Review of Economic Studies (2014) **81**, 1229–1262 doi:10.1093/restud/rdt050

Advance access publication 15 December 2013

Semiparametric Inference in Dynamic Binary Choice Models

A. NORETS
University of Illinois at Urbana-Champaign
and
X. TANG
University of Pennsylvania

First version received July 2011*; final version accepted November* 2013 (*Eds.*)

We introduce an approach for semiparametric inference in dynamic binary choice models that does not impose distributional assumptions on the state variables unobserved by the econometrician. The proposed framework combines Bayesian inference with partial identification results. The method is applicable to models with finite space of observed states. We demonstrate the method on Rust's model of bus engine replacement. The estimation experiments show that the parametric assumptions about the distribution of the unobserved states can have a considerable effect on the estimates of per-period payoffs. At the same time, the effect of these assumptions on counterfactual conditional choice probabilities can be small for most of the observed states.

Key words: Dynamic discrete choice models, Markov decision processes, semiparametric inference, identification, Bayesian estimation, MCMC.

JEL Codes: C25, C35, C33, C14, C11

1. INTRODUCTION

1.1. *Background*

A dynamic discrete choice model is a dynamic programme with discrete controls. These models have been used widely in various fields of economics, including labour economics, health economics, and industrial organization. See Eckstein and Wolpin (1989), Rust (1994), Pakes (1994), Miller (1997), Aguirregabiria and Mira (2010) and Keane *et al.* (2011) for surveys of the literature. In such models, a forward-looking decision-maker chooses an action from a finite set in each time period. The actions affect decision-makers' per-period payoffs and the evolution of state variables. The decision-maker maximizes the expected sum of current and discounted future per-period payoffs. Structural estimation of dynamic discrete choice models is especially useful for evaluating the effects of counterfactual changes in the decision environment. The main objective of this article is to provide a robust method for inference about counterfactuals and structural parameters.

In estimable models, some state variables might be unobserved by econometricians. Introduction of these variables into the model is motivated by the fact that individuals always

have more information about their preferences than econometricians. Also, unobserved state variables play an important operational role in estimation as they help make the model capable of rationalizing observed data (see Section 3.1 in Rust (1994)). To our knowledge, previous work estimating dynamic discrete choice models assumed specific parametric forms of the distribution of unobserved state variables. For example, normally distributed utility shocks are mostly used in applications of the interpolation simulation method of Keane and Wolpin (1994) and the Bayesian estimation methods of Imai *et al.* (2009) and Norets (2009); in the methods of Rust (1987) and Hotz and Miller (1993), extreme value independently identically distributed (i.i.d.) unobserved states are often used to alleviate the computational burden of solving and estimating the dynamic programme.

It is well-known that imposing distributional assumptions can have substantial effect on inference in economic models (a discussion of this can be found, for example, in Manski (1999)). Therefore, it is desirable to provide estimation methods that employ restrictions implied by economic theory such as monotonicity, concavity, and independence and avoid strong distributional assumptions on unobserved states. This has been done for static binary choice models; see, for example, Manski (1975), Cosslett (1983), Han (1987), and Matzkin (1992). We provide a semiparametric approach for inference in dynamic binary choice models (DBCMs). The approach can be used as a set of tools for evaluating robustness of existing parametric estimation methods with respect to distributional assumptions on unobserved states.

1.2. *Model, objects of interest, and identification*

We consider models with conditionally independent and additively separable unobserved states as in Rust (1987) and Hotz and Miller (1993). The observed states are assumed to take only a finite number of possible values. The per-period payoffs can be specified parametrically or non-parametrically with optional shape restrictions such as monotonicity or concavity. Using data on individual actions and transitions for the observed state variables, the econometrician can estimate non-parametrically the transition probabilities for the observed states and conditional choice probabilities (CCPs), which are the probabilities of choosing an action conditional on the observed states.

Applied researchers are mainly interested in inference procedures for model primitives such as per-period payoffs and model predictions resulting from counterfactual changes in model primitives. Counterfactual changes we consider include changes in per-period payoffs and changes in transition probabilities for observed states. In a model of job search, an example of the former would be an increase in per-period unemployment insurance payment and an example of the latter would be a change in duration of unemployment insurance. Model predictions for counterfactual experiments can be summarized by the resulting CCPs, which we will call the counterfactual CCPs in contrast to the actual CCPs corresponding to the data-generating process. Results of counterfactual experiments seem to be of most interest in applications. Therefore, we emphasize the counterfactual CCPs as the main object of interest and treat the distribution of unobserved states and the per-period payoffs as nuisance parameters. We develop a separate set of results for parameters of per-period payoffs as sometimes they are of interest as well.

As a starting point for inference, we provide identification results for per-period payoffs and counterfactual CCPs under known and unknown distributions of the unobserved states. Magnac and Thesmar (2002) showed that per-period payoffs are non-parametrically not identified even under a known distribution of unobserved states. First, we show that exogenous variation in transitions for the observed states can lead to nonparametric point identification of the per-period payoffs under a known distribution of unobserved states. Second, we derive conditions under which normalizations on per-period payoffs, which are sufficient for point identification, affect

or do not affect counterfactual predictions under a known distribution of unobserved states. Third, we show that when the distribution of the unobserved states is not assumed to be known, per-period payoffs and counterfactual CCPs are only set identified even under parametric or shape restrictions on the per-period payoffs. Next, we show that even when per-period payoffs are non-parametrically not point-identified under known distribution of unobserved states, the identified set for the counterfactual CCPs under an unknown distribution of unobserved states can still be informative. The size of the identified set decreases with additional shape or parametric restrictions on the per-period payoffs. Finally, we provide characterizations of the identified sets for the per-period payoffs and counterfactual CCPs under an unknown distribution of unobserved states, which are convenient for numerical construction of the identified sets and for use in inferential procedures.

1.3. *Inference*

We show that in our framework the model can be reparameterized so that the observed state transition probabilities, the actual CCPs, and the counterfactual CCPs can be treated as the parameters. The counterfactual CCPs do not enter the likelihood function directly. They are only partially identified by the restrictions the model places on all the parameters jointly. These model restrictions require the actual and counterfactual CCPs to be consistent with some distribution of unobserved states, counterfactual primitives, actual observed state transition probabilities, and some actual per-period payoffs that satisfy (optional) shape restrictions.

We choose the Bayesian approach to inference, which has the following advantages in our settings. First, even under an unknown distribution of unobserved states, DBCMs impose strong restrictions on the actual CCPs. These restrictions should be exploited in estimation. It is conceptually straightforward to incorporate them into the Bayesian estimation procedure through the restrictions on the prior support. Second, the parameters are very high-dimensional and the model restrictions are complicated. In these settings, Markov Chain Monte Carlo (MCMC) methods are instrumental in making inference procedures computationally feasible.

To simplify the specification of the prior and the construction of the MCMC algorithm, we do not treat the per-period payoffs as parameters explicitly in our estimation procedure for the counterfactual CCPs. The per-period payoffs are only implicitly present in verification of model restrictions on the prior support for the observed state transition probabilities, the actual CCPs, and the counterfactual CCPs. We can recover the identified set for the per-period payoffs separately from the estimation of counterfactual CCPs. The MCMC output from the estimation procedure can be used for construction of frequentist confidence sets for partially identified counterfactual CCPs and parameters of per-period payoffs.

1.4. *Application*

We illustrate our method using a model of bus engine replacement (Rust, 1987). We find that assuming a specific parametric distribution for unobserved states can have a large impact on the estimation of parameters of the per-period payoffs. In particular, without the distributional assumptions on the unobserved states, the identified set for the parameters of the linear per-period payoffs in Rust's model includes values that are 5 times larger than the values used in the data-generating process with the extreme value distributed unobserved states. Moreover, if the linearity of the payoff function is not imposed, then the identified set for payoffs in Rust's model includes values that are more than 3 orders of magnitude different from the DGP values. On the other hand, we find that the identified set of the counterfactual CCPs can be small in most dimensions relative to the sampling variation in the actual CCPs for realistic sample sizes. Thus,

in our example parametric assumptions about the distribution of the unobserved states have a small effect on the counterfactual CCPs for most but not all of the observed states.

We also demonstrate that our inference framework can be supplemented with optional restrictions on the quantiles of the unobserved state distribution. This can be used for incorporating information about these quantiles if it is available to researchers. Alternatively, one can use this to do robustness checks on how deviations from the assumed distributions of unobserved states affect estimation results.

The rest of the article is organized as follows. Section 2 describes identification results for the per-period payoffs and the counterfactual CCPs under known and unknown distributions of unobserved states. In Section 3, we discuss the Bayesian approach to inference, its relation to other approaches, and ways to conduct frequentist inference. The MCMC estimation algorithm and prior specification are discussed in the context of the application in Section 4. Section 5 presents estimation results and the identified sets for the per-period payoffs, the time discount factor, and the counterfactual CCPs for Rust (1987) model. Proofs and algorithm implementation details are delegated to appendices. Estimation algorithms and results for a firm entry and exit model are presented in a Supplementary Appendix, Norets and Tang (2013). Computer codes are available on the website of the *Review of Economic Studies*.

2. IDENTIFICATION

2.1. *Model setup*

In an infinite-horizon dynamic binary choice model, the agent maximizes the expected discounted sum of the per-period payoffs

$$V(x_t,\epsilon_t)=\max_{d_t,d_{t+1},\cdots} E_t\Big(\sum_{j=0}^{\infty}\beta^j u(x_{t+j},d_{t+j},\epsilon_{t+j})\Big),$$

where $d_t\in D=\{0,1\}$ is the control variable, $x_t\in X$ are state variables observed by the econometrician, $\epsilon_t=(\epsilon_{t0},\epsilon_{t1})\in\mathbb{R}^2$ are state variables unobserved by the econometrician, β is the time discount factor, and $u(x_t,d_t,\epsilon_t)$ is the per-period payoff. The state variables evolve according to a controlled first-order Markov process. Under mild regularity conditions (see Bhattacharya and Majumdar, 1989) that are satisfied under the assumptions we make below, the optimal lifetime utility of the agent has a recursive representation:

$$V(x_t,\epsilon_t)=\max_{d_t\in D}[u(x_t,d_t,\epsilon_t)+\beta E\{V(x_{t+1},\epsilon_{t+1})|x_t,\epsilon_t,d_t\}]. \tag{2.1}$$

Hereafter we make the following assumptions.

Assumption 1. *The state space for the observed states is finite and denoted by* $X=\{1,\ldots,K\}$.

Assumption 2. *The per-period payoff is* $u(x_t=i,d_t=j,\epsilon_t)=u_{ji}+\epsilon_{tj}$; ϵ_{tj} *is integrable and* $E(\epsilon_{tj}|x)=0$ *for any* $x\in X$ *and* $j\in D$.

Assumption 3. $Pr(x_{t+1}=i|x_t=k,\epsilon_t,d_t=j)=G^j_{ki}$ *is independent of* ϵ_t. *The distribution of* ϵ_{t+1} *given* $(x_{t+1},x_t,\epsilon_t,d_t)$ *depends only on* x_{t+1} *and is denoted by* $H(\cdot|\cdot)$.

Assumption 4. *The distribution of* $\epsilon_{t0}-\epsilon_{t1}$ *given* $x_t=x$ *is denoted by* $F(\cdot|x)$ *and has a positive density on* $\mathbb{R}^1$ *with respect to (w.r.t.) the Lebesgue measure for any* x *in* X.

Assumptions 1–4 are standard in the literature. Assumption 3 of conditional independence is, perhaps, the strongest one. However, it seems hard to avoid. First, it is a sufficient condition for non-degeneracy of the model (see Rust, 1994). Second, without Assumption 3 it is not clear whether the expected value functions are differentiable with respect to parameters (Norets, 2010). Finally, the assumption is also very convenient for computationally feasible classical (Rust, 1994; Hotz and Miller, 1993) and Bayesian (Norets, 2009) estimation of parametrically specified models.

Except for Section 5.2, the discount factor β is assumed to be fixed and known in what follows. This is a common assumption in the literature on estimation of dynamic discrete choice models and also dynamic stochastic general equilibrium models. The values of β can be taken from macroeconomic calibration literature (Kydland and Prescott, 1982) or studies estimating time discount rates from experimental data (see Section 6 in Frederick *et al.* (2002) for an extensive list of references).

In what follows, it is convenient to use the following notation. $G^j = [G^j_{ki}]$ denotes the Markov transition matrix for the observed states conditional on $d_t = j$ and $G = (G^1, G^0)$. A vector of stacked deterministic parts of the per-period payoffs with $d_t = j$ is denoted by $u_j = (u_{j1}, \ldots, u_{jK})'$ and $u = (u_1, u_0)$. A vector of CCPs is defined by $p = (p_1, \ldots, p_K)'$, $p_i = Pr(d_t = 1 | x_t = i)$, $i = 1, \ldots, K$.

2.2. *Identification under known distributions of unobserved states*

Following existing literature on semi- and non-parametric identification (Roehrig, 1988 and Matzkin, 2007 among others), we refer to parameters (u, G, β, F) as a (model) structure.[1] According to the model, transition probabilities G and CCPs p completely determine the distribution of observables. They can be consistently estimated from data and, thus, (p, G) are assumed to be known in the analysis of identification. In what follows, we assume that the model is correctly specified.

In this subsection, we assume that (β, F) are known to the econometrician. The main results of this subsection include a convenient characterization of the relationship between structure (u, G, β, F) and CCPs p in Lemma 1 and the following implications of Lemma 1 for identification: (i) Corollary 3 shows that exogenous variation in transitions for the observed states can lead to non-parametric point identification of the per-period payoffs under a known F; (ii) Lemma 2 shows that normalizations on per-period payoffs can lead to incorrect counterfactual predictions. Readers who are only interested in semiparametric inference for the counterfactual CCPs with unknown F may skip the identification results after Lemma 1 and proceed to Section 2.2.2, which defines the framework and notation for the inference for counterfactual CCPs.

Under Assumptions 1–3, the Bellman equation (2.1) can be rewritten in vector notation as follows:

$$v_0 = u_0 + \beta G^0 \int \max\{v_0 + \epsilon_0, v_1 + \epsilon_1\} dH(\epsilon | X) \tag{2.2}$$

$$v_1 = u_1 + \beta G^1 \int \max\{v_0 + \epsilon_0, v_1 + \epsilon_1\} dH(\epsilon | X),$$

where $v_j = (v_{j1}, \ldots, v_{jK})'$ is a vector of stacked deterministic parts of the alternative specific lifetime utilities $v_{ji} = u_{ji} + \beta E\{V(x_{t+1}, \epsilon_{t+1}) | x_t = i, d_t = j\}$. We also adopt a Matlab-like convention to simplify notation: for scalar ϵ_j and vector v_j, $v_j + \epsilon_j = (v_{j1} + \epsilon_j, \ldots, v_{jK} + \epsilon_j)'$; for a function/expression

1. As we show below, under Assumptions 2–3, the distribution of unobserved states affects CCPs only through the distribution of the difference in utility shocks, F.

$f(x)$ mapping from X to $\mathbb{R}^1$, we use $f(x_1,\ldots,x_K)$ as the short-hand notation for $(f(x_1),\ldots,f(x_K))'$, which is a mapping into $\mathbb{R}^K$. In this notation,

$$\int \max\{v_0+\epsilon_0, v_1+\epsilon_1\}dH(\epsilon|X)=\begin{pmatrix} \int \max\{v_{01}+\epsilon_0, v_{11}+\epsilon_1\}dH(\epsilon|x=1) \\ \cdots \\ \int \max\{v_{0K}+\epsilon_0, v_{1K}+\epsilon_1\}dH(\epsilon|x=K) \end{pmatrix}.$$

Let us rewrite the Bellman equations (2.2) in a form convenient for analysing identification,

$$\begin{aligned} v_0 &= u_0+\beta G^0[v_0+\int \max\{0, v_1-v_0-(\epsilon_0-\epsilon_1)\}dH(\epsilon|X)] \\ v_1 &= u_1+\beta G^1[v_1+\int \max\{0, \epsilon_0-\epsilon_1-(v_1-v_0)\}dH(\epsilon|X)], \end{aligned} \tag{2.3}$$

where we used $E(\epsilon_j|x)=0$ for any $x\in X$ and $j\in D$. Let $\Delta\epsilon=\epsilon_0-\epsilon_1$; then, (2.3) implies

$$\begin{aligned} v_1-v_0 = (I-\beta G^1)^{-1}[u_1+\beta G^1 \int_{v_1-v_0}^{\infty} (s-(v_1-v_0))\}dF(s|X)] \\ -(I-\beta G^0)^{-1}[u_0+\beta G^0 \int_{-\infty}^{v_1-v_0} (v_1-v_0-s)\}dF(s|X)]. \end{aligned} \tag{2.4}$$

Since $d_t=1$ at $x_t=i$ when $v_{1i}-v_{0i}\geq\Delta\epsilon_i$, we have $p=(p_1,\ldots,p_K)'=F(v_1-v_0|X)$. In the following lemma, we give necessary and sufficient conditions for some p to be the CCPs for a given model structure.

Lemma 1. *A vector p is a vector of CCPs implied by structure (u,G,β,F) if and only if*

$$\begin{aligned} F^{-1}(p|X) = (I-\beta G^1)^{-1}\Big[u_1+\beta G^1\big[\int_{F^{-1}(p|X)}^{\infty} sdF(s|X)-(I-\mathrm{diag}(p))F^{-1}(p|X)\big]\Big] \\ -(I-\beta G^0)^{-1}\Big[u_0+\beta G^0\big[\int_{F^{-1}(p|X)}^{\infty} sdF(s|X)+\mathrm{diag}(p)F^{-1}(p|X)\big]\Big]. \end{aligned} \tag{2.5}$$

The necessity follows immediately from (2.4) by substituting in $v_1-v_0=F^{-1}(p|X)$ and the zero expectation for $\Delta\epsilon$ given X. The proof of sufficiency is given in Appendix C. The system of equations (2.5) is convenient for analysing identification and developing inference procedures for two reasons. First, it is linear in the per-period payoffs u. Second, it involves only conditional choice probabilities p and structural parameters (u,G,β,F), but not other non-primitive objects such as optimal continuation values. A result similar to Lemma 1 can also be established for a finite-horizon model.

To our knowledge, Magnac and Thesmar (2002) were the first to provide a formal analysis of identification in the model we consider here.[2] They used related but different system of equations. Their system involves optimal continuation values, and thus, the links between the CCPs and structural parameters are not explicit in their characterizations. Their positive identification results were about certain differences in future value functions and not about parameters in per-period payoffs. Aguirregabiria (2005) derived a system similar to (2.5) as

2. Rust (1994) discusses lack of point identification in a simplified model without unobserved states.

necessary conditions for a vector p to be the CCPs in single-agent models. He used it to identify choice probabilities under counterfactual changes in u when F is assumed to be known. Aguirregabiria (2010) derived a related representation for the finite-horizon case. A result equivalent to Lemma 1 can also be obtained as a special case of the results derived in Aguirregabiria and Mira (2007) and Pesendorfer and Schmidt-Dengler (2008) for dynamic discrete choice games. Heckman and Navarro (2007) analysed identification while assuming that an outcome variable in each period is observable. Kasahara and Shimotsu (2009) studied a model in which a choice probability is a finite mixture of unobservable component CCPs, which are also conditional on unobserved heterogeneity. Hu and Shum (2012) studied the identification of a dynamic discrete choice model in which observed states and unobserved heterogeneity contain a lot of information about each other. Both of these papers showed how to identify both the CCPs conditional on unobserved heterogeneity and the mixture probabilities, but did not analyse identification of the per-period payoffs.

The following remarks, lemmas, and corollaries summarize the implications of Lemma 1 for identification when F is known (p, G, and β are also considered to be known and fixed).

Remark 1. *Because the number of equations in* (2.5)*, K, is smaller than the number of unknowns, $2K$, u cannot be jointly identified without further restrictions. This was first noted by Magnac and Thesmar (2002). In comparison, here we note that* (2.5) *identifies the difference between the discounted total expected payoffs from two trivial policies of clinging to one of the two actions forever: $(I-\beta G^1)^{-1}u_1-(I-\beta G^0)^{-1}u_0$ (note $(I-\beta G^j)^{-1}=I+\sum_{t=1}^{\infty}(\beta G^j)^t$). In the static case, $\beta=0$, this is reduced to the identification of u_1-u_0.*

Even though per-period payoffs are not point-identified, the linear system in (2.5) defines the identified set for u, which is a lower dimensional subset of $\mathbb{R}^{2K}$. Economic theory often provides shape restrictions on u such as linearity, monotonicity, or concavity. Let us denote the set of feasible values of u by U. Any shape restriction obviously reduces the identified set for u. The following corollaries to Lemma 1 demonstrate how shape restrictions can lead to set and point identification.

Corollary 1. *Suppose per-period payoffs satisfy shape restrictions given by strict inequalities (such as strict monotonicity or concavity). Then payoffs are not point-identified.*

Corollary 2. *Suppose per-period payoffs are linear in parameters, $U=\{u: u_j=Z_j\theta \text{ for } j=0,1\}$, where Z_j is a known $K\times d$ matrix and d is the dimension of θ. Then, θ is point- (over-)identified if the rank of $(I-\beta G^1)^{-1}Z_1-(I-\beta G^0)^{-1}Z_0$ is equal to (strictly greater than) d.*

The corollaries follow immediately from Lemma 1.

2.2.1. Exogenous variation in the transition probabilities of the observed states. In this subsection, we consider an alternative way to point identify per-period payoffs when they are specified non-parametrically. Suppose there are $N\geq 2$ observed types of decision-makers in the data (indexed by $n=1,2,\ldots,N$ respectively), for whom the observed state transition probabilities are different but the per-period payoffs are identical. Denote these transition probabilities by $G^{j,n}$ for $j=1,0$ and $n=1,\ldots,N$.

There are lots of applications where this condition can be satisfied. For example, in models of retirement decisions, transition of income will differ for private and public pension plans. Thus, we have two types of agents: those with private pensions and those with public ones. The condition holds if people with different pension plans have the same preference for income and

leisure. Another example is health care utilization decisions such as women's decision to take mammography. The medical history of patient's parents affects her probability of developing breast cancer, but is likely not to affect per-period payoffs, see Fang and Wang (2008).

With F known, the CCPs for all N types (denoted by $p^1, p^2, \ldots, p^N$ respectively) are now characterized by a system of $2K$ unknowns in u and NK equations. Hence, non-parametric identification of per-period payoffs is possible up to an appropriate normalization, provided there is sufficient rank in the coefficient matrix for u. This is formalized in the following corollary proved in Appendix C.

Corollary 3. *Suppose there are N types of decision-makers with different observed state transition probabilities but the same per-period payoffs. (i) For any $N \geq 2$ and any CCPs $(p^1, p^2, \ldots, p^N)$, u is identified up to $2K - r$ normalizations, where r is the rank of the NK-by-$2K$ matrix:*

$$r = rank \begin{bmatrix} (I - \beta G^{1,1})^{-1}, & -(I - \beta G^{0,1})^{-1} \\ \vdots & \\ (I - \beta G^{1,N})^{-1}, & -(I - \beta G^{0,N})^{-1} \end{bmatrix}. \tag{2.6}$$

(ii) When $N = 2$, r defined in part (i) is equal to

$$r = rank \begin{bmatrix} I - \beta G^{1,1}, & I - \beta G^{1,2} \\ I - \beta G^{0,1}, & I - \beta G^{0,2} \end{bmatrix} and \tag{2.7}$$

$$r = K + rank \left[(I - \beta G^{1,1})(I - \beta G^{0,1})^{-1} - (I - \beta G^{1,2})(I - \beta G^{0,2})^{-1} \right]. \tag{2.8}$$

(iii) When $N = 2$ and $G^{0,1} = G^{0,2}$, $r = K + rank\left[G^{1,1} - G^{1,2}\right]$.

For point-identification of the per-period payoffs (up to a location normalization), we need the highest possible rank in Corollary 3, $r = 2K - 1$. Parts (ii) and (iii) of the corollary provide easy-to-interpret sufficient conditions for when point-identification does and does not hold. It can be seen from part (ii) of the corollary that if one generates $G^{i,j}$ independently from continuous distributions then $r = 2K - 1$ with probability 1. At the same time, part (iii) of the corollary illustrates that when only few elements of the transition matrices change exogenously then point identification of the per-period payoffs might fail. Specifically, when G^0 and at least two rows of G^1 do not change exogenously then the rank of $G^{1,1} - G^{1,2}$ is at most $K - 2$ and, thus, by part (iii) r is at most $2K - 2$.

Some earlier papers have used weaker forms of exclusion restrictions to identify various features of DBCMs. Magnac and Thesmar (2002) show that exclusion restrictions can help identify the difference between current value functions defined as the sum of current static payoffs and future value functions. They show that if there exists a pair of states that yield the same current value functions, then current value functions can be identified for all states with knowledge of the unobserved states distribution. Fang and Wang (2008) use a related but different assumption of exclusion restrictions in which there exists a pair of observable states under which the per-period payoffs are the same while transition probabilities to future states are different. This helps identify per-period payoffs in DBCMs with hyperbolic discounting in which one of the actions yields per-period payoffs that are independent from the observed states. In comparison, the assumption of exogenous variation in observed state transition probabilities in the present article is slightly more restrictive but leads to completely non-parametric identification of u under a known F. In addition, our results also provide a transparent relationship between the number of normalizations required for identification of u and easily verifiable rank conditions on observed state transitions.

2.2.2. Identification of counterfactuals. Structural models are useful in the analysis of counterfactual changes in structural parameters. Lemma 1 can be used to set up a framework for analysing identification of the counterfactual CCPs. Suppose we are interested in the CCPs when the per-period payoffs and the observed state transition probabilities are changed to some counterfactual values.

We consider counterfactual experiments described by a pair $(\tilde{U}, \tilde{G})$, where counterfactual transition probabilities for observed states $\tilde{G} = (\tilde{G}^0, \tilde{G}^1)$ are fixed to some known values and correspondence $\tilde{U} : U \to \mathbb{R}^{2K}$ defines a set of permissible values of counterfactual payoffs $\tilde{u}$ given actual payoffs u. For example, when the per-period payoff of choosing alternative 0 is unchanged by the counterfactual and the per-period payoff of choosing alternative 1 goes up by 10% for all observed states, $\tilde{U}(u) = \{\tilde{u} : \tilde{u}_0 = u_0, \tilde{u}_1 = 1.1 \cdot u_1\}$. We do not consider counterfactual experiments that change β or F.

Analysis of counterfactuals is routinely performed in applications. Consider the following examples of counterfactual changes in u: changes in unemployment insurance benefits in a job search model; changes in entry costs resulting from changes in local taxes in firms' entry/exit model; and changes in new engine prices in bus engine replacement model (Rust, 1987). Examples of counterfactual changes in G also abound: changes in social security pension rules in a model of retirement decisions (Rust and Phelan, 1997; Aguirregabiria, 2010); changes in bus routes and, thus, mileage transitions in the engine replacement example; changes in the evolution of determinants of the demand in entry/exit models (Collard-Wexler, 2013); and changes to bankruptcy laws in mortgage default models.

The identified set of the counterfactual CCPs consists of all $\tilde{p}$ implied by structure $(\tilde{u}, \tilde{G}, \beta, F)$, where $\tilde{u} \in \tilde{U}(u)$ with $u \in U$ and structure (u, G, β, F) implies the CCPs p in the DGP.

Remark 2. *Even if u is not point identified, the identified set of counterfactual CCPs can be a proper subset of $(0,1)^K$. For example, consider a special case of counterfactual analysis, in which $\tilde{U}(u) = \{u\}$ but the observed state transition probabilities are changed from G to $\tilde{G}$. Suppose the set of feasible per-period payoffs U is compact and cdf $F(\cdot|x)$ is continuous for any $x \in X$, which implies uniform continuity of p as a function of (u, G) (the continuity follows by standard arguments, see, for example, Norets (2010); the uniformity follows by the compactness assumption). Then, given any $\epsilon_p > 0$ there exists $\epsilon_G > 0$ such that whenever $||\tilde{G} - G|| \leq \epsilon_G$, any $\tilde{p}$ with $||\tilde{p} - p|| > \epsilon_p$ is not in the identified set, where $||\cdot||$ is a Euclidean norm. This follows immediately from the uniform continuity of p as a function of (u, G).*

Of course, any additional shape restrictions in U will reduce the identified set of the counterfactual CCPs. Also, as we discuss in the previous subsection, the per-period payoffs, and thus the counterfactual CCPs, can be point identified under exogenous variation in the transition probabilities of the observed states.

Lemma 1 implies that the per-period payoffs are not identified. For this reason, one might think that setting u_0 equal to any vector of constants c (*e.g.* $c = 0 \in \mathbb{R}^K$) is a necessary normalization for identifying u_1 non-parametrically. However, such an assignment of u_0 is not "innocuous" in that it can lead to errors in predicting the CCPs under counterfactual transition probabilities of the observed states. The next two lemmas give conditions under which normalizing u_0 affects and does not affect the predicted counterfactual outcomes.

Lemma 2. *Consider a counterfactual experiment where G is changed to $\tilde{G}$ and per-period payoffs are unchanged. Denote the true u_0 in the DGP by u_0^*. Suppose researchers set u_0 to some $c \in \mathbb{R}^K$ in order to estimate u_1. Then the predicted $\tilde{p}$ based on $u_0 = c$ differs from the true*

counterfactual CCPs based on $u_0 = u_0^*$ *unless*

$$\left[(I-\beta G^1)(I-\beta G^0)^{-1} - (I-\beta \tilde{G}^1)(I-\beta \tilde{G}^0)^{-1}\right](c-u_0^*) = 0. \tag{2.9}$$

Lemma 2 suggests that setting u_0 to an arbitrary constant in general affects the predicted counterfactual outcome. To see this, suppose the rank of the matrix in (2.9) multiplying $(c-u_0^*)$ is equal to the largest possible value $K-1$. Corollary 3 with $N=2$, $G^{0,2}=\tilde{G}^0$, and $G^{1,2}=\tilde{G}^1$ provides easy-to-understand sufficient conditions for the rank condition to hold as the matrices in (2.8) and (2.9) have the same form. Under this rank condition, (2.9) holds if and only if $c-u_0^* = \text{const}\cdot(1,1,\ldots,1)'$. Hence, setting u_0 to be a vector with equal coordinates (such as the zero vector) when the actual u_0^* is not independent from observed states (*i.e.* $u_0^* \neq \text{const}\cdot(1,1,\ldots,1)'$) leads to incorrect counterfactual predictions.

On the other hand, the following lemma shows that setting u_0 to an arbitrary vector can serve as an innocuous normalization if the goal is to predict counterfactual outcomes under linear changes in the per-period payoffs.

Lemma 3. *Consider a counterfactual experiment* $\tilde{G}=G$ *and* $\tilde{U}(u)=\{\tilde{u}=\alpha u+(\Delta_1,\Delta_0)\}$, *where* Δ_k*'s are known* K*-vectors and* α *is a known scalar. Setting* u_0 *to an arbitrary vector does not affect the predicted counterfactual CCPs.*

To our knowledge, Lemmas 2 and 3 present the first formal discussion in the literature about the impact of normalizations of per-period payoffs on various types of counterfactual analyses. Proofs of these lemmas are included in Appendix C.

2.3. *Identification under unknown distributions of unobserved states*

In this subsection, we assume that the distribution of unobserved states is unknown to the econometrician. The main results of this subsection include a convenient characterization of the relationship between (u,G,β) and p when the distribution of unobserved states is unknown (Theorem 1) and the following implications of Theorem 1 for identification: Corollaries 4 and 5 show that the identified sets for per-period payoffs and counterfactual CCPs can be explored with the help of linear programming algorithms as long as the shape restrictions on per-period payoffs are defined by linear equalities and inequalities.

Hereafter, we maintain the following assumption.

Assumption 5. *The distribution of* $\Delta\epsilon$ *is independent from* x*:* $F(\cdot|x)=F(\cdot)$.[3]

Equations in (2.5) suggest that the CCPs depend on F only through a finite number of quantiles $F^{-1}(p_k)$ and corresponding integrals $\int_{F^{-1}(p_k)}^{+\infty} s dF(s)$. Lemma 4 below characterizes the relations between such quantiles and corresponding truncated integrals for a generic F. Theorem 1 then combines Lemma 4 with Lemma 1 to characterize the CCPs when F is not known.

Lemma 4. *Given any positive integer* L *and a triple of* L*-vectors* p, δ, *and* e, *where components of* p *are labelled so that* $1>p_1>p_2>\ldots>p_L>0$ *without loss of generality, there exists a distribution*

3. A characterization of the CCPs without this assumption can be obtained by arguments similar to those we employ in Lemma 4 below. However, such a characterization seems to be too weak to be useful in empirical work.

F such that

$$F \text{ has a density } f>0 \text{ on } R \text{ w.r.t. the Lebesgue measure,} \tag{2.10}$$

$$\int sdF(s)=0 \tag{2.11}$$

$$F^{-1}(p_i)=\delta_i,\ i\in\{1,\dots,L\} \tag{2.12}$$

$$e_i=\int_{\delta_i}^{\infty} sdF(s),\ i\in\{1,\dots,L\} \tag{2.13}$$

if and only if

$$\frac{e_1}{1-p_1}>\delta_1>\cdots \tag{2.14}$$

$$\cdots>\delta_i>\frac{e_{i+1}-e_i}{p_i-p_{i+1}}>\delta_{i+1}>\cdots \tag{2.15}$$

$$\cdots>\delta_L>-\frac{e_L}{p_L}. \tag{2.16}$$

Conditions in the lemma are easy to understand geometrically. For example, condition (2.15) simply requires the expectation of $\Delta\epsilon$ conditional on $\Delta\epsilon\in(\delta_{i+1},\delta_i)$ to be in (δ_{i+1},δ_i),

$$\frac{\int_{\delta_{i+1}}^{\delta_i} sdF}{F(\delta_i)-F(\delta_{i+1})}\in(\delta_{i+1},\delta_i).$$

The lemma is proved in Appendix C. The following theorem follows immediately from Lemmas 1 and 4.

Theorem 1. *For a given triple (u,G,β), a vector p in $(0,1)^K$ is the CCPs implied by structure (u,G,β,F) for some F satisfying* (2.10)–(2.11) *if and only if there exist e in $\mathbb{R}^K$ and δ in $\mathbb{R}^K$ such that: (i)*

$$\delta=(I-\beta G^1)^{-1}\Big[u_1+\beta G^1\big[e-(I-\operatorname{diag}(p))\delta\big]\Big] \tag{2.17}$$
$$-(I-\beta G^0)^{-1}\Big[u_0+\beta G^0\big[e+\operatorname{diag}(p)\delta\big]\Big];$$

(ii) After relabelling the K coordinates in p and their corresponding coordinates in δ and e so that $1>p_1\geq p_2\geq\cdots\geq p_K>0$, the unique (strictly ordered) components in p and their corresponding coordinates in δ and e satisfy (2.14)–(2.16)*, and $e_i=e_j$, $\delta_i=\delta_j$ whenever $p_i=p_j$.*

For example, suppose $K=6$, $p=(p_1,p_2,\dots,p_6)=(\frac{1}{3},\frac{2}{5},\frac{3}{4},\frac{1}{3},\frac{1}{10},\frac{2}{5})$, and let $\delta=(\delta_1,\delta_2,\dots,\delta_6)$ and $e=(e_1,e_2,\dots,e_6)$. Then the restrictions in (ii) of Theorem 1 are summarized as: $\delta_1=\delta_4$, $\delta_2=\delta_6$, $e_1=e_4$, $e_2=e_6$, and

$$\frac{e_3}{1-p_3}>\delta_3>\frac{e_2-e_3}{p_3-p_2}>\delta_2>\frac{e_1-e_2}{p_2-p_1}>\delta_1>\frac{e_5-e_1}{p_1-p_5}>\delta_5>-\frac{e_5}{p_5}.$$

When F is unknown, (2.17) implies that the scale of F can be normalized without a loss of generality as long as U is a linear cone, which we assume hereafter. Let $\delta_m = F^{-1}(p_m)$ denote the unrestricted median of F. A convenient normalization is to fix the value of

$$e_m = \int_{F^{-1}(p_m)}^{\infty} s dF(s) = \log(2), \text{ where } p_m = 0.5, \tag{2.18}$$

as in the logistic distribution (the location of F is normalized in Assumption 2 and equation (2.11)).

2.3.1. Identified set for per-period payoffs. By definition, the identified set for per-period payoffs under an unknown F consists of all u in U such that for some F satisfying (2.10)–(2.11) and a scale normalization, structure (u, G, β, F) implies the CCPs p in the DGP. The following corollary to Theorem 1 provides a computationally convenient characterization of the identified set of per-period payoffs.

Corollary 4. *(i) For a given pair (p, G), the identified set of per-period payoffs, $\mathcal{U}(p, G)$, consists of all u in U for which there exists $(\delta, e, \delta_m) \in \mathbb{R}^{2K+1}$ satisfying the following conditions. (a) $(u, G, \beta, \delta, e, p)$ satisfy* (2.17). *(b) Let (p_m, e_m) be defined as in* (2.18). *Define $K+1$-vectors $p^* = (p_1^*, \ldots, p_{K+1}^*)'$, e^*, and δ^* by stacking (p_m, p), (e, e_m), and (δ, δ_m) respectively and relabelling the coordinates so that $1 > p_1^* \geq p_2^* \geq \cdots \geq p_{K+1}^* > 0$. Then, the unique (strictly ordered) coordinates in p^* and the corresponding coordinates in δ^* and e^* satisfy* (2.14)–(2.16), *and $e_i^* = e_j^*$ and $\delta_i^* = \delta_j^*$ whenever $p_i^* = p_j^*$.*

(ii) If U is convex then $\mathcal{U}(p, G)$ is also convex.

The dependence of the identified set $\mathcal{U}(p, G)$ on the time discount factor β and restrictions on the per-period payoffs U is suppressed in the notation for simplicity.

The scale normalization employed in the corollary, which is described in the previous subsection, is convenient for computing the identified sets as it implies the linearity of the equalities and inequalities in the unknowns. A linear programming algorithm can be used to verify whether a given vector of per-period payoffs is in $\mathcal{U}(p, G)$.

If the per-period payoffs are parameterized then an approximation to the identified set of the parameters can be computed on a grid: for each point in the grid, we can check if the corresponding per-period payoffs are in $\mathcal{U}(p, G)$. If the restrictions on u_j are linear ($U = \{u : u_j = Z_j\theta\}$), then a more efficient algorithm described in Section 4.2 and Appendix B.3 can be used to compute the identified set. These strategies are computationally feasible when the dimension of θ is not high. Even when the dimension of θ is high it is feasible to compute the identified sets for lower-dimensional sub-vectors of θ under linear u_j as we describe at the end of Appendix B.3.

The definition of the identified set in Corollary 4 assumes a known time discount factor. It is possible to include β in the vectors of parameters to be identified. We give an example of the joint identified set for β and payoff parameters in our application (Section 5.2).

2.3.2. Identified set for counterfactual CCPs. The notation and examples of counterfactual changes in primitives (u, G) are described in Section 2.2.2. When F is unknown to the econometrician, the identified set of the CCPs under counterfactual experiment $(\tilde{U}, \tilde{G})$ consists of all $\tilde{p}$ implied by any structure $(\tilde{u}, \tilde{G}, \beta, F)$, in which F satisfies (2.10)–(2.11), $\tilde{u} \in \tilde{U}(u)$ with $u \in U$, and structure (u, G, β, F) implies the CCPs p in the DGP. It is important to note that unknown F is assumed to be unchanged by the counterfactual. The following corollary to Theorem 1 provides a computationally convenient characterization of the identified set of counterfactual CCPs.

Corollary 5. *For a given pair (p,G) and a counterfactual experiment $(\tilde{U},\tilde{G})$, the identified set for counterfactual CCPs, $\mathcal{P}(p,G)$, consists of all $\tilde{p}$ for which there exist $u\in U$, $\tilde{u}\in\tilde{U}(u)$, and $(\delta,e,\tilde{\delta},\tilde{e},\delta_m)\in\mathbb{R}^{4K+1}$ such that the following conditions hold. (a) (u,G,β,δ,e,p) and $(\tilde{u},\tilde{G},\beta,\tilde{\delta},\tilde{e},\tilde{p})$ satisfy* (2.17). *(b) Let (p_m,e_m) be as defined in* (2.18). *Define $2K+1$-vectors $p^*=(p_1^*,\ldots,p_{2K+1}^*)'$, e^*, and δ^* by stacking $(p_m,\tilde{p},p)$, $(e,\tilde{e},e_m)$, and $(\delta,\tilde{\delta},\delta_m)$ respectively and relabelling the coordinates so that $1>p_1^*\geq p_2^*\geq\cdots\geq p_{2K+1}^*>0$. Then, the unique (strictly ordered) coordinates in p^* and the corresponding coordinates in δ^* and e^* satisfy* (2.14)–(2.16)*, and $e_i^*=e_j^*$ and $\delta_i^*=\delta_j^*$ whenever $p_i^*=p_j^*$.*

The dependence of the identified set $\mathcal{P}(p,G)$ on the time discount factor β, restrictions on the per-period payoffs U, and counterfactual $(\tilde{U},\tilde{G})$ is suppressed in the notation for simplicity. In the corollary, we use a scale normalization on F defined in (2.18). Note, however, that a scale normalization does not affect $\mathcal{P}(p,G)$.

Given the lack of non-parametric identification for u even under a known F (Remark 1), it is clear that the shape and/or functional form restrictions, U, play an important role in partially identifying the counterfactual CCPs under an unknown F. However, as the following lemma demonstrates even without restrictions on u, the identified set, $\mathcal{P}(p,G)$, can be a proper subset of $(0,1)^K$.

Lemma 5. *Consider a counterfactual change in u such that $(I-\beta G^1)^{-1}(\tilde{u}_1-u_1)+(I-\beta G^0)^{-1}(\tilde{u}_0-u_0)\neq 0$. Then, $\mathcal{P}(p,G)$, is a proper subset of $(0,1)^K$.*

Since the characterization of $\mathcal{P}(p,G)$ is rather involved it seems hard to determine analytically what affects the size of this set. Therefore, we suggest using numerical algorithms to learn about $\mathcal{P}(p,G)$. To verify that a candidate vector $\tilde{p}$ belongs to $\mathcal{P}(p,G)$, one needs to check the feasibility of equalities and inequalities described in Corollary 5. This can be done by a linear programming algorithm described in Appendix B.1 as long as $\tilde{U}(u)$ is defined by linear equalities and inequalities. As we demonstrate in the application (Section 5.3), this linear programming algorithm can be combined with MCMC to estimate the identified set.

3. INFERENCE

The identification results from the previous section demonstrate that per-period payoffs, u, and counterfactual CCPs, $\tilde{p}$, are only partially identified when F is unknown. There is a growing literature in econometrics on inference for partially identified parameters. See for example, Manski (2003), Imbens and Manski (2004), Chernozhukov *et al.* (2007), Rosen (2008), Stoye (2009), Beresteanu and Molinari (2008), Romano and Shaikh (2010), Andrews and Soares (2010), Canay (2010), Bugni (2010), and Moon and Schorfheide (2012).

In principle, one can apply a criterion function approach of Chernozhukov *et al.* (2007) or related approaches to construct confidence sets for the identified sets of $\tilde{p}$ and parameters of u (see Appendix E in Norets and Tang (2013) for a description of a criterion function). In our high-dimensional settings, a criterion function approach seems computationally challenging. Therefore, it is essential to develop an inference procedure that can handle high-dimensional problems.

3.1. *Bayesian approach*

Bayesian inference in partially identified models is conceptually straightforward. In these models, the likelihood function can be represented as a function of point identified parameters ((p,G) in our case). Partially identified parameters and structural model restrictions can enter the econometric

model through the restrictions on the prior distribution for $(p, G, \tilde{p})$. An important advantage of the Bayesian approach is that Bayesian MCMC methods perform well in high-dimensional problems. On the other hand, possible dependence of Bayesian estimation results on the prior is an important issue that needs to be carefully addressed.

Next, we describe data typically used for estimating a DBCM and construct the likelihood function. We then give a detailed description of our Bayesian procedure. The following subsection describes the frequentist properties of the procedure.

DBCMs are usually estimated from panel data on individual choices and observed states, $(x_1^i, d_1^i, \ldots, x_{T_i}^i, d_{T_i}^i)$, where i is an index for individuals in the sample and T_i is the number of time periods in which i is observed to make decisions. Given a vector of CCPs, p, and Markov transition matrices for observed states, G, the distribution of the observables $P\left(\{d_1^i, x_2^i, d_2^i, \ldots, x_{T_i}^i, d_{T_i}^i\}_{i=1}^n | p, G, \{x_1^i\}_{i=1}^n\right)$ is given by

$$p_k^{n_k^1}(1-p_k)^{n_k^0} \cdot \prod_{k,l=1}^{K} (G_{kl}^1)^{\nu_{kl}^1} (G_{kl}^0)^{\nu_{kl}^0}, \tag{3.19}$$

where n is the number of individuals in the sample, n_k^j is the number of observed decisions $d_t^i = j$ at state $x_t^i = k$, ν_{kl}^j is the number of observed transitions from $x_t^i = k$ to $x_{t+1}^i = l$ given the decision $d_t^i = j$. The distribution of the observables above is conditional on the initial observed states x_1^i. This is appropriate in the Rust (1987) model that we consider in Sections 4–5. An alternative that might be appropriate in other applications is to assume that the process for (x_t, d_t) is stationary and combine (3.19) with the implied stationary distribution for x_1^i, which would be a function of (p, G).

In a standard estimation procedure for DBCMs (Rust, 1994; Keane and Wolpin, 1994), (u, G, F) are parameterized. After solving for value functions in the model, one can replace the CCPs p in the likelihood in (3.19) with functions of the parameters. The parameters are then estimated by the maximum likelihood. Estimates of the counterfactual CCPs are obtained by solving the model under the counterfactual changes in the estimated parameters.

When F is unknown, there are multiple values of u and $\tilde{p}$ that can be consistent with the structural model and given values of (p, G). Thus, $(p, G, \tilde{p}, u)$ can all be treated as parameters for estimation. Under this parameterization, the likelihood function is given by (3.19). It depends only on (p, G). As we described in the identification section, the structural model does restrict (p, G) (Theorem 1). It also restricts $(\tilde{p}, u)$ for given (p, G) (Corollaries 4 and 5). It is natural to incorporate these restrictions into the econometric model via the prior distribution.

First, suppose that the primary interest is in $\tilde{p}$ and u can be treated as a nuisance parameter. Then one can define a joint prior for $(p, G, \tilde{p})$ as a distribution truncated to $\{(p, G, \tilde{p}) : \tilde{p} \in \mathcal{P}(p, G)\}$, where $\mathcal{P}(p, G)$ is the identified set for $\tilde{p}$ defined in Corollary 5. The posterior distribution of $(p, G, \tilde{p})$ is proportional to the product of this prior and the likelihood in (3.19). In this case, we can treat (u, δ, e) as nuisance parameters. The prior for them is not specified and (u, δ, e) appear only implicitly in verification that $\tilde{p} \in \mathcal{P}(p, G)$. This reduces the dimension of the problem and considerably simplifies specification of the prior and construction of the MCMC algorithm for exploring the posterior distribution. Properties that researchers would like to impose on u a priori can be included in this approach through the shape or parametric restrictions U. Similarly, certain restrictions on F can also be incorporated (see Section 4.6).

If the primary interest is in u and no counterfactual experiments are considered then the approach of the previous paragraph can be modified (a prior for (p, G, u) is truncated to $\{(p, G, u) : u \in \mathcal{U}(p, G)\}$, where $\mathcal{U}(p, G)$ is the identified set for u defined in Corollary 4). Alternatively, one

can consider the posterior distribution of (p,G) only. The prior for (p,G) can be truncated to the restrictions described in Theorem 1 and u can be treated as a nuisance parameter in the estimation of (p,G) (u only shows up implicitly in the verification of the linear restrictions in Theorem 1). Then, credible and confidence sets for $\mathcal{U}(p,G)$ or $\mathcal{P}(p,G)$ can be constructed after estimation of (p,G). We implement the latter approach in our application (Section 5.5) as it does not require developing an additional MCMC algorithm for exploring the posterior of (p,G,u).

Specifying an uninformative prior for $(p,G,\tilde{p})$ (or (p,G,u) or (p,G)) is not trivial because the support of the prior can be rather complicated. For example, a uniform prior for $(p,G,\tilde{p})$ truncated to $\{(p,G,\tilde{p}):\tilde{p}\in\mathcal{P}(p,G)\}$ can be very informative as we demonstrate in Section 4.3. Flexible hierarchical priors, which allow for a priori dependence in components of $(p,G,\tilde{p})$, seem to provide a general solution to this problem. In Section 4, we provide further motivation and details for the prior specification and implementation of the MCMC algorithm in the context of Rust's model.

3.2. *Frequentist inference based on MCMC output*

In this subsection, we describe frequentist properties of the Bayesian estimation procedure described in the previous subsection. We also present ways to use Bayesian estimation output for construction of classical confidence sets.

To be specific let us consider estimation of $(p,G,\tilde{p})$. Let us assume that the prior density is positive and continuous in an open neighbourhood of the data-generating values of (p,G). By the Bernstein-von Mises theorem (see, for example, Chapter 10 in van der Vaart (1998)), Bayesian credible sets for point-identified parameters (p,G) are asymptotically equivalent to the corresponding confidence sets based on the maximum likelihood estimator (MLE) for (p,G), which is a simple frequency estimator under discrete X.

For partially identified parameters the Bernstein-von Mises theorem does not hold. Moon and Schorfheide (2012) show that credible sets for partially identified parameters are strictly smaller than the corresponding confidence sets asymptotically. Their results apply to our settings. To understand these results suppose that the prior of $\tilde{p}$ conditional on (p,G) is a uniform distribution on $\mathcal{P}(p,G)$. As the sample size increases, the posterior for (p,G) concentrates around the DGP values and the posterior for $\tilde{p}$ converges to the uniform distribution on the identified set for $\tilde{p}$. Thus, a Bayesian credible set for $\tilde{p}$ excludes about 5% of the volume of the identified set. In contrast, a 95% classical confidence set for $\tilde{p}$ typically includes the identified set.

The conceptual differences between classical and Bayesian inference for partially identified parameters can also be described as follows. Bayesian inference does not distinguish between the uncertainty from the lack of point identification and that from the sampling variability. In contrast, a standard classical 95% confidence set allows for errors in 5% of hypothetical repeated samples; however, the lower bound on the coverage rate is imposed at *all* parameter values in the parameter space.

Classical and Bayesian approaches can be reconciled if the identified set is the object of interest. In this case, the posterior for the identified parameters (p,G) implies the posterior distribution for $\mathcal{P}(p,G)$ on a space of sets. Let us denote a $100(1-\alpha)\%$ Bayesian credible set for (p,G) by $B^{p,G}_{1-\alpha}$. Then,

$$B^{\mathcal{P}}_{1-\alpha}=\bigcup_{(p',G')\in B^{p,G}_{1-\alpha}}\mathcal{P}(p',G') \tag{3.20}$$

is a $100(1-\alpha)\%$ credible set for $\mathcal{P}(p,G)$. In settings with multiple prior distributions for partially identified parameters considered in Kitagawa (2011), sets in (3.20) have posterior lower

probability at least $1-\alpha$ (Appendix E.2 in Norets and Tang (2013) provides more details on Kitagawa (2011)'s approach).

If $B^{p,G}_{1-\alpha}$ has $100(1-\alpha)\%$ frequentist coverage then the set in (3.20) also has $100(1-\alpha)\%$ frequentist coverage. One could go further and consider confidence sets of the form

$$C^{\mathcal{P}}_{1-\alpha} = \bigcup_{(p',G')\in C^{p,G}_{1-\alpha}} \mathcal{P}(p',G'), \tag{3.21}$$

where $C^{p,G}_{1-\alpha}$ is a $100(1-\alpha)\%$ frequentist confidence set for (p,G). In a search of confidence sets for $\mathcal{P}(p,G)$ that satisfy any reasonable optimality criterion such as smallest weighted expected volume, one can restrict attention to sets satisfying (3.21) because any confidence set for the identified set can be represented as a superset of (3.21) (Lemma C9 in Appendix C).

In our application, we use approximations to the highest posterior density credible sets as $B^{p,G}_{1-\alpha}$ (and $C^{p,G}_{1-\alpha}$, which is justified by the Bernstein-von Mises theorem). This could lead to conservative confidence sets. However, the problem of finding $C^{p,G}_{1-\alpha}$ so that $C^{\mathcal{P}}_{1-\alpha}$ in (3.21) satisfies some optimality properties seems to be hard to solve analytically or numerically in our high-dimensional settings. Most of the literature on confidence sets for partially identified parameters also does not consider optimality properties.[4] Thus we leave this issue to future research.

Approximations to sets in (3.21) (or (3.20)) for any particular $C^{p,G}_{1-\alpha}$ (or $B^{p,G}_{1-\alpha}$) can be easily obtained from MCMC estimation output, $(p^t, G^t, \tilde{p}^t, t=1,2,\ldots)$, of the Bayesian approach described in the previous subsection. Specifically, sets in (3.21) (or (3.20)) can be estimated by the the support of $\{\tilde{p}^t : (p^t, G^t) \in C^{p,G}_{1-\alpha}\}$ (or $\{\tilde{p}^t : (p^t, G^t) \in B^{p,G}_{1-\alpha}\}$). Thus, the Bayesian approach described in the previous subsection can be used for implementing frequentist inference on $\mathcal{P}(p,G)$ and, in a similar fashion, on $\mathcal{U}(p,G)$.

For applied work, we recommend reporting the whole posterior distributions for $\tilde{p}$ since they are more informative than credible sets. They also have a decision theoretic justification. As we demonstrate in Sections 4–5, posterior distributions can be compared with prior distributions and estimates of the identified sets to evaluate the extent of uncertainty from the set identification, the prior shape, and the sampling variation.

4. APPLICATION: METHODOLOGY

We illustrate our methodology using Rust (1987) model of bus engine replacement. First, we describe the model. Second, we discuss how to construct the identified sets for the per-period payoff parameters. Third, we discuss the prior specification and the MCMC algorithm. Fourth, we show how additional restrictions of F can be incorporated in our framework. Section 5 presents the results for simulated and real data.

4.1. *Rust (1987) model of optimal bus engine replacement*

Rust (1987) model of optimal bus engine replacement is a standard example in the literature. Several papers used it for testing new methodologies for estimation of dynamic discrete choice models (see Aguirregabiria and Mira, 2002; Bajari *et al.*, 2007; and Norets, 2009).

4. One exception to this is Chiburis (2009) who considers optimality in testing moment inequality models.

In the model, a transportation company manager decides in each time period t whether to replace ($d_t=1$) or maintain ($d_t=0$) the engines of each bus in the company's fleet. The observed state variable is the cumulative mileage of a bus engine at time t (denoted by x_t) since the last engine replacement. Some additional factors that can affect the replacement or maintenance costs, denoted by $\epsilon_t=(\epsilon_{t0},\epsilon_{t1})$, are observed by the manager but not the econometrician. The mileage is discretized into $K=90$ intervals $X=\{1,\dots,K\}$. The costs of engine replacement and maintenance at time t are given respectively by $u(x_t=k,d_t=1,\epsilon_t)=u_{1k}+\epsilon_{t1}$ and $u(x_t=k,d_t=0,\epsilon_t)=u_{0k}+\epsilon_{t0}$. u_{1k} is constant across k and it captures the deterministic one-time replacement costs; u_{0k} is the deterministic maintenance cost for an engine in mileage interval k. For the rest of the section, we normalize u_{1k} to 0 for all k. For $x_t\le 88$, the change in mileage $(x_{t+1}-x_t)$ follows a multinomial distribution on $\{0,1,2\}$ with parameters $\pi=(\pi_0,\pi_1,\pi_2)$. For x_t equal to 89 and 90, the multinomial distributions are respectively given by $(\pi_0,1-\pi_0,0)$ and $(1,0,0)$. Buses are assumed to start with a new engine, so the likelihood can be conditional on $x_1=1$ for new buses. The Markov transition matrices for x_t, G, can be easily constructed from π.

Rust (1987) assumes an extreme value distribution for ϵ_{t0} and ϵ_{t1} and exploits this assumption to estimate parameters in u_0. In comparison, we do not rely on assumptions about the parametric form of the distribution of ϵ_t. We assume only that ϵ_t is independent of x_t.

It is computationally convenient to define the set of feasible per-period payoffs U by a linear system of equations or inequalities. This can accommodate all parametric cases considered in Rust (1987) where cost functions are linear in the unknown parameters. To fix ideas, we adopt a simple linear index specification: $U=\{u\colon u_{1k}=0 \text{ and } u_{0k}=\theta_0+\theta_1 k, \text{ for some } \theta_1<0,\theta_0\in\mathbb{R}^1\}$. If F were known to econometricians, θ_0 and θ_1 would be over-identified by Corollary 2 in Section 2. Since U is defined by a linear system of equations, whether $\tilde p\in\mathcal{P}(p,\pi)$ can be verified by checking the feasibility of a system of linear equalities and inequalities (see a characterization of $\mathcal{P}(p,\pi)$ in Corollary 5). A linear programming algorithm for checking the feasibility of the system is described in Appendix B.1.

The counterfactual experiments we consider involve changes only in transition probabilities for the observed state, $\tilde\pi\neq\pi$, where $\tilde\pi$ is known. The per-period payoffs are left unchanged, $\tilde U(u)=\{u\}$.

The following properties of the model are useful for developing prior specification and the MCMC algorithm.

Lemma 6. *If $\theta_1<0$, then $p=(p_1,\dots,p_k,\dots,p_K)$ is increasing in k.*

Lemma 7. *For counterfactual experiments that only change observed state transition probabilities in π, the CCP given $x_t=1$ does not change: $p_1=\tilde p_1$.*

Lemma 8. *Consider the same counterfactual experiment as in Lemma 7. If there are $(p,\tilde p,\pi)$ such that $\tilde p\in\mathcal{P}(p,\pi)$ and no two coordinates in $(p,\tilde p_2,\dots,\tilde p_K)$ are identical, then for any $(p_2^\star,\dots,p_K^\star,\tilde p_2^\star,\dots,\tilde p_K^\star,\pi^\star)$ sufficiently close to $(p_2,\dots,p_K,\tilde p_2,\dots,\tilde p_K,\pi)$, $\tilde p^\star=(p_1,\tilde p_2^\star,\dots,\tilde p_K^\star)\in\mathcal{P}((p_1,p_2^\star,\dots,p_K^\star),\pi^\star)$.*

4.2. *Computing identified sets for per-period payoff parameters*

We developed two algorithms for computing the identified set for per-period payoff parameters. The algorithms are based on solving multiple linear programmes. One algorithm exploits the convexity of the identified set (Corollary 4 (ii)) and the supporting hyperplane theorem, which implies that the boundary of the set can be fully characterized by supporting hyperplanes. To explain the idea of the algorithm, let us consider the case where $\theta\in\mathbb{R}^2$. Any point on the boundary

of a convex set in $\mathbb{R}^2$ is uniquely associated with a tangent line, which is completely characterized by a slope and an intercept. Thus, the task of recovering a convex set in $\mathbb{R}^2$ amounts to recovering the set of pairs of slopes and intercepts that define the boundary points. To implement this idea we define a grid on the slopes. For each slope in the grid, we obtain the intercepts of the two lines with this slope that are tangent to the identified set. The intercepts are obtained by solving a linear programme. Appendix B.3 formally describes the algorithm. The appendix also describes the other algorithm, which performs a grid search in the original parameter space.

Note that if the researcher is interested in a lower-dimensional sub-vector of θ, then the grid search over the slopes or the original parameter space needs to be performed only over the lower-dimensional part of the parameter space.[5]

4.3. *Prior*

Specification of an uninformative prior for $(p,\tilde{p},\pi)$ requires some care. First of all, such a prior must give probability 1 to $\{(p,\tilde{p},\pi):\tilde{p}\in\mathcal{P}(p,\pi)\}$. Furthermore, by Lemma 7 the prior must imply $p_1=\tilde{p}_1$. Thus, from now on we exclude the first coordinate from $\tilde{p}$ (it is implicitly given by p_1). Lemma 8 suggests that it is reasonable to construct the prior for $(p,\tilde{p},\pi)$ by specifying a density with respect to the Lebesgue measure on $\mathbb{R}^{2K+1}$ that is truncated to $\{(p,\tilde{p},\pi):\tilde{p}\in\mathcal{P}(p,\pi)\}$ (the dimension of π is 2, and the dimension of $(p,\tilde{p})$ is $2K-1$ since $p_1=\tilde{p}_1$). At first sight, one might suggest using a product of uniform (or uninformative Beta) densities on $(0,1)$ for each component of $(p,\tilde{p})$ and a Dirichlet density for π. However, such a specification results in a strongly informative prior that can dominate the likelihood even for moderate sample sizes. A short explanation for this is that a priori independence of components of p or $\tilde{p}$ is unreasonable. To get more insight in the context of Rust's model note that Lemma 6 implies monotone coordinates in p and $\tilde{p}$. A uniform distribution for coordinates of p truncated to monotonicity restrictions $p_1<p_2<\cdots<p_K$ results in the marginal distribution for p_K equal to the distribution of the K^{th} order statistic, which is far from uniform for $K=90$. This example is not exactly equivalent to uniform truncated to $\{(p,\tilde{p},\pi):\tilde{p}\in\mathcal{P}(p,\pi)\}$ but it illustrates the problem well.

A general solution to this problem, which is likely to work for other models as well, is to allow for dependence of coordinates of $(p,\tilde{p})$ in the distribution that is to be truncated to $\{(p,\tilde{p},\pi):\tilde{p}\in\mathcal{P}(p,\pi)\}$. Prior dependence in the Bayesian framework can be introduced through hierarchical modelling (see Sims (2006) for an insightful discussion of an example where hierarchical modelling solves a somewhat similar problem; for a textbook treatment of hierarchical models, see Geweke (2005) and Gelman *et al.* (2003)). To illustrate this idea, suppose components of p are i.i.d. Beta$(ms,(1-m)s)$ with location and spread parameters (m,s) truncated to $p_1<p_2<\cdots<p_K$ and (m,s) have a flexible prior distribution. For any fixed (m,s) we would have the same problem of rather dogmatic marginal distributions for components of p when K is large. However, when (m,s) can vary, the marginal distributions of components of p can have considerably larger variances.

To obtain more prior flexibility in conditional prior distributions, we use a finite mixture of beta distributions truncated to $\{(p,\tilde{p},\pi):\tilde{p}\in\mathcal{P}(p,\pi)\}$ as the prior for $(p,\tilde{p})$. Mixtures of beta distributions can approximate and consistently estimate large non-parametric classes of densities; see, for example, Rousseau (2010). Let M denote the number of mixture components, $z_k\in\{1,\ldots,M\}$ (and $\tilde{z}_k$) denote a latent mixture component allocation variable for p_k (and $\tilde{p}_k$), and $Pr(z_k=j)=\alpha_j$ be the mixing probability for component j. Then, before truncation to

5. We thank the editor for pointing this out.

$\{(p,\tilde{p},\pi):\tilde{p}\in\mathcal{P}(p,\pi)\}$,

$$P(p_k|z_k=j,m,s)=\text{Beta}(p_k;m_js_j,(1-m_j)s_j)=\frac{\Gamma(s_j)p_k^{m_js_j}(1-p_k)^{(1-m_j)s_j}}{\Gamma(m_1s_1)\Gamma((1-m_1)s_1)}.$$

Introduction of such allocation variables is standard in MCMC estimation of mixture models; see Diebolt and Robert (1994). Parameterization of beta distribution in terms of location m_j and spread s_j is also convenient for implementation of MCMC estimation algorithms. With this notation, the prior distribution up to a normalizing constant is given by

$$\begin{aligned}
&P(p,\tilde{p},\pi,z,\tilde{z},\alpha,m,s)\propto \qquad (4.22)\\
&\prod_{k:z_k=1}\text{Beta}(p_k;m_1s_1,(1-m_1)s_1)\cdots\prod_{k:z_k=M}\text{Beta}(p_k;m_Ms_M,(1-m_M)s_M)\\
&\cdot\prod_{k:\tilde{z}_k=1}\text{Beta}(\tilde{p}_k;m_1s_1,(1-m_1)s_1)\cdots\prod_{k:\tilde{z}_k=M}\text{Beta}(\tilde{p}_k;m_Ms_M,(1-m_M)s_M)\\
&\cdot\prod_{j=1}^{M}\text{Beta}(m_j;\underline{N}_{m0},\underline{N}_{m1})\cdot\text{Gamma}(s_j;\underline{\gamma}_{s0},\underline{\gamma}_{s1})\\
&\cdot\prod_{k=1}^{K}\alpha_{z_k}\cdot\prod_{k=2}^{K}\alpha_{\tilde{z}_k}\cdot\prod_{j=1}^{M}\alpha_j^{\underline{a}-1}\\
&\cdot\pi_0^{\underline{b}-1}\pi_1^{\underline{b}-1}(1-\pi_0-\pi_1)^{\underline{b}-1}\cdot 1_{\mathcal{P}(p,\pi)}(\tilde{p}),
\end{aligned}$$

where values for the hyperparameters $\underline{N}_{m0}$, $\underline{N}_{m1}$, $\underline{\gamma}_{s0}$, $\underline{\gamma}_{s1}$, $\underline{a}$, and $\underline{b}$ are chosen by the researcher. A standard Dirichlet prior for π is suitable as π is low dimensional and the data contain a lot of information about π.

4.4. *Overview of MCMC algorithm*

The posterior distribution of $(p,\tilde{p},\pi,z,\tilde{z},\alpha,m,s)$ is proportional to the product of the prior in (4.22) and the likelihood in (3.19) with G_{kl}^j replaced by the corresponding elements of π. Its density can be computed up to a normalizing constant from (3.19) and (4.22). Therefore, a Metropolis-Hastings MCMC algorithm can in principle be used for exploring the posterior distribution.[6] To achieve good performance in practice, the proposal transition density in a Metropolis-Hastings algorithm should mimic the posterior distribution. In actual applications, the dimension of p can be high and constructing good proposal distributions for a Metropolis-Hastings algorithm can be challenging. In this case, an MCMC algorithm that uses the Gibbs sampler, which updates only one or a few coordinates of the parameter vector at a time, can be much more effective.[7]

6. To produce draws from some target distribution, a Metropolis-Hastings MCMC algorithm needs only values of a kernel of the target density. The draws are simulated from a transition density and they are accepted with probability that depends on the values of the target density kernel and the transition density. If a new draw is not accepted, the previous draw is recorded as the current draw from the Markov chain. The sequence of draws from this Markov chain converges to the target distribution. For more details, see, for example, Chib and Greenberg (1995) or Geweke (2005).

7. The Gibbs sampler divides the parameter vector in blocks and sequentially produces draws from the distribution of one block conditional on the other blocks and data. For example, to explore $P(\theta_1,\theta_2)$ on iteration r the sampler produces $\theta_1^{(r)}\sim P(\theta_1|\theta_2^{(r-1)})$ and $\theta_2^{(r)}\sim P(\theta_2|\theta_1^{(r)})$. The sequence of draws $(\theta_1^{(r)},\theta_2^{(r)})$ from this Markov chain converges to $P(\theta_1,\theta_2)$. For more details; see, Tierney (1994) or Geweke (2005).

Also, different variations of the Metropolis-Hastings algorithm can be used together with the Gibbs sampler to construct robust hybrid MCMC algorithms, see Tierney (1994). We use these ideas along with particular properties of the model such as Lemma 6 and Lemma 7 to develop an MCMC sampler that performs well in the application and has required theoretical properties. Appendix A provides a detailed description of the algorithm.

The main computational burden of our algorithm is the solution of the linear program for verification of $\tilde{p} \in \mathcal{P}(p,\pi)$ on every iteration of the MCMC algorithm. The number of constraints and variables in the linear program increases only linearly in the size of p. Nevertheless, our semi-parametric estimation algorithm is more computationally intensive than Rust's algorithm for parametric models because the MCMC algorithm typically requires more iterations for convergence than the likelihood maximization in Rust's algorithm. We report approximate computing times for identification and estimation exercises in Section 5.

In Norets and Tang (2013), we consider a firm's entry and exit model and demonstrate that practical MCMC algorithms can be constructed even if the researcher does not have a priori information about properties of CCPs implied by the model restrictions such as monotonicity in Rust's model proved in Lemma 6. To construct a practical MCMC algorithm for a new model, the researcher may need to further experiment with MCMC blocks and proposal densities described in Appendix A and Norets and Tang (2013). The performance of MCMC algorithms depends on the dimension and complexity of the model and it might be more difficult to handle models that are more complex than the ones we consider.

4.5. *Uses of MCMC algorithm*

In addition to exploring the posterior distribution of $(p,\pi,\tilde{p})$ for Bayesian inference in Section 5.4, we use the MCMC algorithm for two other purposes. First, the confidence sets for θ presented in Section 5.5 are constructed as follows. Similar to a confidence set for $\tilde{p}$ discussed in Section 3.2, a valid $1-\alpha$ confidence set for θ can be defined by

$$B^{\theta}_{1-\alpha} = \bigcup_{(p',\pi') \in B^{p,\pi}_{1-\alpha}} \mathcal{U}(p',\pi'), \tag{4.23}$$

where $B^{p,\pi}_{1-\alpha}$ is a $1-\alpha$ confidence set for (p,π) and $\mathcal{U}(p,\pi)$ is the identified set for θ defined in Corollary 4. To construct an asymptotically valid approximation to a confidence set $B^{p,\pi}_{1-\alpha}$ we use a $1-\alpha$ Bayesian credible set. To construct this credible set we estimate mean, $\hat{\mu}$, and variance–covariance matrix, $\hat{\Sigma}$, of the posterior for $(\log p,\pi)$ using MCMC draws from the posterior. The posterior of $\log p$ is better approximated by a normal distribution than the posterior of p, especially for coordinates for which we do not have many observations. $B^{p,\pi}_{1-\alpha}$ is approximated by posterior draws satisfying $[(\log p,\pi)-\hat{\mu}]'\Sigma^{-1}[(\log p,\pi)-\hat{\mu}] \leq c$, where critical value c is chosen so that $100(1-\alpha)\%$ per cent of posterior draws are in the set. Thus, set $B^{\theta}_{1-\alpha}$ has Bayesian and asymptotic frequentist interpretations. For each posterior draw of (p',π') in $B^{p,\pi}_{1-\alpha}$, we compute $\mathcal{U}(p',\pi')$ by the algorithm described in Section 4.2.

Second, we use the MCMC algorithm for recovering the identified set of counterfactual CCPs in Section 5.3 (it is impractical to do a grid search to recover this set in a space with dimension $K=90$). The MCMC algorithm is the same as the one described in Apppendix A, except that we keep (p,π) fixed. The support of the posterior explored by this MCMC is then used as an approximation of the identified set of counterfactual CCPs.

4.6. *Additional restrictions on F*

So far we have taken an agnostic approach to the distribution of the unobserved states. We do not assume anything about F other than the existence of positive density and independence from observed states. Nonetheless, in practice, researchers might have some idea about the magnitude of shocks. It is possible to include researchers' knowledge about F in the form of restrictions on the quantiles into our estimation procedure. For example, in experiments we use the following restrictions,

$$|\delta_i - F^{-1}_{\text{logistic}}(p_i)| < bd \cdot \max\{|F^{-1}_{\text{logistic}}(p_i) - F^{-1}_{\text{normal}}(p_i)|, \sigma_{\text{logistic}}\}, \tag{4.24}$$

where σ_{logistic} is the standard deviation of the logistic distribution and bd is a parameter. Since the distance between the quantiles of normal and logistic distributions around 0.5 is very small, the presence of σ_{logistic} in the bound allows for somewhat bigger deviations from the logistic distribution around 0.5. The parameter bd controls the size of the allowed deviations. Using logistic and normal quantiles as benchmarks is sensible as most of the applications use these distributions for unobserved states. Performing estimation and identification exercises with different values of bd can shed light on sensitivity of results with respect to parametric assumptions about F. At the same time, any degree of flexibility can be attained by setting appropriate values for bd. In experiments below we use $bd \in \{0.25, 1, \infty\}$. To impose these additional restrictions on F in the model, one can just add inequality (4.24) to inequalities in Theorem 1. More generally, any linear restrictions on δ can be included in the model in a similar fashion.

5. APPLICATION: RESULTS

In this section, we present the identified sets for payoff parameters, discount factor, and counterfactual CCPs, estimation results for actual and counterfactual CCPs, and confidence sets for per period payoff parameter in Rust's model. Except for Section 5.5, we use simulated data below. This allows us to compare estimation results with the identified sets corresponding to the DGP.

To simulate the data, we solve the dynamic programming problem to find the actual CCPs as described in Rust (1987). We use the following DGP for simulating the data: logistic F, $\theta_0 = 5.0727$, $\theta_1 = -0.002293$, $\pi_0 = 0.3919$, $\pi_1 = 0.5953$ and the discount factor $\beta = 0.999$. These parameter values correspond to Rust's estimates for group 4 except that we decreased θ_0 by 5 and decreased β by 0.0009 in order to increase engine replacement probabilities for low mileage (without this change simulated data contained no replacement observations for very low x). Section 5.5 presents the application of the methodology for estimating payoff parameters to real data.

5.1. *Identified sets for per-period payoff parameters*

This subsection recovers the population identified set of per-period payoff parameters (θ_0, θ_1) and visualizes the identifying power of additional restrictions on F. The three nested sets in Figure 1 correspond to the identified sets of (θ_0, θ_1) under different quantile restrictions on F as in (4.24) with $bd = \infty, 1, 0.25$. The smaller bd is, the closer F is required to be to the logistic distribution. The largest identified set in Figure 1 corresponds to unrestricted F and includes values of (θ_0, θ_1) that differ from the DGP values by five times. The figure shows that stronger restrictions on quantiles of F considerably reduce the size of the identified set.

Recovering the identified set of u_0 without assuming linearity is challenging because it is impractical to perform a grid search in a 90-dimensional parameter space. However, one can

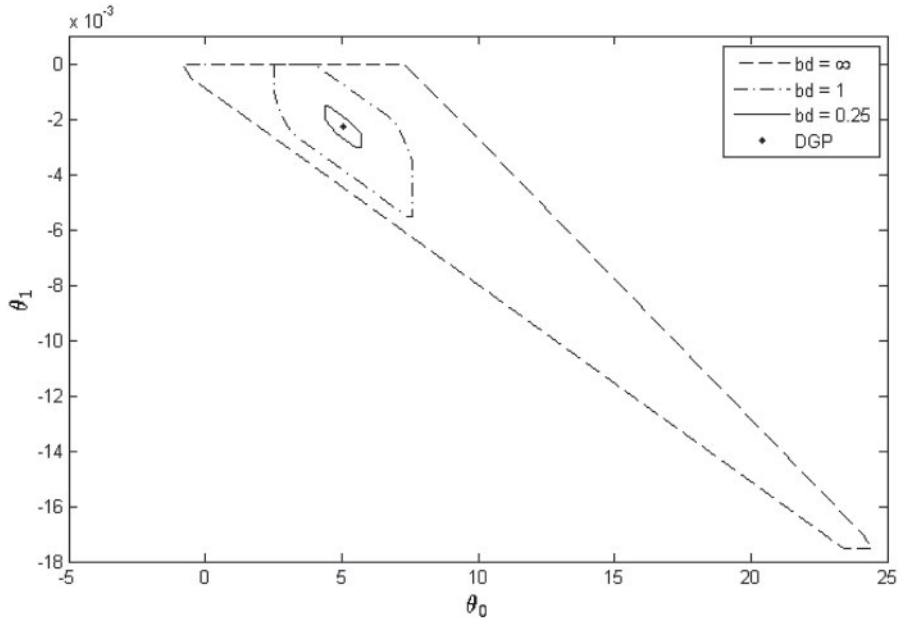

FIGURE 1
Identified sets of cost parameters with $bd = \infty, 1, 0.25$

guess and verify that the identified set of u_0 includes values that are more than three orders of magnitude larger than the DGP values. To understand this, note that by inequalities (2.14)–(2.16), the range of quantiles δ consistent with given CCPs p is large if the smallest CCP, p_1, is small and the largest CCP, p_K, is large. In our parameterization of the GDP, the smallest CCP, p_1, is of the order 10^{-4} and thus, according to (2.16), δ_1 can reach the order 10^4. By (2.17), this can lead to large values in u_0.

5.2. *Joint identification of time discount factor and per-period payoffs*

In this subsection, we relax the assumption that the time discount factor is known and describe the joint identified set for the payoff parameters and the time discount factor. The DGP is described in the beginning of Section 5. We define a fine grid for β and run the algorithm for recovering the identified sets for θ for every value of β in the grid. For all values of β in the grid the identified set for θ is non-empty. Figure 2 shows identified sets of θ computed for several different values of β. The figure shows that even without additional restrictions on F, the joint identified set of (β, θ) is informative. However, the projection of this identified set on the time discount factor dimension is $(0, 1)$. With additional restrictions on F, the projection of the identified set on the time discount factor dimension can be a proper subset of $(0, 1)$. For example, with F restricted by $bd = 0.25$ (see Section 4.6), the projection is $[0.52, 1)$ or, in other words, the identified set for θ is empty for $\beta < 0.52$.

5.3. *Identified sets of counterfactual CCPs*

In this subsection, we examine the identified set of counterfactual CCPs when the transition probabilities are changed to $\tilde{\pi}_0 = 0.6$ and $\tilde{\pi}_1 = 0.3$ and $(\beta, \theta_0, \theta_1)$ are unchanged. Figure 3 presents

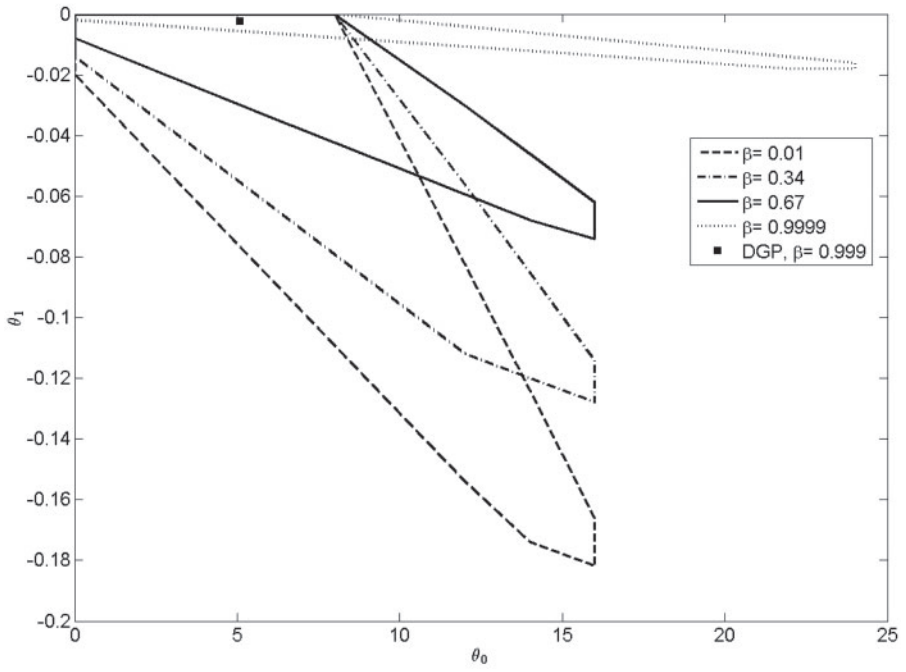

FIGURE 2
Identified sets of θ for different β ($bd=\infty$)

the marginal "posteriors" for counterfactual CCPs with (p,G) fixed at the DGP values. The supports of these distributions are the projections of the K-dimensional identified set for the counterfactual CCPs onto single dimensions corresponding to each coordinate in $\tilde{p}$. Figure 4 depicts these supports for each coordinate k under no additional restrictions on F. As can be seen from the figure, the projections of the identified set for $x \geq 85$ are much larger than the projections for $x < 85$. From Figure 3, we see that as F is restricted to be increasingly closer to the logistic distribution, the identified sets for the counterfactual CCPs in most dimensions change little, even though the identified sets for the per-period payoff parameters presented in Figure 1 are reduced a lot.

5.4. *Estimation of counterfactuals with simulated data*

Using parameter values described in the beginning of Section 5, we simulate a data set for 5000 buses. For each bus, we simulate data starting with $x=1$ until the engine is replaced. The goal is to estimate the counterfactual CCPs without relying on the distributional assumption about ϵ_{t0} and ϵ_{t1} when the transition probabilities are changed to $\tilde{\pi}_0=0.6$ and $\tilde{\pi}_1=0.3$ and $(\beta,\theta_0,\theta_1)$ are unchanged.

We estimate the model for three different numbers of mixture components in the prior, $M=3,6,9$. The results are similar and we report them only for $M=6$. The values for prior hyperparameters are $\underline{N}_{m0}=2$, $\underline{N}_{m1}=2$, $\underline{\gamma}_{s0}=10$, $\underline{\gamma}_{s1}=10$, $\underline{a}=3$, and $\underline{b}=3$. Estimation results are robust to reasonable changes in the prior hyperparameters such as $\underline{\gamma}_{s0}=2$ and $\underline{\gamma}_{s1}=50$. The length of all MCMC runs is about 3 million draws. Using MATLAB on a PC with Intel 2.7 GHz processor and 8 GB RAM, it takes about 38 seconds to obtain 100 MCMC draws. Since the draws are highly serially correlated, we thin the MCMC sample keeping only every 100-th draw. We

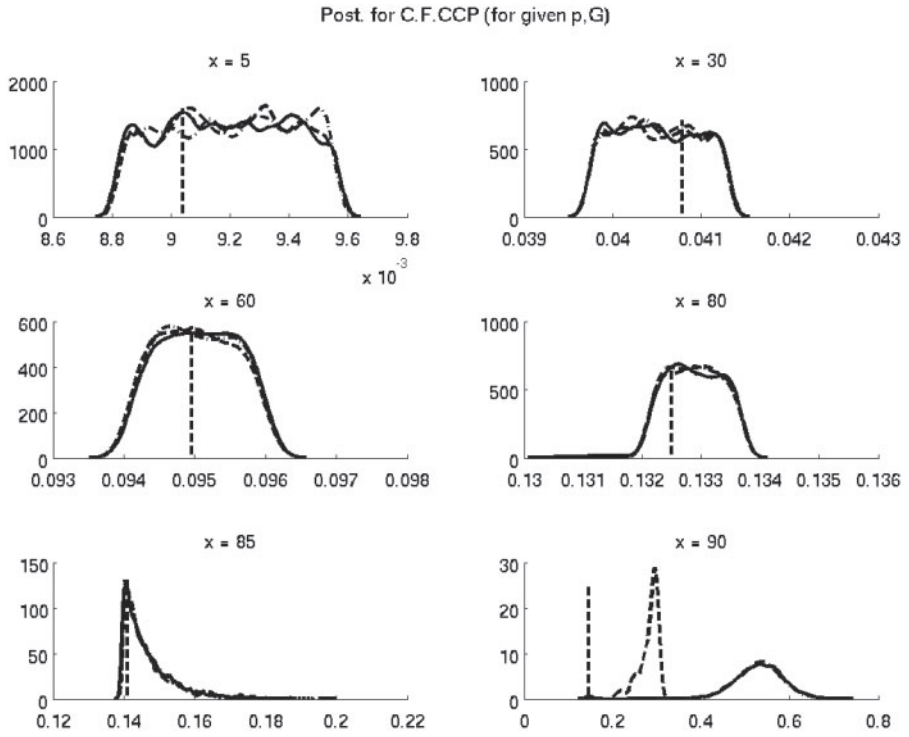

FIGURE 3

Posteriors of c.f. CCPs with p, G fixed as in DGP with $bd = \infty$ (solid), 0.25 (dash), 1 (dash-dot). Vertical lines: "true" c.f. CCPs with ϵ_i extreme value i.i.d

report estimation results using the thinned samples. Trace plots of MCMC draws from several simulator runs suggest that the MCMC algorithm converges.[8] Figure 5 displays posteriors for the actual CCPs and the counterfactual CCPs together. Compared with the posterior of the actual CCPs, the posterior of the counterfactual CCPs $\tilde{p}$ appears to be shifted slightly to the right. Let us provide an intuitive explanation for this increase in counterfactual CCPs. Let $V(x)$ denote the expected continuation value when the current mileage is x and the engine is not replaced. When an engine is replaced the bus in the next period has mileage $x = 1 + j$ with probability π_j. This means that the expected continuation value when engine is replaced is $V(1)$ (it is the same as the one for not replacing the engine at $x = 1$). Then, the choice probability is given by $p(x) = F(u_1(x) - u_0(x) + \beta[V(1) - V(x)])$ and the effect of a change in the transition probabilities on $p(x)$ depends on how the change in $V(1)$ compares to the change in $V(x)$. For different specifications of $(u_1(x), u_0(x))$, the effect of a change in π on $(V(1) - V(x))$ can be either positive or negative. Moreover, the effect can be positive at some x and negative at other x. Intuitively, when transitions to lower mileage become more likely both the expected continuation values of engine replacement and maintenance will go up. Which one goes up more seems to depend on the behaviour of per-period payoff functions and the rest of the structural parameters. For our parameter values, the change is positive for all x.

8. Trace plots and prior and posteriors comparisons are presented in the Supplementary Materials (Norets and Tang (2013)).

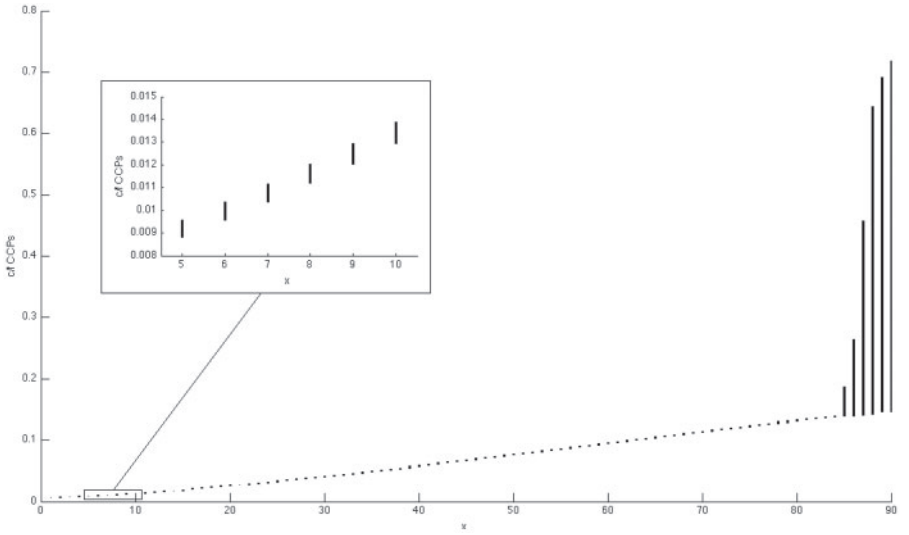

FIGURE 4

Identified set projections for counterfactual CCPs with $bd = \infty$

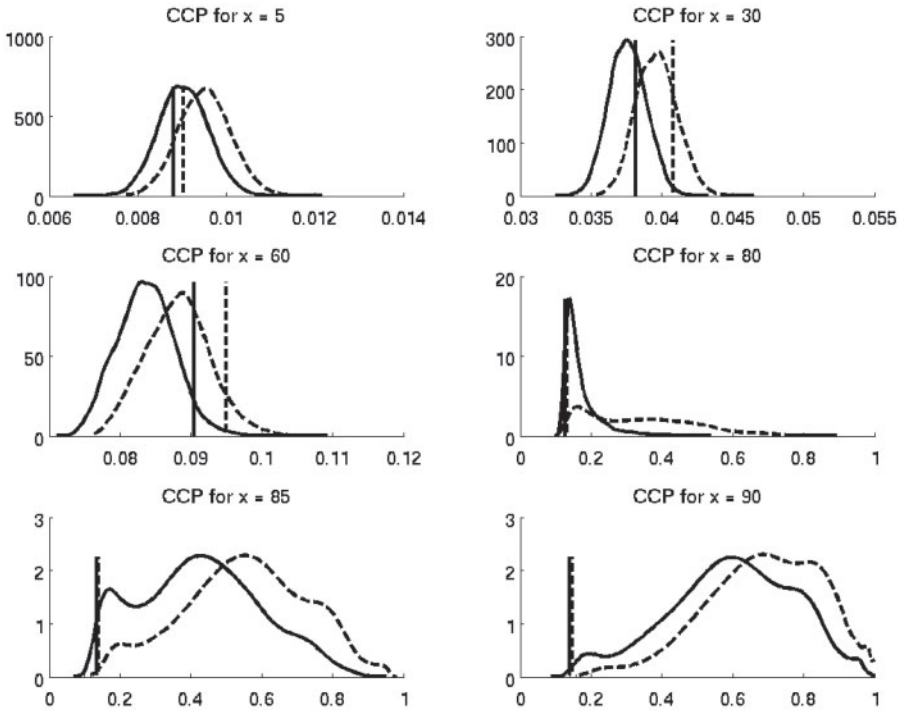

FIGURE 5

Posteriors for actual CCPs (solid) and counterfactual CCPs (dashed). No bounds on δ. Vertical lines show "true" values for actual and counterfactual CCPs

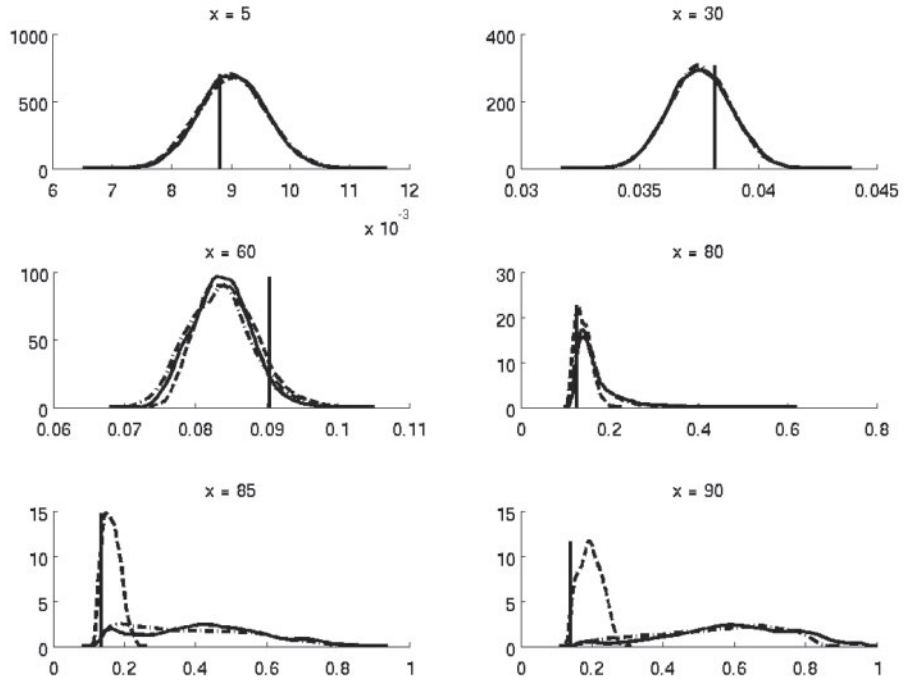

FIGURE 6
Posteriors for actual CCPs with restrictions on F: $bd = \infty$ (solid), $bd = 1$ (dash-dot), $bd = 0.25$ (dashed). Vertical lines are "true" actual CCPs

Figure 5 also reveals that the posteriors for the CCPs flatten out as the mileage increases. The height of posterior densities for the CCP is over 500 at $x = 5$ and close to 20 at $x = 80$. This happens because in our simulated data, most of the engines are replaced at lower or medium mileages. Thus, there are few observations for high mileage and the CCPs for high mileage are estimated less precisely. Figures 6 and 7 present estimation results with additional restrictions on quantiles of F introduced in Section 4.6. While changing bd from infinity to 1 has no visible effect on the actual CCPs posteriors, setting $bd = 0.25$ decreases the heavy right tails of the CCPs posteriors for large mileage. These effects are even more pronounced for the counterfactual CCPs: the almost flat posterior for the CCP at $x = 80, 85, 90$ (with $bd = 1, \infty$) becomes much more informative with $bd = 0.25$. The marginal posteriors for higher coordinates $k = 80, 85, 90$ in Figure 7 should also be interpreted as evidence that a specification of the unobserved state distribution only has a substantial impact on the counterfactual CCPs for a few observed states. Comparing the posterior distributions with the identified sets for the counterfactual CCPs in Section 5.3, we can assess the contribution of the lack of point identification to the posterior distributions. A comparison of Figures 3 and 7 suggests that the identified sets for the counterfactual CCPs are small relative to the posterior support for most of the coordinates of $\tilde{p}$.

5.5. *Inference for payoff parameters using real data*

In this subsection, we compare parameter estimation results from Rust (1987) and our semi-parametric procedure using Rust's data on the buses from group 4. Rust's estimates for (θ_0, θ_1) from Table IX are $(10.075, -0.002293)$ with the standard errors correspondingly

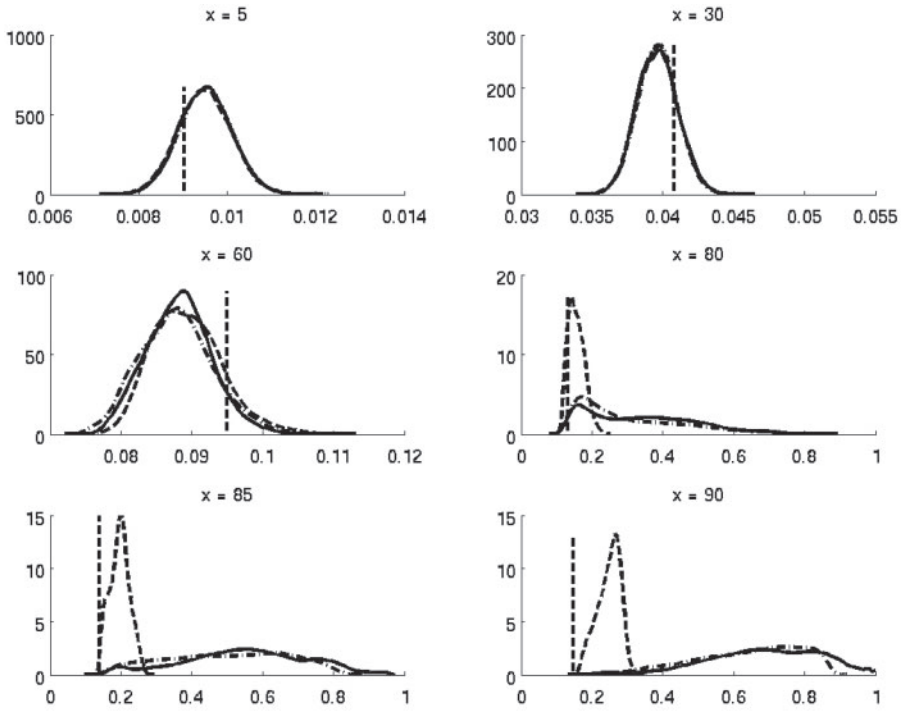

FIGURE 7

Posteriors for c.f. CCPs with restrictions on F: $bd=\infty$ (solid), $bd=1$ (dash-dot), $bd=0.25$ (dashed). Vertical lines are "true" c.f. CCPs

(1.582, 0.000639). Figure 8 depicts a 90% confidence set for (θ_0, θ_1) obtained without any parametric assumptions about F. The confidence set is the union of the sets shown in Figure 8. The algorithm for its construction is described in Section 4.2. A comparison of the confidence set with Rust's results is consistent with the identification results of Section 5.1 based on simulated data: distributional assumptions about unobserved states can have an enormous effect on parameter estimates in DBCMs.

6. EXTENSIONS AND FUTURE WORK

Our method can be extended to dynamic binary choice games of incomplete information. In such games, the individual player's problem is similar to the single agent's problem we described above, see Aguirregabiria and Mira (2007) and Pesendorfer and Schmidt-Dengler (2008). The essential difference is that the agent's per-period payoff and the observed states transition probabilities depend on the actions of other players. As Aguirregabiria and Mira (2007) and Pesendorfer and Schmidt-Dengler (2008), we can consider Markov perfect equilibria and assume that the observed data correspond to a single equilibrium (or that observations can be divided into groups or markets so that every group corresponds to a single equilibrium). Then, the identified sets for counterfactual CCPs and per-period payoffs can be characterized in the same way as in the single-agent case considered above.

Our methodology can also be extended to models with a finite number of unobserved agent types. In these models, CCPs for each agent type can be non-parametrically point

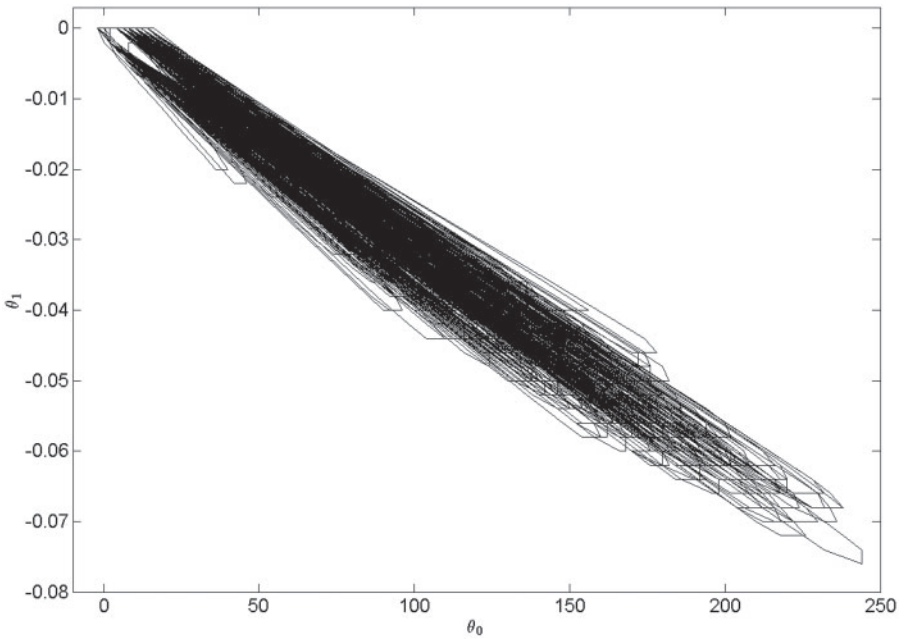

FIGURE 8
90% confidence set for θ

identified. Kasahara and Shimotsu (2009) provide testable sufficient conditions for that. Under point identification of CCPs for each agent type, we can apply Corollaries 4 and 5 to characterize identified sets for per-period payoffs and counterfactual CCPs for each type. Finite number of unobserved agent types can also be accommodated in MCMC algorithms by using standard data augmentation techniques for finite mixture models (Diebolt and Robert, 1994; Fruhwirth-Schnatter, 2006). It also seems possible to extend the MCMC algorithm to estimation of models with time-variant unobserved heterogeneity considered by Arcidiacono and Miller (2011). Generalizing our framework to multinomial choice models is an important direction for future research. Finally, tuning and extending the proposed MCMC algorithms to more complex DBCMs with high-dimensional parameters could be a difficult task, and this is another important area for future work.

APPENDIX

A. MCMC ALGORITHM

This appendix provides a description of the Metropolis-within-Gibbs algorithm for exploring the posterior distribution of $(p, \tilde{p}, \pi, z, \tilde{z}, \alpha, m, s)$ defined in Section 4.4. The algorithm is implemented in Matlab. The computer code is available on the website of the *Review of Economic Studies*. The algorithm consists of the following blocks.

1 Stack current draws of p and $\tilde{p}$ in one vector $(p, \tilde{p})$. Sort $(p, \tilde{p})$ in an ascending order. Draw a candidate for each even coordinate of the sorted vector from a beta distributions proportional to the product of (4.22) and (3.19) and truncated to be between the adjacent odd coordinates. Accept the draw with probability 1 if the candidate does not violate the model restrictions ($\tilde{p} \in \mathcal{P}(p, G)$) and reject otherwise. This block is introduced because draws that do not preserve the order and change many coordinates of $(p, \tilde{p})$ at the same time are rarely accepted.

2 The same as block 1 but for odd coordinates. Markov transition in blocks 1and 2 does preserve the stationary distribution of the Markov chain. However, since the model restrictions in Rust's model do not imply a particular

order of coordinates in $(p,\tilde{p})$ (they only imply an order within p and $\tilde{p}$ separately) we need to add Markov transitions that would allow a change in the order. The following three blocks achieve this.

3 Pick an index $k \in \{1,\dots,K\}$ randomly or deterministically. Draw a candidate for p_k from a beta distribution proportional to product of (4.22) and (3.19) and truncated to (p_{k-1}, p_{k+1}). Accept the draw with probability 1 if the candidate does not violate the model restrictions and reject otherwise.

4 The same as block 3 but for $\tilde{p}_k$.

5 Metropolis-Hastings random walk algorithm for $(p,\tilde{p},\pi)$ with a normal proposal distribution. This block ensures that any part of the posterior support can be reached by the algorithm, which is required for MCMC convergence; see, for example, Tierney (1994) or Geweke (2005). The variance of the proposal distribution has to be small in this block to get any accepted draws.

6 Draw a candidate for π from a Dirichlet distribution proportional to the product of (4.22) and (3.19). Accept the candidate with probability 1 if it does not violate the model restrictions and reject otherwise.

7 Blocks for sampling beta mixture prior parameters $(z,\tilde{z},\alpha,m,s)$ are described in Norets and Tang (2010)

Since it is computationally expensive to verify the model restrictions, we combine block 6 with all other blocks except 5. We start the algorithm from a solution to the problem corresponding to extreme value distribution for the shocks (we know it satisfies the model restrictions). We check the correctness of the algorithm implementation by joint distribution tests; see Geweke (2004).

B. LINEAR PROGRAMMING ALGORITHMS

In this section it assumed that the per-period payoff is defined parametrically ($u_j = Z_j\theta$).

B.1. *Verifying that a vector of counterfactual CCPs belongs to the identified set*

To be specific let us describe the algorithm when the counterfactual does not change u. To verify the model restrictions (whether $\tilde{p} \in \mathcal{P}(p,G)$) we do the following: (i) express δ as a linear function of (θ,e) and $\tilde{\delta}$ as a linear function of $(\theta,\tilde{e})$ using (2.17), where the linear coefficients depend on (p,G) and $(\tilde{p},\tilde{G})$ respectively; (ii) substitute these linear functions into strict inequalities described by Corollary 5; (iii) the result of (i)–(ii) is a system of strict inequalities $Ax < b$, where $x = (\theta', e', \tilde{e}', \delta_m)'$ and matrix A and vector b are computed using $(p,\tilde{p},G,\tilde{G},\beta)$; (iv) solve the following linear programming problem: $\min_{x,t} t$ subject to $Ax - t \le b$. If the resulting optimal $t^*(p,G,\tilde{p})$ is non-positive then $\tilde{p} \in \mathcal{P}(p,G)$.

B.2. *Verifying that a vector of actual CCPs is consistent with model restrictions*

To verify the model restrictions from Theorem 1 we do the following: (i) express δ as a linear function of (θ,e) using (2.17), where the linear coefficients depend on (p,G); (ii) substitute this linear function into strict inequalities (2.14)–(2.16); (iii) the result of (i)–(ii) is a system of strict inequalities $Ax < b$, where $x = (\theta', e', \delta_m)'$ and matrix A and vector b are computed using (p,G,β); (iv) solve the following linear programming problem: $\min_{x,t} t$ subject to $Ax - t \le b$. If the resulting optimal $t^*(p,G)$ is non-positive then p is consistent with the model.

B.3. *Computing the identified set for per-period payoffs*

B.3.1. Algorithm I. This algorithm for recovering the identified set of θ, denoted by $\Theta_I(p,G)$, exploits the convexity of $\Theta_I(p,G)$ and the supporting hyperplane theorem, which implies that the boundary of $\Theta_I(p,G)$ can be fully characterized by a set of supporting hyperplanes. More specifically, to approximate $\Theta_I(p,G)$, we define a finite grid on normal vectors of hyperplanes. A normal vector of a hyperplane is a unit vector orthogonal to the hyperplane. The grid, $w_1,\dots,w_M$, is constructed using a grid on $d-1$ angles in a spherical coordinate system, where d is the dimension of θ.

For each w_i in the grid, we use a linear programing algorithm to find a pair of hyperplanes orthogonal to w_i that are tangent to the boundary of the identified set. Specifically, we solve two linear programing problems that respectively minimize and maximize the index $w_i'\theta$, subject to the constraint that θ is in the identified set with some additional small slack $\eta > 0$. More formally, we minimize and maximize $w_i'\theta$ subject to $A^\star y \le b^\star - \eta$, where $y = (\theta, e, \delta_m)$, and matrix $A^\star$ and vector $b^\star$ are such that $A^\star y < b^\star$ represent the system of equalities and inequalities from Corollary 4 characterizing the identified set (the equalities are used to substitute out δ in the strict inequalities). The slackness parameter η is introduced so that we can use weak inequalities $A^\star x \le b^\star - \eta$ in the constraints of the optimization problem (there would exist no solution under strict inequalities).

For each i let us denote the solutions to the minimization and maximization problems, respectively, by $\underline{y}_i^*$ and $\overline{y}_i^*$. The sub-vectors in $\underline{y}_i^*$ and $\overline{y}_i^*$ that correspond to θ are approximately on the boundary of $\Theta_I(p,G)$ along the direction w_i.

If the researcher is interested only in a sub-vector $\theta_1 \in \mathbb{R}^{d_1}$ of $\theta=(\theta_1,\theta_2)$ then the normal vector w_i is defined as a unit vector in $\mathbb{R}^{d_1}$ and the above algorithm is applied with the objective $w_i'\theta_1$ in the linear program. Note that when $d_1=1$, the dimension of the linear program does not change but the grid on the normal vectors consists only of 1 element.

B.3.2. Algorithm II. This algorithm performs grid search in the original parameter space. It partitions the parameter space into hypercubes (for example, grid rectangles in $\mathbb{R}^2$) and collects all hypercubes that contain at least one parameter value consistent with the actual CCPs.

The algorithm starts with a hypercube known to contain θ in the identified set, which is obtained from the solution of the linear program described in Section B.2.

Once an initial hypercube is located, we check whether the adjacent hypercubes have a non-empty intersection with the identified set. Specifically, when θ is in $\mathbb{R}^2$, we need to check the eight adjacent rectangles. To check a given rectangle we combine linear inequalities that define the rectangle with the linear inequalities from Section B.2 and check if they are jointly feasible. If yes, we include the rectangle into the approximation of the identified set. We continue until all the boundary rectangles of the identified set approximation have no unchecked adjacent rectangles.

This grid search algorithm has to solve more LP programs than Algorithm I described above. However, the linear programs in Algorithm II are numerically more stable in Matlab in the case of Rust's model. Therefore, Algorithm II is used to obtain results in Figures 1, 2, and 8. The results for the entry/exit model in Norets and Tang (2013) are obtained by Algorithm I.

Matlab solution of one linear program in Algorithms I or II applied to Rust's model takes about 1 second on a PC with Intel 2.7 GHz processor and 8 GB RAM.

C. PROOFS

Proof (Sufficiency in Lemma 1)

Suppose p satisfies (2.5). We will show that it is a vector of CCPs. Define vectors y_0 and y_1 as follows,

$$y_1=(I-\beta G^1)^{-1}\Big[u_1+\beta G^1\Big[\int_{F^{-1}(p|X)}^{\infty} sdF(s|X)-(I-\mathrm{diag}(p))F^{-1}(p|X))\Big]\Big] \tag{C.1}$$

$$y_0=(I-\beta G^0)^{-1}\Big[u_0+\beta G^0 big[\int_{F^{-1}(p|X)}^{\infty} sdF(s|X)+\mathrm{diag}(p)F^{-1}(p|X)\Big]\Big].$$

By (2.5) and definition of (y_0,y_1), we have $F^{-1}(p|X)=y_1-y_0$ and $p=\int_{-\infty}^{y_1-y_0} dF(s|X))$. Using these two equations we can get rid of p in (C.1),

$$y_1=(I-\beta G^1)^{-1}\Big[u_1+\beta G^1\Big[\int_{y_1-y_0}^{\infty} sdF(s|X)-\int_{y_1-y_0}^{\infty} dF(s|X)(y_1-y_0))\Big]\Big] \tag{C.2}$$

$$y_0=(I-\beta G^0)^{-1}\Big[u_0+\beta G^0\Big[\int_{(y_1-y_0)}^{\infty} sdF(s|X)+\int_{(y_1-y_0)}^{\infty} dF(s|X)(y_1-y_0)\Big]\Big].$$

From (C.2) one can reverse the steps leading from (2.2) to (2.4) in Section 2.2 to show that (y_0,y_1) have to satisfy the Bellman equation (2.2). Since the solution of the Bellman equation is unique, $(y_0,y_1)=(v_0,v_1)$ and $p=\int_{-\infty}^{y_1-y_0} dF(s|X)$ is a vector of CCPs. ‖

Proof (Corollary 3)

(i) The system of NK equations characterizing CCPs for N agent types is

$$\begin{bmatrix} M_U^1 u=[M_D^1,-M_E^1][d_1',e_1']' \\ \vdots \\ M_U^N u=[M_D^N,-M_E^N][d_N',e_N']' \end{bmatrix}, \tag{C.3}$$

where d_n and e_n are K-vectors with coordinates $d_{n,k}=F^{-1}(p_k^n|x=k)$, $e_{n,k}=\int_{d_{n,k}}^{\infty} sdF(s|x=k)$; and

$$M_U^n=[\,(I-\beta G^{1,n})^{-1},-(I-\beta G^{0,n})^{-1}\,]$$

$$M_D^n=(I-\beta G^{1,n})^{-1}[I-\mathrm{diag}(p^n)]+(I-\beta G^{0,n})^{-1}\mathrm{diag}(p^n)$$

$$M_E^n=(I-\beta G^{1,n})^{-1}\beta G^{1,n}-(I-\beta G^{0,n})^{-1}\beta G^{0,n}.$$

Since $(I-\beta G^{j,n})^{-1}=I+\beta G^{j,n}+\beta^2(G^{j,n})^2+\beta^3(G^{j,n})^3+...$, the sum of all columns in $(I-\beta G^{j,n})^{-1}$ must be proportional to $(1,1,\ldots,1)'$. It then follows that the maximum rank possible is $2K-1$ for the NK-by-$2K$ matrix of coefficients in front

of u. When the rank of $[(M_U^1)',\ldots,(M_U^N)']'$ is equal to r and $2K-r$ components of u are normalized to some values, the linear system in (C.3) has a unique solution for the rest of the components of u.

(ii) Multiplication of the matrix in (2.6) by

$$\begin{bmatrix} -I+\beta G^{1,1}, & 0_{K\times K} \\ 0_{K\times K}, & -I+\beta G^{0,2} \end{bmatrix} \text{ and } \begin{bmatrix} 0_{K\times K}, & -I+\beta G^{1,2} \\ I-\beta G^{0,1}, & 0_{K\times K} \end{bmatrix} \tag{C.4}$$

from the left and right correspondingly delivers the matrix in (2.7). Since matrices in (C.4) have full rank, the multiplication does not affect the rank and (2.7) is proved.

To prove (2.8) it suffices to show that the dimensions of the null spaces of the matrices in (2.6) and (2.8) are the same (the dimension of the matrix is equal to the sum of its rank and the dimension of its null space). Suppose $x_1,\ldots,x_l \in \mathbb{R}^K$ are a basis of the null space of the matrix in (2.8). Define $y_i=(I-\beta G^{1,1})(I-\beta G^{0,1})^{-1}x_i=(I-\beta G^{1,2})(I-\beta G^{0,2})^{-1}x_i$, $i=1,\ldots,l$, where the second equality follows from the definition of x_i. Vectors $(y_i',x_i')'$, $i=1,\ldots,l$, are independent since x_i's are independent. For any $(y',x')'$ in the null space of the matrix in (2.6), $y=(I-\beta G^{1,1})(I-\beta G^{0,1})^{-1}x=(I-\beta G^{1,2})(I-\beta G^{0,2})^{-1}x$ and, thus, x is in the null space of the matrix in (2.8) and $x=\sum_{i=1}^l \alpha_i x_i$ for some scalars α_i's. It follows that $y=\sum_{i=1}^l \alpha_i y_i$. Thus, $(y_i',x_i')'$, $i=1,\ldots,l$, form a basis of the null space of the matrix in (2.6) with $N=2$, and (2.8) is proved.

(iii) This part is an immediate implication of (2.8) since multiplication by a square matrix of full rank (for example, $I-\beta * G^{0,1}$) does not affect the rank.

‖

Proof (Lemma 2)

Construct a linear system of $2K$ equations in u by stacking (2.5) for p and (u,G,β,F) and (2.5) for $\tilde{p}$ and $(u,\tilde{G},\beta,F)$. The system can be simplified as

$$A_1u_1-A_0u_0=B[Q(p)',\kappa(p)']' \tag{C.5}$$

$$\tilde{A}_1u_1-\tilde{A}_0u_0=\tilde{B}[Q(\tilde{p})',\kappa(\tilde{p})']', \tag{C.6}$$

where $A_j=\left(I-\beta G^j\right)^{-1}$ and $\tilde{A}_j=\left(I-\beta \tilde{G}^j\right)^{-1}$ are K-by-K matrices constructed from the observed G and the counterfactual $\tilde{G}$, respectively; $B=[A_1,A_0-A_1]$ and $\tilde{B}=[\tilde{A}_1,\tilde{A}_0-\tilde{A}_1]$ are K-by-$2K$; and Q and κ are functions that map from $(0,1)^K$ to $\mathbb{R}^K$ with coordinates

$$Q_k(p)=F^{-1}(p_k|x_k),\ \kappa_k(p)=\int_{-\infty}^{Q_k(p)}(Q_k(p)-s)dF(s|x_k),\ k=1,\ldots,K.$$

The form of these two functions depend on the unobserved state distribution F. The vectors $(p,\tilde{p})$ denote observed and counterfactual CCPs respectively. Suppose we set $u_0=c$ while the truth is $u_0=u_0^*$. Given this assignment of u_0, the remaining K parameters in u_1 are recovered as

$$u_1=A_1^{-1}\left\{B[Q(p)',\kappa(p)']'+A_0c\right\}. \tag{C.7}$$

The counterfactual analysis then amounts to recovering the $\tilde{p}$ that satisfies

$$\tilde{B}[Q(\tilde{p})',\kappa^0(\tilde{p})']'=\tilde{A}_1A_1^{-1}B[Q(p)',\kappa(p)']'+\left(\tilde{A}_1A_1^{-1}A_0-\tilde{A}_0\right)c. \tag{C.8}$$

With F assumed known and p identified from the DGP, this implies that whenever $\left(\tilde{A}_1A_1^{-1}A_0-\tilde{A}_0\right)(c-u_0^*)\neq 0$, the choice of c has an impact on $\tilde{p}$ predicted as the solution to the equation above. ‖

Proof (Lemma 3)

In this case, the counterfactual CCPs, denoted $\hat{p}$, are characterized by

$$A_1(\alpha u_1+\Delta_1)-A_0(\alpha u_0+\Delta_0)=B[Q(\hat{p})',\kappa(\hat{p})']', \tag{C.9}$$

where A_j, Q, and κ are defined as in the proof of Lemma 2. Suppose the truth in DGP is $u_0=u_0^*$ but we set u_0 equal to some arbitrarily chosen vector c in order to estimate u_1. With u_1 recovered as in (C.7), identifying counterfactual CCPs amounts to finding $\hat{p}$ such that

$$B[Q(\hat{p})',\kappa(\hat{p})']'=\alpha B[Q(p)',\kappa(p)']'+A_1\Delta_1-A_0\Delta_0. \tag{C.10}$$

It then follows that the choice of c has no impact on the characterization of $\hat{p}$ in (C.10). ‖

Proof (Lemma 4)

Suppose $1>p_1>p_2>\cdots>p_L>0$, $\delta_1,\ldots,\delta_L$, $e_1,\ldots,e_L$, and F satisfy (2.10)–(2.13). Then,

$$e_1=\int_{\delta_1}^{\infty} sdF>\delta_1\int_{\delta_1}^{\infty} dF=\delta_1(1-p_1)$$

imply (2.14). Inequalities in (2.15) follow since

$$\frac{e_{i+1}-e_i}{p_i-p_{i+1}}=\frac{\int_{\delta_{i+1}}^{\delta_i} sdF}{F(\delta_i)-F(\delta_{i+1})}\in(\delta_{i+1},\delta_i).$$

for F satisfying (2.10). By (2.11),

$$e_k=\int_{\delta_k}^{\infty} sdF=-\int_{-\infty}^{\delta_k} sdF>-\delta_k p_k$$

and (2.16) follows.

To prove the other direction of the lemma, suppose (2.14)–(2.16) hold. Let us construct a particular density $f(s)>0$ that satisfies (2.10)–(2.13). For $s\in(-\infty,\delta_k]$ let $f(s)=c_k\exp(b_k s)$, where $b_k=p_k/(\delta_k p_k+e_k)$ and $c_k=b_k p_k\exp(-b_k\delta_k)$. For $s\in[\delta_1,\infty)$ let $f(s)=c_1\exp(b_1 s)$, where $b_1=-(1-p_1)/(e_1-(1-p_1)\delta_1)$ and $c_1=-b_1(1-p_1)\exp(-b_1\delta_1)$. For $s\in(\delta_i,\delta_{i-1})$ let $f(s)=h_{i1}1_{(\delta_i,r_i)}(s)+h_{i2}1_{(r_i,\delta_{i-1})}(s)$, where $r_i=(e_i-e_{i-1})/(p_{i-1}-p_i)$, $h_{i1}=[(p_{i-1}-p_i)\delta_{i-1}-(e_i-e_{i-1})]/[(r_i-\delta_i)(\delta_{i-1}-\delta_i)]$, and $h_{i2}=[(e_i-e_{i-1})-(p_{i-1}-p_i)\delta_i]/[(\delta_{i-1}-r_i)(\delta_{i-1}-\delta_i)]$. It is easy to verify by direct calculation that such f satisfies (2.10)–(2.13). ‖

Proof (Corollary 4)

Part (i) is an immediate implication of Theorem 1. To prove part (ii) suppose $u^1,u^2\in\mathcal{U}(p,G)$. By definition, there exist $(e^1,\delta^1,\delta^1_m)$ and $(e^2,\delta^2,\delta^2_m)\in\mathbb{R}^{2K+1}$ such that $(p,\delta^1,\delta^1_m,e^1,u^1)$ and $(p,\delta^2,\delta^2_m,e^2,u^2)$ satisfy conditions (a) and (b) that define $\mathcal{U}(p,G)$ in Corollary 4. Let $\delta^\alpha=\alpha\delta^1+(1-\alpha)\delta^2$ for a generic $\alpha\in(0,1)$, and likewise define $u^\alpha,e^\alpha,\delta^\alpha_m$ also as convex combinations. By convexity of U, $u^\alpha\in U$. With β,G,p fixed, the linear equalities and inequalities in conditions (a) and (b) hold for $(p,u^\alpha,\delta^\alpha,e^\alpha,\delta^\alpha_m)$. Thus $u^\alpha\in\mathcal{U}(p,G)$ for any $\alpha\in(0,1)$. ‖

Proof (Lemma 5)

To prove the lemma it suffices to show that the actual CCP p cannot belong to the identified set of counterfactual CCPs $\mathcal{P}(p,G)$ under the conditions of the lemma.

Suppose $p\in\mathcal{P}(p,G)$. Then $\tilde{\delta}=\delta$ and $\tilde{e}=e$. However, in this case, the condition assumed in the lemma implies that (2.17) cannot hold simultaneously for actual and counterfactual environments. This contradicts the supposition that $p\in\mathcal{P}(p,G)$. ‖

Proof (Lemma 6) If $\theta_1<0$, then $u(x,\epsilon,d)$ is non-increasing in x. Also, (G^0,G^1) are monotone non-increasing Markov transition matrices. Therefore, by a standard argument for value function monotonicity (Stokey and Lucas (1989)), v_1 and v_0 from (2.2) are non-increasing in x. Moreover, v_1 does not depend on x and v_0 is strictly decreasing in x because u_0 is strictly decreasing. Thus, $p(x)=F(v_1(x)-v_0(x))$ is strictly increasing in x. ‖

Proof (Lemma 7)

In Rust's model, at $x=1$ the future expected value functions are equal as the observed state transition probabilities are the same for $d=1$ and $d=0$ at $x=1$. Thus, the choice probability at $x=1$ is determined only by the per-period payoff functions and the distribution of ϵ. Therefore, if a counterfactual experiment involves changes only in the observed state transition probabilities π, then the coordinate of the CCP vector corresponding to x_1 does not change: $p_1=\tilde{p}_1$. ‖

Proof (Lemma 8) Under the conditions of the lemma, the characterization of $\{(p,\tilde{p},\pi):\tilde{p}\in\mathcal{P}(p,\pi)\}$ can be given in terms of the feasibility of a system of strict inequalities (see Appendix B.1). Since the inequalities are strict, they have to be satisfied in an open neighbourhood of the original feasible point. ‖

Lemma 9. *Any confidence set for $\mathcal{P}(p,G)$ can be represented as a superset of a set in the form* (3.21).

Proof To see this formally consider an arbitrary $100(1-\alpha)\%$ confidence set for $\mathcal{P}(p,G)$ denoted by $A^{\mathcal{P}}_{1-\alpha}(\omega)$, where ω denotes data and possibly randomization variables. Define $C^{p,G}_{1-\alpha}(\omega)=\{p',G':\mathcal{P}(p',G')\subset A^{\mathcal{P}}_{1-\alpha}(\omega)\}$. Denote the DGP values by (p_0,G_0). By definition of $C^{p,G}_{1-\alpha}(\omega)$, $[\omega:\mathcal{P}(p_0,G_0)\subset A^{\mathcal{P}}_{1-\alpha}(\omega)]\subset[\omega:(p_0,G_0)\in C^{p,G}_{1-\alpha}(\omega)]$. Thus, if $A^{\mathcal{P}}_{1-\alpha}(\omega)$ has $100(1-\alpha)\%$ coverage for $\mathcal{P}(p_0,G_0)$ then $C^{p,G}_{1-\alpha}(\omega)$ has at least $100(1-\alpha)\%$ coverage for (p_0,G_0). Also, $C^{\mathcal{P}}_{1-\alpha}(\omega)$ in (3.21) defined by these $C^{p,G}_{1-\alpha}(\omega)$ is a $100(1-\alpha)\%$ confidence set for $\mathcal{P}(p_0,G_0)$, and $C^{\mathcal{P}}_{1-\alpha}(\omega)\subset A^{\mathcal{P}}_{1-\alpha}(\omega),\forall\omega$. ‖

Acknowledgments. We are grateful to Hanming Fang, Han Hong, Rosa Matzkin, Ulrich Müller, Bernard Salanie, Frank Schorfheide, Chris Sims, Elie Tamer, Ken Wolpin, and participants in seminars at Copenhagen, Harvard-MIT,

Indiana, Penn, Penn State, Princeton, Wisconsin, Brown, ASSA 2011, Columbia, and Cowless 2011 Summer Conference for helpful discussions. We also thank the editor and anonymous referees for useful comments and suggestions.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

AGUIRREGABIRIA, V. (2005), "Nonparametric Identification of Behavioral Responses to Counterfactual Policy Interventions in Dynamic Discrete Decision Processes", *Economics Letters*, **87**, 393–398.
AGUIRREGABIRIA, V. (2010), "Another Look at the Identification of Dynamic Discrete Decision Processes: An Application to Retirement Behavior", *Journal of Business and Economic Statistics*, **28**, 201–218.
AGUIRREGABIRIA, V. and MIRA, P. (2002), "Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models", *Econometrica*, **70**, 1519–1543.
AGUIRREGABIRIA, V. and MIRA, P. (2007), "Sequential Estimation of Dynamic Discrete Games", *Econometrica*, **75**, 1–53.
AGUIRREGABIRIA, V. and MIRA, P. (2010), "Dynamic Discrete Choice Structural Models: A Survey", *Journal of Econometrics*, **156**, 38–67.
ANDREWS, D. W. K. and SOARES, G. (2010), "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection", *Econometrica*, **78**, 119–157.
ARCIDIACONO, P. and MILLER, R. A. (2011), "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity", *Econometrica*, **79**, 1823–1867.
BAJARI, P., BENKARD, L. and LEVIN, J. (2007), "Estimating Dynamic Models of Imperfect Competition", *Econometrica*, **75**, 1331–1370.
BERESTEANU, A. and MOLINARI, F. (2008), "Asymptotic Properties for a Class of Partially Identified Models", *Econometrica*, **76**, 763–814.
BHATTACHARYA, R. and MAJUMDAR, M. (1989), "Controlled Semi-Markov Models - The Discounted Case", *Journal of Statistical Planning and Inference*, **21**, 365–381.
BUGNI, F. A. (2010), "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set", *Econometrica*, **78**, 735–753.
CANAY, I. A. (2010), "El Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity", *Journal of Econometrics*, **156**, 408–425.
CHERNOZHUKOV, V., HONG, H. and TAMER, E. (2007), "Estimation and Confidence Regions for Parameter Sets in Econometric Models", *Econometrica*, **75**, 1243–1284.
CHIBURIS, R. (2009), "Approximately Most Powerful Tests for Moment Inequalities. Unpublished Manuscript, The University of Texas at Austin.
CHIB, S. and GREENBERG, E. (1995), "Understanding the Metropolis-Hastings Algorithm", *The American Statistician*, **49**, 327–335.
COLLARD-WEXLER, A. (2013), "Demand Fluctuations in the Ready-Mix Concrete Industry", *Econometrica*, **81**, 1003–1037.
COSSLETT, S. R. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model", *Econometrica*, **51**, 765–782.
DIEBOLT, J. and ROBERT, C. P. (1994), "Estimation of Finite Mixture Distributions Through Bayesian Sampling", *Journal of the Royal Statistical Society. Series B (Methodological)*, **56**, 363–375.
ECKSTEIN, Z. and WOLPIN, K. (1989), "The Specification and Estimation of Dynamic Stochastic Discrete Choice Models: A Survey", *Journal of Human Resources*, **24**, 562–598.
FANG, H. and WANG, Y. (2008), "Estimating Dynamic Discrete Choice Models with Hyperbolic Discounting, with an Application to Mammography Decisions. Unpublished Manuscript, University of Pennsylvania.
FREDERICK, S., LOEWENSTEIN, G. and O'DONOGHUE, T. (2002), "Time Discounting and Time Preference: A Critical Review", *Journal of Economic Literature*, **40**, 351–401.
FRUHWIRTH-SCHNATTER, S. (2006, 8), *Finite Mixture and Markov Switching Models (Springer Series in Statistics)* (1st edn), (Springer).
GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2003), *Bayesian Data Analysis*, 2nd edn (*Chapman & Hall/CRC Texts in Statistical Science*) New York: Chapman and Hall/CRC.
GEWEKE, J. (2004), "Getting it Right: Joint Distribution Tests of Posterior Simulators", *Journal of the American Statistical Association*, **99**, 799–804.
GEWEKE, J. (2005), *Contemporary Bayesian Econometrics and Statistics* (Wiley-Interscience).
HAN, A. (1987), "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator", *Journal of Econometrics*, **35**, 303–316.
HECKMAN, J. J. and NAVARRO, S. (2007), "Dynamic Discrete Choice and Dynamic Treatment Effects", *Journal of Econometrics*, **136**, 341–396.
HOTZ, J. and MILLER, R. (1993), "Conditional Choice Probabilities and the Estimation of Dynamic Models", *Review of Economic Studies*, **60**, 497–530.

HU, Y. and SHUM, M. (2012), "Nonparametric Identification of Dynamic Models with Unobserved State Variables", *Journal of Econometrics*, **171**, 32–44.
IMAI, S., JAIN, N. and CHING, A. (2009), "Bayesian Estimation of Dynamic Discrete Choice Models", *Econometrica*, **77**, 1865–1899.
IMBENS, G. W. and MANSKI, C. F. (2004), "Confidence Intervals for Partially Identified Parameters", *Econometrica*, **72**, 1845–1857.
KASAHARA, H. and SHIMOTSU, K. (2009), "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices", *Econometrica*, **77**, 135–175.
KEANE, M. P., TODD, P. E. and WOLPIN, K. I. (2011), *The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications*, Volume 4 of *Handbook of Labor Economics*, Chapter 4. (Elsevier). 331–461.
KEANE, M. and WOLPIN, K. (1994), "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence", *Review of Economics and Statistics*, **76**, 648–672.
KITAGAWA, T. (2011), "Estimation And Inference For Set-Identified Parameters Using Posterior Lower Probability", Unpublished manuscript, UCL.
KYDLAND, F. E. and PRESCOTT, E. C. (1982), "Time to Build and Aggregate Fluctuations", *Econometrica*, **50**, 1345–1370.
MAGNAC, T. and THESMAR, D. (2002), "Identifying Dynamic Discrete Decision Processes", *Econometrica*, **70**, 801–816.
MANSKI, C. F. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, **3**, 205–228.
MANSKI, C. F. (1999), *Identification Problems in the Social Sciences* (Harvard University Press).
MANSKI, C. F. (2003), *Partial Identification of Probability Distributions* (NY, USA: Springer).
MATZKIN, R. (1992), "Non-parametric and Distribution-Free Estimation of the Binary Choice and the Threshold Crossing Models", *Econometrica*, **60**, 239–270.
MATZKIN, R. L. (2007), "Nonparametric Identification. in Heckman, J. and Leamer, E. (eds) *Handbook of Econometrics*, Volume 6b of *Handbook of Econometrics*, Chapter 73, (Elsevier), 5307–5368.
MILLER, R. A. (1997), "Estimating Models of Dynamic Optimization with Microeconomic Data", in Pesaran, M. and Schmidt, P. (eds), *Handbook of Applied Econometrics*, Volume 2 of *Handbook of Econometrics*, Chapter 6, (Basil Blackwell) 246–299.
MOON, H. R. and SCHORFHEIDE, F. (2012), "Bayesian and Frequentist Inference in Partially Identified Models", *Econometrica*, **80**, 755–782.
NORETS, A. (2009), "Inference in Dynamic Discrete Choice Models with Serially Correlated Unobserved State Variables", *Econometrica*, **77**, 1665–1682.
NORETS, A. (2010), "Continuity and Differentiability of Expected Value Functions in Dynamic Discrete Choice Models", *Quantitative economics*, **1**.
NORETS, A. and TANG, X. (2010), "A Note on MCMC Estimation of a Finite Beta Mixture". Unpublished manuscript, University of Illinois at Urbana-Champaign and University of Pennsylvania.
NORETS, A. and TANG, X. (2013), "Supplementary Materials for Semiparametric Inference in Dynamic Binary Choice Models". Wen Appendix.
PAKES, A. (1994), "The Estimation of Dynamic Structural Models: Problems and Prospects", in Lafont, J. and Sims, C. (eds), *Advances in Econometrics: Proceedings of the 6th World Congress of the Econometric Society*.
PESENDORFER, M. and SCHMIDT-DENGLER, P. (2008, 07), "Asymptotic Least Squares Estimators for Dynamic Games", *Review of Economic Studies*, **75**, 901–928.
ROEHRIG, C. S. (1988), "Conditions for Identification in Nonparametric and Parametric Models", *Econometrica*, **56**, 433–447.
ROMANO, J. P. and SHAIKH, A. M. (2010), "Inference for the Identified Set in Partially Identified Econometric Models", *Econometrica*, **78**, 169–211.
ROSEN, A. M. (2008), "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities", *Journal of Econometrics*, **146**, 107–117.
ROUSSEAU, J. (2010), "Rates of Convergence for the Posterior Distributions of Mixtures of Betas and Adaptive Nonparametric Estimation of the Density", *The Annals of Statistics*, **38**, 146–180.
RUST, J. (1987), "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher", *Econometrica*, **55**, 999–1033.
RUST, J. (1994), "Structural Estimation of Markov Decision Processes", in Engle, R. and McFadden, D. (eds), *Handbook of Econometrics* (North Holland).
RUST, J. and PHELAN, C. (1997), "How Social Security and Medicare Affect Retirement Behavior in a World of Incomplete Markets", *Econometrica*, **65**, 781–831.
SIMS, C. (2006), "On an Example of Larry Wasserman". Unpublished manuscript, Princeton University.
STOKEY, N. and LUCAS, R. (1989), *Recursive Methods in Economic Dynamics* (Harvard University Press).
STOYE, J. (2009), "More on Confidence Intervals for Partially Identified Parameters", *Econometrica*, **77**, 1299–1315.
TIERNEY, L. (1994), "Markov Chains for Exploring Posterior Distributions", *The Annals of Statistics*, **22**, 1758–1762.
VAN DER VAART, A. (1998), *Asymptotic Statistics* (Cambridge University Press).