0.909)	0.10 (0.090)	0.00 (0.000)	0.00 (0.001)
$s=1$	0.36 (0.350)	0.04 (0.043)	0.06 (0.068)	0.54 (0.539)

$r(s,a)$	a		
	0	1	2
$s=0$	10^*	6 (6.08)	3 (2.95)
$s=1$	3^*	5 (4.99)	7 (6.99)

Table 4 Model parameters of a randomly generated numerical example and the values fitted by our model (in the parentheses). The maximal element-wise difference: 0.026 in dynamics $\{P(z', s'|s, a)\}$, 0.08 in reward $\{r(s, a)\}$.

We also fit the same data set using the Rust's model with result listed in Table 5. The true transition probabilities $P(s'|s, a)$ are obtained by $P(s'|s, a) = \sum_{z'} P(z', s'|s, a)$ and their values are listed in the table without parentheses. The estimation result (in the parentheses) shows that the maximal deviation (element-wise) are 0.275 and 2.06 in transition probabilities and reward, respectively. Thus, our hidden state model performs better than the Rust's model in the partially observable case. Furthermore, the chi-squared statistic of the log-likelihood ratio test is 1124.46 with degree of

$a=0$	$s'=0$	$s'=1$
$s=0$	0.80 (0.704)	0.20 (0.296)
$s=1$	0.00 (0.275)	1.00 (0.725)

$a=1$	$s'=0$	$s'=1$
$s=0$	0.90 (0.735)	0.10 (0.265)
$s=0$	0.00 (0.263)	1.00 (0.737)

$a=2$	$s'=0$	$s'=1$
$s=0$	1.00 (0.836)	0.00 (0.164)
$s=1$	0.40 (0.500)	0.60 (0.500)

$r(s,a)$	a		
	0	1	2
$s=0$	10^*	6 (6.64)	3 (4.49)
$s=1$	3^*	5 (3.75)	7 (4.94)

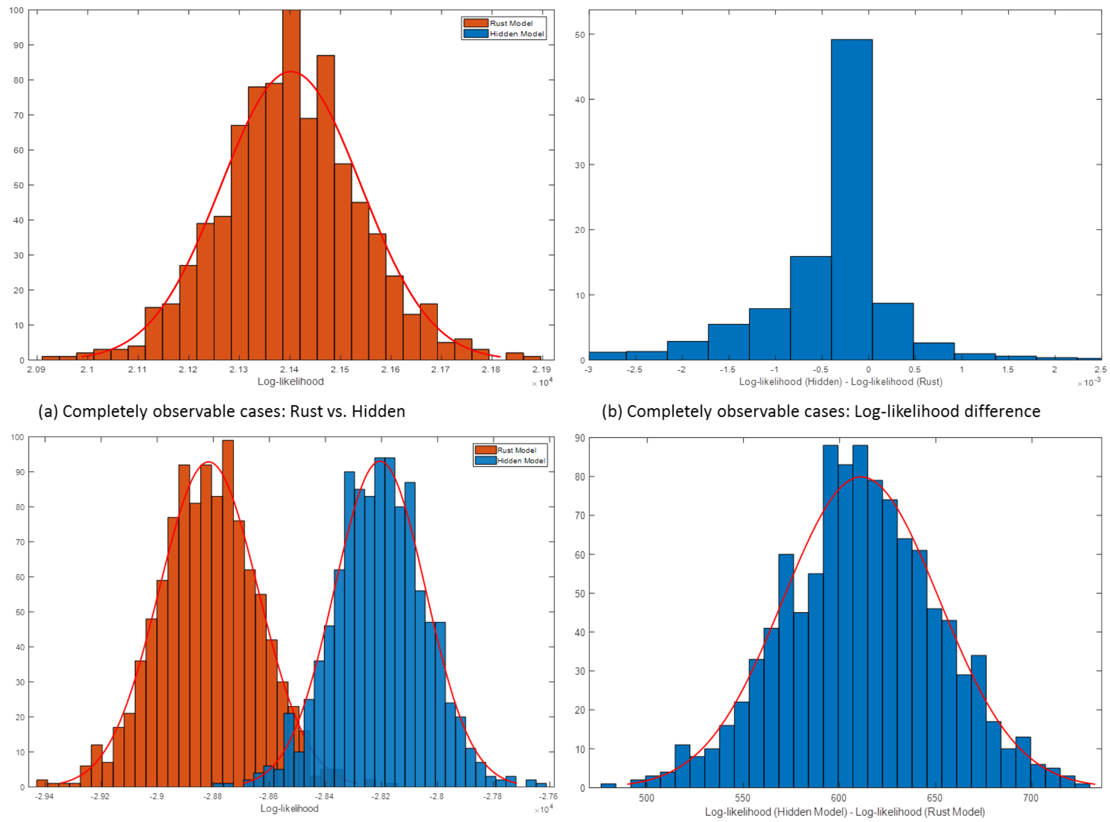
Table 5 Parameter values fitted by the Rust's model. The maximal element-wise difference: 0.275 in transition probabilities $\{P(s'|s,a)\}$, 2.06 in reward $\{r(s,a)\}$.

freedom 12, rejecting the null hypothesis of no significant difference between the two models with a very small p value close to zero. See Table 6.

	Log-likelihood	Degree of Freedom	Statistic	p Value
Hidden state model	-27642.11	22	1124.46	$< 10^{-10}$
Rust's model	-28204.34	10		

Table 6 Log-Likelihood ratio test for our hidden model and the Rust's model.

In addition, Fig. 2 illustrates the distributions of log-likelihood functions fitted by both models under various randomly generated synthetic data sets. If the data is generated in a completely observable environment, the developed hidden state model will generate the same result as what are estimated by the Rust's model; see Fig. 2(a)(b). However, if the data is generated in a partially observable environment, the log-likelihood function fitted by our model ℓ is significantly greater than the maximal log-likelihood function produced by the Rust's model ℓ^{Rust}; see Fig. 2(c)(d). In fact, $\ell \geq \ell^{Rust}$ almost surely as the Rust's model is a special case of our model where the observation matrix is identity. Furthermore, the likelihood ratio test rejects the Rust's model with high confidence intervals and results generated by the Rust's model can be misleading and/or erroneous.



(a) Completely observable cases: Rust vs. Hidden
(b) Completely observable cases: Log-likelihood difference
(c) Partially observable cases: Rust vs. Hidden
(d) Partially observable cases: Log-likelihood difference

Figure 2 The log-likelihood distributions of both the Rust's model and our model for a thousand synthetic datasets

7. Conclusions

In this paper, we developed a dynamic discrete choice model for the case where the underlying system state and the associated system dynamics are hidden for the decision-making agent. At each decision epoch, the agent infers the system state from possibly noise-corrupted observations on which to base action selection. We formulated the decision making process of the agent on the basis of partially observable Markov decision processes subject to independent and identically distributed random shocks, generalizing the existing dynamic discrete choice models to partially observable settings. We analyzed the structural properties of the proposed hidden state model and proved that the model is still identifiable from sample trajectories if the discount factor, the distribution of the random shock, the reward structure in a reference action, and the possible number of hidden states are known. As the (completely observable) dynamic discrete choice models are widely used in problems where the relevant state variables may only be partially observed, we analyzed the potential model misspecification errors when the Rust's model is used in a partially observable setting. The possible estimation discrepancies were also demonstrated via numerical examples. Finally, we compared our hidden state model to the Rust's model on the bus engine dataset in Rust (1987). The significant improvement in the log-likelihood function and the log-likelihood ratio test strongly suggested that our hidden state model outperforms the Rust's model. Furthermore, our model also revealed economically meaningful features of Mr. Zurcher's behavior ignored by the Rust's model.

As this research represents a first effort on developing dynamic discrete choice models for partially observable systems, future research directions are numerous. For example, our current model assumes that the distribution of the random shock is i.i.d.. Many studies have examined more complicated structures in completely observable dynamic choice models. Thus, extending these analyses to the partially observable case will be value-added. Also, computational challenges of our model are obviously not trivial. Both continuous-state MDP and POMDP suffer from the well-known curse of dimensionality, meaning that to achieve a certain level of accuracy, the number of discretized points have to grow exponentially. Moreover, observations in many real applications can be high dimensional. Thus, another research direction is to address computational challenges of high dimensional hidden state models. Application wise, it will be interesting to apply this model to various relevant applications, indicatively, building predictive models of human control when a relevant psychological trait (e.g., fatigue, attention) is hidden.

Appendix

[1]The proof of Theorem 5 essentially implies that $P_{\theta_2}(z,a)$ can be obtained by $\sigma_{\theta_2}(z, \lambda_{\theta_2}(z_{t+1}, x_t, a_t), a_{t+1}) = P(z|\zeta_{t+1}, a_{t+1})$, assuming we know the belief at time t. In many applications, it is reasonable to expect that we could find such belief points. For example, in the engine replacement example, it is acceptable to reason that the state of a newly replace engine is good.

Proof of Theorem 1. The proof is by induction. Assume $V_{t+1}(\zeta_{t+1},\epsilon_{t+1}) = V_{t+1}(x_{t+1},\epsilon_{t+1})$, then

$$\begin{aligned} V_t(\zeta_t,\epsilon_t) &= \max_{a_t\in A}\left\{\sum_{s_t} P(s_t|\zeta_t)r(s_t,a_t) + \epsilon_t(a_t) + \beta\sum_{z_{t+1}}\int P(z_{t+1}|\zeta_t,a_t)V_{t+1}(x_{t+1},\epsilon_{t+1})d\mu(\epsilon_{t+1}|z_{t+1})\right\} \\ &= \max_{a_t\in A}\left\{r(x_t,a_t) + \epsilon_t(a_t) + \beta\sum_{z_{t+1}}\int \sigma(z_{t+1},x_t,a_t)V_{t+1}(\lambda(z_{t+1},x_t,a_t),\epsilon_{t+1})d\mu(\epsilon_{t+1}|z_{t+1})\right\} \\ &= V_t(x_t,\epsilon_t) \end{aligned}$$

where the last equality is due to the fact that both σ and λ are functions of x_t. $\square$

Proof of Theorem 2. Eq. (20) is directly from the definition of G and

$$G[\{u(x_t,a_t), a_t\in A\}|x_t,z_t] = \int\sum_{a_t\in A}[u(x_t,a_t)+\epsilon_t(a_t)]1_{\{Y=a\}}d\mu(\epsilon_t|z_t).$$

$(i)-(iii)$ follow the same idea as in Rust (1994). $\square$

Proof of Theorem 3. Under the modified regularity conditions, the proof follows the exactly same procedure as in Rust (1994) for Theorems 3.2-3.3. The controlled process $\{z_{t+1},x_t,a_t\}$ is Markovian because the conditional probability of a_t is $P(a_t|x_t)$, the conditional probability of x_{t+1} is provided by $\lambda(z_{t+1},x_t,a_t)$, and the conditional probability of z_{t+1} is given by $\sigma(z_{t+1},x_t,a_t)$. $\square$

Proof of Theorem 4. The result follows by Theorem 3 and Train (2002). $\square$

Proof of Proposition 1. The result is obvious by induction, $\sigma\ge 0$, the maximal of convex functions is still convex, and Theorem 2(i). $\square$

Proof of Theorem 5. Given two system dynamics $P_1(z,a)$, $P_2(z,a)$. Since the dataset contains x_0,a_0, we can obtain $\sigma_0^1(z_1,x_0,a_0) = x_0P_1(z_1,a_0)1$ and $\sigma_0^2(z_1,x_0,a_0) = x_0P_2(z_1,a_0)1$. Note that $\sigma_0(z_1,x_0,a_0) = P(z_1|\zeta_0,a_0)$, where $P(z_1|\zeta_0,a_0)$ can be obtained from the first period of the data. Thus, σ can be obtained from the data and $\sigma_0^1=\sigma_0^2$ if and only if $P_1(z,a)1 = P_2(z,a)1$. If $\sigma_0^1\neq\sigma_0^2$, we are done. Otherwise, update belief by Eq. (14), $x_1^1 = \lambda_1(z_1,x_0,a_0) = \frac{x_0P_1(z_1,a_0)}{\sigma_0^1(z_1,x_0,a_0)}, x_1^2 = \lambda_2(z_1,x_0,a_0) = \frac{x_0P_2(z_1,a_0)}{\sigma_0^2(z_1,x_0,a_0)}$. Then $x_1^1=x_1^2$ if and only if $P_1(z,a)=P_2(z,a)$. Now, $\sigma_1^1(z_2,x_1^1,a_1) = x_1^1P_1(z_2,a_1)1$ and $\sigma_1^2(z_2,x_1^2,a_1) = x_1^2P_2(z_2,a_1)1$, and σ_1 is obtainable from the two period of data as $\sigma_1(z_2,x_1,a_1) = P(z_2|\zeta_1,a_1)$. Thus, $\sigma_1^1=\sigma_1^2$ if and only if $x_1^1=x_1^2$, indicating $P_1(z,a)=P_2(z,a)$ assuming $P(z,a)$ is not rank-1. $\square$

Proof of Theorem 6. By Theorem 1 and Theorem 5, both CCP $\pi(a|x_t) = P(a|\zeta_t)$ and hidden dynamics can be identified from the data. Treating belief as the state and since $|\zeta_t| < \infty$, Proposition 1 in Hotz and Miller (1993) and Magnac and Thesmar (2002) show that there is a one-to-one mapping $q(X): R^{|A|} \to R^{|A|}$, only depending on μ, which maps the choice probability set $\{\pi(a|x)\}$ to the set of the difference in action-specific value function $\{Q(x,a) - Q(x,a=0)\}$, namely,

$$Q(x,a) - Q(x,a=0) = q_a(\{\pi(a'|x)\};\mu), \tag{41}$$

$q_0(.) = 0$, and $q = (q_0, ..., q_{|A|-1})$. Thus, if we know $Q(x,a=0)$, we can recover $Q(x,a), \forall a \in A$. Note that

$$\begin{aligned}
Q(x,a) &= r(x,a) + \beta E_{z|x,a}\left[E_{\epsilon|z} \max_{a' \in A}\left\{Q(\lambda(z,x,a),a') + \epsilon'(a')\right\}\right] \\
&= r(x,a) + \beta E_{z|x,a}\left[E_{\epsilon|z}\left[\max_{a' \in A}\left\{Q(\lambda(z,x,a),a') - Q(\lambda(z,x,a),0) + \epsilon'(a')\right\} + Q(\lambda(z,x,a),0)\right]\right] \\
&= r(x,a) + \beta E_{z|x,a}\Big[G\big[\{Q(\lambda(z,x,a),a') - Q(\lambda(z,x,a),0), a' \in A\}\,|\lambda(x,z,a),z\big]\Big] \\
&\quad + \beta E_{z|x,a}\left[Q(\lambda(z,x,a),0)\right]
\end{aligned} \tag{42}$$

Because of the mapping q, and that the hidden dynamic $\{P(z,a)\}$ is known, the quantity

$$\beta E_{z|x,a}\Big[G\big[\{Q(\lambda(z,x,a),a') - Q(\lambda(z,x,a),0), a' \in A\}\,|\lambda(x,z,a),z\big]\Big]$$

is known. Under the assumption of $r(s,0) = 0$, we have

$$\begin{aligned}
Q(x,0) = {}& \beta E_{z|x,a}\Big[G\big[\{Q(\lambda(z,x,a),a') - Q(\lambda(z,x,a),0), a' \in A\}\,|\lambda(x,z,a),z\big]\Big] \\
& + \beta E_{z|x,a}\left[Q(\lambda(z,x,a),0)\right]
\end{aligned} \tag{43}$$

with only unknown $Q(X,0)$. It is easy to see there is a unique solution to Eq. (43) due to the contraction mapping theorem. Consequently, all $Q(X,a), a \in A$ can be recovered by Eq. (41). Lastly,

$$\begin{aligned}
r(x,a) = {}& Q(x,a) \\
& - \beta E_{z|x,a}\Big[G\big[\{Q(\lambda(z,x,a),a') - Q(\lambda(z,x,a),0), a' \in A\}\,|\lambda(x,z,a),z\big] - Q(\lambda(z,x,a),0)\Big].
\end{aligned}$$

As $r(x,a) = xr(a)$, $\{r(s,a) : s \in S, a \in A\}$ can be uniquely determined. $\square$

Proof of Corollary 1. When $Q_T(X,0)$ and $r(a=0)$ are known, we can obtain $Q_t(X,0)$ via

$$Q_t(x,0)=r(x,a=0)+E_z\Big[G\big[\{Q(\lambda(z,x,a),a')-Q(\lambda(z,x,a),0),a'\in A\}\,|\lambda(x,z,a),z\big]\Big]$$
$$+E_z\left[Q_{t+1}(\lambda(z,x,a),0)\right]. \tag{44}$$

The rest follows exactly as in the proof of Theorem 6. $\square$

Proof of Proposition 2.

$$P(z'|z,a)-\sigma(z',a,x_*)=\max_{\tilde{x}\in X,\tilde{a}\in A}\sigma(z',\lambda(z,\tilde{x},\tilde{a}),a)-\sigma(z',\lambda(z,\tilde{x}_*,\tilde{a}_*),a)$$
$$\geq\max_{j\in S}\left[\lambda(z,e_j,\tilde{a}_*)-\lambda(z,\tilde{x}_*,\tilde{a}_*)\right]P(z',a)1.$$

Assume by contradiction that

$$\max_{j\in S}\left[\lambda(z,e_j,\tilde{a}_*)-\lambda(z,\tilde{x}_*,\tilde{a}_*)\right]P(z',a)1<0.$$

Then

$$\lambda(z,e_j,\tilde{a}_*)P(z',a)1<\lambda(z,\tilde{x}_*,\tilde{a}_*)P(z',a)1,\forall j\in S,$$

which is

$$\frac{e_jP(z,\tilde{a}_*)P(z',a)1}{e_jP(z,\tilde{a}_*)1}<\frac{\tilde{x}_*P(z,\tilde{a}_*)P(z',a)1}{\tilde{x}_*P(z,\tilde{a}_*)1},\forall j\in S.$$

Note that $\tilde{x}_*=\sum_j\tilde{x}_{*,j}e_j$, thus

$$\frac{\tilde{x}_*P(z,\tilde{a}_*)P(z',a)1}{\tilde{x}_*P(z,\tilde{a}_*)1}=\frac{\sum_j\tilde{x}_{*,j}e_jP(z,\tilde{a}_*)P(z',a)1}{\sum_j\tilde{x}_{*,j}e_jP(z,\tilde{a}_*)1}$$

Recall that if $a,b,c,d>0$, then $a/b\geq c/d\Leftrightarrow a/b\geq(a+c)/(b+d)$. Thus, we have

$$\frac{\sum_j\tilde{x}_{*,j}e_jP(z,\tilde{a}_*)P(z',a)1}{\sum_j\tilde{x}_{*,j}e_jP(z,\tilde{a}_*)1}\leq\max_{j\in S}\frac{e_jP(z,\tilde{a}_*)P(z',a)1}{e_jP(z,\tilde{a}_*)1},$$

a contradiction. Hence, $\max_{j\in S}\left[\lambda(z,e_j,\tilde{a}_*)-\lambda(z,\tilde{x}_*,\tilde{a}_*)\right]P(z',a)1\geq 0.$ $\square$

Proof of Lemma 1.

$$\frac{\sigma(z,x,a)}{\sigma(z,x',a)}=\sum_s\frac{\sum_{s'}P(z,s'|s,a)x(s)}{\sigma(z,x',a)}=\sum_s\frac{\sum_{s'}P(z,s'|s,a)x'(s)}{\sigma(z,x',a)}\frac{x(s)}{x'(s)}$$

Note that $f(s)=\frac{\sum_{s'}P(z,s'|s,a)x'(s)}{\sigma(z,x',a)}\geq 0$ and $\sum_s f(s)=\sum_s\frac{\sum_{s'}P(z,s'|s,a)x'(s)}{\sigma(z,x',a)}=\frac{\sigma(z,x',a)}{\sigma(z,x',a)}=1$. Thus,

$$\frac{\sigma(z,x,a)}{\sigma(z,x',a)}\geq\min_s\frac{x(s)}{x'(s)},\forall z$$

leading to inequality (36). The proof for inequality (37) is similar. $\square$

Proof of Proposition 3. Note that

$$D_1(\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),\sigma(\cdot,a,x_*)) = D_1(\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),\sigma(\cdot,a,\lambda(z,\tilde{x}_*,\tilde{a}_*)))$$
$$\leq D_1(\lambda(z,\tilde{x},\tilde{a}),\lambda(z,\tilde{x}_*,\tilde{a}_*)) \leq D(z,\tilde{a},\tilde{a}_*)$$

where the second to the last inequality is by Lemma 1, and the last inequality is from the fact that $D_1(x,\rho x' + (1-\rho)x'') \leq \max\{D_1(x,x'),D_1(x,x'')\}, x,x',x'' \in X, 0 \leq \rho \leq 1$,(A.3 in Platzman 1980) and $\lambda(z,x,a) = \sum_i \frac{x_i[e_i P(z,a)1]}{[xP(z,a)1]}\lambda(z,e_i,a)$ in A.4 Platzman (1980). □

Proof of Proposition 4. For the first part,

$$\begin{aligned}
&\sigma(z'=i,x,a) - P(s'=i|s,a)\\
&= \sum_{\tilde{s}}\sum_{s''} P(z'=i|s'')P(s''|\tilde{s},a)x(\tilde{s}) - P(i|s,a)\\
&= \sum_{\tilde{s}}[P(i|\tilde{s},a) - P(i|s,a)]x(\tilde{s}) + \eta\sum_{\tilde{s}}(\sum_{s''}\kappa_{s'',z=i}P(s''|\tilde{s},a) - P(i|\tilde{s},a))x(\tilde{s}),
\end{aligned}$$

by the definition of $\mathscr{Q}$. For the second part, note that $\sigma(z'=i,x,a)$ can also be written as

$$\sigma(z'=i,x,a) = (1-\eta)\sum_{\tilde{s}} P(i|\tilde{s},a)x(\tilde{s}) + \eta\sum_{\tilde{s}}\sum_{s''}\kappa_{s'',z=i}P(s''|\tilde{s},a)x(\tilde{s}).$$

Let $\pi^1(i) = \sum_{\tilde{s}} P(i|\tilde{s},a)x(\tilde{s})$ and $\pi^2(i) = \sum_{\tilde{s}}\sum_{s''}\kappa_{s'',z=i}P(s''|\tilde{s},a)x(\tilde{s})$, then $\pi^1,\pi^2 \in X$. The result follows from (A.3) in Platzman (1980). □

Proof of Theorem 7. Note that $P(z'=i|z,a) - P(s'=i|s,a) = P(z'=i|z,a) - \sigma(z'=i,a,\lambda(z,\tilde{x}_*,\tilde{a}_*)) + \sigma(z'=i,a,\lambda(z,\tilde{x}_*,\tilde{a}_*)) - P(s'=i|s,a)$. The results follow by Proposition 2 and Proposition 4. In addition, since D_1 is a metric on X, we have

$$D_1(\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),P(.|s,a)) \leq D_1[\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),\sigma(\cdot,x_*,a)] + D_1[\sigma(\cdot,x_*,a),P(\cdot|s,a)].$$

Hence, Propositions 3 - 4 lead to the result. □

Proof of Corollary 2. If $\eta = 0$, then $x = e_s$, $D_1[\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),\sigma(\cdot,x_*,a)] = 0$, and $D_1[\sigma(\cdot,x_*,a),P(\cdot|s,a)] = 0$. Hence, $D_1[\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a}),P(.|s,a)] = 0$, implying the result. □

Proof of Theorem 8.

$$
\begin{aligned}
\rho_*(z,a) &\le |Q(x_*,a) - Q^R(z,a)| \\
&\le |r(x_*,a) - r^R(z,a)| \\
&+ \beta|\sum_{z'} \sigma(z',a,x_*)\log(\sum_{a'}\exp(Q(\lambda(z',x_*,a),a')) - \sum_{z'} P(z'|z,a)\log(\sum_{a'}\exp(Q^R(z',a')))| \\
&\le |x_* r(a) - x_* r^R(a) + x_* r^R(a) - r^R(z,a)| \\
&+ \beta|\sum_{z'} \sigma(z',a,x_*)[\log(\sum_{a'}\exp(Q(\lambda(z',x_*,a),a')) - \log(\sum_{a'}\exp(Q^R(z',a')))]| \\
&+ \beta|\sum_{z'}[\sigma(z',a,x_*) - P(z'|z,a)]\log(\sum_{a'}\exp(Q^R(z',a')))| \\
&\le ||r - r^R|| + \triangle r^R \\
&+ \beta \max_{z',a'}(|Q(\lambda(z',x_*,a),a')) - Q^R(z',a')|) + \beta\log|A| + 2\beta||\sigma - P^R||_1^*(||Q^R|| + \log|A|) \\
&\le [||r - r^R|| + \triangle r^R](1+\beta) \\
&+ \beta^2 \max_{z',a'}\max_{z'',a''}(|Q(\lambda(z'',a',\lambda(z',a,x_*)),a'')) - Q^R(z'',a'')|) + (\beta+\beta^2)\log|A| \\
&+ 2(\beta+\beta^2)||\sigma - P^R||_1^*(||Q^R|| + \log|A|) \\
&\le [||r - r^R|| + \triangle r^R]\frac{1}{1-\beta} + \frac{\beta}{1-\beta}\log|A| + 2\frac{\beta}{1-\beta}||\sigma - P^R||_1^*(||Q^R|| + \log|A|),
\end{aligned}
$$

where the fourth inequality is by $\max_i\{x_i\} \le \log\sum_{i=1}^N \exp(x_i) \le \max_i\{x_i\} + \log|N|$, the fifth inequality is by induction, and the last inequality is because $Q \in \mathscr{B}$ is bounded. The result follows by rearranging the terms.

On the other hand,

$$
\begin{aligned}
&|r^R(z,a) - r(x_*,a)| \\
&\le |Q(x_*,a) - Q^R(z,a)| \\
&+ \beta|\sum_{z'} \sigma(z',a,x_*)\log(\sum_{a'}\exp(Q(\lambda(z',x_*,a),a')) - \sum_{z'} P(z'|z,a)\log(\sum_{a'}\exp(Q^R(z',a')))|
\end{aligned}
$$

Let $a^*(x) \in \arg\max_{a\in A} Q(x,a)$, and $a_R^*(z) \in \arg\max_{a\in A} Q^R(z,a)$. Then, we have

$$|r^R(z,a) - r(x_*,a)|$$

$$
\begin{aligned}
&\leq \rho^*(z,a) \\
&+\beta|\sum_{z'}\sigma(z',a,x_*)Q(\lambda(z',x_*,a),a^*[\lambda(z',x_*,a)]) - \sum_{z'}P(z'|z,a)Q^R(z',a^*[\lambda(z',x_*,a)])| \\
&+\beta|\sum_{z'}\sigma(z',a,x_*)\log(\sum_{a'}\exp\left(Q(\lambda(z',x_*,a),a') - Q(\lambda(z',x_*,a),a^*[\lambda(z',x_*,a)])\right) \\
&-\sum_{z'}P(z'|z,a)\log(\sum_{a'}\exp\left(Q^R(z',a') - Q^R(z',a^*[\lambda(z',x_*,a)])\right))|,
\end{aligned}
$$

since $\log\sum_{i=1}^N \exp(x_i) = a + \log\sum_{i=1}^N \exp(x_i - a), \forall a \in R$. Now let

$$
\begin{aligned}
h(z,a,x_*) = &\sum_{z'}\sigma(z',a,x_*)\log(\sum_{a'}\exp\left(Q(\lambda(z',x_*,a),a') - Q(\lambda(z',x_*,a),a^*[\lambda(z',x_*,a)])\right) \\
&-\sum_{z'}P(z'|z,a)\log(\sum_{a'}\exp\left(Q^R(z',a') - Q^R(z',a^*[\lambda(z',x_*,a)])\right))
\end{aligned}
$$

Then,

$$
\begin{aligned}
&|r^R(z,a) - r(x_*,a)| \\
&\leq \rho^*(z,a) \\
&+\beta|\sum_{z'}\sigma(z',a,x_*)Q(\lambda(z',x_*,a),a^*[\lambda(z',x_*,a)]) - \sum_{z'}\sigma(z',a,x_*)Q^R(z',a^*[\lambda(z',x_*,a)])| \\
&+\beta|\sum_{z'}\sigma(z',a,x_*)Q^R(z',a^*[\lambda(z',x_*,a)]) - \sum_{z'}P(z'|z,a)Q^R(z',a^*[\lambda(z',x_*,a)])| + \beta|h(z,a,x_*)| \\
&\leq (1+\beta)\max_{z,a}\rho^*(z,a) + 2\beta||\sigma - P^R||_1^*||Q^R|| + \beta|h(z,a,x_*)|.
\end{aligned}
$$

Taking the maximum on both sides gives

$$
\max_{z,a,x_*}|r^R(z,a) - r(x_*,a)| \leq (1+\beta)\max_{z,a}\rho^*(z,a) + 2\beta||\sigma - P^R||_1^*||Q^R|| + \beta\max_{z,a,x_*}|h(z,a,x_*)|,
$$

indicating the result. Lastly, it is easy to see that $h(z,a,x_*) \to 0$ as $Q \to I$, where I is the identity matrix. Furthermore,

$$
\begin{aligned}
h(z,a,x_*) &\leq \log|A| - \sum_{z'}P(z'|z,a)[Q^R(z',a_R^*(z')) - Q^R(z',a^*[\lambda(z',x_*,a)])] \\
&\leq \log|A| + \max_{z,a,a'}|Q^R(z,a) - Q^R(z,a')|,
\end{aligned}
$$

and

$$h(z,a,x_*) \geq -\log|A| - \sum_{z'} P(z'|z,a)[Q^R(z',a_R^*(z')) - Q^R(z',a^*[\lambda(z',x_*,a)])]$$
$$\geq -\log|A| - \max_{z,a,a'} |Q^R(z,a) - Q^R(z,a')|.$$

Hence, the result follows. □

References

Abbring JH, Daljord Ø (2020) Identifying the discount factor in dynamic discrete choice models. *Quantitative Economics* 11(2):471–501.

Aguirregabiria V, Mira P (2002) Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models. *Econometrica* 70:1519–1543.

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

Arcidiacono P, Jones JB (2003) Finite mixture distributions, sequential likelihood and the EM algorithm. *Econometrica* 71(3):933–946.

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

Blevins JR (2016) Sequential Monte Carlo methods for estimating dynamic microeconomic models. *Journal of Applied Econometrics* 31(5):773–804.

Boularias A, Kober J, Peters J (2011) Relative entropy inverse reinforcement learning. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, 182–189 (PMLR).

Caliendo L, Dvorkin M, Parro F (2019) Trade and labor market dynamics: general equilibrium analysis of the China trade shock. *Econometrica* 87(3):741–835.

Choi J, Kim K (2011) Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research* 12:691–730, ISSN 1532-4435, URL `http://dl.acm.org/citation.cfm?id=1953048.2021028`.

Cirillo C, Xu R, Bastin F (2016) A dynamic formulation for car ownership modeling. *Transportation Science* 50(1):322–335.

Diermeier D, Keane M, Merlo A (2005) A political economy model of congressional careers. *American Economic Review* 95(1):347–373.

Duflo E, Hanna R, Ryan SP (2012) Incentives work: getting teachers to come to school. *American Economic Review* 102(4):1241–1278.

Eckstein Z, Wolpin K (1999) Why youth drop out of high school: The impact of preferences, opportunities and abilities. *Econometrica* 67(6):1295–1339.

Fang H, Wang Y (2015) Estimating dynamic discrete choice models with hyperbolic discounting, with an application to mammography decisions. *International Economic Review* 56:565–596.

Finn C, Levine S, Abbeel P (2016) Guided cost learning: deep inverse optimal control via policy optimization. *International conference on Machine learning.*

Hendel I, Nevo A (2006) Measuring the implications of sales and consumer inventory behavior. *Econometrica* 74(6):1637–1673.

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

Hotz VJ, Miller RA, Sanders S, Smith J (1994) A simulation estimator for dynamic models of discrete choice. *The Review of Economic Studies* 61:265–289.

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

Kasahara H, Shimotsu K (2018) Estimation of discrete choice dynamic programming models. *Journal of Applied Econometrics* 69(1):28–58.

Keane M, Wolpin K (1997) The career decisions of young men. *Journal of Political Economy* 105(3):473–522.

Magnac T, Thesmar D (2002) Identifying dynamic discrete choice processes. *Econometrica* 70(2):801–816.

McFadden D (1981) Econometric models of probabilistic choice. *Structural Analysis of Discrete Data* (MIT Press).

Platzman LK (1980) Optimal infinite-horizon undiscounted control of finite probabilistic systems. *SIAM Journal on Control and Optimization* 18:362–380.

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* 999–1033.

Rust J (1994) Structural estimation of Markov decision processes. *Handbook of econometrics* 4:3081–3143.

Rust J, Phelan C (1997) How social security and medicare affect retirement behavior in a world of incomplete markets. *Econometrica* 65(4):781–831.

Sanchez M, Klutke GA (2016) *Reliability and Life-Cycle Analysis of Deteriorating Systems* (Springer).

Simoes A, Viegas JM, Farinha JT, Fonseca I (2017) The state of art of hidden Markov models for predictive maintenance of diesel engines. *Quality and Reliability Engineering International* 33:2765–2779.

Steimle LN, Denton BT (2017) Markov decision processes for screening and treatment of chronic diseases. *Markov Decision Processes in Practice* (International Series in Operations Research & Management Science, vol 248. Springer, Cham).

Su CL, Judd KL (2012) Constrained optimization approaches to estimation of structural models. *Econometrica* 80(5):2213–2230.

Train K (2002) *Discrete choice methods with simulation* (Cambridge University Press).

White CC (1991) A survey of solution techniques for the partially observed Markov decision process. *Annals of Operations Research* 32:215–230.

White CC, Scherer WT (1994) Finite-memory suboptimal design for partially observed Markov decision processes. *Operations Research* 42(3):439–455.

Ziebart BD, Maas A, Bagnell JA, Dey AK (2008) Maximum entropy inverse reinforcement learning. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* 1433–1438.