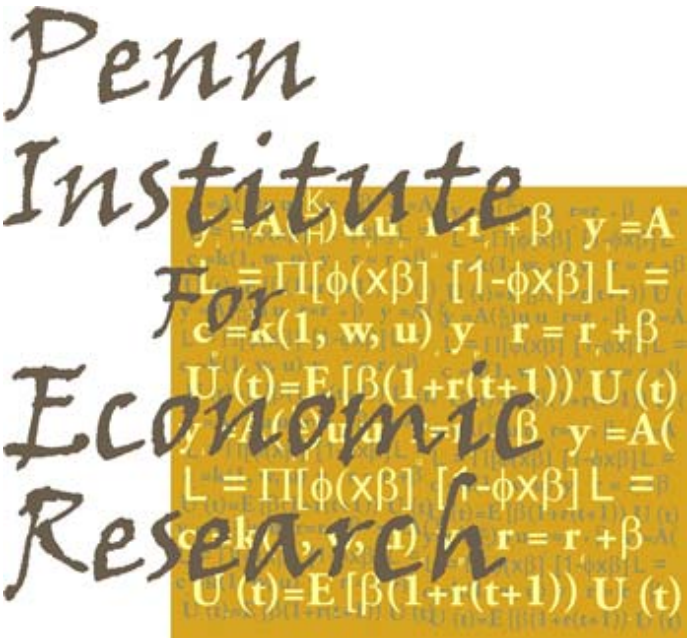

Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper 12-017

"Semiparametric Inference in Dynamic Binary Choice Models
Second Version"

by

Andriy Norets and Xun Tang

http://ssrn.com/abstract=2046145

[1] [2]

Semiparametric Inference in Dynamic Binary Choice Models

Andriy Norets and Xun Tang

April 25, 2012

SEMIPARAMETRIC INFERENCE IN DYNAMIC BINARY CHOICE MODELS

Andriy Norets and Xun Tang

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 for 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.

JEL code: C14, C15, C25

Keywords: Dynamic discrete choice models, Markov decision processes, dynamic games, semiparametric inference, identification, Bayesian estimation, MCMC.

1. INTRODUCTION

1.1. *Background*

A dynamic discrete choice model is a dynamic program with discrete controls. These models have been used widely in various fields of economics, including labor eco-

[1]First version: April 14, 2010, current version: April 25, 2012.

[2]We are grateful to Hanming Fang, Han Hong, Bo Honore, Rosa Matzkin, Ulrich Muller, 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.

Department of Economics, Princeton University; anorets@princeton.edu

Department of Economics, University of Pennsylvania; xuntang@sas.upenn.edu

nomics, health economics, and industrial organization. See **?**, **?**, **?**, **?** and **?** 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 paper is to provide a robust method for inference about counterfactuals.

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 **?**). 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 unobservables are mostly used in applications of the interpolation simulation method of **?** and the Bayesian estimation methods of **?** and **?**; in the methods of **?** and **?**, extreme value independently identically distributed (i.i.d.) unobservables are often assumed to alleviate the computational burden of solving and estimating the dynamic program.

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 **?**). Therefore, it is desirable to provide estimation methods that employ restrictions implied by economic theory such as monotonicity and concavity and avoid strong distributional assumptions on unobservables. This has been done for static binary choice models; see, for example, **?**, **?**, **?**, and **?**. 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 unobservables.

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

We consider models with conditionally independent and additively separable unobserved states as in **?** and **?**. 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 nonparametrically 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 distribution of the unobserved states. **?** showed that per-period payoffs are nonparametrically not identified

even under known distribution of unobservables. First, we show that exogenous variation in transitions for the observed states suffices for nonparametric point identification of the per-period payoffs under known distribution of unobservables. Second, we derive conditions under which normalizations on per-period payoffs, which are sufficient for point identification, affect and do not affect counterfactual predictions under known distribution of unobservables. Third, we show that when the distribution of the unobservables 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 nonparametrically not point-identified under known distribution of unobservables, the identified set for the counterfactual CCPs under unknown distribution of unobservables 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 unknown distribution of unobservables, 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. Beyond philosophical considerations, the Bayesian approach has the following advantages. First, even under unknown distri-

bution 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 estimation. The MCMC output from the estimation procedure can be used for construction of valid 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 (**?**). 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 and the sampling uncertainty can be large relative to the

identified set. Thus, in our example parametric assumptions about the distribution of the unobserved states can 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 unobservables can affect estimation results.

The rest of the paper is organized as follows. Section 2 describes identification results for the per-period payoffs and the counterfactual CCPs under known and unknown distribution of unobserved states. In Section 3, we discuss the Bayesian approach to inference, its relation to other approaches, and ways to conduct valid frequentist inference. The MCMC estimation algorithm and prior specification are discussed in the context of the application in Section 4. In addition to estimation results, this section presents identified sets for the per-period payoffs, the time discount factor, and the counterfactual CCPs for **?** model. Proofs and algorithm implementation details are delegated to appendices.

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

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

where $d_t \in D = \{0, 1\}$ is the control variable, $x_t \in X$ are state variables observed by the econometrician, $\epsilon_t = (\epsilon_{0t}, \epsilon_{1t}) \in \mathbb{R}^2$ are state variables unobserved by the econometrician, β is the time discount factor, and $u(x_t, \epsilon_t, d_t)$ is the per-period payoff.

The state variables evolve according to a controlled first-order Markov process. Under mild regularity conditions (see **?**) that are satisfied under the assumptions we make below, the optimal lifetime utility of the agent has a recursive representation:

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

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* $u(x_t, \epsilon_t, d_t) = u_{ji} + \epsilon_j$ *when* $(x_t, d_t) = (i, j)$; ϵ_j *is integrable and* $E(\epsilon_j | x) = 0$ *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* $F_{\epsilon|x}$.

ASSUMPTION 4 *The distribution of* $\epsilon_0 - \epsilon_1$ *given any* x *has a positive density on* $\mathbb{R}^1$ *with respect to (w.r.t.) the Lebesgue measure.*

The assumptions 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 **?**). Second, without Assumption 3 it is not clear whether the expected value functions are differentiable with respect to parameters (**?**). Finally, the assumption is also very convenient for computationally feasible classical (**?**, **?**) and Bayesian (**?**) estimation of parametrically specified models.

Except for Section 4.3, 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 (**?**) or studies

estimating time discount rates from experimental data (see Section 6 in **?** for an extensive list of references).

2.2. *Identification under known distributions of unobservables*

In this subsection, we characterize the CCPs assuming that the unobserved state distribution is known to econometricians (Lemma 1). Later in Section 2.3, we extend these results to obtain a characterization of the CCPs under unknown $F_{\epsilon|X}$. After presenting Lemma 1, we discuss its key implications for the identification of the per-period payoffs. While some of these implications are not exploited in our semiparametric inference procedure, they do shed light on key questions in structural estimation such as the identifying power of exogenous variation in the transition probabilities for the observed states and the choice of appropriate payoff normalizations for inferring counterfactuals. To the best of our knowledge, these results have not been reported in the previous literature. Readers who are only interested in semiparametric inference for the counterfactual CCPs with unknown $F_{\epsilon|X}$ may skip the remarks and corollaries 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 (1) can be rewritten in vector notation as follows,

$$
\begin{aligned}
(2) \qquad v_0 &= u_0 + \beta G^0 \int \max\{v_0 + \epsilon_0, v_1 + \epsilon_1\} dF_{\epsilon|x}(\epsilon|X) \\
v_1 &= u_1 + \beta G^1 \int \max\{v_0 + \epsilon_0, v_1 + \epsilon_1\} dF_{\epsilon|x}(\epsilon|X),
\end{aligned}
$$

where $u_j = (u_{j1}, \ldots, u_{jK})'$ is a vector of stacked deterministic parts of the per-period payoffs with $d_t = j$; $v_j = (v_{j1}, \ldots, v_{jK})'$ is a vector of stacked deterministic parts of the alternative specific lifetime utilities $v_{ji} = u_{ji} + \beta E\{V(x_{t+1}, \epsilon_{t+1})|x_t = i, d_t = j\}$; and $G^j = [G^j_{ki}]$ is the Markov transition matrix for the observed states conditional on $d_t = j$. 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 $f(x)$ 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,

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

Let us rewrite the Bellman equations (2) in a form convenient for analyzing identification,

$$\begin{aligned} (3) \qquad v_0 &= u_0 + \beta G^0[v_0 + \int \max\{0, v_1 - v_0 - (\epsilon_0 - \epsilon_1)\} dF_{\epsilon|x}(\epsilon|X)] \\ v_1 &= u_1 + \beta G^1[v_1 + \int \max\{0, \epsilon_0 - \epsilon_1 - (v_1 - v_0)\} dF_{\epsilon|x}(\epsilon|X)], \end{aligned}$$

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, (3) implies

$$\begin{aligned} (4) \qquad v_1 - v_0 = {} & (I - \beta G^1)^{-1}[u_1 + \beta G^1 \int_{v_1 - v_0}^{\infty} (s - (v_1 - v_0))\} dF_{\Delta\epsilon|x}(s|X)] \\ & - (I - \beta G^0)^{-1}[u_0 + \beta G^0 \int_{-\infty}^{v_1 - v_0} (v_1 - v_0 - s)\} dF_{\Delta\epsilon|x}(s|X)]. \end{aligned}$$

Let $p_i = Pr(d_t = 1|x_t = i)$ denote the CCP implied by the model. Since $d_t = 1$ at $x_t = i$ when $v_{1i} - v_{0i} \geq \Delta\epsilon_i$, we have $p = (p_1, \ldots, p_K)' = F_{\Delta\epsilon|x}(v_1 - v_0|X)$. In the following lemma, we give necessary and sufficient conditions for some p to be the CCP given certain structural parameters. Let us denote $u = (u_1, u_0)$ and $G = (G^1, G^0)$.

LEMMA 1 *A vector p is the CCP for a DBCM given $(u, G, \beta, F_{\Delta\epsilon|x})$ if and only if*

$$\begin{aligned} (5) \qquad & F^{-1}_{\Delta\epsilon|x}(p|X) = \\ & (I - \beta G^1)^{-1}\left[u_1 + \beta G^1\big[\int_{F^{-1}_{\Delta\epsilon|x}(p|X)}^{\infty} s dF_{\Delta\epsilon|x}(s|X) - (I - diag(p))F^{-1}_{\Delta\epsilon|x}(p|X)\big]\right] \\ & - (I - \beta G^0)^{-1}\left[u_0 + \beta G^0\big[\int_{F^{-1}_{\Delta\epsilon|x}(p|X)}^{\infty} s dF_{\Delta\epsilon|x}(s|X) + diag(p)F^{-1}_{\Delta\epsilon|x}(p|X)\big]\right] \end{aligned}$$

The necessity follows immediately from (4) by substituting in $v_1 - v_0 = F^{-1}_{\Delta\epsilon|X}(p)$ and the zero expectation for ϵ_j (and $\Delta\epsilon$) given X. The proof of sufficiency is given in Appendix B. The system of equations (5) is convenient for analyzing 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 $(u, G, \beta, F_{\Delta\epsilon|x})$, 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, **?** were the first to provide a formal analysis of identification in the model we consider here. They used related but different system of equations. Their system involve 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. **?** derived a system similar to (5) as necessary conditions for a vector p to be the CCP in single-agent models. He used it to identify choice probabilities under counterfactual changes in u_j when $F_{\Delta\epsilon|X}$ is assumed to be known. **?** 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 **?** and **?** for dynamic discrete choice games. **?** analyzed identification while assuming that an outcome variable in each period is observable. **?** studied a model in which a choice probability is a finite mixture of unobservable component CCPs, which are also conditional on unobserved heterogeneity. **?** 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 analyze identification of the per-period payoffs.

The following remarks and lemmas summarize the implications of Lemma 1 for identification when $F_{\Delta\epsilon|X}$ is known. In the analysis of the identification of u, the vector of

CCP p and the observed state transition probabilities G are considered to be known and fixed. For now, we also assume that β is known.

REMARK 1 *Because the number of equations in (5), K, is smaller than the number of unknowns, $2K$, u cannot be jointly identified without further restrictions. This was first noted by* **?**. *In comparison, here we note that (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 (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. However, the following corollaries to Lemma 1 demonstrate how shape restrictions can lead to set and point identification.

COROLLARY 1 *Suppose the model is correctly specified, and 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 (a) per-period payoffs are linear in parameters, $U = \{u : u_j = Z_j\theta$ for $j = 0, 1\}$, where Z_j is a known $K \times d$ matrix and d is the dimension of θ; and (b) the model is correctly specified. 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 without invoking parametric restrictions. 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 **?**.

With $F_{\Delta\epsilon|X}$ 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 B.

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

$$\begin{bmatrix} (I - \beta G^{1,1})^{-1}, -(I - \beta G^{0,1})^{-1} \\ \vdots \\ (I - \beta G^{1,N})^{-1}, -(I - \beta G^{0,N})^{-1} \end{bmatrix}$$

For point-identification of the per-period payoffs (up to a location normalization), we need the highest possible rank for the matrix in Corollary 3, $r = 2K - 1$. This is a mild restriction since even for $N = 2$ only in very special cases we can expect $r < 2K - 1$. For example, if one generates $G^{i,j}$ from a continuous distribution then $r = 2K - 1$ with probability 1.

Some earlier papers have used weaker forms of exclusion restrictions to identify various features of DBCMs. **?** 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. **?** 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 current paper is slightly more restrictive but leads to completely non-parametric identification of per-period payoffs.

2.2.2. *Identification of counterfactuals*

Structural models are useful in analysis of counterfactual changes in structural parameters. Lemma 1 can be used to set up a framework for analyzing 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 $\tilde{u} = (\tilde{u}_0, \tilde{u}_1)$ and $\tilde{G} = (\tilde{G}^0, \tilde{G}^1)$ while the distribution of the unobserved states $F_{\Delta\epsilon|X}$ remains the same. The counterfactual transition probabilities and per-period payoffs may either be assigned known numerical values or be known functions of the primitives G and u in

the data-generating process (DGP). For example, the transition probabilities and the per-period payoff of choosing alternative 0 could be unchanged by the counterfactual, $\tilde{G} = G$ and $\tilde{u_0} = u_0$, and the per-period payoff of choosing alternative 1 can go up by 10%, $\tilde{u}_1 = 1.1 \cdot u_1$.

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. Examples of counterfactual changes in G also abound: changes in social security pension rules, e.g., changes in how pension income depends on retirement age; 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; and changes in bankruptcy laws in mortgage default models.

Let p denote the actual CCPs in the DGP given by (u, G, β, F). Define a system of K equations similar to (5) that characterizes the CCPs in the counterfactual context $(\tilde{G}^0, \tilde{G}^1, \tilde{u}_0, \tilde{u}_1, \beta, F)$,

$$
\begin{aligned}
(6) \qquad & F^{-1}_{\Delta\epsilon|x}(\tilde{p}|X) = \\
& (I - \beta\tilde{G}^1)^{-1}\left[\tilde{u}_1 + \beta\tilde{G}^1\big[\int_{F^{-1}_{\Delta\epsilon|x}(\tilde{p}|X)}^{\infty} s dF_{\Delta\epsilon|x}(s|X) - (I - diag(\tilde{p}))F^{-1}_{\Delta\epsilon|x}(\tilde{p}|X))\big]\right] \\
& - (I - \beta\tilde{G}^0)^{-1}\left[\tilde{u}_0 + \beta\tilde{G}^0\big[\int_{F^{-1}_{\Delta\epsilon|x}(\tilde{p}|X)}^{\infty} s dF_{\Delta\epsilon|x}(s|X) + diag(\tilde{p})F^{-1}_{\Delta\epsilon|x}(\tilde{p}|X)\big]\right],
\end{aligned}
$$

where $\tilde{p}$ is the vector of CCP implied by the counterfactual primitives. The identified set of the counterfactual CCPs consists of all $\tilde{p}$ that render the equations in (5) and (6) solvable in u.

REMARK 2 *Even if u is not point identified the set of counterfactual CCPs that can be rationalized by the model is restricted for certain counterfactuals. For example, consider a special case of counterfactual analysis, in which $\tilde{u} = u$ 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 $F_{\Delta\epsilon|X}$ is continuous, which implies uniform continuity of p as a function of (u, G) (the continuity follows by standard arguments, see, for example, ?, and 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, are 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 counterfactual contexts where G is changed to $\tilde{G}$. 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 CCP based on $u_0 = u_0^*$ unless*

$$\left[(I - \beta G^1)(I - \beta G^0)^{-1} - (I - \beta \tilde{G}^1)(I - \beta \tilde{G}^0)^{-1}\right](c - u_0^*) = 0. \tag{7}$$

By construction, the rank of the matrix in (7) cannot be higher than $K - 1$ (see the proof of Corollary 3). It is smaller than $K - 1$ only in special cases. 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 (7) multiplying $(c-u_0^*)$ is equal to $K-1$. Under this rank condition, (7) 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 where u is changed to $\tilde{u} = \alpha u + (\Delta_1, \Delta_0)$ (where Δ_k are K-vectors and α is a scalar) while G is unchanged. 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 u_0 on various types of counterfactual analyses. Proofs of these lemmas are included in Appendix B.

2.3. *Identification under unknown distributions of unobservables*

We now extend Lemma 1 to characterizes the CCPs when the distribution of the unobserved state is unknown to the econometrician. Hereafter, we maintain the following assumption.

ASSUMPTION 5 *The distribution of $\Delta\epsilon$ is independent from x.*[1]

Let us denote the distribution of $\Delta\epsilon$ by F. Equations in (5) suggest that the CCPs depend on F only through a finite number of quantiles $(F^{-1}(p_k))$ and corresponding integrals $\int_{F^{-1}(p_k)}^{+\infty} s dF$. Lemma 4 below characterizes the relations between such

[1] 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.

quantiles and corresponding truncated integrals for any 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 labeled so that $1 > p_1 > p_2 > ... > p_L > 0$ without loss of generality, there exists a distribution F such that*

(8) *F has a density $f > 0$ on R w.r.t. the Lebesgue measure,*

(9) $$\int s dF = 0$$

(10) $$F^{-1}(p_i) = \delta_i,\ i \in \{1, \ldots, L\}$$

(11) $$e_i = \int_{\delta_i}^{\infty} s dF,\ i \in \{1, \ldots, L\}$$

if and only if

(12) $$\frac{e_1}{1 - p_1} > \delta_1 > \cdots$$

(13) $$\cdots > \delta_i > \frac{e_{i+1} - e_i}{p_i - p_{i+1}} > \delta_{i+1} > \cdots$$

(14) $$\cdots > \delta_L > -\frac{e_L}{p_L}.$$

Conditions in the lemma are easy to understand geometrically. For example, condition (13) 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),

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

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

THEOREM 1 *For a given triple (β, u, G), a $p \in (0,1)^K$ is the CCP for some F*

satisfying (8)-(9) *if and only if there exist* $e \in \mathbb{R}^K$ *and* $\delta \in \mathbb{R}^K$ *such that: (i)*

$$\delta = (I - \beta G^1)^{-1}\Big[u_1 + \beta G^1\big[e - (I - diag(p))\delta\big]\Big] - (I - \beta G^0)^{-1}\Big[u_0 + \beta G^0\big[e + diag(p)\delta\big]\Big]; \tag{15}$$

(ii) After relabeling 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* (12)-(14)*, and* $e_i = e_j$*,* $\delta_i = \delta_j$ *whenever* $p_i = p_j$*.*

For example, suppose $K = 6$, $p = (p_1, p_2, \ldots, p_6) = (\frac{1}{3}, \frac{2}{5}, \frac{3}{4}, \frac{1}{3}, \frac{1}{10}, \frac{2}{5})$, and let $\delta = (\delta_1, \delta_2, \ldots, \delta_6)$ and $e = (e_1, e_2, \ldots, e_6)$. 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, (15) implies that the scale of F can be normalized without a loss of generality as long as U is a linear cone. Let $\delta_m = F^{-1}(p_m)$ denote the unrestricted median of F. A convenient normalization is to fix the value of

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

as in the logistic distribution (the location of F is normalized in Assumption 2 and equation (9)).

2.3.1. *Identified set for per-period payoffs*

Theorem 1 can be used to characterize the identified set of per-period payoffs. By definition, such a set consists of all $u \in U$ that are consistent with the known discount factor β, (G, p) identified by the data, and some F satisfying (8)-(9). The following corollary also takes into account the scale normalization of F described in the previous subsection.

COROLLARY 4 *The identified set of per-period payoffs, $\mathcal{U}(p,G)$, consists of all $u \in U$ for which there exists $(\delta, e, \delta_m) \in \mathbb{R}^{2K+1}$ such that (i) (p, δ, e, u) solves (15) and (ii) after relabeling coordinates of $p^* = (p, p_m)$ and the corresponding coordinates in $e^* = (e, e_m)$, $\delta^* = (\delta, \delta_m)$ so that $1 > p_1^* \geq p_2^* \geq \cdots \geq p_{K+1}^* > 0$, the unique (strictly-ordered) coordinates in p^* and the corresponding coordinates in δ^*, e^* satisfy (12)-(14), and $e_i^* = e_j^*$, $\delta_i^* = \delta_j^*$ whenever $p_i^* = p_j^*$.*

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 is convenient for computing the identified sets as it implies the linearity of the equalities and inequalities in the unknowns.[2] However, it is easier to get insight into what determines the size of the identified set under a scale normalization that bounds the truncated integrals: $\int_0^\infty s dF(s) = \log(2)$, which implies $0 < e_i \leq \log(2)$ for all i. For simplicity, suppose that u_0 does not depend on x and normalized to $(0, \ldots, 0)$ and u_1 is unrestricted. Then, it can be deduced from (15) that the size of the identified set for u_1 is determined to a large extent by the range of values that δ can take. Next, let us examine how the range of δ depends on p. Lemma 4 with $L = 1$ implies that for every i, $e_i/(1-p_i) > \delta_i > -e_i/p_i$. This inequality suggests that as $p_i \to 1$ an upper bound for δ_i converges to infinity and as $p_i \to 0$ a lower bound for δ_i converges to minus infinity. Applying the same reasoning to all the inequalities in (12)-(14) and selecting e so that $e_1 < e_2 < \cdots < e_i$, one can deduce that the least upper bound for δ_i is large if $(1 - p_1, p_1 - p_2, \ldots, p_{i-1} - p_i)$ are small. Similarly, the least upper bound for δ_i is small if $(p_i - p_{i+1}, \ldots, p_{K-1} - p_K, p_K)$ are small. Of course, the identified set for u is also determined by G and optional parametric or shape restrictions U. Nevertheless,

[2] If the restrictions in U are also linear (in general, they do not have to be linear) then a linear programming algorithm similar to the one described in Appendix A can be used to compute the identified set.

the size and the spacing of the coordinates of the CCPs do seem to contain a lot of information about the identified set for u, especially in extreme cases. For example, if all the components of p are clustered near 0 (or 1) then the range of δ solving (12)-(14) is very large and the identified set for u can potentially be very large. Conversely, if the coordinates of p are clustered around 0.5 then the identified set should not be too big. The identified sets we compute in our application (Section 4.2) confirm these arguments.

The definition of the identified set in Corollary 4 assumes 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 4.3).

2.3.2. *Identified set for counterfactual CCPs*

The following corollary uses Theorem 1 to characterize the identified set for CCPs under a counterfactual change in model primitives (u, G). The notation and examples of counterfactuals are described in Section 2.2.2. It is important to note that the distribution of unobserved states, F, is assumed to be unchanged by the counterfactual.

COROLLARY 5 *For given (p, G) identified by the data and counterfactual primitives $(\tilde{u}, \tilde{G})$, the identified set for counterfactual CCPs is given by*
$\mathcal{P}(p, G) = \{\tilde{p} : \exists u \in U$ *and* $(\delta, e, \tilde{\delta}, \tilde{e}, \delta_m) \in \mathbb{R}^{4K+1}$ *such that (i) (p, δ, e, u) satisfy (15), and $(\tilde{p}, \tilde{\delta}, \tilde{e}, \tilde{u})$ satisfy*

$$\tilde{\delta} = (I - \beta\tilde{G}^1)^{-1}\left[\tilde{u}_1 + \beta\tilde{G}^1\left[\tilde{e} - (I - diag(\tilde{p}))\tilde{\delta}\right]\right] - (I - \beta\tilde{G}^0)^{-1}\left[\tilde{u}_0 + \beta\tilde{G}^0\left[\tilde{e} + diag(\tilde{p})\tilde{\delta}\right]\right]; \tag{16}$$

and (ii) after relabeling coordinates of $p^* = (p, \tilde{p}, p_m)$ and their corresponding coordinates in $e^* = (e, \tilde{e}, e_m)$, $\delta^* = (\delta, \tilde{\delta}, \delta_m)$ so that $1 > p_1^* \geq p_2^* \geq \cdots \geq p_{2K+1}^* > 0$,

the unique (strictly-ordered) components in p^* *and their corresponding coordinates in* δ^*, e^* *satisfy* (12)-(14)*, and* $e_i^* = e_j^*$*,* $\delta_i^* = \delta_j^*$ *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.

The identified set for $\tilde{p}$ usually has a non-empty interior even if u is known. To see this suppose that p and $\tilde{p}$ implied by some F and $u \in U$ have no common coordinates $(\tilde{p}_i \neq p_j, \forall i, j)$. If for some $\tilde{p}'$, $\tilde{p}' - \tilde{p}$ is sufficiently small then the equalities in (16) can be satisfied for $\tilde{p}'$ with $\tilde{e}$ unchanged and $\tilde{\delta}$ changed by a small amount. Because all the inequalities in (12)-(14) are strict they still are satisfied if $\tilde{p}$ and $\tilde{\delta}$ are changed by sufficiently small amounts. Thus, $\tilde{p}' \in \mathcal{P}(p, G)$ if $\tilde{p}' - \tilde{p}$ is sufficiently small. In some special cases, the coordinates of p and $\tilde{p}$ can be the same for any F. For example, in Rust's model considered in Section 4, $\tilde{p}_K = p_K$ under a counterfactual change in G. Such coordinates of $\tilde{p}$ are of course point identified.

Given the lack of nonparametric identification for u even under 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 unknown F. However, as the following lemma demonstrates even without restrictions 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. If $(\tilde{u}, \tilde{G})$ are fixed or $\tilde{u}$ is linear in u and $\tilde{G}$ is fixed then the equalities and

inequalities are linear and their feasibility can be verified by a linear programming algorithm described in Appendix A. As we demonstrate in the application (Section 4.7), this linear programming algorithm can be combined with MCMC to estimate the identified set.

3. INFERENCE

3.1. *Frequentist inference under partial identification*

The identification results from the previous section demonstrate that per-period payoffs, u, and counterfactual CCPs, $\tilde{p}$, are only partially identified when the distribution of unobservables is unknown. There is a growing literature in econometrics on inference for partially identified parameters. See for example, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, and **?**.

In principle, it seems possible to apply a criterion function approach of **?** to construct confidence sets for the identified sets of $\tilde{p}$ and parameters of u. Let us denote the parameter of interest ($\tilde{p}$ or parameters of u) by θ and the corresponding identified set by Θ_I. Using characterizations of Θ_I in Corollaries 4 or 5, we can define a criterion function $Q(p, G, \theta)$ that is minimized if and only if θ belongs to Θ_I.[3] The identified set Θ_I can be estimated by a contour set $\hat{\Theta}_I = \{\theta : a_n Q(\hat{p}, \hat{G}, \theta) < c_n\}$, where $\hat{p}$ and $\hat{G}$ are estimators of p and G, a_n is a normalizing sequence, and c_n is a possibly data dependent sequence. An asymptotically valid confidence set of level α for Θ_I is given by $\{\theta : a_n Q(\hat{p}, \hat{G}, \theta) < c_\alpha\}$, where critical value c_α is αth quantile of $a_n \sup_{\theta \in \Theta_I} Q(\hat{p}, \hat{G}, \theta)$. This critical value can be estimated by subsampling or bootstrap procedures.

When the problem is very high-dimensional computation of an estimate of the identified set $\hat{\Theta}_I$ and, especially, repeated maximization of $Q(\hat{p}, \hat{G}, \theta)$ subject to $\theta \in \hat{\Theta}_I$ needed for the bootstrap or subsampling procedure seems to be computationally infeasible. Implementations of **?** and related approaches in the literature only involve low (mostly one or two) dimensional problems. In most applications of DBCM the

[3] See the end of Appendix A for one possible definition of $Q(p, G, \theta)$.

dimension of $\tilde{p}$ is expected to be very high. Even if the object of interest is not $\tilde{p}$ itself but a low dimensional function of $\tilde{p}$ the computational challenges remain the same. Therefore, it is essential to develop an inference procedure that can handle high-dimensional problems.

3.2. *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 review standard maximum likelihood estimation (MLE) of the model when the unobserved state distribution F is assumed to be known. We then give a detailed description of our Bayesian procedure when F is unknown. The following subsections describe the frequentist properties of the procedure and an extension that incorporates additional restrictions on the unknown F.

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 is given by

$$P\left(\{d_1^i, x_2^i, d_2^i, \ldots, x_{T_i}^i, d_{T_i}^i\}_{i=1}^n | p, G, \{x_1^i\}_{i=1}^n\right) = \prod_{k=1}^K p_k^{n_k^1}(1-p_k)^{n_k^0} \cdot \prod_{k,l=1}^K (G_{kl}^1)^{\nu_{kl}^1}(G_{kl}^0)^{\nu_{kl}^0}, \tag{17}$$

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 **?** model that we use for experiments in Section 4. An alternative that might be appropriate in other applications is to assume that the process for (x_t, d_t) is stationary and combine (17) with the implied stationary distribution for x_1^i, which would be a function of (p, G).

In a standard MLE procedure for DBCMs (**?**, **?**), (u, G, F) are parameterized. After solving for value functions in the model, one can replace the CCPs p in the likelihood in (17) with functions of the parameters. The parameters are then estimated by the MLE. 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 (17). 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 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 (17). 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 3.4).

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)). Then, credible and confidence sets for $\mathcal{U}(p, G)$ can be constructed after estimation of (p, G) (u can be treated as a nuisance parameter at the estimation stage). We illustrate the latter approach in Section 4.8.

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. 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.

3.3. *Relationship between frequentist and Bayesian approaches*

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 $(p, G, \tilde{p})$. By the Bernstein-von Mises theorem, Bayesian credible sets for point-identified parameters (p, G) are asymptotically equivalent to the corresponding confidence sets obtained from the MLE for (p, G) (assuming the prior density is positive and continuous in an open neighborhood of the data-generating values of (p, G), see, for example, Chapter 10 in **?**).

For partially identified parameters the Bernstein-von Mises theorem does not hold.

? 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 $\tilde{p}$ is likely to include 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,

$$B^{\mathcal{P}}_{1-\alpha} = \bigcup_{(p',G') \in B^{p,G}_{1-\alpha}} \mathcal{P}(p', G') \tag{18}$$

is a $100(1-\alpha)\%$ credible set for $\mathcal{P}(p, G)$. In settings with multiple prior distributions for partially identified parameters considered in **?**, sets in (18) have posterior lower probability at least $1-\alpha$.

If $B^{p,G}_{1-\alpha}$ has $100(1-\alpha)\%$ frequentist coverage then the set in (18) also has $100(1-\alpha)\%$ frequentist coverage. Not all credible sets have the frequentist coverage property even when the Bernstein-von Mises theorem holds. For example, a credible set constructed to exclude a ball around a fixed parameter value has zero coverage at that parameter value. Examples of credible sets that do have the coverage property include one-

sided, highest posterior density, and equal tail probability credible sets. Credible sets $B^{p,G}_{1-\alpha}$ that minimize the volume of $B^{\mathcal{P}}_{1-\alpha}$ in (18) in general do not have the coverage property but can have it under certain conditions on the shape of the identified set (see Proposition 5.2(ii) in **?** for a set of sufficient conditions for one-dimensional partially identified parameters).

One could go further and consider confidence sets of the form

$$C^{\mathcal{P}}_{1-\alpha} = \bigcup_{(p',G')\in C^{p,G}_{1-\alpha}} \mathcal{P}(p',G'), \tag{19}$$

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 (19) because any confidence set for the identified set can be represented as a superset of (19). 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 (19) defined by these $C^{p,G}_{1-\alpha}(\omega)$ is a $100(1-\alpha)\%$ confidence set for $\mathcal{P}(p_0,G_0)$, and $C^{\mathcal{P}}_{1-\alpha}(\omega) \subset A^{\mathcal{P}}_{1-\alpha}(\omega)$, $\forall\omega$.

The problem of finding $C^{p,G}_{1-\alpha}$ so that $C^{\mathcal{P}}_{1-\alpha}$ in (19) satisfies some optimality properties seems to be very hard to solve analytically or numerically in our high-dimensional settings. Therefore, in our application, we just 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). Most of the literature on confidence sets for partially identified parameters also does not consider optimality properties.[4]

Approximations to sets in (19) (or (18)) 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

[4]One exception to this is **?** who considers optimality in testing moment inequality models.

approach described in the previous subsection. Specifically, sets in (19) (or (18)) 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 since they are more informative than credible sets. They also have a decision theoretic justification. As we demonstrate in the application section, 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 from the sampling variation.

3.4. *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,

$$|\delta_i - F^{-1}_{logistic}(p_i)| < bd \cdot \max\{|F^{-1}_{logistic}(p_i) - F^{-1}_{normal}(p_i)|, \sigma_{logistic}\}, \tag{20}$$

where $\sigma_{logistic}$ 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 $\sigma_{logistic}$ 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. In experiments below we use $bd \in \{0.25, 1, \infty\}$. Using logistic and normal quantiles as benchmarks is sensible as most of the applications use these distributions for unobservables. At the same time, any degree of flexibility can be attained by setting appropriate values for bd. To impose these addi-

tional restrictions on F in the model one can just add inequality (20) to inequalities in Theorem 1. More generally, any restrictions on δ can be included in the model in a similar fashion.

4. APPLICATION

We illustrate our inference framework using **?** model of bus engine replacement. First, we describe the model. Second, we construct the identified sets for the per-period payoff parameters and time discount factor. Third, we discuss the prior specification and the MCMC algorithm for estimation. Fourth, we estimate the counterfactual CCPs using simulated data. We compare the estimation results with the identified sets for the counterfactual CCPs. Finally, we estimate the model using real data from Rust and compare our confidence sets for the parameters of the per-period payoffs with Rust's estimates and confidence sets.

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

? 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 **?**, **?**, and **?**).

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 $\epsilon_t = (\epsilon_{t1}, \epsilon_{t0})$, are observed by the manager but not the econometrician. The mileage is discretized into $K = 90$ intervals $X = \{1, \ldots, K\}$. The costs of engine replacement and maintenance at time t are given respectively by $u(x_t = k, d_t = 1, \epsilon_t) = u_{1,k} + \epsilon_{t1}$ and $u(x_t = k, d_t = 0, \epsilon_{t0}) = u_{0,k} + \epsilon_{t0}$. $u_{1,k}$ is constant across k and it captures the deterministic one-time replacement costs; $u_{0,k}$ is the deterministic maintenance cost for an

engine in mileage interval k. For the rest of the section, we normalize $u_{1,k}$ to 0 for all k. For $t \leq 88$, 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 $t = 89$ 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.

? 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_{t0} and ϵ_{t1}. 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 **?** where cost functions are linear in the unknown parameters. To fix ideas, we adopt a simple linear index specification: $U = \{u : \ u_{1k} = 0 \text{ and } u_{0k} = \theta_0 + \theta_1 k, \text{ for some } \theta_1 < 0, \theta_0 \in \mathbb{R}^1\}$. If F were known to econometricians, θ_0 and θ_1 would be over-identified (see 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 the definition of $\mathcal{P}(p, \pi)$ in Corollary 5). A linear programming algorithm for checking the feasibility of the system is briefly described at the end of Appendix A.

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 ($\tilde{\theta}_0 = \theta_0$ and $\tilde{\theta}_1 = \theta_1$).

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

LEMMA 6 *If $\theta_1 < 0$, then $p = (p_1, \ldots, p_k, \ldots, p_K)$ 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 *If there are $(p', \tilde{p}', \pi')$ such that $\tilde{p}' \in \mathcal{P}(p', \pi')$ and no two coordinates in $(p', \tilde{p}'_2, \ldots, \tilde{p}'_K)$ are identical, then for any $(p, \tilde{p}_2, \ldots, \tilde{p}_K, \pi_0, \pi_1)$ sufficiently close to $(p', \tilde{p}'_2, \ldots, \tilde{p}'_K, \pi'_0, \pi'_1)$ $\tilde{p} \in \mathcal{P}(p, \pi)$.*

Except for Section 4.8, we use simulated data below. This allows us to avoid possible misspecification issues and 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 **?**. We use the following DGP for simulating the data: logistic F, $\theta_0 = 5.0727$, $\theta_1 = -.002293$, $\pi_0 = .3919, \pi_1 = .5953$ and the discount factor $\beta = .999$. 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 4.8 demonstrates the application of the methodology to real data.

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

This subsection recovers the identified set of per-period payoff parameters (θ_0, θ_1) and visualizes the identifying power of additional restrictions on F. By definition, the identified set consists of all (θ_0, θ_1) that make $u_{0k} = \theta_0 + \theta_1 k$ and $u_{1k} = 0$ consistent with (π, β, p) in the DGP. We compute this two-dimensional set by doing a grid search in the parameter space of (θ_0, θ_1) and collecting all pairs that make (15) feasible with solutions in (e, δ) that satisfy (12)-(14). This is done by a linear programming algorithm similar to the one we use in checking that $\tilde{p} \in \mathcal{P}(p, \pi)$ (as described in Appendix A). The set that we find by grid search is a valid approximation to the identified set since the identified set is convex (convexity follows immediately from

Theorem 1 and the linearity of per-period payoffs). Using MATLAB medium scale linear programming algorithm on a PC with Intel 2.7GHz processor and 8GB RAM, it takes about 44 seconds to compute an approximation to the identified set on a grid with step sizes $(1, 0.0005)$ for (θ_0, θ_1).

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 (20) 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 5 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 guess and verify that the identified set of u_0 includes values that are more than 3 orders of magnitude larger than the DGP values. To understand this, note that by inequalities (12)-(14), 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.

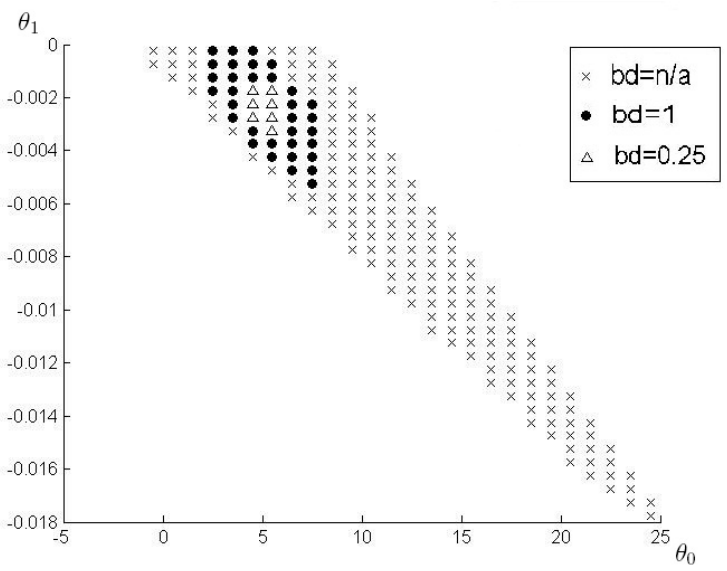

FIGURE 1.— Identified sets of cost parameters with $bd = \infty, 1, 0.25$

In our parameterization that roughly corresponds to Rust's estimates, the smallest CCP, p_1, is of the order 10^{-4} and thus, according to (14), δ_1 can reach the order 10^4. By (15), this can lead to large values in u_0.

4.3. *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 Section 4.1. 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 gird. 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.

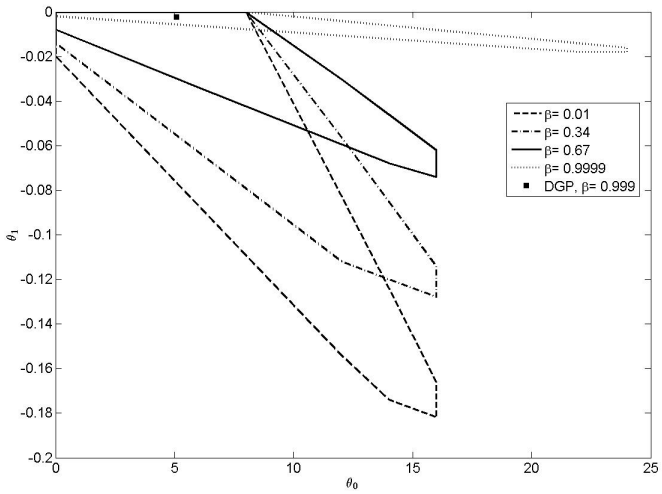

FIGURE 2.— Identified sets θ for different β ($bd = \infty$)

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 3.4), the projection is $[0.52, 1)$ or, in other words, the identified set for θ is empty for $\beta < 0.52$.

4.4. *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 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 modeling (see **?** for an insightful discussion of an example where hierarchical modeling solves a somewhat similar problem; for a textbook treatment of hierarchical models, see **?** and **?**). 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 nonparametric classes of densities; see, for example, **?**. 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 $\{(p,\tilde{p},\pi) : \tilde{p} \in \mathcal{P}(p,\pi)\}$,

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

Introduction of such allocation variables is standard in MCMC estimation of mixture models; see **?**. 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

$$\begin{aligned}(21)\qquad & P(p,\tilde{p},\pi,z,\tilde{z},\alpha,m,s) \propto \\ & \prod_{k:z_k=1} \text{Beta}(p_k; m_1 s_1, (1-m_1)s_1) \cdots \prod_{k:z_k=M} \text{Beta}(p_k; m_M s_M, (1-m_M)s_M) \\ & \cdot \prod_{k:\tilde{z}_k=1} \text{Beta}(\tilde{p}_k; m_1 s_1, (1-m_1)s_1) \cdots \prod_{k:\tilde{z}_k=M} \text{Beta}(\tilde{p}_k; m_M s_M, (1-m_M)s_M) \\ & \cdot \prod_{j=1}^{M} \text{Beta}(m_j; \underline{N}_{m0}, \underline{N}_{m1}) \cdot \text{Gamma}(s_j; \underline{\gamma}_{s0}, \underline{\gamma}_{s1}) \\ & \cdot \prod_{k=1}^{K} \alpha_{z_k} \cdot \prod_{k=2}^{K} \alpha_{\tilde{z}_k} \cdot \prod_{j=1}^{M} \alpha_j^{\underline{a}-1} \\ & \cdot \pi_0^{\underline{b}-1} \pi_1^{\underline{b}-1} (1-\pi_0-\pi_1)^{\underline{b}-1} \cdot 1_{\mathcal{P}(p,\pi)}(\tilde{p}),\end{aligned}$$

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.

4.5. *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 (21) and the likelihood in (17) with G_{kl}^j replaced by the corresponding elements of π. Its density can be computed up to a normalizing constant from (17) and (21). Therefore, a Metropolis-Hastings MCMC algorithm can in principle be used

for exploring the posterior distribution.[5] 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.[6] Also, different variations of the Metropolis-Hastings algorithm can be used together with the Gibbs sampler to construct robust hybrid MCMC algorithms, see **?**. 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 practice and has required theoretical properties. It is worth noting that Lemma 6 is by no means indispensable to our approach. An MCMC algorithm can be implemented regardless of whether the ordering of CCPs in the DGP is known. On the other hand, knowledge of any restriction on parameters that can be derived from primitive conditions (such as the ordering of CCPs derived from shape restrictions on per-period payoffs in Rust's model) can help construct efficient proposal distributions for MCMC algorithms. Thus, we incorporate order restrictions from Lemma 6 in proposal densities in some (but not all) of our Gibbs sampler blocks. Appendix A provides a detailed description of the algorithm.

[5]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, **?** or **?**.

[6]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, **?** or **?**.

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 requires more iterations for convergence than the number of iterations required for the likelihood maximization in Rust's algorithm. We report approximate computing times for identification and estimation exercises in Sections 4.2 and 4.6.

4.6. *Estimation of counterfactuals on simulated data*

The DGP for simulating the data used in this subsection is described in Section 4.1. We simulate a data set for 5,000 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 $\tilde{\pi}_0 = .6$ and $\tilde{\pi}_1 = .3$ 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 2.7GHz processor and 8GB 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 report estimation results using the thinned samples. Trace plots of MCMC draws from several simulator runs suggest that the MCMC algorithm converges. See Figure 8 and Figure 9 in Appendix C for plots from one of these simulator runs.[7]

[7]These figures show the trace plots for p_k and $\tilde{p}_k$, $k = 50, 60, 70, 80, 85, 90$. Lower coordinates converge faster.

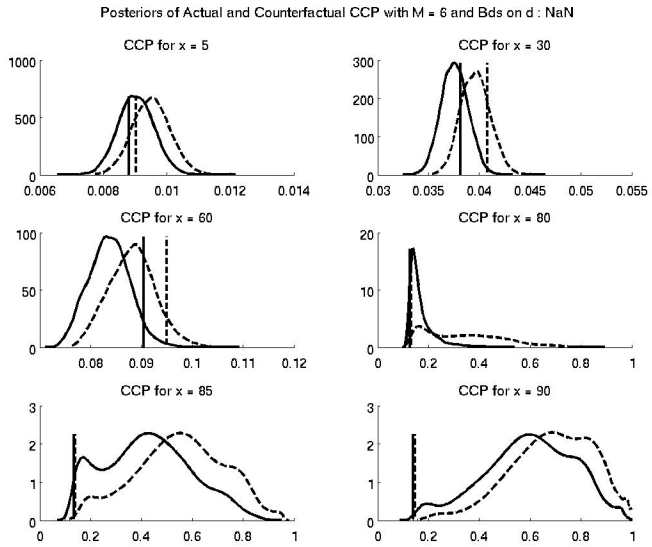

FIGURE 3.— Posteriors for actual CCPs (solid) and counterfactual CCPs (dashed). No bounds on δ. Vertical lines show "true" values for actual and counterfactual CCPs.

To compare prior and posterior, we plot marginal prior and posterior distributions for actual and counterfactual CCPs in Figure 10 and Figure 11 in Appendix C. Comparing marginal priors and posteriors in Figure 11 demonstrates that the model restrictions and the data can be quite informative about CCPs in the counterfactual experiment.

Figure 3 displays posteriors for the CCPs and the counterfactual CCPs together. Compared with the posterior of 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 $p(x) = F_{\Delta\epsilon}(u_1(x) - u_0(x) + \beta[V(1) - V(x)])$ 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 behavior of utility functions and the rest of the structural parameters. For our parameter values, the change is positive for all x.

Figure 3 also reveals two features in the estimation results. First, 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. Second, the uncertainty about the actual and counterfactual CCPs is comparable for most of the coordinates because the identified sets for the counterfactual CCPs are small as we show in Section 4.7 below.

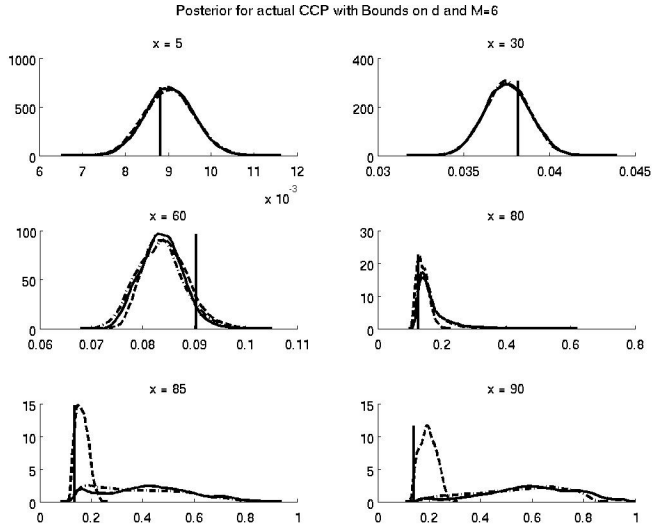

FIGURE 4.— Posteriors for actual CCPs with restrictions on F: $bd = \infty$ (solid), $bd = 1$ (dash-dot), $bd = 0.25$ (dashed). Vertical lines are "true" actual CCPs.

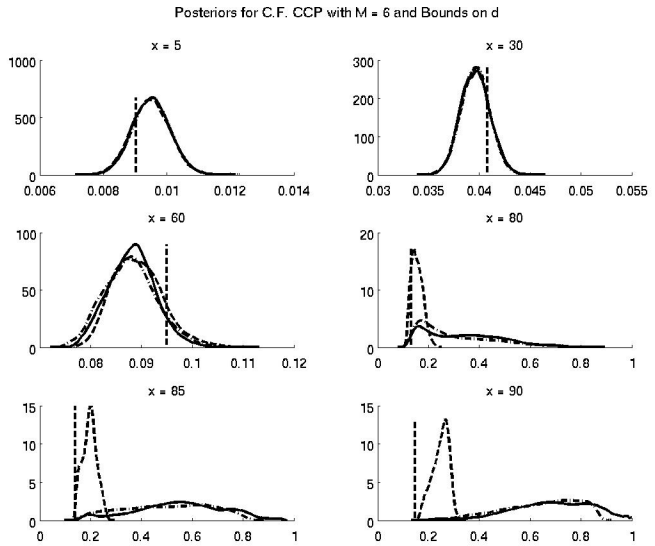

FIGURE 5.— Posteriors for c.f. CCPs with restrictions on F: $bd = \infty$ (solid), $bd = 1$ (dash-dot), $bd = 0.25$ (dashed). Vertical lines are "true" c.f. CCPs.

Figures 4 and 5 present estimation results with additional restrictions on quantiles of F introduced in Section 3.4. While changing bd from infinity to 1 has no visible effect on the 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 5 should also be interpreted as evidence that specifications of the unobserved state distribution can have a substantial impact on the counterfactual CCPs for some observed states.

4.7. *Identified sets of counterfactual CCPs*

In this subsection, we examine the identified set of counterfactual CCPs. By comparing these identified sets with posteriors in Figure 5 we can assess the contribution of the estimation and identification uncertainty to the posterior distributions. It is impractical to use a grid search to recover the identified set for counterfactual CCPs in a space with dimension $K = 90$. We use an MCMC algorithm for recovering this set instead. The algorithm is the same as the one we used in estimation except we keep (p, π) fixed at the DGP values used in the estimation experiments. Figure 6 presents the resulting "posteriors" for counterfactual CCPs. 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 $\tilde{p}$.

From Figure 6, we see that the additional restrictions on F visibly reduce the size of the identified set only for the counterfactual CCPs corresponding to higher mileage. A comparison of Figures 5 and 6 suggests that the identified sets for the counterfactual CCPs are small relative to estimation uncertainty for most of the coordinates of $\tilde{p}$. Also, as F is restricted to be increasingly closer to the logistic distribution, the identified sets for the counterfactual CCPs in most dimensions change little, while the identified sets for the per-period payoff parameters are reduced a lot. Overall, the

results of this section suggest that the effect of distributional assumptions about F on counterfactual CCPs can be small for most observed states even though the effect on the estimated per-period payoffs is large.

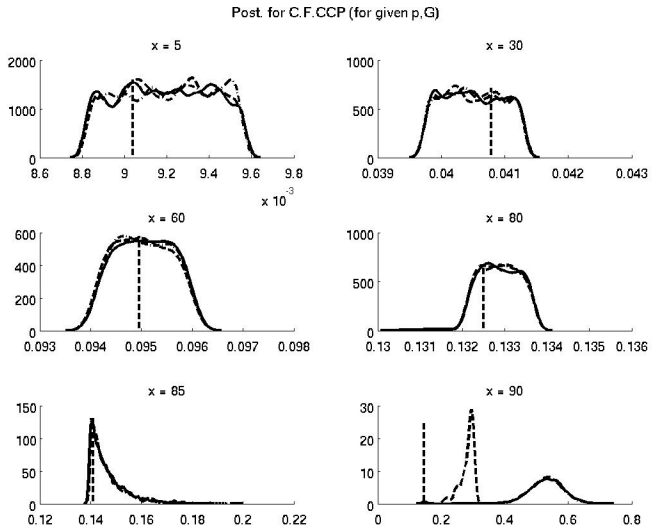

FIGURE 6.— Posteriors of c.f. CCPs with p, G fixed as in DGP with $bd = \infty$ (solid), 0.25 (dash), 1 (dash-dot). Vertical lines: c.f. CCPs with $\epsilon_i \sim$ extreme value i.i.d.

4.8. *Estimation with real data*

In this subsection, we compare parameter estimation results from **?** and our semiparametric 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 $(1.582, 0.000639)$. Figure 7 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 the figure. As discussed in Section 3.3, a valid $1-\alpha$ confidence for θ is given by

$$B^{\theta}_{1-\alpha} = \bigcup_{(p',\pi') \in B^{p,\pi}_{1-\alpha}} \mathcal{U}(p', \pi'), \tag{22}$$

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 $[(\log p, G)-\hat{\mu}]'\Sigma^{-1}[(\log p, G)-\hat{\mu}] \leq c$, where critical value c is chosen so that $100(1-\alpha)\%$ percent of posterior draws are in the set. Thus, set $B^{\theta}_{1-\alpha}$ has Bayesian and asymptotic frequentist interpretations. Comparison of the confidence set with Rust's results confirms findings from Section 4.2 based on artificial data: distributional assumptions about unobservables can have an enormous effect on parameter estimates in DBCMs.

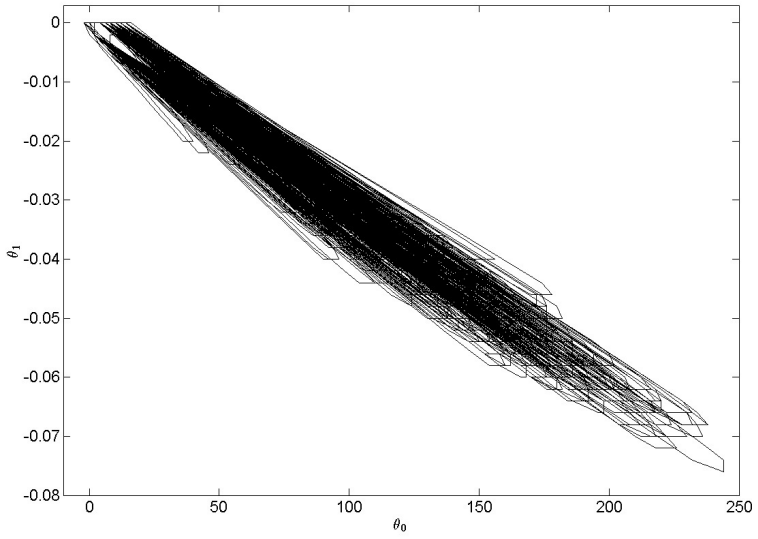

FIGURE 7.— Confidence set for θ

5. 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 **?** and **?**. 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 **?** and **?**, 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 time-invariant unobserved heterogeneity in per-period payoffs. Specifically, to models with finite number of unobserved agent types that have different per-period payoffs. In these models, CCPs for each agent type can be non-parametrically point identified. **?** 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 the MCMC estimation algorithm by using standard data augmentation techniques for finite mixture models (**?**, **?**). Specifically, we can introduce additional variables into the MCMC algorithm, which specify the agent type for each observation in the sample. Conditional on these type variables, the Gibbs sampler blocks for CCPs and other variables pertaining to different types are the same as those in the algorithm for no heterogeneity case. The Gibbs sampler blocks for the type variables are simple multinomial distributions. Time-invariant unobserved heterogeneity in observed state transition probabilities and distributions of unobserved state variables can also be accommodated as long as the sufficient conditions for point

identification in **?** are satisfied. Of course, the estimation algorithm for models with unobserved heterogeneity will be more computationally demanding.

It seems possible to extend the MCMC algorithm to estimation of models with time-variant unobserved heterogeneity considered by **?**. However, more work is required to understand identification in that framework.

Extensions of our framework to multinomial choice models is another important direction for future research.

APPENDIX A: MCMC ALGORITHM

This appendix provides a description of the Metropolis-within-Gibbs algorithm for exploring the posterior distribution of $(p, \tilde{p}, \pi, z, \tilde{z}, \alpha, m, s)$. The algorithm is implemented in Matlab and the code is available upon request (we plan to make it available on the web in the near future). The algorithm consists of the following blocks.

1. Stack current draws of p and $\tilde{p}$ in one vector $(p, \tilde{p})$ and sort it in an ascending order and draw a candidate for even coordinates of the sorted vector from a beta distributions proportional to the product of (21) and (17) and truncated to have the same order as current $(p, \tilde{p})$. Accept the draw with probability 1 if the candidate does not violate restrictions $\tilde{\mathcal{M}}$ and reject otherwise. This block was 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 1 and 2 does preserve the stationary distribution of the Markov chain. However, since $\tilde{\mathcal{M}}$ does not imply a particular order of coordinates in $(p, \tilde{p})$ (it only implies an order within p and $\tilde{p}$ separately) we need to add Markov transitions that would allow change in the order. The following three blocks achieve this.
3. Pick an index $k \in \{1, \ldots, K\}$ randomly or deterministically. Draw a candidate for p_k from a beta distribution proportional to product of (21) and (17) and

truncated to (p_{k-1}, p_{k+1}). Accept the draw with probability 1 if the candidate does not violate restrictions $\tilde{\mathcal{M}}$ 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 support, $\tilde{\mathcal{M}}$, can be reached by the algorithm, which is required for MCMC convergence; see, for example, **?** or **?**. 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 (21) and (17). Accept the draw with probability 1 if the candidate does not violate restrictions $\tilde{\mathcal{M}}$ and reject otherwise.
7. Blocks for sampling beta mixture prior parameters $(z, \tilde{z}, \alpha, m, s)$ are described in **?**

Since it is computationally expensive to verify restrictions $\tilde{\mathcal{M}}$ 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 $\tilde{\mathcal{M}}$). We check the correctness of the algorithm implementation by joint distribution tests; see **?**.

To verify whether $(p, \tilde{p}, \pi) \in \tilde{\mathcal{M}}$ we substitute expressions for δ and $\tilde{\delta}$ from (15) into (12)-(14) and rewrite the resulting inequalities in the following form $Ax < 0$, where x includes θ_0, θ_1, e, and e_m and matrix A is computed from $(p, \tilde{p}, \pi, \tilde{\pi})$. Then we use the Matlab simplex method to solve the following linear programming problem: $\min_{x,t:Ax<t} t$. If the resulting optimal $t^*(p, G, \tilde{p})$ is non-positive then $(p, \tilde{p}, \pi) \in \tilde{\mathcal{M}}$. This linear program can also be used to define the criterion function for **?** approach to inference on $\tilde{p}$: $Q(p, G, \tilde{p}) = \max(0, t^*(p, G, \tilde{p}))$.

APPENDIX B: PROOFS

PROOF: (Sufficiency in Lemma 1)

Suppose p satisfies (5). We will show that it is a vector of CCPs. Define vectors y_0 and y_1 as follows,

$$y_1 = (I - \beta G^1)^{-1}\Big[u_1 + \beta G^1\big[\int_{F^{-1}_{\Delta\epsilon|x}(p|X)}^{\infty} s dF_{\Delta\epsilon|x}(s|X) - (I - diag(p))F^{-1}_{\Delta\epsilon|x}(p|X))\big]\Big] \tag{23}$$
$$y_0 = (I - \beta G^0)^{-1}\Big[u_0 + \beta G^0\big[\int_{F^{-1}_{\Delta\epsilon|x}(p|X)}^{\infty} s dF_{\Delta\epsilon|x}(s|X) + diag(p)F^{-1}_{\Delta\epsilon|x}(p|X)\big]\Big]$$

By (5) and definition of (y_0, y_1), we have $F^{-1}_{\Delta\epsilon|x}(p|X) = y_1 - y_0$ and $p = \int_{-\infty}^{y_1-y_0} dF_{\Delta\epsilon|x}(s|X))$. Using these two equations we can get rid of p in (23),

$$y_1 = (I - \beta G^1)^{-1}\Big[u_1 + \beta G^1\big[\int_{y_1-y_0}^{\infty} s dF_{\Delta\epsilon|x}(s|X) - \int_{y_1-y_0}^{\infty} dF_{\Delta\epsilon|x}(s|X)(y_1 - y_0))\big]\Big] \tag{24}$$
$$y_0 = (I - \beta G^0)^{-1}\Big[u_0 + \beta G^0\big[\int_{(y_1-y_0)}^{\infty} s dF_{\Delta\epsilon|x}(s|X) + \int_{(y_1-y_0)}^{\infty} dF_{\Delta\epsilon|x}(s|X)(y_1 - y_0)\big]\Big]$$

From (24) one can reverse the steps leading from (2) to (4) in Section 2.2 to show that (y_0, y_1) have to satisfy the Bellman equation (2). Since the solution of the Bellman equation is unique, $(y_0, y_1) = (v_0, v_1)$ and $p = \int_{-\infty}^{y_1-y_0} dF_{\Delta\epsilon|x}(s|X)$ is a vector of CCPs.

Q.E.D.

PROOF: (Corollary 3)

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

$$\begin{bmatrix} M^1_U u = [M^1_D, -M^1_E][d_1', e_1']' \\ \vdots \\ M^N_U u = [M^N_D, -M^N_E][d_N', e_N']' \end{bmatrix}, \tag{25}$$

where d_n and e_n are K-vectors with coordinates $d_{n,k} = F^{-1}_{\Delta\epsilon|X=k}(p^n_k)$, $e_{n,k} = \int_{d_{n,k}}^{\infty} s dF_{\Delta\epsilon|X=k}(s)$; and

$$M^n_U = [\ (I - \beta G^{1,n})^{-1}, -(I - \beta G^{0,n})^{-1}\ ]$$
$$M^n_D = (I - \beta G^{1,n})^{-1}[I - diag(p^n)] + (I - \beta G^{0,n})^{-1} diag(p^n)$$
$$M^n_E = (I - \beta G^{1,n})^{-1}\beta G^{1,n} - (I - \beta G^{0,n})^{-1}\beta G^{0,n}.$$

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 (25) has a unique solution for the rest of the components of u.

Q.E.D.

PROOF: (Lemma 2)

Construct a linear system of $2K$ equations by stacking (6) and (5) and imposing $\tilde{u}=u$. The system can be simplified as:

$$A_1u_1 - A_0u_0 = B[Q(p)',\kappa(p)']' \tag{26}$$

$$\tilde{A}_1u_1 - \tilde{A}_0u_0 = \tilde{B}[Q(\tilde{p})',\kappa(\tilde{p})']' \tag{27}$$

where $A_j = (I-\beta G^j)^{-1}$ 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

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

The form of these two functions depend on the unobserved state distribution $F_{\Delta\epsilon|X}$. 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

$$u_1 = A_1^{-1}\left\{B[Q(p)',\kappa(p)']' + A_0c\right\} \tag{28}$$

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

$$\tilde{B}[Q(\tilde{p})',\kappa^0(\tilde{p})']' = \tilde{A}_1A_1^{-1}B[Q(p)',\kappa(p)']' + \left(\tilde{A}_1A_1^{-1}A_0-\tilde{A}_0\right)c \tag{29}$$

With $F_{\Delta\epsilon|X}$ assumed known and p identified from the DGP, this implies that whenever $\left(\tilde{A}_1 A_1^{-1} A_0 - \tilde{A}_0\right)(c - u_0^*) \neq 0$, the choice of c has an impact on $\tilde{p}$ predicted as the solution to the equation above. *Q.E.D.*

PROOF: (Lemma 3)
In this case, the counterfactual CCPs, denoted $\hat{p}$, are characterized by

$$A_1(\alpha u_1 + \Delta_1) - A_0(\alpha u_0 + \Delta_0) = B[Q(\hat{p})', \kappa(\hat{p})']', \tag{30}$$

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 (28), identifying counterfactual CCPs amounts to finding $\hat{p}$ such that

$$B[Q(\hat{p})', \kappa(\hat{p})']' = \alpha B[Q(p)', \kappa(p)']' + A_1\Delta_1 - A_0\Delta_0. \tag{31}$$

It then follows that the choice of c has no impact on the characterization of $\hat{p}$ in (31). *Q.E.D.*

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 (8)-(11). Then,

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

imply (12). Inequalities in (13) follow since

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

for F satisfying (8). By (9),

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

and (14) follows.

To prove the other direction of the lemma, suppose (12)-(14) hold. Let us construct a particular density $f(s) > 0$ that satisfies (8)-(11). 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 (8)-(11). *Q.E.D.*

PROOF: (Lemma 5)

We now show for any $(\tilde{u}, u)$ satisfying conditions of the lemma, the actual CCP p cannot belong to the identified set of counterfactual CCPs $\tilde{p}$. Suppose $p \in \mathcal{P}(p, G)$ then $\tilde{\delta} = \delta$ and $\tilde{e} = e$. It then follows that (15) and (16) cannot both hold simultaneously when $\tilde{p}$ in (16) is replaced by p. This contradicts the supposition that $p \in \mathcal{P}(p, G)$. *Q.E.D.*

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 (**?**), v_1 and v_0 from (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. *Q.E.D.*

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 utility 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: $p_1 = \tilde{p}_1$. *Q.E.D.*

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 the end of the description of the MCMC algorithm in Appendix A). Since the inequalities are strict, they have to be satisfied in an open neighborhood of the original feasible point. *Q.E.D.*

APPENDIX C: EXTRA FIGURES

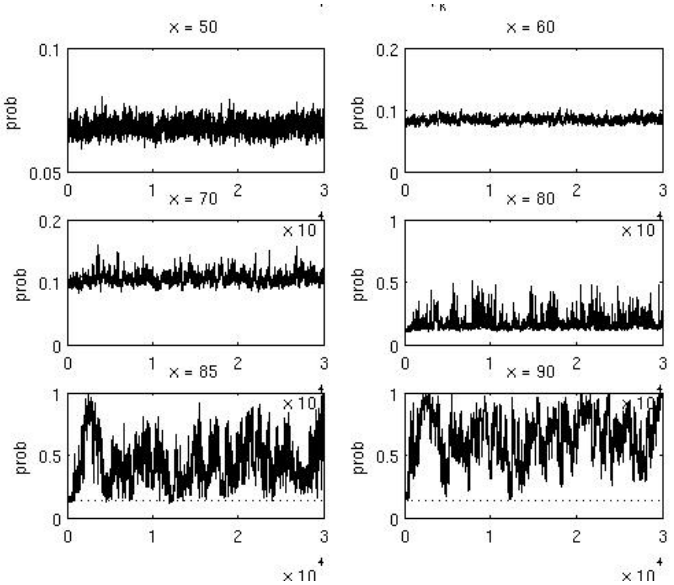

FIGURE 8.— Trace plots of posterior draws for coordinates in actual CCPs p.

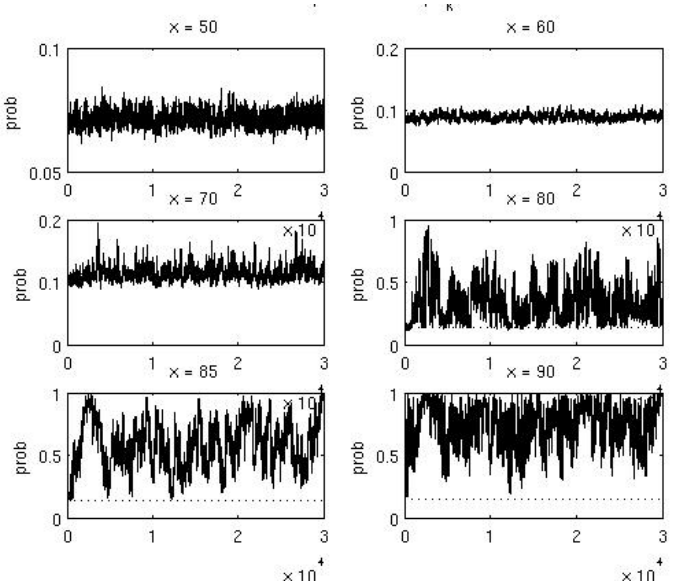

FIGURE 9.— Trace plots of posterior draws for counterfactual CCPs $\tilde{p}$.

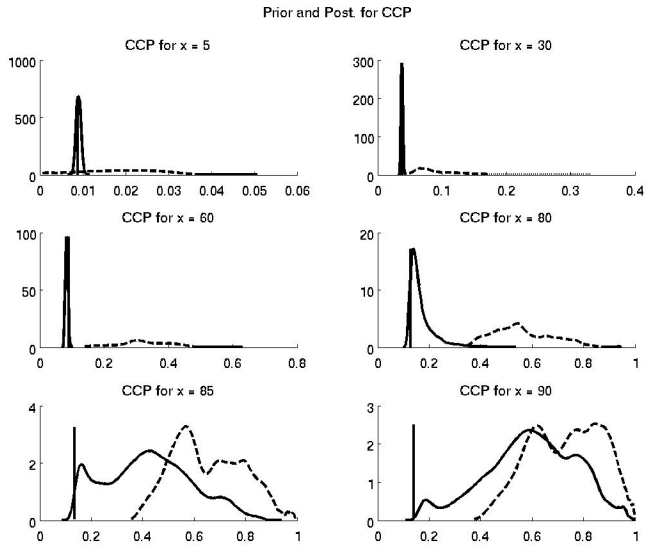

FIGURE 10.— Marginal priors and posteriors for actual CCPs p. Prior - dashed. Posterior - solid. Vertical - "true" actual CCPs.

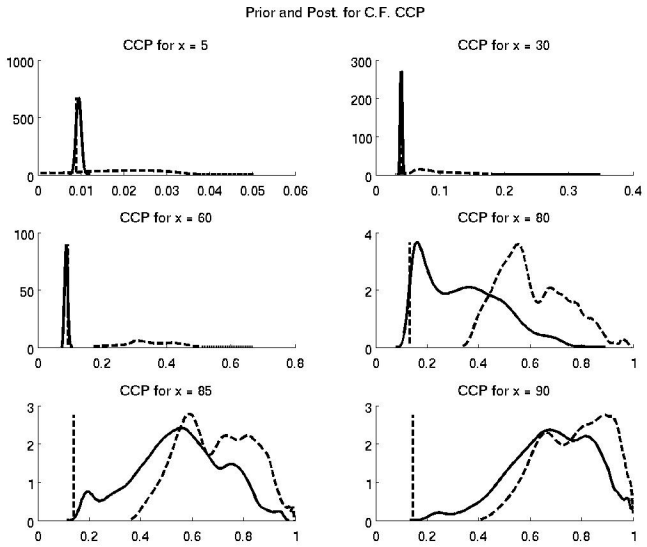

FIGURE 11.— Marginal priors and posteriors for counterfactual CCPs $\tilde{p}$. Prior - dashed. Posterior - solid. Vertical - "true" counterfactual CCPs.