SEMIPARAMETRIC BAYESIAN ESTIMATION OF DYNAMIC DISCRETE CHOICE MODELS

ANDRIY NORETS [1] AND KENICHI SHIMIZU [2] [3]

We propose a tractable semiparametric estimation method for dynamic discrete choice models. The distribution of additive utility shocks is modeled by location-scale mixtures of extreme value distributions with varying numbers of mixture components. Our approach exploits the analytical tractability of extreme value distributions and the flexibility of the location-scale mixtures. We implement the Bayesian approach to inference using Hamiltonian Monte Carlo and an approximately optimal reversible jump algorithm from Norets (2021). For binary dynamic choice model, our approach delivers estimation results that are consistent with the previous literature. We also apply the proposed method to multinomial choice models, for which previous literature does not provide tractable estimation methods in general settings without distributional assumptions on the utility shocks. In our simulation experiments, we show that the standard dynamic logit model can deliver misleading results, especially about counterfactuals, when the shocks are not extreme value distributed. Our semiparametric approach delivers reliable inference in these settings. We develop theoretical results on approximations by location-scale mixtures in an appropriate distance and posterior concentration of the set identified utility parameters and the distribution of shocks in the model.

KEYWORDS: Dynamic Discrete choice, Bayesian nonparametrics, set identification, location-scale mixtures, MCMC, Hamiltonian Monte Carlo, reversible jump.

1. INTRODUCTION

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

---

[1]Department of Economics, Brown University; andriy_norets@brown.edu

[2]Adam Smith Business School, University of Glasgow; kenichi.shimizu@glasgow.ac.uk.

economics, and industrial organization. See, for example, Rust (1994) and Aguirregabiria and Mira (2010) for literature surveys. In such models, a forward-looking decision-maker chooses an action from a finite set in each time period. The actions affect decision-maker's per-period payoff and the evolution of state variables. The decision-maker maximizes the expected sum of current and discounted future per-period payoffs.

Some state variables in these models are usually assumed to be unobserved by the econometrician. The introduction of the unobserved state variable is motivated by the fact that the decision maker usually has more information than the econometrician. If the econometrician observes all state variables, then economic theory implies that the pair of the control and the state variables should obey some deterministic relationship which is never the case with real data (see, for example, page 1008 of Rust (1987) for further discussion).

Most of the previous work on estimation of dynamic discrete choice models imposes specific parametric assumptions on the distribution of the unobserved states or utility shocks. The most commonly used parametric assumption is that the unobserved states are extreme value independently identically distributed (i.i.d.). This assumption alleviates the computational burden of solving the dynamic program and computing the likelihood function. At the same time, it is well known in the literature that imposing parametric distributional assumptions can cause problems, see, for example, Manski (1999). The researcher would not know a priori if such an assumption indeed is problematic and therefore it is better to relax it if possible.

There are several previous papers that treat the unobserved state distribution nonparametrically for the binary choice case. Aguirregabiria (2010) shows the nonparametric identification of the shock distribution under particular assumptions on the per-period payoffs. Norets and Tang (2013) show that under unknown distribution of the unobserved state and discrete observed states, the utility parameters and the unobserved state distribution are only set-identified. They also show how to compute the identified sets. Buchholz et al. (2020) provide idenitifcation results for the per-period payoffs when the observed state is continuous.

For the mutinomial choice case, Chen (2017) uses exclusion restrictions (a subset of the state variables affects only current utility, but not state transition probabilities) to obtain identification and estimation results. In settings without exclusion restrictions, Norets (2011)

shows that it is in principle possible to extend the method from Norets and Tang (2013) to compute the identified set in multinomial case, but it is computationally infeasible.

In this paper, we propose a tractable semiparametric estimation method applicable to the multinomial choice case. It is based on modeling the unknown distribution of shocks by a finite mixture of extreme value distributions with a varying number of mixture components. Our approach exploits the analytical tractability of extreme value distributions and the flexibility of the location-scale mixtures. The unobserved utility shocks can be integrated out analytically in the likelihood function and the expected value functions, similarly to the case with extreme value distributed shocks. At the same time, we show that the location-scale mixtures can approximate densities from a large nonparametric class in an appropriate distance and that for any given distribution of utility shocks, a finite mixture of extreme value distributions can deliver exactly the same conditional choice probabilities. Posterior concentration on the identified sets of utility parameters and the distribution of shocks is an implication of these results. We implement the Bayesian approach to inference for the model using Hamiltonian Monte Carlo and an approximately optimal reversible jump algorithm from Norets (2021). We apply our framework to binary and multinomial choice models. For the binary dynamic choice model from Rust (1987), our approach delivers estimation results that are consistent with the previous literature on semiparametric estimation (Norets and Tang (2013)). For the multinomial choice model of medical care use and work absence from Gilleskie (1998), we demonstrate how uncertainty about model parameters and counterfactuals increases when the distributional assumptions on the shocks are relaxed. Moreover, we show that the standard dynamic logit model can deliver misleading results, especially about counterfactuals, when the shocks are not extreme value distributed. Our semiparametric approach delivers reliable inference in these settings.

Even when the distribution of the utility shocks is assumed to be known, parameters and counterfactuals in dynamic discrete choice models could still be only set identified under a variety of scenarios such as very flexible specifications of utility functions, lack of exclusion restrictions or variation in transitions for the observed state variables, and unknown time discount factors; see, for example, Rust (1994), Magnac and Thesmar (2002), Norets and

Tang (2013), Abbring and Daljord (2020), and Kalouptsidi et al. (2021). Our estimation framework does not require any special adjustments to accommodate such scenarios since parameters and counterfactuals are already set identified under nonparametric specification of the distribution of shocks.

The rest of the paper is organized as follows. Section 2 describes the general model setup. In Section 3, we introduce our semiparametric framework. In Section 4, we describe the Bayesian estimation method. Section 5 presents theoretical results. Sections 6 and 7 contain the applications. Derivations, proofs, and implementation details are given in appendices.

2. GENERAL MODEL SETUP

In the infinite-horizon version of the model, the decision maker maximizes the expected discounted sum of the per-period payoffs

$$(1) \qquad \max_{d_t, d_{t+1}, \dots} E_t \left(\sum_{j=0}^{\infty} \beta_j u \left(x_{t+j}, d_{t+j}, \epsilon_{t+j} \right) \right),$$

where $d_t \in \{0, 1, \dots, J\} = D$ is the control variable, $x_t \in X$ is the state variable observed by the econometrician, $\epsilon_t = (\epsilon_{t0}, \epsilon_{t1}, \dots, \epsilon_{tJ})^T \in R^{J+1}$ is the state variable unobserved by the econometrician, β is the time discount factor, and $u(x_t, d_t, \epsilon_t)$ is the per-period payoff. The decision-maker observes both x_t and ϵ_t at time t before making the decision.

Following Rust (1987) and the subsequent literature, we assume that (i) the per-period payoffs are additively separable in ϵ_t, $u(x_t, d_t, \epsilon_t) = u(x_t, d_t) + \epsilon_{td_t}$; (ii) ϵ_t's are independent of other variables and independently identically distributed (i.i.d.) over time according to a distribution F with zero mean; (iii) the observed states evolve according to a controlled Markov chain $G_x^j = Pr(x_{t+1}|x_t = x, d_t = j)$. In most applications, the utility functions are assumed to depend on a vector of unknown parameters, $\theta \in \mathbb{R}^{d_\theta}$, that are estimated. Below, we often omit θ in $u(x_t, d_t; \theta)$ and related objects such as value functions for notation brevity.

Under mild regularity conditions (Bhattacharya and Majumdar (1989)), the decision problem

in (1) admits the following Bellman representation

$$(2) \qquad Q(x) = \int \max_{j=0,1,\ldots,J} \left[u(x,j) + \beta G_x^j Q + \epsilon_j \right] dF(\epsilon),$$

where Q is called the *Emax* function and $G_x^j Q$ denotes $E(Q(x_{t+1})|x_t = x, d_t = j)$. The conditional choice probability (CCP) can be expressed as

$$(3) \qquad p(d|x) = \int 1 \left\{ u(x,d) + \beta G_x^d Q + \epsilon_d \geq u(x,j) + \beta G_x^j Q + \epsilon_j, \forall j \right\} dF(\epsilon).$$

For a panel of observations, $\{x_{it}, d_{it}, \ i = 1, \ldots, n, \ t = 1, \ldots, T\}$, of n decision makers over T time periods, the partial likelihood function (with the fixed G_x^j pre-estimated from the observed transitions as is commonly done in practice) can be expressed as

$$(4) \qquad \log L(x_{it}, d_{it}, \ i = 1, \ldots, n, \ t = 1, \ldots, T) = \sum_{i,t} \log p(d_{it}|x_{it}).$$

Rust (1987) proposed to solve the dynamic optimization problem by first iterating on the Bellman equation (2) to get close to the fixed point Q and then using a Newton method that quickly converges to the fixed point from a close starting point. With Q at hand, one can compute the CCPs in (3) and evaluate the likelihood function at a given $(u; \beta; G_x^j, j \in D, x \in X)$. Alternatively, Su and Judd (2012) proposed to use constrained optimization to maximize the likelihood function subject to (2). In either scenario, assuming that ϵ_{tj}'s are i.i.d. Gumbel (or extreme value type I) delivers analytical expressions for the integrals in (2) and (3),

$$(5) \qquad p(d|x) = \frac{e^{u(x,d)+\beta G_x^d Q}}{\sum_{j=0}^J e^{u(x,j)+\beta G_x^j Q}}, \quad Q(x) = \log \sum_{d=0}^J e^{u(x,d)+\beta G_x^d Q}.$$

In the resulting dynamic logit specification, the computational burden of the model solution and estimation is considerably alleviated. Hence, the dynamic logit is predominantly used in applications of the estimable dynamic discrete choice models. At the same time, the econometrics literature suggests that the distributional assumptions could be problematic in

general, see, for example, Manski (1999). In the following section, we specify a non-parametric model for the distribution of shocks for the general multinomial choice case that provides analytical simplifications comparable to those of the dynamic logit.

3. SEMIPARAMETRIC MODEL

Rather than making a particular parametric assumption, we model the distribution of unobserved states using a flexible mixture specification. In order to reduce the number of parameters, we use an innocuous normalization $\epsilon_{t0} = 0$ (the agent's decisions and value functions do not change if ϵ_{t0} is subtracted from the per-period payoff $u(x_t, d_t, \epsilon_t)$ for all $d_t \in \{0, 1, \ldots, J\}$).

For $\mu \in \mathbb{R}^J$ and $\sigma > 0$, let us define a multivariate Gumbel density by

$$(6) \qquad \phi(z; \mu, \sigma) = \prod_{j=1}^{J} \frac{1}{\sigma} \phi\left(\frac{z_j - \mu_j}{\sigma}\right), \text{ where } \phi(z_j) = e^{-z_j - \gamma - e^{-z_j - \gamma}}$$

is the univariate Gumbel density and γ is the Euler-Mascheroni constant. Some relevant properties of the Gumbel distribution are outlined in Appendix D.2.

For $\psi = \{\mu_k \in \mathbb{R}^J, \sigma_k \in \mathbb{R}_+, \omega_k \in (0, 1), k = 1, 2, \ldots\}$, we model the unknown density by a location-scale mixture of Gumbel densities

$$(7) \qquad \epsilon_t \sim \sum_{k=1}^{m} \omega_k \phi(\cdot; \mu_k, \sigma_k),$$

with a variable number of mixture components m for which a prior distribution on the set of positive integers is specified. Mixture models are extensively used in econometrics and statistics literature, see monographs by McLachlan and Peel (2000) and Fruhwirth-Schnatter (2006) for references. It is well known that location-scale mixtures with a variable or infinite number of components can approximate any continuous or smooth density arbitrarily well. For example, Bayesian models based on normal mixtures deliver optimal up to a log factor posterior contraction rates in adaptive estimation of smooth densities (Rousseau (2010) and Shen et al. (2013)). To develop intuition for this type of results note that the standard nonparametric density estimator based on kernel ϕ is a special case of (7), or, alternatively

and more in line with the actual proofs, the expectation of the standard kernel density estimator is a continuous mixture that can be discretized into a special case of (7). Thus, it is reasonable to expect that the specification (7) is very flexible. Indeed, in Section 5, we show that it can approximate smooth multivariate densities arbitrarily well in an appropriate distance so that the conditional choice probabilities and the *Emax* function implied by the model with (7) approximate those from the model with an arbitrary smooth density for ϵ_t.

The model specification with (7) also possesses attractive analytical properties. If a normalization $\epsilon_{t0} = 0$ is not imposed and $(J + 1)$-dimensional version of (7) is used, then Q and p could be expressed as mixtures of the appropriately recentered and rescaled expressions from the logit model (5). However, even if the normalization $\epsilon_{t0} = 0$ is imposed, which is preferred as it reduces the dimension of the distribution we model nonparametrically, closed form expressions for Q and p are still available. They are presented in the following lemma.

LEMMA 1   *Suppose* $\epsilon \sim \sum_{k=1}^{m} \omega_k \phi(\cdot; \mu_k, \sigma_k)$. *Then,*

$$(8) \qquad p(d|x) = \begin{cases} \sum_{k=1}^{m} \omega_k \exp\left[-e^{-a_{kx}}\right], & if \ \ d = 0 \\ \sum_{k=1}^{m} \omega_k \exp\left[\frac{u(x,d)+\beta G_x^d Q + \mu_{dk}}{\sigma_k} - A_{kx}\right] \left\{1 - \exp\left[-e^{-a_{kx}}\right]\right\}, & if \ \ d = 1, ..., J; \end{cases}$$

$$(9) \qquad Q(x) = \sum_{k=1}^{m} \omega_k \sigma_k \left[A_{kx} + E1(e^{-a_{kx}})\right], \ \ where \ E1(z) = \int_z^{\infty} e^{-t}/t dt,$$

$$A_{kx} = \log \sum_{j=1}^{J} \exp\left(\frac{u(x,j) + \beta G_x^j Q + \mu_{jk}}{\sigma_k}\right) \ \ and \ a_{kx} = \frac{u(x,0) + \beta G_x^0 Q}{\sigma_k} + \gamma - A_{kx}.$$

The derivations of (8) and (9) can be found in Appendix A. The derivatives of (8) and (9) that are useful for the model solution and estimation are given in Appendix D.4. Similarly to Rust (1987), we obtain the solution of the Belman equation (9) by a Newton-Kantorovich method described in Appendix D.1.

4. INFERENCE

4.1. *Motivation and Overview of the Bayesian Approach*

In estimation of models based on location-scale mixtures with a variable number of components, the econometrician faces several problems. First, the scale parameters need to be bounded away from zero; otherwise, the likelihood function is unbounded. Second, the likelihood function is a rather complex function of parameters with multiple modes. Third, the number of mixture components needs to be selected in the estimation procedure. Finally, there is usually considerable uncertainty about the estimated parameter values and it should be taken into account in model predictions and counterfactual analysis.

The Bayesian approach to inference and the associated simulation methods are well suited for solving these problems. Prior distributions can provide soft constraints for the scale parameters and an appropriate penalization for the number of mixture components or model complexity. MCMC methods can successfully explore very complex posterior or likelihood surfaces. Posterior predictive distributions for objects of interest automatically incorporate the uncertainty about parameter values including the number of mixture components.

Our MCMC algorithm for simulating from the model posterior distribution combines Hamiltonian Monte Carlo (HMC) for simulating parameters conditional on the number of mixture components and an approximately optimal reversible jump algorithm from Norets (2021) for simulating the number of mixture components. HMC is a very popular and efficient way of constructing proposal distributions for the Metropolis-Hastings algorithm. See, for example, Neal (2012) for an introduction. It requires only evaluation of the likelihood and the prior and their derivatives. The proposals are obtained following the Hamiltonian dynamics on the parameter space that describe the movement of a puck on a frictionless surface with some initial random momentum. For implementing the HMC step of the algorithm we utilize the HMC sampler from the Matlab Statistics and Machine Learning toolbox.

The reversible jump algorithm from Norets (2021) also uses the values of the likelihood, the prior and their derivatives with respect to a part of the parameter vector. As noted in Section 3, for our specification, the derivatives are available in closed form given a solution to the simplified Bellman equation (9). A detailed description of the MCMC algorithm is given in

Appendix B.

4.2. *Normalizations*

In addition to the normalization $\epsilon_{t0} = 0$, the scale of ϵ_t can be innocuously normalized. Instead, to simplify the MCMC algorithm we impose a location and scale normalization on the parameters of the per-period payoffs and keep the location and scale of (7) unrestricted. Specifically, for a linear in parameters utility specification

$$u(x_t, j, \epsilon_t; \theta) = \theta_j + z_j(x_t)'\theta_{J+1:d_\theta} + \epsilon_{tj},$$

the intercepts θ_j, $j = 0, \dots, J$, can be fixed to arbitrary values as long as the locations of ϵ_{tj}, $j = 1, \dots, J$, are unrestricted.

In applications, we set $\theta_0 = 0$ and θ_j, $j = 1, \dots, J$ to $\hat{\theta}_j^{dl}$, the estimates obtained under the dynamic logit specification. To normalize the scale of ϵ_t, we assume that the sign of one of the coefficients, say θ_{J+1}, is known and we keep this coefficient fixed (to the corresponding dynamic logit estimate, $\hat{\theta}_{J+1}^{dl}$). Thus, the MCMC algorithm produces draws of $\theta_{J+2:d_\theta}$ and the mixture parameters $(\psi_{1:m}, m)$ in (7). For comparisons of the estimation results with the dynamic logit estimates and the identified sets in Norets and Tang (2013), the parameter draws are renormalized for reporting as follows

$$s(\psi_{1:m}, m) \cdot \left(\hat{\theta}_{1:J}^{dl} + \sum_{k=1}^{m} \omega_k \mu_k, \hat{\theta}_{J+1}^{dl}, \theta_{J+2:d_\theta} \right),$$

where the addition of $\sum_{k=1}^{m} \omega_k \mu_k$ to the intercepts corresponds to the zero mean for shocks and the scale factor is defined by

$$s(\psi_{1:m}, m) = \log 2 \big/ E[\tilde{\epsilon}_{t1} 1(\tilde{\epsilon}_{t1} \geq M_{\tilde{\epsilon}_{t1}})],$$

where $\tilde{\epsilon}_{t1} = \epsilon_{t1} - \sum_{k=1}^{m} \omega_k \mu_{1k}$, $M_{\tilde{\epsilon}_{t1}}$ denotes the median of $\tilde{\epsilon}_{t1}$, and $\epsilon_{t1} \sim \sum_{k=1}^{m} \omega_k \phi(\cdot; \mu_{1k}, \sigma_k)$. There are many possible scale normalizations. The particular scale normalization we use here reduces to the one introduced by Norets and Tang (2013) for the binary choice case.

Let us emphasize that the normalizations discussed above are innocuous for estimation and counterfactual analysis as long as the assumed sign of θ_{J+1} is correct.

4.3. *Priors*

Let us introduce the prior distributions for the parameters of the mixture in (7). We use the following prior distributions on the number of mixture components and the mixing weights,

$$\Pi(m) \propto e^{-\underline{\mathrm{A}}_m m (\log m)^\tau}, \tag{10}$$

$$\Pi(\omega_1, ..., \omega_m | m) = \mathrm{Dirichlet}(\underline{\mathrm{a}}/m, ..., \underline{\mathrm{a}}/m), \tag{11}$$

where the hyperparameters $\underline{\mathrm{a}}$, $\underline{\mathrm{A}}_m$, and τ are specified in the applications below. For the theoretical results obtained in the present paper, we only need $\Pi(m) > 0, \forall m$ and full support on the simplex for $\Pi(\omega_1, ..., \omega_m | m)$. Nevertheless, the functional forms in (10) and (11) perform well in applications and deliver optimal posterior contraction rates in nonparametric multivariate density estimation by mixtures of normal distributions, see, for example, Shen et al. (2013) and Norets and Pelenis (2021). We allow the scale parameter σ_k to have a multiplicative part σ that is common across the mixture components: $\sigma_k = \tilde{\sigma}_k \cdot \sigma$. This multiplicative specification performs well in a variety of applications of location-scale mixture models (see, for example, Geweke (2005)) and is also important for the aforementioned optimal posterior concentration results for mixtures of normals. In the applications, we use finite mixtures of normals as flexible priors for $\log \sigma$, $\log \tilde{\sigma}_k$ and the location parameters μ_{kj}.

5. APPROXIMATION RESULTS AND ASYMPTOTICS

In this section, we show that location-scale mixtures of Gumbel densities can arbitrarily well approximate densities from a large nonparametric class. These approximation results combined with the Schwartz (1965)'s theorem imply a posterior consistency result for the set identified model parameters. We also show that a model with a finite mixture of Gumbels can exactly match the CCPs from a model with an arbitrary distribution of shocks.

5.1. *Approximation Results*

Let us first define a distance for distributions of utility shocks: for F_i with density f_i, $i = 1, 2$,

$$\rho(F_1, F_2) = \int (1 + \sum_{j=0}^{J} |\epsilon_j|) |f_1(\epsilon) - f_2(\epsilon)| d\epsilon.$$

This distance is appropriate for our purposes as the *Emax* function and the conditional choice probabilities are continuous in that distance as shown in the following lemma.

LEMMA 2 *Suppose (i) $|u(x, j)| \leq \bar{u} < \infty$ for all $x \in X$ and $j = 0, 1, ..., J$; (ii) under F, the density for $\epsilon_j - \epsilon_d$ is bounded for all $j \neq d$; (iii) under F, $E(|\epsilon_j|)$ is finite for all j. Then, the Emax function and the conditional choice probabilities are locally Lipschitz continuous in F,*

$$\sup_{x} |Q(x; F) - Q(x; \tilde{F})| \leq C \cdot \rho(F, \tilde{F}),$$

$$\sup_{d,x} |p(d|x; F) - p(d|x; \tilde{F})| \leq C' \cdot \rho(F, \tilde{F}),$$

where constants C and C' depend on β, $\bar{u}$ and the bounds on the densities and moments in conditions (ii)-(iii).

The lemma holds irrespective of whether the innocuous normalization $\epsilon_{t0} = 0$ is imposed. Its proof is given in Appendix C.

The following lemma shows that densities satisfying smoothness and finite moment conditions can be approximated by mixtures of Gumbels in distance ρ.

LEMMA 3 *Let f be a density on $\mathbb{R}^I$ satisfying a moment existence condition*

$$\int ||\mu||_2 f(\mu) d\mu < \infty,$$

and a smoothness condition

(12) $|f(z + h) - f(z)| \leq ||h||_2 L_f(z) e^{\tau ||h||_2},$

for some $\tau > 0$ and an envelope function $L_f(\cdot)$ such that

$$(13) \qquad \int (1 + |z_i|) L_f(z) dz < \infty, i = 1, ..., I.$$

Then, for any $\delta > 0$, there exist (m, ω, μ, σ) where $m \in \mathbb{Z}^+$, $\omega_j \in [0,1]$ with $\sum_{j=1}^m \omega_j = 1$, $\mu_j \in \mathbb{R}^I$, $j = 1, ..., m$, and $\sigma > 0$ such that

$$\rho \left(f(\cdot), \sum_{j=1}^m \omega_j \phi(\, \cdot \, ; \mu_j, \sigma) \right) < \delta.$$

We conjecture that the smoothness and tail conditions on f in the lemma can be weakened at the expense of the proof simplicity. The lemma is proved in Appendix C. The proof uses only smoothness and tail conditions on ϕ that are shown to hold for Gumbel densities in Lemmas 6 and 7 in Appendix C. Thus, Lemma 3 holds for more general location-scale mixtures. These generalizations do not seem essential and we do not elaborate on them here for brevity.

The final intermediate result that we need for establishing posterior consistency is the continuity of finite Gumbel mixtures in parameters in distance ρ, which we present in the following lemma.

LEMMA 4   *Let F^1 and F^2 denote two mixtures of Gumbel densities on $\mathbb{R}^J$ with densities $f^i(\epsilon) = \sum_{k=1}^m \omega_k^i \phi \left(\epsilon; \mu_k^i, \sigma^i \right)$. Then, for a given $\delta > 0$ and F_1, there exists $\tilde{\delta} > 0$ such that for any F_2 with parameters satisfying: $|\sigma^1 - \sigma^2| < \tilde{\delta}$, $|\omega_k^1 - \omega_k^2| < \tilde{\delta}$, and $|\mu_k^1 - \mu_k^2| < \tilde{\delta}$, $k = 1, \ldots, m$, we have $\rho(F^1, F^2) < \delta$.*

5.2. *Posterior Consistency*

Let us denote the short panel dataset by $D^n = \{d_{it}, x_{it}, t = 1, \ldots, T, i = 1, \ldots, n\}$; the observations are assumed to be independently identically distributed over i, with a small T and a large n. The utility function is parameterized by a vector $\theta \in \mathbb{R}^{d_\theta}$, $u(x, d; \theta)$. Let $P(\theta, F) = \{p(d|x; \theta, F), x \in X, d = 0, \ldots, J\}$ denote the collection of the CCPs for the distribution of shocks F and parameters θ.

THEOREM 1 *Suppose (i) The observed state space is finite, $X = \{1, \ldots, K\}$; (ii) G^d, $d = 0, \ldots, J$ and the distribution of the initial observed state x_{i1} are known and fixed; (iii) $\forall x \in X$, $\exists t \in \{1, \ldots, T\}$, such that $Pr(x_{it} = x) > 0$; (iv) $u(x, d; \theta)$ is continuous in θ; (v) (θ_0, F_0) are the data generating values of parameters, F_0 satisfies the conditions of Lemma 3; (vi) For any $\delta > 0$, $\Pi(B_\delta(\theta_0)) > 0$, where $B_\delta(\theta_0)$ is a ball with radius δ and center θ_0; (vii) For any $\delta > 0$, positive integer m, $\mu_k \in \mathbb{R}^J$, $\sigma_k > 0$, $w_k \geq 0$, $k = 1, \ldots, m$, $\sum_{k=1}^m w_k = 1$, $\Pi(B_\delta(\mu_1, \sigma_1, \ldots, \mu_m, \sigma_m, w_1, \ldots, w_{k-1})|m) > 0$. Then, for any $\delta > 0$,*

$$\Pi\big(\theta, F : ||P(\theta_0, F_0) - P(\theta, F)|| > \delta \big| D^n\big) \to 0 \ almost \ surely.$$

The theorem shows that the posterior concentrates on the set of parameters and distributions of shocks (θ, F) such that their implied CCPs $P(\theta, F)$ are arbitrarily close to the data generating CCPs $P(\theta_0, F_0)$. To prove this result we use Schwartz (1965) posterior consistency theorem: if the prior puts positive mass on any Kullback-Leibler neighborhood of the data generating distribution then the posterior puts probability converging to 1 on any weak neighborhood of the data generating distribution. Since X is finite, the convergence in weak topology and Kullback-Leibler divergence for distributions on $\{d_{it}, x_{it}, t = 1, \ldots, T\}$ are equivalent to convergence for vectors $\{p(d|x), x \in X, d = 0, \ldots, J\}$ in a euclidean metric when G^d and the distribution of the initial x_{i1} are fixed and satisfy our theorem condition (iii). Thus, to obtain the conclusion of the theorem we only need to establish that the prior puts positive probability on any euclidean neighborhood of $P(\theta_0, F_0)$. First, note that when $u(x, d; \theta)$ is continuous in θ, $P(\theta, F)$ is also continuous in θ in our settings, see, for example Norets (2010); and, thus, Lipschitz continuity of $P(\theta, F)$ in F from Lemma 2 delivers continuity of $P(\theta, F)$ in (θ, F). The finite mixture approximation result in Lemma 3, the continuity of $P(\theta, F)$ in (θ, F), the continuity of finite mixtures in parameters in Lemma 4, and the theorem conditions (vi) and (vii) on the priors, imply a positive prior probability for any neighborhood of $P(\theta_0, F_0)$, and thus, the theorem conclusion.

The finiteness of the observed state space is important for our posterior consistency argument even though our approximation and continuity results in lemmas above do not require it. The finiteness assumption is not limiting in practice as in most applications even continuous

observed states are discretized. The assumption of known G^d can be easily relaxed at the expense of a more involved notation.

Theorem 1 characterizes the support of the posterior in the limit but not its shape, which can also be of interest. Note that the data depend on (θ, F) only through CCPs $P(\theta, F)$ and the posterior for CCPs concentrates at $P(\theta_0, F_0)$. Therefore, the posterior for (θ, F) converges to the conditional prior $\Pi(\theta, F|P)$ at $P = P(\theta_0, F_0)$ under continuity conditions on $\Pi(\theta, F|P)$, see, for example, Plagborg-Møller (2019). As the distribution of shocks is an infinite dimensional object and the solution to the dynamic program does not have a simple explicit form, it appears difficult to characterize the conditional prior $\Pi(\theta, F|P)$, which is implied by the map $P(\theta, F)$ and the prior on (θ, F) . Nevertheless, we can deduce from our approximation and continuity results that under the conditions of Theorem 1, for $\delta > 0$ there exists $\tilde{\delta} > 0$ such that $\theta \in B_{\tilde{\delta}}(\theta_0)$ and $F \in B_{\tilde{\delta}}(F_0)$ imply $P(\theta, F) \in B_\delta(P(\theta_0, F_0))$ and

$$
\Pi\left(\theta \in B_{\tilde{\delta}}(\theta_0), F \in B_{\tilde{\delta}}(F_0) \middle| P \in B_\delta(P(\theta_0, F_0))\right) = \frac{\Pi(\theta \in B_{\tilde{\delta}}(\theta_0), F \in B_{\tilde{\delta}}(F_0))}{\Pi\left(P \in B_\delta(P(\theta_0, F_0))\right)}
$$
$$
\geq \Pi\left(\theta \in B_{\tilde{\delta}}(\theta_0), F \in B_{\tilde{\delta}}(F_0)\right) > 0,
$$

which suggests that the conditional prior would not rule out the data generating parameter values.

5.3. *Exact Matching of CCPs*

In this subsection, we show that for a finite observed state space, our model formulation based on finite mixtures can exactly match the CCPs from a model with an arbitrary distribution of shocks.

LEMMA 5   *Suppose (i) The observed state space is finite, $X = \{1, \ldots, K\}$; (ii) (θ_0, F_0) are the data generating values of parameters; (iii) F_0 has a density that is positive on $\mathbb{R}^J$. Then there exists a finite mixture of Gumbels F such that $P(\theta_0, F_0) = P(\theta_0, F)$.*

The result in Lemma 5 holds not only for mixtures of Gumbels but more generally for location-scale mixtures of distributions with finite first moments, which is evident from the

proof presented in Appendix C. Lemma 5 can be used to weaken the smoothness assumptions on F_0 in the posterior consistency results of Theorem 1 (the approximation results in Lemma 3 can be replaced by exact CCPs matching in Lemma 5). Nevertheless, the approximation results in Lemma 3 have independent value. First, they hold for infinite X. Furthermore, they imply that the prior on the distribution of shocks is flexible in a sense that it puts positive probability on any metric ρ neighborhood in a large nonparametric class of distributions, which suggests that the conditional prior for the distribution of shocks and parameters given CCPs, $\Pi(\theta, F|P)$, is also flexible as discussed at the end of Section 5.2.

5.4. *Inference for Identified Sets*

Norets and Tang (2013) and generalizations of their results to the multinomial case in Norets (2011) show that the utility parameters θ and the distribution of shocks F, and, thus, functions of (θ, F) such as results of counterfactual experiments, are set identified in the present settings. Moon and Schorfheide (2012) show that in contrast to the point identified regular settings, the Bayesian credible sets for set identified parameters do not have frequentist coverage properties and are too small from the classical perspective. Norets and Tang (2013) point out in their Section 3.2 that Bayesian and classical inference results can be reconciled if inference is performed on the identified sets. Kline and Tamer (2016) further study this approach and Kitagawa (2011) obtain related results under multiple priors for set identified parameters. In this subsection, we describe how credible and confidence sets for identified sets can be defined and computed from the output of our MCMC algorithm.

Suppose the data generating values of parameters are (θ_0, F_0) and we are interested in $\eta_0 = g(\theta_0, F_0)$. The data generating values of CCPs, $P_0 = P(\theta_0, F_0)$, can be consistently estimated from the observed data, and, thus, are considered known in the identification analysis. The identified set for η_0 is defined by

$$I_\eta(P_0) = \{\eta = g(\theta, F), \ \forall (\theta, F) \ s.t. \ P(\theta, F) = P_0\}.$$

Following Norets and Tang (2013) and Kline and Tamer (2016), we can define a posterior distribution on the space of identified sets $I_\eta(P)$ using the marginal posterior distribution

on the CCPs P. Then, a $1 - \alpha$-credible set for $I_\eta(P)$ can be defined from a $1 - \alpha$-credible set for CCPs, $B_{1-\alpha}^P$, by

$$(14) \qquad B_{1-\alpha}^{I_\eta} = \bigcup_{P \in B_{1-\alpha}^P} I_\eta(P).$$

When $B_{1-\alpha}^P$ is a $1 - \alpha$ highest posterior density set and the Bernstein - von Mises theorem holds for the point identified CCPs P, then $B_{1-\alpha}^P$ asymptotically has a $1 - \alpha$ frequentist coverage probability for $P(\theta_0, F_0)$ and, thus, $B_{1-\alpha}^{I_\eta}$ has at least a $1 - \alpha$ frequentist coverage probability for the identified set $I_\eta(P_0)$ and η_0.

The sets in (14) might be conservative; nevertheless, they should be reported in applications as in the limit they do not depend on the shape of the prior and possess both frequentist and Bayesian properties.

For a scalar η_0, an approximation to $B_{1-\alpha}^{I_\eta}$ can be computed from a sample of MCMC posterior draws $\{\theta^{(l)}, F^{(l)}, P^{(l)} = P(\theta^{(l)}, F^{(l)}), \ l = 1, \ldots, L\}$ as follows. First, we obtain an approximation to $B_{1-\alpha}^P$

$$\hat{B}_{1-\alpha}^P = \left\{ P : \ (P - \bar{P})' \hat{\Sigma}^{-1} (P - \bar{P}) \leq \chi_{1-\alpha}^2(JK) \right\},$$

where $\bar{P} = \sum_{l=1}^L P^{(l)}/L$, $\hat{\Sigma} = \sum_{l=1}^L (P - \bar{P})(P - \bar{P})'/L$, and $\chi_{1-\alpha}^2(JK)$ is the $1 - \alpha$ quantile of the χ^2 distribution with JK degrees of freedom. Then, $B_{1-\alpha}^{I_\eta}$ is approximated by

$$\hat{B}_{1-\alpha}^{I_\eta} = \left[\min_{l: \ P^{(l)} \in \hat{B}_{1-\alpha}^P} \eta\left(\theta^{(l)}, F^{(l)}\right), \ \max_{l: \ P^{(l)} \in \hat{B}_{1-\alpha}^P} \eta\left(\theta^{(l)}, F^{(l)}\right) \right].$$

We report these sets along with the standard HPD sets in the application in Section 7.

6. APPLICATION I: RUST'S BINARY CHOICE MODEL

Norets and Tang (2013) propose a method for computing identified sets for parameters in dynamic binary choice models and apply their method to the Rust (1987)'s model. In this section, we show that our semiparametric model can also recover the identified set for that model.

6.1. *Rust (1987)'s Optimal Bus Engine Replacement Problem*

In each time period t, the agent decides whether to replace the bus engine ($d_t = 1$) or not ($d_t = 0$) given the current mileage x_t of the bus. Replacing an engine costs θ_0. If $d_t = 0$, then the agent conducts a regular maintenance which costs $-\theta_1 x$. The utility function of the agent is $u(x, 0) = \theta_0 + \theta_1 x$ and $u(x, 1) = \epsilon$. Rust (1987) assumes that ϵ follows the logistic distribution. The mileage $x_t \in \{1, \dots, K = 90\}$ evolves over time following the transition probabilities: $Pr(x_{t+1}|x_t, d_t = 0; \theta) = \theta_2$ for $x_{t+1} - x_t = 0$; $Pr(x_{t+1}|x_t, d_t = 0; \theta) = \theta_3$ for $x_{t+1} - x_t = 1$; $Pr(x_{t+1}|x_t, d_t = 0; \theta) = 1 - \theta_2 - \theta_3$ for $x_{t+1} - x_t = 2$; and $Pr(x_{t+1}|x_t, d_t = 0; \theta) = 0$ otherwise. When the engine is replaced ($d_t = 1$), the mileage restarts at $x_t = 1$.

As in Norets and Tang (2013), we use the following data generating process: logistic distribution for ϵ, $\theta_0 = 5.0727, \theta_1 = -0.002293, \theta_2 = 0.3919, \theta_3 = 0.5953$ and the discount factor $\beta = 0.999$.

At the data generating parameters, we solve the dynamic program to obtain the vector of CCPs, $(p(d|1), \dots, p(d|K))$ for $d = 0, 1$. Rather than using simulated observations in this exercise, we use the true CCPs as the sample frequencies and report the results for different sample sizes. Specifically, for a given $N \in Z^+$, we set n_{dx}, the number of times d was chosen at each state x as follows, $n_{0x} = p(0|x) \times N$ and $n_{1x} = p(1|x) \times N$ for $x = 1, \dots, K$. In this way, we can check if our MCMC algorithm for the semiparametric model specification can recover the identified set computed by the algorithm from Norets and Tang (2013) for a given fixed vector of CCPs. Estimation results for simulated data are presented for the multinational choice application in Section 7.

Since we do not impose a location and scale normalization on the mixture specification for the distribution of ϵ and Rust's model has only two utility parameters that are defined by the location and the scale, the values of (θ_0, θ_1) corresponding to the location and scale of the logistic distribution (used in Rust (1987) and Norets and Tang (2013)) can be computed from the values of $(\sigma, \omega_k, \mu_k, \sigma_k, \ k = 1, \dots, m)$ as described in Appendix E. We use the following flexible prior distributions that are tuned to spread the prior probability over a large region

for (θ_0, θ_1) that includes the identified set.

$$\pi(m = k) \propto e^{-\underline{A}_m k (\log k)^\tau}, \ \underline{A}_m = 0.05, \ \tau = 5,$$

$$(\omega_1, ..., \omega_m) \sim \text{Dirichlet}(\underline{a}/m, ..., \underline{a}/m), \ \underline{a} = 10,$$

$$\mu_k \sim 0.5 N(2.5, 1^2) + 0.5 N(-3, 7^2),$$

$$\log \sigma_k \sim 0.4 N(0, 1^2) + 0.6 N(-6, 1^2), \ \log \sigma \sim N(0, 0.01^2).$$

Below, we report the draws of (θ_0, θ_1) obtained from the draws of $(\sigma, \omega_k, \mu_k, \sigma_k, \ k = 1, \ldots, m)$ for the prior and the posterior for $N \in \{3, 10\}$ and compare them to the true identified set. Panel (a) of Figure 1 shows the prior draws of the utility parameters. First note that the prior draws are not uniformly distributed on the utility parameter space. In practice, it is difficult to come up with a prior for the distribution parameters that implies a uniform prior in the θ space. Second, many prior draws are outside of the identified set. The other two panels in Figure 1 show posterior draws of utility parameters with different number of observations $N \in \{3, 10\}$. The posterior concentrates more on the identified set as N increases. To assess the convergence of the MCMC algorithm, consider Figure 2 showing the

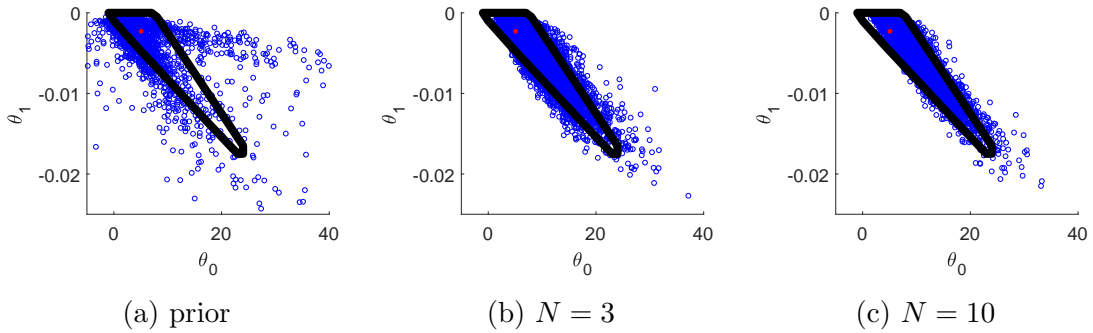| (a) prior | (b) $N = 3$ | (c) $N = 10$ |

Figure 1: Utility parameter draws. Prior (left). Posterior draws for $N = 3$ (center) and $N = 10$ (right). 500,000 MCMC iterations. The first 100,000 draws discarded as burn-in. Every 10th draw is shown. The true identified set is shown by solid black lines. The red dot corresponds to the point-identified utility parameter values under the logistic distribution for ϵ.

draws of utility parameters for 500,000 MCMC iterations. As can be seen from the figure, the chain sweeps through the identified set repeatedly during the MCMC run. Thus, we

conclude that our approach can be used to recover the identified sets of utility parameters. Figure 6 in Appendix E shows additional evidence of MCMC convergence including trace plots of m, $\sum_{k=1} \omega_k \mu_k$, and other parameters.
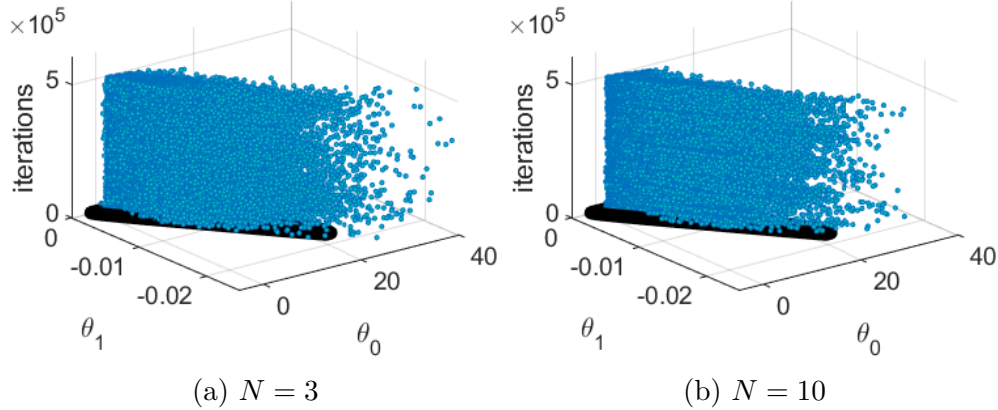

(a) $N = 3$ \qquad (b) $N = 10$

Figure 2: Utility parameter draws. The z-axis represents the MCMC iterations. The true identified set is shown in black.

7. APPLICATION II: GILLESKIE'S MULTINOMIAL MODEL OF MEDICAL CARE USE AND WORK ABSENCE

In this section, we illustrate our methodology using a multinomial choice model of medical care use and work absence from Gilleskie (1998). An extension of Norets and Tang (2013) to the multinomial case seems computationally infeasible and, hence, in this application we provide comparisons of our method only with a dynamic logit specification.

7.1. *The model*

In the model, individuals occupy one of $K + 1$ distinct health states; well, $k = 0$, or sick with an illness type $k \in \{1, \ldots, K\}$. An individual receives the utility associated with being well until contracting an illness of a specific type. We make the simplifying assumption that there is only one illness type, $(K = 1)$, although Gilleskie (1998) works with two illness types, $K = 2$. An illness episode lasts for T periods and enumerated by $t = 1, \ldots, T$. $t = 0$ corresponds to the state of being well, $k = 0$.

7.1.1. *Alternatives*

An individual who became sick makes decisions about doctor visits and and work absences. In each period t of an illness, alternatives available to an employed individual who is sick are: $d_t = 0$ - work and don't visit a doctor, $d_t = 1$ - work and visit a doctor, $d_t = 2$ - don't work and don't visit a doctor, and $d_t = 3$ - don't work and visit a doctor. The utility of the agent depends on the elapsed length of the current illness t, the accumulated number of physician visits v_t, and the accumulated number of work absences a_t. The state variables observed by the econometrician and the agent at t are $x_t = (t, v_t, a_t)$. Note that $k = 1$ if and only if $t > 0$, so k does not appear in the definition of x_t.

7.1.2. *State variable transitions*

The state variables evolve in the following way. An individual always starts with the state of being well, $x_0 = (0, 0, 0)$. The individual contracts an illness and moves to the state $x = (1, 0, 0)$ with probability $\pi^S(H)$, where H is a vector of exogenous indicators for health status and being between 45-64 years of age.

The accumulated number of physician visits v_t and the accumulated number of illness-related absence from work a_t both take values in $\{0, 1, \ldots, T-1\}$. They start from $v_1 = a_1 = 0$ and evolve in the following way: $v_{t+1} = v_t + 1(d_t = 1 \text{ or } 3)$ and $a_{t+1} = a_t + 1(d_t = 2 \text{ or } 3)$.

In each illness period $t \in \{1, \ldots, T\}$, the individual recovers and returns to the state of being well with probability $\pi^W(x_t, d_t)$. Gilleskie parametrizes and estimates $\pi^W(x_t, d_t)$ and $\pi^S(H)$ prior to estimating the preference parameters. We use those estimate in our application.

7.1.3. *Utility*

The per-period consumption is defined as $C(x_t, d_t) = Y - \big[PC1(d_t = 1 \text{ or } 3) + Y\big(1 - L\Phi(x_t, d_t)\big)1(d_t = 2 \text{ or } 3)\big]1(t > 0)$, where Y is the per-period labor income, PC is the cost of a medical visit, and $\Phi(x_t, d_t) = \exp(\phi_1 + \phi_2 a'(x_t, d_t))/[1 + \exp(\phi_1 + \phi_2 a'(x_t, d_t))]$ is the portion of income that sick leave coverage replaces, where $a'(x_t, d_t)$ is the value of a_{t+1} given (x_t, d_t). $L \in (0, 1)$ is the sick leave coverage rate.

The per-period utilities can be expressed in the following form.

$$u(d_t = 1, x_t, \epsilon_t) = \theta_1 + \theta_4 1(t = 0) + \theta_6 C(x_t, 1) 1(t > 0) + \epsilon_{t1},$$
$$u(d_t = 2, x_t, \epsilon_t) = \theta_2 + \theta_4 1(t = 0) + \theta_6 C(x_t, 2) 1(t > 0) + \epsilon_{t2},$$
$$u(d_t = 3, x_t, \epsilon_t) = \theta_3 + \theta_4 1(t = 0) + \theta_6 C(x_t, 3) 1(t > 0) + \epsilon_{t3},$$
$$u(d_t = 0, x_t, \epsilon_t) = \qquad \theta_5 1(t = 0) + \theta_6 C(x_t, 0) 1(t > 0) + \epsilon_{t0},$$

where $\theta_4 = -\infty$ so that when $t = 0$ the decision $d_t = 0$ is always chosen. Since we do not restrict the location and scale of ϵ_t, the values θ_p, $p = 1, 2, 3$ can be set to arbitrary values and θ_5 can be set to an arbitrary positive value in our semiparametric estimation procedure. See Appendix F.2 for more detail on per-period utilities.

7.2. *Estimation*

For data generation we use parameter values based on estimates in Gilleskie (1998) for Type 2 illness with some adjustments so that the expected number of doctor visits and work absences roughly match with Gilleskie's sample. We let the data generating distribution of the utility shocks to be a two-component mixture of extreme value distributions: $\sum_{k=1}^{2} \omega_k \phi(\epsilon; \mu_k, \sigma_k)$. See Appendix F.3 for the data-generating parameter values. The panel data $\{x_{it}, d_{it}, i = 1, \ldots, n, t = 1, \ldots, T\}$ is sequentially simulated for $n = 100$ individuals and $T = 8$ periods.

The priors are specified as follows, $\underline{a} = 10$, $A_m = 0.05$, and $\tau = 5$, $\mu_{jk} \sim N(0, 2^2)$, $\log \sigma_k \sim N(0, 1)$, $\log \sigma \sim N(0, 0.01^2)$, and $\theta_6 \sim N(0, 4^2)$. This gives normal prior on μ_{jk}'s and θ_6 with large variances. The log-normal prior on the component specific scale parameters also implies sufficiently large prior probabilities for large values of σ_k's. Prior sensitivity checks presented in Appendix F.5 show that the obtained estimation results are not substantively affected by moderate changes in the prior. Appendix F.6 shows results when extreme value distribution of shocks is used for the data generation.

7.2.1. *Estimation Results and Counterfactuals*

We use 20,000 MCMC iterations to explore the posterior distribution. Figure 3 shows a trace plot and a p.m.f. of the number of mixture components m. The posterior has its peak at $m = 2$ and the posterior probability of $m = 1$ is small.
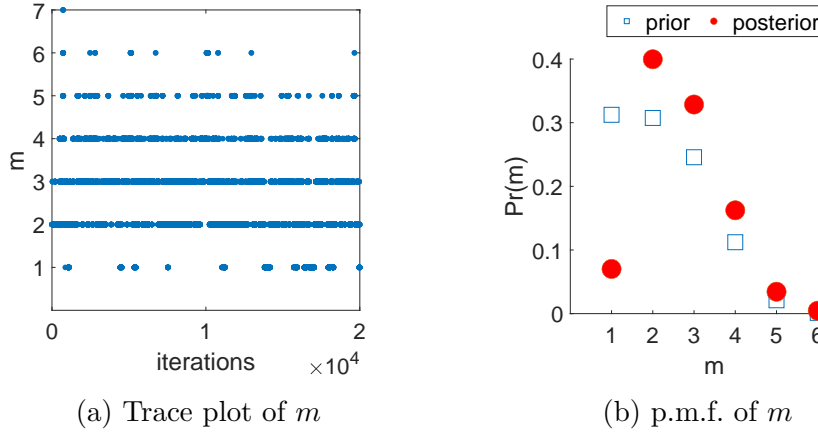

(a) Trace plot of m (b) p.m.f. of m

Figure 3: Trace plot and p.m.f. of m

Figure 4 shows the posterior densities of the utility function parameters in the location and scale normalization corresponding to the original model in Gilleskie (1998) described in Section 4.2. The corresponding trace plots presented in Figure in Appendix F.4 provide evidence that the MCMC algorithm converged.

The standard parametric approach that Gilleskie takes is to assume that the shocks are extreme value i.i.d. and to estimate the model by the maximum likelihood method. In Figure 5 and Table I, we compare estimation results from our semiparametric method to those from the MLE. For the comparison, we use the expected number of doctor visits $E(v)$ implied by the model. It is a function of the model parameters and is of interest in the application. One of the main advantages of structural estimation is that it provides an attractive framework for counterfactual experiments. We consider a counterfactual experiment presented in Section 6.2 of Gilleskie's paper (Experiment 1). In this experiment, we are interested in the behavior of individuals when the coinsurance rate paid out of pocket is set to zero. Thus, we examine the counterfactual model solution when $PC = 0$ and the transition probabilities and (θ, F)
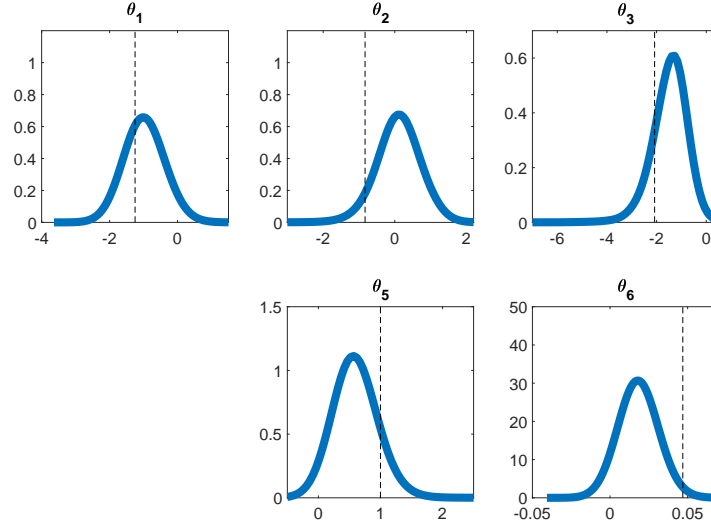
Figure 4: Posterior densities of utility parameters (solid) and the data generating values (dashed).

are unchanged. The last row of Table I and panel (b) in Figure 5 display the estimation results for $E(v)$ in the counterfactual environment.

In addition to the posteriors, the true values, the HPD credible intervals, and the MLE confidence intervals computed by the Delta method, Figure 5 and Table I present credible intervals for the identified sets of $E(v)$, $\hat{B}_{0.95}^{I_{E(v)}}$, that are introduced in Section 5.4. While the point estimates of the increase in $E(v)$ are similar for both approaches, the 95% Bayesian credible interval for the identified set of the counterfactual $E(v)$ in the semiparametric model is up to 10 times wider than the 95% confidence interval for the MLE in the parametric settings. Importantly, the confidence interval by far misses the true counterfactual value of $E(v)$, while the credible interval for the identified set includes the true value and the HPD credible interval gets very close to it.

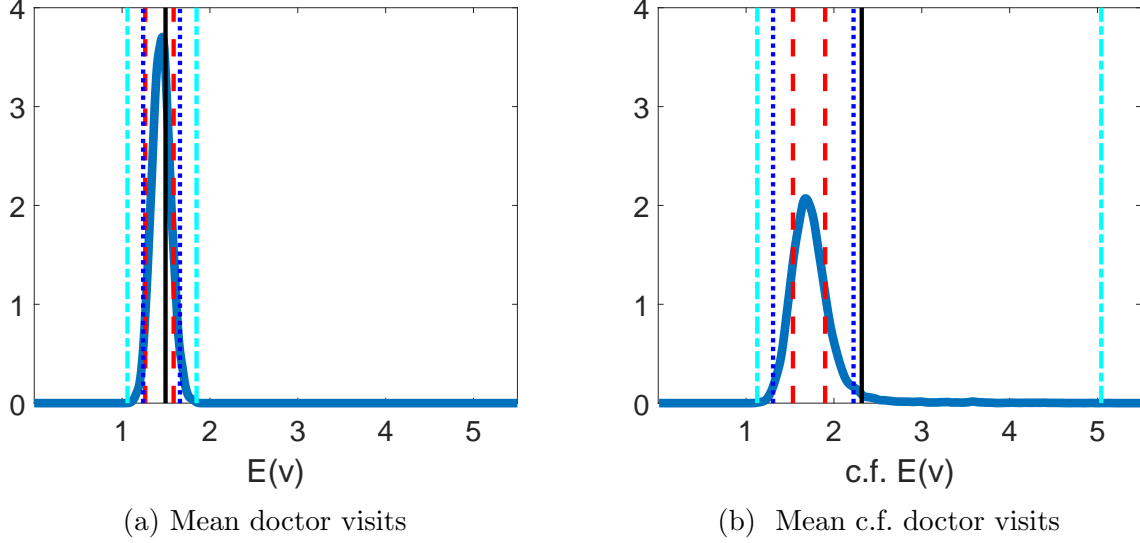(a) Mean doctor visits   (b) Mean c.f. doctor visits

Figure 5: Actual and counterfactual (c.f.) expected number of doctor visits. Posterior density (blue solid), 95% confidence interval (red dashed), 95% HPD interval (blue dotted), $\hat{B}_{0.95}^{I_{E(v)}}$ (light blue dash-dotted), and data generating value (black solid).

		MLE under logit			Semiparametric Bayes	
	True	Est.	95%CI (length)	Est.	95%HPD (length)	$\hat{B}_{0.95}^{I_{E(v)}}$ (length)
$E(v)$	1.49	1.42	[1.26, 1.58] (0.32)	1.44	[1.23, 1.65] (0.41)	[1.06, 1.84] (0.78)
$c.f.E(v)$	2.31	1.71	[1.53, 1.89] (0.36)	1.75	[1.30, 2.22] (0.91)	[1.12, 5.04] (3.91)

Table I: Estimated (Est.) $E(v)$ and $c.f.E(v)$: the MLE with its 95% confidence interval and the posterior mean with the 95% HPD interval and $\hat{B}_{0.95}^{I_{E(v)}}$.

These results illustrate the following general observations. First, the dynamic logit MLE can deliver misleading point and set estimates, especially for counterfactuals, when the shocks are not extreme value distributed. Second, as is well know from Moon and Schorfheide (2012), the HPD credible intervals for set identified parameters might be too short at least from the classical perspective. Third, the credible intervals for the identified sets introduced in Section 5.4 are more conservative than the standard HPD intervals and might be preferred under set identification. Finally, and most importantly, the whole posterior distribution (and not just

point and set estimators) should be reported and taken into account in making decisions or policy recommendations.

8. CONCLUSION

In this paper, we propose and implement a semiparametric Bayesian estimation method for dynamic discrete choice models that uses flexible mixture specifications for modeling the distribution of unobserved state variables. We establish approximation and posterior consistency results that provide frequentist asymptotic guarantees for the method. Our approach is shown to perform well in practice for binary and mutlinomial choice models. The proposed framework is a robust and computationally tractable semiparametric alternative to the standard dynamic logit model; it provides more reliable inference, especially for counterfactuals.

REFERENCES

ABBRING, J. H. AND O. DALJORD (2020): "Identifying the discount factor in dynamic discrete choice models," *Quantitative Economics*, 11, 471–501.

AGUIRREGABIRIA, V. (2010): "Another Look at the Identification of Dynamic Discrete Decision Processes: An Application to Retirement Behavior," *Journal of Business and Economic Statistics*, 28, 201–218.

AGUIRREGABIRIA, V. AND P. MIRA (2010): "Dynamic discrete choice structural models: A survey," *Journal of Econometrics*, 156, 38 – 67, structural Models of Optimization Behavior in Labor, Aging, and Health.

BHATTACHARYA, R. AND M. MAJUMDAR (1989): "Controlled Semi-Markov Models - The Discounted Case," *Journal of Statistical Planning and Inference*, 365–381.

BUCHHOLZ, N., M. SHUM, AND H. XU (2020): "Semiparametric estimation of dynamic discrete choice models," *Journal of Econometrics*.

CHEN, L.-Y. (2017): "Identification of Discrete Choice Dynamic Programming Models with Nonparametric Distribution of Unobservables," *Econometric Theory*, 33, 551–577.

FRUHWIRTH-SCHNATTER, S. (2006): *Finite Mixture and Markov Switching Models (Springer Series in Statistics)*, Springer, 1 ed.

GEWEKE, J. (2005): *Contemporary Bayesian Econometrics and Statistics*, Wiley-Interscience.

——— (2007): "Interpretation and inference in mixture models: Simple MCMC works," *Computational Statistics and Data Analysis*, 51, 3529 – 3550.

GILLESKIE, D. (1998): "A dynamic stochastic model of medical care use and work absence," *Econometrica*, 6, 1–45.

KALOUPTSIDI, M., P. T. SCOTT, AND E. SOUZA-RODRIGUES (2021): "Identification of counterfactuals in dynamic discrete choice models," *Quantitative Economics*, 12, 351–403.

KITAGAWA, T. (2011): "Estimation and Inference for Set-identified Parameters Using Posterior Lower Probability," Unpublished manuscript, UCL.

KLINE, B. AND E. TAMER (2016): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, 7, 329–366.

MAGNAC, T. AND D. THESMAR (2002): "Identifying Dynamic Discrete Decision Processes," *Econometrica*, 70, 801–816.

MANSKI, C. F. (1999): *Identification Problems in the Social Sciences*, Harvard University Press.

MCLACHLAN, G. AND D. PEEL (2000): *Finite Mixture Models*, John Wiley & Sons, Inc.

MOON, H. R. AND F. SCHORFHEIDE (2012): "Bayesian and Frequentist Inference in Partially Identified Models," *Econometrica*, 80, 755–782.

NEAL, R. M. (2012): "MCMC using Hamiltonian dynamics. arXiv e-prints, page," *arXiv preprint arXiv:1206.1901*.

NORETS, A. (2010): "Continuity and Differentiability of Expected Value Functions in Dynamic Discrete Choice Models," *Quantitative economics*, 1.

——— (2011): "Semiparametric identification of dynamic multinomial choice models," Unpublished Manuscript, Princeton University.

——— (2021): "Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension Models," *Econometric Theory*, 37, 49–81.

NORETS, A. AND J. PELENIS (2021): "Adaptive Bayesian Estimation of Mixed Discrete-Continuous Distributions under Smoothness and Sparsity," *Econometrica*, forthcoming.

NORETS, A. AND X. TANG (2013): "Semiparametric inference in dynamic binary choice models," *Review of Economic Studies*, 81, 1229–1262.

PLAGBORG-MØLLER, M. (2019): "Bayesian Inference on Structural Impulse Response Functions," *Quantitative Economics*, 10, 145–184, replication files (Matlab code and data).

ROUSSEAU, J. (2010): "Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density," *The Annals of Statistics*, 38, 146–180.

RUST, J. (1987): "Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher," *Econometrica*, 55, 999–1033.

——— (1994): "Structural Estimation of Markov Decision Processes," in *Handbook of Econometrics*, ed. by R. Engle and D. McFadden, North Holland.

SCHWARTZ, L. (1965): "On Bayes procedures," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4, 10–26.

SHEN, W., S. TOKDAR, AND S. GHOSAL (2013): "Adaptive Bayesian multivariate density estimation with

Dirichlet mixtures," *Biometrika*, 100, 623–640.

SU, C.-L. AND K. L. JUDD (2012): "Constrained Optimization Approaches to Estimation of Structural Models," *Econometrica*, 80, 2213–2230.

APPENDIX A: CLOSED FORM EXPRESSIONS FOR CCPS AND EMAX

PROOF OF LEMMA 1: Note that for $E1(z) = \int_z^\infty e^{-t}/t \, dt$, $-E1(z) = Ei(-z)$ for positive $z > 0$, where $Ei(z) = -\int_{-z}^\infty e^{-t}/t \, dt$ denotes the exponential integral function.

The mixture distribution in (7) can be represented as $\epsilon | Z = k \sim \phi(\epsilon; \mu_k, \sigma_k)$, where the random variable Z indicates a mixture component, $Pr(Z = k) = \omega_k$. Then,

$$p(d) = \sum_{k=1}^m \omega_k p(d|Z = k) \text{ and } E\left[\max_{j=0,1,\ldots,J} v_j + \epsilon_j\right] = \sum_{k=1}^m \omega_k E\left[\max_{j=0,1,\ldots,J} v_j + \epsilon_j | Z = k\right],$$

where the dependence on the observed state x is suppressed to simplify the notation.

For each $k = 1, \ldots, m$, $Pr(d = 0 | Z = k)$ equals to

$$Pr(\epsilon_j \leq v_0 - v_j \forall j = 1, \ldots, J | Z = k) = \prod_{j=1}^J \exp\left[-e^{-\left(\frac{v_0 - v_j - \mu_{jk}}{\sigma_k}\right)} - e^{-\gamma}\right]$$

$$= \exp\left[-e^{-\gamma} e^{-\frac{v_0}{\sigma_k}} \sum_{j=1}^J e^{\frac{v_j + \mu_{jk}}{\sigma_k}}\right] = \exp\left[-e^{-\gamma} e^{-\frac{v_0}{\sigma_k}} e^{A_k}\right] = \exp[-e^{-a_k}].$$

For $d \neq 0$, for each $k = 1, \ldots, m$, note that $P(d|Z = k) = \int p(d|Z = k, \epsilon_d) f(\epsilon_d | Z = k) d\epsilon_d$ and

$$p(d|Z = k, \epsilon_d) = Pr\left(v_d + \epsilon_d \geq v_j + \epsilon_j \forall j \neq d | Z = k, \epsilon_d\right)$$

$$= Pr\left(0 \leq \epsilon_d + v_d - v_0 \text{ and } \epsilon_j \leq \epsilon_d + v_d - v_j \forall j \neq 0, d | Z = k, \epsilon_d\right)$$

$$= Pr\left(\epsilon_j \leq \epsilon_d + v_d - v_j \forall j \neq 0, d | Z = k, \epsilon_d\right) 1\left(0 \leq \epsilon_d + v_d - v_0\right)$$

$$= \prod_{j=1, j\neq d}^J \exp\left[-e^{-\left(\frac{\epsilon_d + v_d - v_j - \mu_{jk}}{\sigma_k}\right)} e^{-\gamma}\right] 1\left(0 \leq \epsilon_d + v_d - v_0\right)$$

$$= \prod_{j=1}^J \exp\left[-e^{-\left(\frac{\epsilon_d + v_d - v_j - \mu_{jk}}{\sigma_k}\right)} e^{-\gamma}\right] \exp\left[e^{-\left(\frac{\epsilon_d - \mu_{dk}}{\sigma_k}\right)} e^{-\gamma}\right] 1\left(0 \leq \epsilon_d + v_d - v_0\right).$$

Hence, $p(d|Z = k)$ equals to

$$
= \int \prod_{j=1}^{J} \exp\left[-e^{-\left(\frac{s+v_d-v_j-\mu_{jk}}{\sigma_k}\right)}e^{-\gamma}\right] \exp\left[e^{-\left(\frac{s-\mu_{dk}}{\sigma_k}\right)}e^{-\gamma}\right] 1\,(0 \le s + v_d - v_0)
$$

$$
\times \frac{1}{\sigma_k}e^{-\left(\frac{s-\mu_{dk}}{\sigma_k}\right)}e^{-\gamma}\exp\left[-e^{-\left(\frac{s-\mu_{dk}}{\sigma_k}\right)}e^{-\gamma}\right]ds
$$

$$
= \int \prod_{j=1}^{J} \exp\left[-e^{-\left(\frac{s+v_d-v_j-\mu_{jk}}{\sigma_k}\right)}e^{-\gamma}\right] 1\,(0 \le s + v_d - v_0) \times \frac{1}{\sigma_k}e^{-\left(\frac{s-\mu_{dk}}{\sigma_k}\right)}e^{-\gamma}ds
$$

$$
= \int_{s=-\infty}^{s=\infty} \prod_{j=1}^{J} \exp\left[-e^{-\left(\frac{s-\mu_{dk}}{\sigma_k}\right)}e^{-\left(\frac{v_d+\mu_{dk}}{\sigma_k}\right)}e^{\left(\frac{u_j+\mu_{jk}}{\sigma_k}\right)}e^{-\gamma}\right] 1\,(0 \le s + v_d - v_0)
$$

$$
\times \frac{1}{\sigma_k}e^{-\left(\frac{s-\mu_{dk}}{\sigma_k}\right)}e^{-\gamma}ds.
$$

Let $t = e^{\frac{s-\mu_{dk}}{\sigma_k}}$. Then $ds = -\sigma_k\frac{1}{t}dt$. Note that $0 \le s + v_d - v_0 \iff \frac{v_0-v_d-\mu_{dk}}{\sigma_k} \le \frac{s-\mu_{dk}}{\sigma_k} \iff e^{-\left(\frac{v_0-v_d-\mu_{dk}}{\sigma_k}\right)} \ge t$ and that $s = \infty \iff t = 0, s = -\infty \iff t = \infty$. Now

$$
p(d|Z = k) = \int_{t=\infty}^{t=0} \prod_{j=1}^{J} \exp\left[-te^{-\left(\frac{v_d+\mu_{dk}}{\sigma_k}\right)}e^{\left(\frac{v_j+\mu_{jk}}{\sigma_k}\right)}e^{-\gamma}\right] 1\left(t \le e^{-\left(\frac{v_0-v_d-\mu_{dk}}{\sigma_k}\right)}\right)
$$

$$
\times \frac{1}{\sigma_k}te^{-\gamma}\left(-\sigma_k\frac{1}{t}dt\right)
$$

$$
= e^{-\gamma}\int_{t=0}^{t=e^{-\left(\frac{v_0-v_d-\mu_{dk}}{\sigma_k}\right)}} \exp\left[-te^{-\left(\frac{v_d+\mu_{dk}}{\sigma_k}\right)}e^{-\gamma}\sum_{j=1}^{J}e^{\left(\frac{v_j+\mu_{jk}}{\sigma_k}\right)}\right]dt
$$

$$
= e^{-\gamma}\frac{\exp\left[-te^{-\left(\frac{v_d+\mu_{dk}}{\sigma_k}\right)}e^{-\gamma}\sum_{j=1}^{J}e^{\left(\frac{v_j+\mu_{jk}}{\sigma_k}\right)}\right]}{-e^{-\left(\frac{v_d+\mu_{dk}}{\sigma_k}\right)}e^{-\gamma}\sum_{j=1}^{J}e^{\left(\frac{v_j+\mu_{jk}}{\sigma_k}\right)}}\Bigg|_{t=0}^{t=e^{-\left(\frac{v_0-v_d-\mu_{dk}}{\sigma_k}\right)}}
$$

$$
= \frac{-e^{\left(\frac{v_d+\mu_{dk}}{\sigma_k}\right)}}{\sum_{j=1}^{J}e^{\left(\frac{v_j+\mu_{jk}}{\sigma_k}\right)}}\left\{\exp\left[-e^{-\left(\frac{v_0-v_d-\mu_{dk}}{\sigma_k}\right)}e^{-\left(\frac{v_d+\mu_{dk}}{\sigma_k}\right)}e^{-\gamma}\sum_{j=1}^{J}e^{\left(\frac{v_j+\mu_{jk}}{\sigma_k}\right)}\right] - 1\right\}
$$

$$
= \frac{e^{\left(\frac{v_d+\mu_{dk}}{\sigma_k}\right)}}{\sum_{j=1}^{J}e^{\left(\frac{v_j+\mu_{jk}}{\sigma_k}\right)}}\left\{1 - \exp\left[-e^{-\gamma}e^{-\left(\frac{v_0}{\sigma_k}\right)}e^{-A_k}\right]\right\} = \frac{e^{\left(\frac{v_d+\mu_{dk}}{\sigma_k}\right)}}{\sum_{j=1}^{J}e^{\left(\frac{v_j+\mu_{jk}}{\sigma_k}\right)}}\left[1 - \exp[-e^{-a_k}]\right].
$$

We see that the probabilities sum to 1,

$$\sum_{d=0}^{J} P(d|Z=k) = \exp[-e^{-a_k}] + \sum_{d=1}^{J} \frac{e^{\left(\frac{v_d+\mu_{dk}}{\sigma_k}\right)}}{\sum_{j=1}^{J} e^{\left(\frac{v_j+\mu_{jk}}{\sigma_k}\right)}} \left[1 - \exp[-e^{-a_k}]\right] = 1.$$

This proves (8). To prove (9), we have for each $k = 1, ..., m$,

$$E\left[\max_{j=0,1,...,J} v_j + \epsilon_j | Z = k\right] =$$

$$E\left[\max_{j=0,1,...,J} v_j + \epsilon_j | \max_{j=1,...,J} v_j + \epsilon_j \le v_0, Z = k\right] Pr\left[\max_{j=1,...,J} v_j + \epsilon_j \le v_0 | Z = k\right]$$

$$+E\left[\max_{j=0,1,...,J} v_j + \epsilon_j | \max_{j=1,...,J} v_j + \epsilon_j > v_0, Z = k\right] Pr\left[\max_{j=1,...,J} v_j + \epsilon_j > v_0 | Z = k\right].$$

First, note that $Pr\left[\max_{j=1,...,J} v_j + \epsilon_j \le v_0 | Z = k\right] = Pr\left[v_j + \epsilon_j \le v_0, \forall j = 1, ..., J | Z = k\right]$

$$= \prod_{j=1}^{J} \exp\left[-e^{-\left(\frac{v_0-v_j-\mu_{jk}}{\sigma_k}\right)} e^{-\gamma}\right] = \exp\left[-e^{-\gamma} \sum_{j=1}^{J} e^{-\left(\frac{v_0-v_j-\mu_{jk}}{\sigma_k}\right)}\right].$$

Note that $\sum_{j=1}^{J} e^{-\left(\frac{v_0-v_j-\mu_{jk}}{\sigma_k}\right)} = e^{-\frac{v_0}{\sigma_k}} \sum_{j=1}^{J} e^{-\frac{v_j+\mu_{jk}}{\sigma_k}} = e^{-\frac{v_0}{\sigma_k}} e^{A_k} = \exp\left[-\frac{v_0}{\sigma_k} + A_k\right]$

$= \exp\left[-\frac{v_0-A_k\sigma_k}{\sigma_k}\right]$. Hence, $Pr\left[\max_{j=1,...,J} v_j + \epsilon_j \le v_0 | Z = k\right] = \exp\left[-e^{\left(-\frac{v_0-A_k\sigma_k}{\sigma_k}\right)} e^{-\gamma}\right] =$

$\exp[-e^{-a_k}]$. This means, $\max_{j=1,...,J} v_j + \epsilon_j \le v_0 | Z = k \sim \phi\left(\cdot; A_k\sigma_k, \sigma_k\right)$. Next, we have

$$E\left[\max_{j=0,1,...,J} v_j + \epsilon_j | \max_{j=1,...,J} v_j + \epsilon_j > v_0, k\right]$$

$$= E\left[\max_{j=1,...,J} v_j + \epsilon_j | \max_{j=1,...,J} v_j + \epsilon_j > v_0, k\right] = \int_{v_0}^{\infty} \frac{y f_{\max_{j=1,...,J} v_j+\epsilon_j}(y|Z=k)}{Pr\left[\max_{j=1,...,J} v_j + \epsilon_j > v_0\right]} dy.$$

Note that $\int_{v_0}^{\infty} y f_{\max_{j=1,...,J} v_j+\epsilon_j}(y|Z=k)dy = \int_{v_0}^{\infty} y \frac{1}{\sigma_k} e^{\left(-\frac{v_0-A_k\sigma_k}{\sigma_k}\right)} e^{-\gamma} \left[-e^{\left(-\frac{v_0-A_k\sigma_k}{\sigma_k}\right)} e^{-\gamma}\right] dy$.

Let $z = \gamma + \frac{y}{\sigma_k} - A_k$. Then $dz = \frac{1}{\sigma_k} dy$. Note that $y = v_0 \to z = \gamma + \frac{v_0}{\sigma_k} - A_k = a_k$ and

$y = \infty \to z = \infty$. Note that $y = \sigma_k(z - \gamma + A_k)$. We have

$$\int_{v_0}^{\infty} y f_{\max_{j=1,...,J} v_j+\epsilon_j}(y|Z=k)dy = \int_{a_k}^{\infty} \sigma_k(z-\gamma+A_k) \frac{1}{\sigma_k} e^{-z} \exp\left[-e^{-z}\right] (\sigma_k dz)$$

$$= \sigma_k \int_{a_k}^{\infty} z e^{-z} \exp\left[-e^{-z}\right] dz + \sigma_k(A_k - \gamma) \int_{a_k}^{\infty} e^{-z} \exp\left[-e^{-z}\right] dz$$

$$= \sigma_k \left[\gamma - a_k \exp\left(-e^{-a_k}\right) + E1(e^{-a_k})\right] + \sigma_k(A_k - \gamma)\left[1 - \exp\left(-e^{-a_k}\right)\right].$$

Finally, $E\left[\max_{j=0,1,\ldots,J} v_j + \epsilon_j | Z = k\right] = v_0 \exp\left(-e^{-a_k}\right) + \int_{v_0}^{\infty} y f_{\max_{j=1,\ldots,J} v_j + \epsilon_j}(y | Z = k) dy$

$$= v_0 \exp\left(-e^{-a_k}\right) +$$

$$\sigma_k\left[\gamma - a_k \exp\left(-e^{-a_k}\right) + E1(e^{-a_k}) + A_k - \gamma - A_k \exp\left(-e^{-a_k}\right) + \gamma \exp\left(-e^{-a_k}\right)\right]$$

$$= \sigma_k \left[\exp\left(-e^{-a_k}\right)\left(\frac{v_0}{\sigma_k} + \gamma - A_k - a_k\right) + E1(e^{-a_k}) + A_k\right]$$

$$= \sigma_k \left[A_k + E1(e^{-a_k})\right].$$

This proves (9). $Q.E.D.$

APPENDIX B: MCMC ALGORITHM

For the application of a reversible jump algorithm, we need to transform the mixing weights into unnormalized weights γ_k, $k = 1, \ldots$, so that conditional on m, $\omega_k = \gamma_k / \sum_{l=1}^{m} \gamma_l$ and the Dirichlet prior on $(\omega_1, \ldots, \omega_m)$ corresponds to a gamma prior for the unnormalized weights: $\gamma_k | m \sim Gamma(\underline{a}/m, 1)$, $k = 1, \ldots, m$. Let $\psi_k = (\mu_k, \tilde{\sigma}_k, \gamma_k)$ and $\psi_{1m} = (\theta, \sigma, \psi_1, \ldots, \psi_m)$, where θ includes model parameters such as coefficients in the utility functions. With this notation, the likelihood function is denoted by $p(D_n | m, \psi_{1m})$. Our MCMC algorithm alternates the following two blocks: (i) HMC for $\Pi(\psi_{1m} | m, D_n)$ and (ii) an optimal reversible jump for m. The following subsections provide details for each block.

B.1. *Optimal Reversible Jump from Norets (2021)*

The following short description of the reversible jump algorithm is adapted from Norets and Pelenis (2021), see Norets (2021) for more details. Denote a proposal distribution for the parameter of a new mixture component $m+1$ by $\tilde{\pi}_{m+1}(\psi_{m+1} | D_n, \psi_{1m})$. The algorithm works as follows. Simulate proposal m^* from $Pr(m^* = m + 1 | m) = Pr(m^* = m - 1 | m) = 1/2$. If $m^* = m + 1$, then also simulate $\psi_{m+1} \sim \tilde{\pi}_{m+1}(\psi_{m+1} | D_n, \psi_{1m})$. Accept the proposal with

probability $\min\{1, \alpha(m^*, m)\}$, where

$$\alpha(m^*, m) = \frac{p(D_n|m^*, \psi_{1m^*})\Pi(\psi_{1m^*}|m^*)\Pi(m^*)}{p(D_n|m, \psi_{1m})\Pi(\psi_{1m}|m)\Pi(m)}$$

(15)
$$\cdot \left(\frac{1\{m^* = m + 1\}}{\tilde{\pi}_m(\psi_{m+1}|\psi_{1m}, Y)} + 1\{m^* = m - 1\}\tilde{\pi}_{m-1}(\psi_m|\psi_{1m-1}, D_n) \right).$$

Norets (2021) shows that an optimal choice of proposal $\tilde{\pi}_m$ is the conditional posterior $p(\psi_{m+1}|D_n, m+1, \psi_{1m})$. The conditional posterior can be evaluated up to a normalization constant; however, it seems hard to directly simulate from it and compute the required normalization constant. Hence, we use a Gaussian approximation to $p(\psi_{m+1}|D_n, m+1, \psi_{1m})$ as the proposal (with the mean equal to the conditional posterior mode, obtained by a Newton method, and the variance equal to the inverse of the negative of the Hessian evaluated at the mode).

B.2. *HMC and Transformation of Parameters*

We utilize a built-in HMC package in Matlab. The package requires the parameters to be unbounded. Recall that the vector of parameters is $\psi_{1m} = \left(\theta, \sigma, \omega_k, \mu_k, ...\mu_{Jk}, \tilde{\sigma}_k, \ k = 1, ...\right)$, where the weights and the scales are not unbounded. For a given fixed m, we denote the transformed parameters by $\chi = (\theta, \log \sigma, \alpha_k, \mu_k, \log \tilde{\sigma}_k, \ k = 1, ..., m)$, where

$$\omega_k = \frac{e^{\alpha_k}}{1 + \sum_{\ell=1}^{m-1} e^{\alpha_\ell}}, k = 1, ..., m-1, \ \text{and} \ \alpha_m = 0$$

so that all the components of χ are unbounded.

We only need to supply the Matlab HMC algorithm with a function that evaluates the log of the likelihood times prior and its gradient. The package can choose HMC's parameters, such as a step size, automatically, and we perform this automatic initialization once for each value of m that we encounter in the MCMC run. The form of the prior for the transformed mixing weights and the derivatives of the likelihood used in the algorithm are reported in Appendix D.

APPENDIX C: PROOFS AND INTERMEDIATE RESULTS

PROOF OF LEMMA 2:   We first show Lipschitz continuity of the Emax function. Suppose there are two distributions F_1 and F_2. Define the corresponding Emax functions $Q_i(x) = Q(x; F_i)$.

$$Q_2(x) - Q_1(x) = \left(Q_2(x) - \int \max_d \left[u(x, d) + \beta G_x^d(Q_1) + \epsilon_d \right] dF_2(\epsilon) \right)$$
$$+ \left(\int \max_d \left[u(x, d) + \beta G_x^d(Q_1) + \epsilon_d \right] dF_2(\epsilon) - Q_1(x) \right).$$

The term in the first parentheses of the right hand side is bounded by

$$\int \max_d \left[\left(u(x, d) + \beta G_x^d(Q_2) + \epsilon_d \right) - \left(u(x, d) + \beta G_x^d(Q_1) + \epsilon_d \right) \right] dF_2(\epsilon)$$
$$= \beta \max_d \left[G_x^d(Q_2) - G_x^d(Q_1) \right]$$
$$\leq \beta \max_d \sum_{x' \in X} |Q_2(x') - Q_1(x')| G_x^d(x')$$
$$\leq \beta ||Q_2 - Q_1||,$$

where $||Q_2 - Q_1|| = \sup_{x \in X} |Q_1(x) - Q_1(x)|$. The term in the second parentheses is

$$\int \max_d \left[u(x, d) + \beta G_x^d(Q_1) + \epsilon_d \right] \left[f_2(\epsilon) - f_1(\epsilon) \right] d(\epsilon)$$
$$\leq \int \left[\bar{u} + \beta \sup_x Q_1(x) + \sum_d |\epsilon_d| \right] |f_2(\epsilon) - f_1(\epsilon)| \, d(\epsilon)$$
$$\leq c \int \left[1 + \sum_d |\epsilon_d| \right] |f_2(\epsilon) - f_1(\epsilon)| d(\epsilon) = c\rho(F_1, F_2)$$

for $c = \max\{1, \bar{u} + \beta(\bar{u} + \sum_d \int |\epsilon_d| dF_1(\epsilon))/(1 - \beta)\}$. Thus,

$$Q_2(x) - Q_1(x) \leq \beta ||Q_2 - Q_1|| + c\rho(F_1, F_2),$$

for each $x \in X$. Finally, we have

$$||Q_2 - Q_1|| \leq \frac{c}{1-\beta}\rho(F_1, F_2) = C\rho(F_1, F_2).$$

This proves the local Lipschitz continuity of the Emax function.

Given some $d^* \in \{0, 1, ..., J\}$ and x, define the set

$$S_1 = \left\{\epsilon : u(x, d^*) + \beta G_x^{d^*}(Q_1) + \epsilon_{d^*} \geq u(x, d) + \beta G_x^d(Q_1) + \epsilon_d, \forall d\right\},$$

on which d^* is optimal under F_1. Similarly, define S_2. With this notation,

$$p(d^*|x; F_i) = \int 1(S_i)dF_i(\epsilon),$$

for $i = 1, 2$, where $1(\cdot)$ is an indicator function. Now,

$$p(d^*|x; F_1) - p(d^*|x; F_2) = \int [1(S_1) - 1(S_2)]dF_1 + \int 1(S_2)d(F_1 - F_2).$$

The second integral is bounded by $\rho(F_1, F_2)$. Note that $1(S_1) - 1(S_2)$ is bounded above by

$$\sum_d 1(\{ d^* \text{ is optimal under } F_1 \text{ but } d \text{ is optimal under } F_2\})$$

$$+1(\{ d \text{ is optimal under } F_1 \text{ but } d^* \text{ is optimal under } F_2\}).$$

If d^* is optimal under F_1 and d is optimal under F_2, then

$$u(x, d) - u(x, d^*) + \beta G_x^d(Q_1) - \beta G_x^{d^*}(Q_1) \leq \epsilon_{d^*} - \epsilon_d \leq u(x, d) - u(x, d^*) + \beta G_x^d(Q_2) - \beta G_x^{d^*}(Q_2).$$

Denote this interval by A_1. Similarly, if d is optimal under F_1 and d^* is optimal under F_2, then

$$u(x, d^*) - u(x, d) + \beta G^{d^*}(Q_1) - \beta G_x^d(Q_1) \leq \epsilon_{d^*} - \epsilon_d \leq u(x, d^*) - u(x, d) + \beta G^{d^*}(Q_2) - \beta G_x^d(Q_2).$$

Denote this interval by A_2. Note that the length of the intervals A_1 and A_2 is bounded by $2\beta||Q_2 - Q_1||$. Also, by condition (ii) of the lemma, the density of the difference $\epsilon_{d^*} - \epsilon_d$ implied by F_1 is bounded by some positive constant $\bar{f} > 0$. Thus,

$$\int [1(S_1) - 1(S_2)]dF_1 \leq \sum_d \int 1(\epsilon_{d^*} - \epsilon_d \in A_1)dF_1(\epsilon) + \sum_d \int 1(\epsilon_{d^*} - \epsilon_d \in A_2)dF_1(\epsilon)$$

$$\leq 4(J+1)\beta||Q_2 - Q_1||\bar{f} \leq 4(J+1)\beta C\rho(F_1, F_2),$$

where the last inequality follows by the proven Lipschitz continuity of the Emax function. In summary, we have, for each $d^* = 0, 1, ..., J$,

$$|p(d^*|x; F_1) - p(d^*|x; F_2)| \leq 4(J+1)\beta C\rho(F_1, F_2) + \rho(F_1, F_2) = C'\rho(F_1, F_2),$$

for all $x \in X$. $Q.E.D.$

PROOF OF LEMMA 3:

$$\rho\left(f(\cdot), \sum_{j=1}^m \omega_j \phi(\cdot; q_j, \sigma)\right) = \int \left(1 + \sum_{i=1}^I |z_i|\right)\left|f(z) - \sum_{j=1}^m \omega_j \phi(z; q_j, \sigma)\right|dz$$

$$\leq \sum_{i=1}^I \int (1 + |z_i|)\left|f(z) - \sum_{j=1}^m \omega_j \phi(z; q_j, \sigma) \pm \int \phi(z; \mu, \sigma) f(\mu)d\mu\right|dz$$

$$(16) \quad \leq \sum_{i=1}^I \int (1 + |z_i|)\left|f(z) - \int \phi(z; \mu, \sigma) f(\mu)d\mu\right|dz$$

$$(17) \quad + \int (1 + |z_i|)\left|\int \phi(z; \mu, \sigma) f(\mu)d\mu - \sum_{j=1}^m \omega_j \phi(z; q_j, \sigma)\right|dz.$$

With the change of variable of μ_i to $\theta_i = \frac{z_i - \mu_i}{\sigma}, i = 1, ..., I$, (16) is bounded by

$$\sum_{i=1}^I \int \int (1 + |z_i|)\left|f(z) - f(z - \sigma\theta)\right|\phi(\theta)d\theta dz$$

$$\leq \sum_{i=1}^I \int \int (1 + |z_i|)L_f(z)e^{\tau\sigma||\theta||_2}\sigma||\theta||_2\phi(\theta)d\theta dz$$

$$= \sigma \left[\int ||\theta||_2 e^{\tau\sigma||\theta||_2} \phi(\theta) d\theta \right] \sum_{i=1}^{d} \left[\int (1 + |z_i|) L_f(z) dz \right],$$

where the inequality follows from the smoothness assumption (12). Lemma 6 shows that the term in the first square brackets is bounded for sufficiently small σ. The last term is finite by assumption (13). Hence, we can choose σ small enough to make the term above arbitrarily small.

To bound (17), let $A_j, j = 0, 1, ..., m$, be a partition of $\mathbb{R}^I$ consisting of adjacent hypercubes $A_1, ..., A_m$ with sides $h_m^{1/I}$ so that they are collectively centered at zero and the rest of the space is A_0. As m increases, the fine part of the partition becomes finer, $h_m \to 0$ and $m \to \infty$. Also, it covers larger and larger parts of $\mathbb{R}^I$; that is, $m \cdot h_m \to \infty$ as $m \to \infty$. Define $\omega_j = \int_{A_j} f(\mu) d\mu$. This implies,

$$\int \phi(z; \mu, \sigma) f(\mu) d\mu - \sum_{j=1}^{m} \omega_j \phi(z; q_j, \sigma)$$

$$= \sum_{j=0}^{m} \int_{A_j} \phi(z; \mu, \sigma) f(\mu) d\mu - \sum_{j=1}^{m} \int_{A_j} \phi(z; q_j, \sigma) f(\mu) d\mu$$

$$= \sum_{j=1}^{m} \int_{A_j} \left[\phi(z; \mu, \sigma) - \phi(z; q_j, \sigma) \right] + \int_{A_0} \phi(z; \mu, \sigma) f(\mu) d\mu.$$

The expression in (17) can be bounded as follows,

$$\sum_{i=1}^{I} \int (1 + |z_i|) \left| \int \phi(z; \mu, \sigma) f(\mu) d\mu - \sum_{j=1}^{m} \omega_j \phi(z; q_j, \sigma) \right| dz$$

$$\leq \sum_{i=1}^{I} \sum_{j=1}^{m} \int \int_{A_j} (1 + |z_i|) \left| \phi(z; \mu, \sigma) - \phi(z; q_j, \sigma) \right| f(\mu) d\mu dz$$

$$+ \sum_{i=1}^{I} \int \int_{A_0} (1 + |z_i|) \phi(z; \mu, \sigma) f(\mu) d\mu dz.$$

Consider the change of variable of z to $y = \frac{z-\mu}{\sigma}$ and define $\delta_j = \frac{q_j - \mu}{\sigma}$, $j = 1, ..., m$ so that

$\frac{z-q_j}{\sigma} = y - \delta_j$ and the above equals to

$$\sum_{i=1}^{I}\sum_{j=1}^{m}\int\int_{A_j}(1+|\mu_i+\sigma y_i|)\Big|\phi(y)-\phi(y-\delta_j)\Big|f(\mu)d\mu dy + \int\int_{A_0}(1+|\mu_i+\sigma y_i|)\phi(y)f(\mu)d\mu dy.$$

As $\mu, q_j \in A_j$, we have $||\delta_j||_\infty \leq \left\|\frac{q_j-\mu}{\sigma}\right\|_1 \leq \frac{I\cdot h_m^{1/d}}{I\cdot\sigma} \equiv \bar{\delta}$, which will be made small. By Lemma 7, $\phi(\cdot)$ is Lipschitz continuous and $|\phi(y)-\phi(y-\delta_j)| \leq \bar{L}_\phi(y)||\delta_j||_\infty$, which implies that $|\phi(y)-\phi(y-\delta_j)| \leq \bar{L}_\phi(y)\bar{\delta}$.

The first term above is bounded by $\bar{\delta}$ times

$$\sum_{i=1}^{I}\left(\sum_{j=1}^{m}c_{1j}\int_{A_j}(1+|\mu_i|)f(\mu)d\mu + c_{2j}\sigma\right),$$

where $c_{1j} = \int \bar{L}_\phi(y)dy$ and c_{2j} is such that $\int |y_i|\bar{L}_\phi(y)dy < c_{2j}$ for $i = 1,...,I$. The bound approaches to zero as $m \to \infty$. Since $\phi(\cdot)$ has bounded first moment, the second integral is bounded by

$$\sum_{i=1}^{I}\int_{A_0}(1+|\mu_i|+\sigma c)f(\mu)d\mu,$$

for some constant $c > 0$. As $mh_m \to \infty$, this term goes to zero.

$$Q.E.D.$$

PROOF OF LEMMA 4: We have

$$\rho\left(F^1, F^2\right) = \int\left(1+\sum_j|\epsilon_j|\right)\Big|f^1(\epsilon)-f^2(\epsilon)\Big|d\epsilon$$

$$= \int\left(1+\sum_j|\epsilon_j|\right)\Big|\sum_{k=1}^{m}\omega_k^1\phi\left(\epsilon;\mu_k^1,\sigma_k^1\right)-\omega_k^2\phi\left(\epsilon;\mu_k^2,\sigma_k^2\right)\pm\omega_k^1\phi\left(\epsilon;\mu_k^2,\sigma_k^2\right)\Big|d\epsilon$$

$$(18) \quad \leq \sum_{k=1}^{m}\omega_k^1\int\left(1+\sum_j|\epsilon_j|\right)\Big|\phi\left(\epsilon;\mu_k^1,\sigma_k^1\right)-\phi\left(\epsilon;\mu_k^2,\sigma_k^2\right)\Big|d\epsilon$$

$$(19) \quad + \sum_{k=1}^{m}|\omega_k^1-\omega_k^2|\int\left(1+\sum_j|\epsilon_j|\right)\phi\left(\epsilon;\mu_k^2,\sigma_k^2\right)d\epsilon.$$

Expression 19 is bounded by a positive constant times $\sum_{k=1}^{m} |\omega_k^1 - \omega_k^2|$.

To bound 18 note that

$$
\int \left(1 + \sum_j |\epsilon_j|\right) \left| \phi\left(\epsilon; \mu_k^1, \sigma_k^1\right) - \phi\left(\epsilon; \mu_k^2, \sigma_k^2\right) \right| d\epsilon
$$

$$
= \int \left(1 + \sum_j |\epsilon_j|\right) \left| \phi\left(\epsilon; \mu_k^1, \sigma_k^1\right) - \phi\left(\epsilon; \mu_k^2, \sigma_k^2\right) \pm \phi\left(\epsilon; \mu_k^2, \sigma_k^1\right) \right| d\epsilon
$$

$$
\leq \int \left(1 + \sum_j |\epsilon_j|\right) \left| \phi\left(\epsilon; \mu_k^1, \sigma_k^1\right) - \phi\left(\epsilon; \mu_k^2, \sigma_k^1\right) \right| d\epsilon
$$

$$
+ \int \left(1 + \sum_j |\epsilon_j|\right) \left| \phi\left(\epsilon; \mu_k^2, \sigma_k^1\right) - \phi\left(\epsilon; \mu_k^2, \sigma_k^2\right) \right| d\epsilon.
$$

With the change of variable $z = \frac{\epsilon - \mu_k^1}{\sigma_k^1}$ and by Lemma 7, the first part can be bounded as follows

$$
\int \left(1 + \sum_j |\sigma_k^1 z_j + \mu_{jk}^1|\right) \left| \phi(z; 0_{J+1}, 1) - \phi\left(z - \frac{\mu_k^2 - \mu_k^1}{\sigma_k^1}; 0_{J+1}, 1\right) \right| dz
$$

$$
\leq \left\| \frac{\mu_k^2 - \mu_k^1}{\sigma_k^1} \right\|_\infty \int \left(1 + \sum_j \left| \sigma_k^1 z_j + \mu_{jk}^1 \right|\right) \bar{L}_\phi(z) dz,
$$

where the integral is bounded. By Lemma 8, the second part can be bounded as follows

$$
\int \left(1 + \sum_j |\epsilon_j|\right) \left| \phi\left(\epsilon; 0_{J+1}, \sigma_k^1\right) - \phi\left(\epsilon; 0_{J+1}, \sigma_k^2\right) \right| d\epsilon \leq \left| \frac{1}{\sigma_k^1} - \frac{1}{\sigma_k^2} \right| \int \left(1 + \sum_j |\epsilon_j|\right) \bar{M}_\phi(\epsilon) d\epsilon,
$$

where the integral is bounded. To summarize, we have

$$
\rho\left(F^1, F^2\right) \leq \sum_{k=1}^{m} \omega_k^1 \left[c_1 \left\| \frac{\mu_k^2 - \mu_k^1}{\sigma_k^1} \right\|_\infty + c_2 \left| \frac{1}{\sigma_k^1} - \frac{1}{\sigma_k^2} \right| \right] + c_3 \sum_{k=1}^{m} |\omega_k^1 - \omega_k^2|,
$$

for some positive finite constants c_i's, which implies the claimed result.   Q.E.D.

PROOF OF LEMMA 5:   Let us denote the set of shock values at which alternative j is opti-

mal at the observed state k by

$$E_{jk} = \{(\epsilon_1, \ldots, \epsilon_J) : \ u(k,j;\theta_0) + \beta G_k^j Q + \epsilon_j \geq u(k,d;\theta_0) + \beta G_k^d Q + \epsilon_d, \forall d = 0, 1, \ldots, J\},$$

where the normalization $\epsilon_0 = 0$ is used. Sets $\{E_{jk}, j = 0, 1, \ldots, J\}$ define a partition of $\mathbb{R}^J$ (up to the overlapping boundaries) for each $k \in \{1, \ldots, K\}$. Consider a refinement of these K partitions,

$$\left\{A : \ A = \bigcap_{k=1}^K E_{j_k k}, \ j_k \in \{0, \ldots, J\}, \ \lambda(A) > 0\right\} = \{A_1, \ldots, A_L\},$$

where λ is the Lebesgue measure. Let us define

$$q_l = \int_{A_l} dF_0(\epsilon_{1J}) \in (0,1),$$

$$r_l = \int_{A_l} \epsilon_{1J} dF_0(\epsilon_{1J}) \in \mathbb{R}^J,$$

where $\epsilon_{1J} = (\epsilon_1, \ldots, \epsilon_J)'$. It follows from a characterization of the identified sets under unknown distribution of shocks in Norets (2011) that any distribution that implies the same $\{q_l, r_l, l = 1, \ldots, L\}$ delivers the same CCPs. The key to this result is that the shocks enter the utility additively and the integrals over the shocks in the Bellman equations can be replaced with expressions depending on the distribution of shocks only through $\{q_l, r_l, l = 1, \ldots, L\}$. In what follows, we construct a finite mixture distribution that delivers the same $\{q_l, r_l, l = 1, \ldots, L\}$ as F_0. Specifically, consider a mixture with $L \cdot (J+1)$ components

$$\sum_{l=1}^L \sum_{j=0}^J \omega_{lj} \phi(\cdot; \mu_{lj}, \sigma),$$

where we fix the locations as follows. Since F_0 is assumed to have a positive density, by Lemma 2 in Norets (2011), $r_l/q_l \in int(A_l)$ for all l. This is an implication of the supporting hyperplane theorem and the closedness and convexity of E_{jk}'s and thus A_l, $l = 1, \ldots, L$

(r_l/q_l is the expectation of ϵ_{1J} conditional on $\epsilon_{1J} \in A_l$). Therefore, $\exists \Delta > 0$ such that

$$\mu_{lj} = r_l/q_l + e_j \cdot \Delta \in int(A_l), \; j = 1, ..., J,$$
$$\mu_{l0} = r_l/q_l - (1, \ldots, 1)' \cdot \Delta \in int(A_l), \; l = 1, ..., L,$$

where e_j is a column vector of length J with 1 in the jth coordinate and 0's in the others. The mixing weights ω_{lj} are chosen as a solution to the following linear system of equations that matches $\{q_l, r_l, l = 1, \ldots, L\}$,

$$q_l = \sum_{\tilde{l}=1}^{L} \sum_{j=0}^{J} \omega_{\tilde{l}j} \int_{A_l} \phi(\epsilon_{1J}; \mu_{\tilde{l}j}, \sigma) d\epsilon_{1J},$$

(20)
$$r_l = \sum_{\tilde{l}=1}^{L} \sum_{j=0}^{J} \omega_{\tilde{l}j} \int_{A_l} \epsilon_{1J} \phi(\epsilon_{1J}; \mu_{\tilde{l}j}, \sigma) d\epsilon_{1J}, \quad l \in \{1, ..., L\}.$$

Let us show that for sufficiently small σ, this linear system has a unique solution that belongs to the $(L(J+1)-1)$-simplex. First, note that since $\mu_{lj} \in int(A_l)$,

$$\lim_{\sigma \to 0} \int_{A_l} \phi(\epsilon_{1J}; \mu_{\tilde{l}j}, \sigma) d\epsilon_{1J} = 1\{l = \tilde{l}\} \text{ and } \lim_{\sigma \to 0} \int_{A_l} \epsilon_{1J} \phi(\epsilon_{1J}; \mu_{\tilde{l}j}, \sigma) d\epsilon_{1J} = \mu_{lj} 1\{l = \tilde{l}\}.$$

Thus, the limiting system corresponding to $\sigma = 0$ in (20) is

(21)
$$q_l = \sum_{j=0}^{J} \omega_{lj}, \quad r_l = \sum_{j=0}^{J} \omega_{lj} \mu_{lj}, \quad l \in \{1, ..., L\}.$$

Plugging the definition of μ_{lj} into (21), we obtain $\omega_{lj} = \omega_{l0}$, $j = 1, \ldots, J$ and, thus, the limiting system (21) has the unique solution

$$\omega_{lj}^* = q_l/(J+1), \; l = 1, \ldots, L, \; j = 0, \ldots, J.$$

It follows that the matrix of the linear coefficients in the limiting system is invertible. Since matrix inversion is continuous and $\omega_{lj}^* > 0$, the system (20) has a unique strictly positive solution $\{\omega_{lj}^{**}, \; l = 1, \ldots, L, \; j = 0, \ldots, J\}$ for all sufficiently small σ. Since $\sum_{l=1}^{L} q_l = 1$,

$\sum_{l,j} \omega_{lj}^{**} = 1$ as well.

<div align="right">Q.E.D.</div>

LEMMA 6   *The term*

$$\int ||z||_2 e^{\tau\sigma \cdot ||z||_2} \phi(z; 0_d, 1) dz$$

is bounded for any $\tau > 0$ and sufficiently small $\sigma > 0$ where 0_d is a column vector of zeros with length d.

PROOF: By Cauchy-Schwartz inequality, the term is bounded by

$$\left(\int ||z||_2^2 \phi(z; 0_d, 1) dz \right)^{1/2} \left(\int e^{2\tau\sigma \cdot ||z||_2} \phi(z; 0_d, 1) dz \right)^{1/2}.$$

The first term in the product is bounded. As 2-norms are bounded by 1-norms, the term in the second parentheses is bounded by

$$\int e^{2\tau\sigma \cdot ||z||_1} \phi(z; 0_d, 1) dz = \prod_{i=1}^{d} \int_{-\infty}^{\infty} e^{2\tau\sigma|z_i|} \phi(z_i) dz_i$$

$$\leq \prod_{i=1}^{d} \int_{-\infty}^{\infty} e^{2\tau\sigma|z_i|} e^{-z_i} dz_i = 2^d \prod_{i=1}^{d} \int_{0}^{\infty} e^{2\tau\sigma z_i - z_i} dz_i,$$

which is bounded for small enough σ.

<div align="right">Q.E.D.</div>

LEMMA 7   *The density $\phi(\,\cdot\,; \mu, \sigma)$ is locally Lipschitz continuous in the location parameter with envelope $\bar{L}_\phi$ in a sense that, for an arbitrary $\bar{\delta}$, if $\delta \in \mathbb{R}^d$ is bounded by $\bar{\delta}$, then*

$$|\phi(z; 0_d, 1) - \phi(z - \delta; 0_d, 1)| \leq \bar{L}_\phi(z) ||\delta||_\infty,$$

where 0_d is a column vector of zeros with length d, $||\delta||_\infty = \max_{i=1,\ldots,d} |\delta_i|$, and

$$\bar{L}_\phi(z) = L_\phi(z_1) \prod_{i=2}^{d} \phi(z_i) + L_\phi(z_2)\phi(z_1 - \delta_1) \prod_{i=3}^{d} \phi(z_i) + \cdots + L_\phi(z_d) \prod_{i=1}^{d-1} \phi(z_i - \delta_i),$$

$$L_\phi(z) = \begin{cases} K_1 e^{-(z+\gamma)} e^{-e^{-K_2(z+\gamma)}}, & \text{if } z > -\gamma \\ K_3 e^{-2(z+\gamma)} e^{-e^{-K_4(z+\gamma)}}, & \text{otherwise,} \end{cases}$$

for some positive K_i's that could depend on $\bar{\delta}$.

PROOF: First, we establish local Lipschitz continuity for the univariate case. We have for $z, \delta \in \mathbb{R}$

$$|\phi(z) - \phi(z - \delta)| = |\phi'(z - \tilde{\delta})| \cdot |\delta|,$$

for some $\tilde{\delta}$ between 0 and $|\delta|$ where $\phi'(z) = \phi(z)(e^{-z-\gamma} - 1)$. Note that if $z > -\gamma$, then $|\phi'(z)| \leq c_1\phi(z)$ and if $z < -\gamma$, then $|\phi'(z)| \leq c_2 e^{-2z-2\gamma-c_3 e^{-z-\gamma}}$ for some constants $c_1, c_2, c_3 > 0$. From this, it follows that for $\tilde{\delta}$ which is bounded, $|\phi'(z - \tilde{\delta})| \leq L_\phi(z)$ where

$$L_\phi(z) = \begin{cases} K_1 e^{-(z+\gamma)} e^{-e^{-K_2(z+\gamma)}}, & \text{if } z > -\gamma \\ K_3 e^{-2(z+\gamma)} e^{-e^{-K_4(z+\gamma)}}, & \text{otherwise,} \end{cases}$$

for some positive K_i's that could depend on $\bar{\delta}$. Now, to show the result for the multivariate case, let $z, \delta \in \mathbb{R}^d$. With $||\delta||_\infty = \max_{i=1,\ldots,d} |\delta_i|$, we have

$$\phi(z; 0_d, 1) - \phi(z - \delta; 0_d, 1) = \prod_{i=1}^{d} \phi(z_i) - \prod_{i=1}^{d} \phi(z_i - \delta_i) \pm \phi(z_1 - \delta_1) \prod_{i=2}^{d} \phi(z_i)$$

$$\pm \phi(z_1 - \delta_1)\phi(z_2 - \delta_2) \prod_{i=3}^{d} \phi(z_i) \cdots \pm \phi(z_1 - \delta_1)\phi(z_2 - \delta_2) \cdots \phi(z_d - \delta_d)$$

$$= [\phi(z_1) - \phi(z_1 - \delta_1)] \prod_{i=2}^{d} \phi(z_i) + \phi(z_1 - \delta_1)[\phi(z_2) - \phi(z_2 - \delta_2)] \prod_{i=3}^{d} \phi(z_i)$$

$$+ \cdots + \prod_{i=1}^{d-1} \phi(z_i - \delta_i)[\phi(z_d) - \phi(z_d - \delta_d)].$$

By the proven locally Lipschitz continuity for the univariate case, this is bounded by

$$L_\phi(z_1)|\delta_1| \prod_{i=2}^{d} \phi(z_i) + L_\phi(z_2)|\delta_2|\phi(z_1 - \delta_1) \prod_{i=3}^{d} \phi(z_i) + \cdots + L_\phi(z_d)|\delta_d| \prod_{i=1}^{d-1} \phi(z_i - \delta_i)$$

$$\leq ||\delta||_\infty \left[L_\phi(z_1) \prod_{i=2}^{d} \phi(z_i) + L_\phi(z_2)\phi(z_1 - \delta_1) \prod_{i=3}^{d} \phi(z_i) + \cdots + L_\phi(z_d) \prod_{i=1}^{d-1} \phi(z_i - \delta_i) \right].$$

$$Q.E.D.$$

LEMMA 8   *The density $\phi(z; \mu, \sigma)$ is locally Lipschitz continuous in the inverse of the scale parameter with envelope $\bar{M}_\phi$ in a sense that given two scale parameters $\sigma^1, \sigma^2 > 0$,*

$$\left| \phi\left(z; 0_d, \sigma^1\right) - \phi\left(z; 0_d, \sigma^2\right) \right| \leq \left| \frac{1}{\sigma^1} - \frac{1}{\sigma^2} \right| \bar{M}_\phi(z), \text{ where}$$

$$\bar{M}_\phi(z) = M_\phi(z_1) \prod_{i=2}^{d} \phi\left(z_i; 0, \sigma^1\right) + M_\phi(z_2)\phi\left(z_1; 0, \sigma^2\right) \prod_{i=3}^{d} \phi\left(z_i; 0, \sigma^1\right) + \cdots + M_\phi(z_d) \prod_{i=1}^{d-1} \phi\left(z_i; 0, \sigma^2\right)$$

and $M_\phi(z_i) = \frac{1}{\sigma^1} L_\phi\left(\frac{z_i}{\sigma^1}\right) |z_i| + \phi\left(\frac{z_i}{\sigma^2}\right)$, $i = 1, ..., d$ for $L_\phi(\cdot)$ defined in Lemma 7.

PROOF: We first show the result for the univariate case. Let $\sigma^1, \sigma^2 > 0$ and $z \in \mathbb{R}$.

$$\left| \phi\left(z; 0, \sigma^1\right) - \phi\left(z; 0, \sigma^2\right) \right| = \left| \frac{1}{\sigma^1}\phi\left(\frac{z}{\sigma^1}\right) - \frac{1}{\sigma^2}\phi\left(\frac{z}{\sigma^2}\right) \pm \frac{1}{\sigma^1}\phi\left(\frac{z}{\sigma^2}\right) \right|$$

$$\leq \left| \frac{1}{\sigma^1}\phi\left(\frac{z}{\sigma^1}\right) - \frac{1}{\sigma^1}\phi\left(\frac{z}{\sigma^2}\right) \right| + \left| \frac{1}{\sigma^1}\phi\left(\frac{z}{\sigma^2}\right) - \frac{1}{\sigma^2}\phi\left(\frac{z}{\sigma^2}\right) \right|$$

$$= \frac{1}{\sigma^1}\left| \phi\left(\frac{z}{\sigma^1}\right) - \phi\left(\frac{z}{\sigma^1} - \left(\frac{z}{\sigma^1} - \frac{z}{\sigma^2}\right)\right) \right| + \phi\left(\frac{z}{\sigma^2}\right)\left| \frac{1}{\sigma^1} - \frac{1}{\sigma^2} \right|$$

$$\leq \frac{1}{\sigma^1} L_\phi\left(\frac{z}{\sigma^1}\right)\left| \frac{1}{\sigma^1} - \frac{1}{\sigma^2} \right| |z| + \phi\left(\frac{z}{\sigma^2}\right)\left| \frac{1}{\sigma^1} - \frac{1}{\sigma^2} \right| = M_\phi(z)\left| \frac{1}{\sigma^1} - \frac{1}{\sigma^2} \right|,$$

for $L_\phi(\cdot)$ defined in Lemma 7. In the multivariate case,

$$\phi\left(z; 0_d, \sigma^1\right) - \phi\left(z; 0_d, \sigma^2\right) = \prod_{i=1}^{d} \phi\left(z_i; 0, \sigma^1\right) - \prod_{i=1}^{d} \phi\left(z_i; 0, \sigma^2\right)$$

$$\pm \phi\left(z_1; 0, \sigma^2\right) \prod_{i=2}^{d} \phi\left(z_i; 0, \sigma^1\right) \pm \phi\left(z_1; 0, \sigma^2\right) \phi\left(z_2; 0, \sigma^2\right) \prod_{i=3}^{d} \phi\left(z_i; 0, \sigma^1\right)$$

$$\pm \cdots \pm \prod_{i=1}^{d} \phi\left(z_i; 0, \sigma^2\right)$$

$$= \left[\phi\left(z_1; 0, \sigma^1\right) - \phi\left(z_1; 0, \sigma^2\right)\right] \prod_{i=2}^{d} \phi\left(z_i; 0, \sigma^1\right)$$

$$+ \phi\left(z_1; 0, \sigma^2\right) \left[\phi\left(z_2; 0, \sigma^1\right) - \phi\left(z_2; 0, \sigma^2\right)\right] \prod_{i=3}^{d} \phi\left(z_i; 0, \sigma^1\right)$$

$$+ \cdots + \prod_{i=1}^{d-1} \phi\left(z_i; 0, \sigma^2\right) \left[\phi\left(z_d; 0, \sigma^1\right) - \phi\left(z_d; 0, \sigma^2\right)\right].$$

With the result for the univariate case, the last expression is bounded by

$$\left|\frac{1}{\sigma^1} - \frac{1}{\sigma^2}\right| \left\{ M_\phi(z_1) \prod_{i=2}^{d} \phi\left(z_i; 0, \sigma^1\right) + M_\phi(z_2) \phi\left(z_1; 0, \sigma^2\right) \prod_{i=3}^{d} \phi\left(z_i; 0, \sigma^1\right) \right.$$

$$\left. + \cdots + M_\phi(z_d) \prod_{i=1}^{d-1} \phi\left(z_i; 0, \sigma^2\right) \right\}.$$

$$Q.E.D.$$

APPENDIX D: SUPPLEMENT. AUXILIARY RESULTS AND DETAILS

D.1. *Newton-Kantorovich algorithm for DP solution*

For a given value of the model parameter ψ, the Emax function Q_ψ is the fixed point of the operator $T_\psi(\cdot)$ defined by

$$
\begin{aligned}
T_\psi(\mathbf{Q})[x] &= \sum_{k=1}^{m} \omega_k \sigma_k \left[A_{kx} + E1(e^{-a_{kx}}) \right] \\
&= \sum_{k=1}^{m} \omega_k \sigma_k \Bigg\{ \\
&\log \sum_{j=1}^{J} \exp \left(\frac{u(x,j,;\psi) + \beta \sum_{y=1}^{K} G_{xy}^j \mathbf{Q}(y) + \mu_{jk}}{\sigma_k} \right) + \\
&E1 \Bigg(\exp \Bigg[- \frac{u(0,j,;\psi) + \beta \sum_{y=1}^{K} G_{xy}^j \mathbf{Q}(y)}{\sigma_k} - \gamma \\
&\qquad - \log \sum_{j=1}^{J} \exp \left\{ \frac{u(x,j,;\psi) + \beta \sum_{y=1}^{K} G_{xy}^j \mathbf{Q}(y) + \mu_{jk}}{\sigma_k} \right\} \Bigg] \Bigg) \Bigg\}.
\end{aligned}
$$

To find the fixed point of T_ψ, following Rust (1987), we use the Newton-Kantorovich algorithm, which is essentially a Newton method for solving the nonlinear system of equations, because it is more efficient than the iterations on T_ψ.

LEMMA 9 *The Newton-Kantorovich algorithm has the update rule*

$$
(22) \qquad Q^{(n+1)} = Q^{(n)} - \left[I - T_\psi'(Q^{(n)}) \right]^{-1} \left[I - T_\psi \right] (Q^n),
$$

where

$$
T_\psi'(Q) = \beta \begin{pmatrix} \sum_{d=0}^{J} G_{11}^d P(d|x=1) & \cdots & \sum_{d=0}^{J} G_{1K}^d P(d|x=1) \\ \vdots & & \vdots \\ \sum_{d=0}^{J} G_{K1}^d P(d|x=K) & \cdots & \sum_{d=0}^{J} G_{KK}^d P(d|x=K) \end{pmatrix}.
$$

PROOF OF LEMMA 9:   Note that

$$T'_\psi(Q) = \begin{pmatrix} \frac{\partial}{\partial Q(1)} T_\psi(Q)(1) & \cdots & \frac{\partial}{\partial Q(K)} T_\psi(Q)(1) \\ \vdots & & \vdots \\ \frac{\partial}{\partial Q(1)} T_\psi(Q)(K) & \cdots & \frac{\partial}{\partial Q(K)} T_\psi(Q)(K) \end{pmatrix}.$$

Recall that $T_\psi(Q)(x) = \sum_{k=1}^m \omega_k \sigma_k \left[A_{kx} + E1(e^{-a_{kx}}) \right]$. Denoting $f(a_{kx}) = E1(e^{-a_{kx}})$, we have

$$\frac{\partial}{\partial Q(1)} T_\psi(Q)(x) = \sum_{k=1}^m \omega_k \sigma_k \left[\frac{\partial}{\partial Q(1)} A_{kx} + \frac{\partial}{\partial Q(1)} f(a_{kx}) \right], \quad \text{for } x = 1, \dots, K.$$

Recall that

$$A_{kx} = \log \sum_{j=1}^J \exp \left[\frac{v(x,j) + \mu_{jk}}{\sigma_k} \right] = \log \sum_{j=1}^J \exp \left[\frac{u(x,j) + \beta \sum_{y=1}^K G_{xy}^j Q(y) + \mu_{jk}}{\sigma_k} \right],$$

$$a_{kx} = \frac{v_0}{\sigma_k} + \gamma - A_k = \frac{1}{\sigma_k} \left[u(x,0) + \beta \sum_{y=1}^K G_{xy}^j Q(y) \right] + \gamma - A_{kx}.$$

Hence, for example,

$$\frac{\partial A_{k1}}{\partial Q(1)} = \frac{1}{\sum_{j=1}^J \exp \left[\frac{v(1,j) + \mu_{jk}}{\sigma_k} \right]} \sum_{d=1}^J \frac{\partial \exp \left[\frac{v(1,d) + \mu_{dk}}{\sigma_k} \right]}{\partial \left[\frac{v(1,d) + \mu_{dk}}{\sigma_k} \right]} \frac{\partial \left[\frac{v(1,d) + \mu_{dk}}{\sigma_k} \right]}{\partial Q(1)}$$

$$= \frac{1}{\sum_{j=1}^J \exp \left[\frac{v(1,j) + \mu_{jk}}{\sigma_k} \right]} \sum_{d=1}^J \exp \left[\frac{v(1,d) + \mu_{dk}}{\sigma_k} \right] \frac{\beta}{\sigma_k} G_{11}^d$$

$$= \frac{\beta}{\sigma_k} \sum_{d=1}^J \frac{\exp \left[\frac{v(1,d) + \mu_{dk}}{\sigma_k} \right]}{\sum_{j=1}^J \exp \left[\frac{v(1,j) + \mu_{jk}}{\sigma_k} \right]} G_{11}^d$$

$$= \frac{\beta}{\sigma_k} \sum_{d=1}^J \exp \left[\frac{v(1,d) + \mu_{dk}}{\sigma_k} - A_{k1} \right] G_{11}^d.$$

Also,

$$\frac{\partial f(a_{k1})}{\partial Q(1)} = \frac{\partial g(e^{-a_{k1}})}{\partial e^{-a_{k1}}} \frac{\partial e^{-a_{k1}}}{\partial(-a_{k1})} \frac{\partial(-a_{k1})}{\partial a_{k1}} \frac{\partial a_{k1}}{\partial Q(1)} = g'(e^{-a_{k1}})e^{-a_{k1}}(-1)\frac{\partial a_{k1}}{\partial Q(1)},$$

where we let $g(y) = E1(y)$. We know that $g'(y) = -\frac{e^{-y}}{y}$. Note that $\frac{\partial a_{k1}}{\partial Q(1)} = \frac{\beta}{\sigma_k}G_{11}^1 - \frac{\partial A_{k1}}{\partial Q(1)}$.
Hence

$$\frac{\partial f(a_{k1})}{\partial Q(1)} = -\exp\left[-e^{-a_{k1}}\right]e^{a_{k1}}e^{-a_{k1}}(-1)\frac{\partial a_{k1}}{\partial Q(1)} = \exp\left[-e^{-a_{k1}}\right]\left[\frac{\beta}{\sigma_k}G_{11}^1 - \frac{\partial A_{k1}}{\partial Q(1)}\right].$$

Now,

$$\frac{\partial}{\partial Q(1)}T_\psi(Q)(1) = \sum_{k=1}^{m}\omega_k\sigma_k\left[\frac{\partial A_{k1}}{\partial Q(1)} + \exp\left[-e^{-a_{k1}}\right]\left[\frac{\beta}{\sigma_k}G_{11}^1 - \frac{\partial A_{k1}}{\partial Q(1)}\right]\right]$$

$$= \sum_{k=1}^{m}\omega_k\sigma_k\left[\frac{\partial A_{k1}}{\partial Q(1)}\left(1 - \exp\left[-e^{-a_{k1}}\right]\right) + \exp\left[-e^{-a_{k1}}\right]\frac{\beta}{\sigma_k}G_{11}^1\right].$$

Note that

$$\sum_{k=1}^{m}\omega_k\sigma_k\frac{\partial A_{k1}}{\partial Q(1)}\left(1 - \exp\left[-e^{-a_{k1}}\right]\right)$$

$$= \sum_{k=1}^{m}\omega_k\sigma_k\frac{\beta}{\sigma_k}\underbrace{\sum_{d=1}^{J}\exp\left[\frac{v(1,d) + \mu_{dk}}{\sigma_k} - A_{k1}\right]G_{11}^d}_{\frac{\partial A_{k1}}{\partial Q(1)}}\left(1 - \exp\left[-e^{-a_{k1}}\right]\right)$$

$$= \beta\sum_{d=1}^{J}G_{11}^d\sum_{k=1}^{m}\omega_k\exp\left[\frac{v(1,d) + \mu_{dk}}{\sigma_k} - A_{k1}\right]\left(1 - \exp\left[-e^{-a_{k1}}\right]\right) = \beta\sum_{d=1}^{J}G_{11}^d P(d|x = 1)$$

$$\sum_{k=1}^{m}\omega_k\sigma_k\exp\left[-e^{-a_{k1}}\right]\frac{\beta}{\sigma_k}G_{11}^1 = \beta\sum_{k=1}^{m}\omega_k\exp\left[-e^{-a_{k1}}\right]G_{11}^1 = \beta G_{11}^1 P(0|x = 1).$$

Hence,

$$\frac{\partial}{\partial Q(1)}T_\psi(Q)(1) = \beta\sum_{d=0}^{J}G_{11}^d P(d|x = 1).$$

We can similarly compute other elements. *Q.E.D.*

D.2. *Properties of univariate extreme value distributions*

Let $Z \sim \phi(\cdot)$. Then it has zero mean and variance $\frac{\pi^2}{6}$. Its density is $e^{-z-\gamma-e^{-z-\gamma}}$ and its cdf is $e^{-e^{-z-\gamma}}$, where γ is the Euler-Mascheroni constant.

Lemma 10 and Lemma 11 show how to compute median and truncated integrals of a random variable that follows a mixture of extreme value distributions, which is helpful for imposing the location and scale normalizations after a MCMC run.

LEMMA 10   *Median*

1. *Let $X \sim \phi(\cdot; \mu, \sigma)$. Then its median is $\mu - \sigma \log \log 2 - \sigma\gamma$.*
2. *Let $X \sim \sum_{k=1}^{m} \omega_k \phi(\cdot; \mu_k, \sigma_k)$. Then there is no closed form for its median. It has to be solved for via a root-finding algorithm.*

PROOF: Let $X \sim \phi(\cdot; \mu, \sigma)$. By definition of median, we want to find M such that $0.5 = Pr(X < M)$. Letting $t(x) = e^{-\left(\frac{x-\mu}{\sigma}\right)}e^{-\gamma}$, this is equivalent to

$$0.5 = \exp\left[-t(M)\right] \iff t(M) = \log 2 \iff e^{-\left(\frac{M-\mu}{\sigma}\right)}e^{-\gamma} = \log 2$$
$$\iff -\left(\frac{M-\mu}{\sigma}\right) - \gamma = \log \log 2 \iff M = \mu - \sigma \log \log 2 - \sigma\gamma$$

Q.E.D.

LEMMA 11   *Truncated Integral*

1. *Let $X \sim \phi(\cdot; \mu, \sigma)$. Then*

$$\int_{M}^{\infty} x\phi(x; \mu, \sigma)dx = \mu - M \exp\left[-e^{-b}\right] + \sigma E1(e^{-b})$$

where $b = \frac{M-\mu}{\sigma} + \gamma$

2. *Let $X \sim p(\cdot|\psi, m) = \sum_{k=1}^{m} \omega_k \phi(\cdot; \mu_k, \sigma_k)$. Then*

$$\int_{M}^{\infty} x p(x|\psi, m) dx = \sum_{k=1}^{m} \omega_k \left\{ \mu_k - M \exp\left[-e^{-b_k}\right] + \sigma_k E1(e^{-b_k}) \right\}$$

where $b_k = \frac{M-\mu_k}{\sigma_k} + \gamma$

PROOF: Let $X \sim \phi(\cdot; \mu, \sigma)$.

$$
\begin{aligned}
\int_{M}^{\infty} x\phi(x; \mu, \sigma) dx &= \int_{M}^{\infty} x \frac{1}{\sigma} e^{-\left(\frac{x-\mu}{\sigma}\right)} e^{-\gamma} \exp\left[-e^{-\left(\frac{x-\mu}{\sigma}\right)} e^{-\gamma}\right] dx \\
&= \int_{b}^{\infty} [\sigma z + (\mu - \sigma\gamma)] \frac{1}{\sigma} e^{-z} \exp\left[-e^{-z}\right] \sigma dz \\
&= \sigma \underbrace{\int_{b}^{\infty} z e^{-z} \exp\left[-e^{-z}\right] dz}_{\text{I}} + (\mu - \sigma\gamma) \underbrace{\int_{b}^{\infty} e^{-z} \exp\left[-e^{-z}\right] dz}_{\text{II}}
\end{aligned}
$$

where we let $z = \left(\frac{x-\mu}{\sigma}\right) + \gamma$ which means that $x = \sigma z + (\mu - \sigma\gamma), dx = \sigma dz$. $x = \infty \implies z = \infty, x = M \implies z = \frac{M-\mu}{\sigma} + \gamma = b$. For computing the first integral, let $y = e^{-z}$ which means that $dz = -\frac{1}{y} dy, z = -\log y$. We have

$$I = \int_{e^{-b}}^{0} (-\log y) y e^{-y} \left(-\frac{1}{y} dy\right) = -\int_{0}^{e^{-b}} \log y e^{-y} dy$$

Recall that $\gamma = -\int_{0}^{\infty} \log y e^{-y} dy = -\left(\int_{0}^{e^{-b}} \log y e^{-y} dy + \int_{e^{-b}}^{\infty} \log y e^{-y} dy\right)$. Hence, $I = \gamma + \int_{e^{-b}}^{0} \log y e^{-y} dy$. By integration by parts, letting $u = \log y, dv = e^{-y} dy \implies du \frac{1}{y} dy, v = -e^{-y}$,

$$
\begin{aligned}
I &= \gamma + \int_{e^{-b}}^{0} \log y e^{-y} dy = \gamma + \int u \, dv - \int v \, du \\
&= \gamma - \left[\log y e^{-y}\right]_{e^{-b}}^{\infty} + \int_{e^{-b}}^{\infty} \frac{1}{y} e^{-y} dy = \gamma - \left[0 - (-b) \exp\left(-e^{-b}\right)\right] + E1\left(e^{-b}\right) \\
&= \gamma - b \exp\left(-e^{-b}\right) + E1\left(e^{-b}\right)
\end{aligned}
$$

where by definition $E1(z) = \int_{z}^{\infty} \frac{1}{t} e^{-t} dt$. For the second integral, letting $y = e^{-z}$ which means

that $dz = -\frac{1}{y}dy$, we have

$$
\begin{aligned}
II &= \int_b^\infty e^{-z} \exp\left[-e^{-z}\right] dz \\
&= \int_{e^{-b}}^0 y e^{-y}\left(-\frac{1}{y}dy\right) = \int_0^{e^{-b}} e^{-y} dy = -\left(\exp\left(-e^{-b}\right) - 1\right) \\
&= 1 - \exp\left(-e^{-b}\right)
\end{aligned}
$$

Finally,

$$
\int_M^\infty x dF(x) = \sigma I + (\mu - \sigma\gamma)II = \mu - M \exp\left[-e^{-b}\right] + \sigma E1(e^{-b})
$$

$$Q.E.D.$$

D.3. *The Prior for the Transformed Mixing Weights*

We want to define a prior on $(\alpha_1, \ldots, \alpha_{m-1})$ which implies the prior $(\omega_1, \ldots, \omega_m) \sim Dir(a_1, \ldots, a_m)$ or equivalently $\gamma_\ell \sim Gamma(a_\ell, 1)$ for $\ell = 1, \ldots, m$. Note that the inverse map $g(\alpha) = \omega$ is defined by $\omega_\ell = \frac{e^{\alpha_\ell}}{1 + \sum_{s=1}^{m-1} e^{\alpha_s}}$, for $\ell = 1, \ldots, m-1$. Denote the Jacobian $A(\alpha) = \frac{dg(\alpha)}{d\alpha}$. By the change of variable formula, we have $f_\alpha(\alpha) = f_\omega(g(\alpha)) det\frac{dg(\alpha)}{d\alpha} = f_\omega(g(\alpha)) det A(\alpha)$.

In order to have $f_\omega(\omega) = Dir(\omega_1, \ldots, \omega_m; \bar\omega_1, \ldots, \bar\omega_m)$, we can use the above expression for the density $f_\alpha(\alpha)$ to determine the desired prior for $\alpha_1, \ldots, \alpha_m$.

In addition, for using HMC, we need the first order derivative $\frac{\partial f_\alpha(\alpha)}{\partial\alpha}$. Note that in general, $\frac{\partial}{\partial\alpha} det A(\alpha) = det A(\alpha) tr\left(A(\alpha)^{-1}\frac{\partial}{\partial\alpha}A(\alpha)\right)$. Hence, $\frac{\partial f_\alpha(\alpha)}{\partial\alpha} = \left[\frac{\partial}{\partial\alpha}f_\omega(g(\alpha))\right] det A(\alpha) + f_\omega(g(\alpha))\left[det A(\alpha) tr\left(A(\alpha)^{-1}\frac{\partial}{\partial\alpha}A(\alpha)\right)\right]$.

D.4. *Derivatives of the Log-likelihood*

Let $L(\chi) = \log p(D_n|\chi, m) = \sum_{d,x} n_{dx} \times \log p(d|x; \chi, m)$ denote the log-likelihood as a function of χ, where n_{dx} is the number of decision makers in the data set that chose d at the observed state x. Below, we present the derivatives of $L(\chi)$ with respect to χ for $J = 1$. The computation for $J > 1$ can be done similarly.

LEMMA 12   *Derivatives of $L(\chi)$ with respect to χ for $J = 1$.*

(1) For $k = 1, ..., m - 1$,

$$\frac{\partial L(\chi)}{\partial \alpha_k} = \sum_{k=1}^{m-1} \frac{\partial L(\chi)}{\partial \omega_\ell} \frac{\partial \omega_\ell}{\partial \alpha_k}, \ where$$

$$\frac{\partial L(\chi)}{\partial \omega_\ell} = \sum_x n_{0x} \frac{p(0|x, \ell) - p(0|x, m) + \sum_k \omega_k \frac{\partial p(0|x,k)}{\partial \omega_\ell}}{\sum_{k'} \omega_{k'} p(0|x, k')}$$

$$+ \sum_x n_{1x} \frac{p(1|x, \ell) - p(1|x, m) + \sum_k \omega_k \frac{\partial p(1|x,k)}{\partial \omega_\ell}}{\sum_{k'} \omega_{k'} p(1|x, k')},$$

$$\frac{\partial p(0|x, k)}{\partial \omega_\ell} = \exp\left[-e^{-a_{xk}} - a_{xk}\right] \frac{\beta}{\sigma_k} \sum_{y=1}^{K} (G_{xy}^0 - G_{xy}^1) \frac{\partial Q(y)}{\partial \omega_\ell},$$

$$\frac{\partial \omega_\ell}{\partial \alpha_k} = \frac{e^{\alpha_\ell}}{(1 + \sum_{s=1}^{m-1} e^{\alpha_s})^2} \left[(1 + \sum_{s \neq k}^{m-1} e^{\alpha_s}) 1(k = \ell) - e^{\alpha_k} 1(k \neq \ell)\right].$$

(2) For $k = 1, ..., m$,

$$\frac{\partial L(\chi)}{\partial \mu_k} = \sum_x n_{0x} \frac{\sum_\ell \omega_\ell \frac{\partial p(0|x,\ell)}{\partial \mu_k}}{\sum_{k'} \omega_{k'} p(0|x, k')} + \sum_x n_{1x} \frac{\sum_\ell \omega_\ell \frac{\partial p(1|x,\ell)}{\partial \mu_k}}{\sum_{k'} \omega_{k'} p(1|x, k')}, \ where$$

$$\frac{\partial p(0|x, \ell)}{\partial \mu_k} = \exp\left[-e^{-a_{x\ell}} - a_{x\ell}\right] \frac{\partial a_{x\ell}}{\partial \mu_k},$$

$$\frac{\partial a_{x\ell}}{\partial \mu_k} = \frac{\beta}{\sigma_\ell} \sum_{y=1}^{K} (G_{xy}^0 - G_{xy}^1) \frac{\partial Q(y)}{\partial \mu_k} - \frac{1}{\sigma_\ell} 1(k = \ell).$$

(3) For $s_k = \log \tilde{\sigma}_k$, $k = 1, ..., m$,

$$\frac{\partial L(\chi)}{\partial s_k} = \frac{\partial L(\chi)}{\partial \sigma_k} \sigma_k, \ where$$

$$\frac{\partial L(\chi)}{\partial \sigma_k} = \sum_x n_{0x} \frac{\sum_\ell \omega_\ell \frac{\partial p(0|x,\ell)}{\partial \sigma_k}}{\sum_{k'} \omega_{k'} p(0|x, k')} + \sum_x n_{1x} \frac{\sum_\ell \omega_\ell \frac{\partial p(1|x,\ell)}{\partial \sigma_k}}{\sum_{k'} \omega_{k'} p(1|x, k')},$$

$$\frac{\partial p(0|x,\ell)}{\partial \sigma_k} = \exp\left[-e^{-a_{x\ell}} - a_{x\ell}\right]\frac{\partial a_{x\ell}}{\partial \sigma_k},$$

$$\frac{\partial a_{x\ell}}{\partial \sigma_k} = \frac{\beta}{\sigma_\ell}\sum_{y=1}^{K}(G_{xy}^0 - G_{xy}^1)\frac{\partial Q(y)}{\partial \mu_k} + \frac{1}{\sigma_\ell}[\gamma - a_{xk}]1(k=\ell).$$

(4) With respect to $s = \log(\sigma)$

$$\frac{\partial L(\chi)}{\partial s} = \sum_{\ell=1}^{m}\frac{\partial L(\chi)}{\partial s_\ell}.$$

(5) The derivatives of the Emax function can be computed by the implicit function theorem:

$$\frac{\partial Q(x)}{\partial \omega_k} = -\frac{\frac{\partial f_x}{\partial \omega_k}}{\frac{\partial f_x}{\partial Q(x)}}, \quad \frac{\partial Q(x)}{\partial \mu_k} = -\frac{\frac{\partial f_x}{\partial \mu_k}}{\frac{\partial f_x}{\partial Q(x)}}, \quad \frac{\partial Q(x)}{\partial \sigma_k} = -\frac{\frac{\partial f_x}{\partial \sigma_k}}{\frac{\partial f_x}{\partial Q(x)}},$$

where $f_x = Q(x) - \sum_{k=1}^{m}\omega_k\sigma_k\left[A_{xk} + E1(e^{-a_{xk}})\right]$ and

$$\frac{\partial f_x}{\partial Q(x)} = 1 - \beta G_{xx}^1 - \beta(G_{xx}^0 - G_{xx}^1)\sum_{k=1}^{m}\omega_k\exp\left[-e^{-a_{xk}}\right],$$

$$\frac{\partial f_x}{\partial \omega_k} = -\sigma_k\left[A_{xk} + E1(e^{-a_{xk}})\right] + \sigma_m\left[A_{xm} + E1(e^{-a_{xm}})\right],$$

$$\frac{\partial f_x}{\partial \mu_k} = -\omega_k\left[1 - \exp(-e^{-a_{xk}})\right],$$

$$\frac{\partial f_x}{\partial \sigma_k} = -\omega_k\left[E1(e^{-a_{xk}}) + \exp(-e^{-a_{xk}})(\gamma - a_{xk})\right].$$

PROOF: For (1),

$$\frac{\partial L(\chi)}{\partial \alpha_k} = \sum_{k=1}^{m-1}\frac{\partial L(\chi)}{\partial \omega_\ell}\frac{\partial \omega_\ell}{\partial \alpha_k},$$

$$\frac{\partial L(\chi)}{\partial \omega_\ell} = \sum_{x}n_{0x}\frac{\frac{\partial}{\partial \omega_\ell}\sum_k\omega_k p(0|x,k)}{\sum_{k'}\omega_{k'}p(0|x,k')} + \sum_{x}n_{1x}\frac{\frac{\partial}{\partial \omega_\ell}\sum_k\omega_k p(1|x,k)}{\sum_{k'}\omega_{k'}p(1|x,k')},$$

$$\sum_{k}^{m}\omega_k p(d|x,k) = \sum_{k}^{m-1}\omega_k p(d|x,k) + (1 - \sum_{k}^{m-1}\omega_k)p(d|x,m),$$

$$\frac{\partial}{\partial \omega_\ell} \sum_k^m \omega_k p(d|x,k) = p(d|x,\ell) + \sum_k^{m-1} \omega_k \frac{\partial p(d|x,k)}{\partial \omega_\ell} - p(d|x,m) + (1 - \sum_k^{m-1} \omega_k) \frac{\partial p(d|x,m)}{\partial \omega_\ell}$$

$$= p(d|x,\ell) - p(d|x,m) + \sum_k^m \omega_k \frac{\partial p(d|x,k)}{\partial \omega_\ell}$$

$$\implies \frac{\partial L(\chi)}{\partial \omega_\ell} = \sum_x n_{0x} \frac{p(0|x,\ell) - p(0|x,m) + \sum_k \omega_k \frac{\partial p(0|x,k)}{\partial \omega_\ell}}{\sum_{k'} \omega_{k'} p(0|x,k')}$$

$$+ \sum_x n_{1x} \frac{p(1|x,\ell) - p(1|x,m) + \sum_k \omega_k \frac{\partial p(1|x,k)}{\partial \omega_\ell}}{\sum_{k'} \omega_{k'} p(1|x,k')},$$

$$\frac{\partial p(0|x,k)}{\partial \omega_\ell} = \exp\left[-e^{-a_{xk}} - a_{xk}\right] \frac{\partial a_{xk}}{\partial \omega_\ell}.$$

For $J = 1$, we have,

$$a_{x\ell} = \frac{1}{\sigma_\ell} \left[v(x,0) - v(x,1) - \mu_\ell\right] + \gamma$$

$$= \frac{1}{\sigma_\ell} \left[u(x,0) - u(x,1) + \beta \sum_{y=1}^K (G_{xy}^0 - G_{xy}^1) Q(y) - \mu_\ell\right] + \gamma.$$

Hence,

$$\frac{\partial p(0|x,k)}{\partial \omega_\ell} = \exp\left[-e^{-a_{xk}} - a_{xk}\right] \frac{\beta}{\sigma_k} \sum_{y=1}^K (G_{xy}^0 - G_{xy}^1) \frac{\partial Q(y)}{\partial \omega_\ell}.$$

Note that

$$\frac{\partial p(1|x,k)}{\partial \omega_\ell} = 1 - \frac{\partial p(0|x,k)}{\partial \omega_\ell}.$$

Moreover, recall that $\omega_\ell = \frac{e^{\alpha_\ell}}{1 + \sum_{s=1}^{m-1} e^{\alpha_s}}$.

If $k = \ell$, then $\frac{\partial \omega_\ell}{\partial \alpha_k} = \frac{e^{\alpha_\ell}(1 + \sum_{s=1}^{m-1} e^{\alpha_s}) - e^{\alpha_\ell} e^{\alpha_\ell}}{(1 + \sum_{s=1}^{m-1} e^{\alpha_s})^2} = \frac{e^{\alpha_\ell}(1 + \sum_{s \neq \ell}^{m-1} e^{\alpha_s})}{(1 + \sum_{s=1}^{m-1} e^{\alpha_s})^2}$.

If $k \neq \ell$, then $\frac{\partial \omega_\ell}{\partial \alpha_k} = \frac{-e^{\alpha_\ell} e^{\alpha_k}}{(1 + \sum_{s=1}^{m-1} e^{\alpha_s})^2}$.

For (2),

$$
\begin{aligned}
\frac{\partial L(\chi)}{\partial \mu_k} &= \sum_x n_{0x} \frac{\frac{\partial}{\partial \mu_k} \sum_k \omega_k p(0|x,k)}{\sum_{k'} \omega_{k'} p(0|x,k')} + \sum_x n_{1x} \frac{\frac{\partial}{\partial \mu_k} \sum_k \omega_k p(1|x,k)}{\sum_{k'} \omega_{k'} p(1|x,k')} \\
&= \sum_x n_{0x} \frac{\sum_\ell \omega_\ell \frac{\partial p(0|x,\ell)}{\partial \mu_k}}{\sum_{k'} \omega_{k'} p(0|x,k')} + \sum_x n_{1x} \frac{\sum_\ell \omega_\ell \frac{\partial p(1|x,\ell)}{\partial \mu_k}}{\sum_{k'} \omega_{k'} p(1|x,k')},
\end{aligned}
$$

$$
\frac{\partial p(0|x,\ell)}{\partial \mu_k} = \exp\left[-e^{-a_{x\ell}} - a_{x\ell} \right] \frac{\partial a_{x\ell}}{\partial \mu_k},
$$

$$
\frac{\partial a_{x\ell}}{\partial \mu_k} =
\begin{cases}
\frac{\beta}{\sigma_\ell} \sum_{y=1}^K (G_{xy}^0 - G_{xy}^1) \frac{\partial Q(y)}{\partial \mu_k} - \frac{1}{\sigma_\ell}, & \text{if } k = \ell \\
\frac{\beta}{\sigma_\ell} \sum_{y=1}^K (G_{xy}^0 - G_{xy}^1) \frac{\partial Q(y)}{\partial \mu_k}, & \text{otherwise.}
\end{cases}
$$

(3) Recall that $\sigma_k = \sigma \tilde{\sigma}_k$. $\tilde{\sigma}_k = e^{s_k}$, so $\frac{\partial \tilde{\sigma}_k}{\partial s_k} = \tilde{\sigma}_k$. We have

$$
\begin{aligned}
\frac{\partial L(\chi)}{\partial s_k} &= \frac{\partial L(\chi)}{\partial \sigma_k} \frac{\partial \sigma \tilde{\sigma}_k}{\partial \tilde{\sigma}_k} \frac{\partial \tilde{\sigma}_k}{\partial s_k} = \frac{\partial L(\chi)}{\partial \sigma_k} \sigma \tilde{\sigma}_k = \frac{\partial L(\chi)}{\partial \sigma_k} \sigma_k, \\
\frac{\partial L(\chi)}{\partial \sigma_k} &= \sum_x n_{0x} \frac{\frac{\partial}{\partial \sigma_k} \sum_k \omega_k p(0|x,k)}{\sum_{k'} \omega_{k'} p(0|x,k')} + \sum_x n_{1x} \frac{\frac{\partial}{\partial \sigma_k} \sum_k \omega_k p(1|x,k)}{\sum_{k'} \omega_{k'} p(1|x,k')} \\
&= \sum_x n_{0x} \frac{\sum_\ell \omega_\ell \frac{\partial p(0|x,\ell)}{\partial \sigma_k}}{\sum_{k'} \omega_{k'} p(0|x,k')} + \sum_x n_{1x} \frac{\sum_\ell \omega_\ell \frac{\partial p(1|x,\ell)}{\partial \sigma_k}}{\sum_{k'} \omega_{k'} p(1|x,k')},
\end{aligned}
$$

$$
\frac{\partial p(0|x,\ell)}{\partial \sigma_k} = \exp\left[-e^{-a_{x\ell}} - a_{x\ell} \right] \frac{\partial a_{x\ell}}{\partial \sigma_k},
$$

$$
\frac{\partial a_{x\ell}}{\partial \sigma_k} =
\begin{cases}
\frac{1}{\sigma_\ell} \left[\gamma - a_{xk} + \beta \sum_{y=1}^K (G_{xy}^0 - G_{xy}^1) \frac{\partial Q(y)}{\partial \sigma_k} \right], & \text{if } k = \ell \\
\frac{\beta}{\sigma_\ell} \sum_{y=1}^K (G_{xy}^0 - G_{xy}^1) \frac{\partial Q(y)}{\partial \sigma_k}, & \text{otherwise.}
\end{cases}
$$

(4) Note that $\sigma = e^s$, so $\frac{\partial \sigma}{\partial s} = \sigma$

$$
\frac{\partial L(\chi)}{\partial s} = \sum_{k=1}^m \frac{\partial L(\chi)}{\partial \sigma_k} \frac{\partial \sigma \tilde{\sigma}_k}{\partial \sigma} \frac{\partial \sigma}{\partial s} = \sum_{k=1}^m \frac{\partial L(\chi)}{\partial \sigma_k} \sigma_k = \sum_{k=1}^m \frac{\partial L(\chi)}{\partial s_k}.
$$

(5) Recall that

$$Q(x) = \sum_{k=1}^{m} \omega_k \sigma_k \big[A_{xk} + E1(e^{-a_{xk}}) \big],$$

$$f_x = Q(x) - \sum_{k=1}^{m} \omega_k \sigma_k \big[A_{xk} + E1(e^{-a_{xk}}) \big],$$

$$a_{xk} = \frac{1}{\sigma_k} \left[u(x,0) - u(x,1) + \beta \sum_{y=1}^{K} (G_{xy}^0 - G_{xy}^1) Q(y) - \mu_k \right] + \gamma,$$

$$A_{xk} = \frac{1}{\sigma_k} \left[u(x,1) + \beta \sum_{y=1}^{K} G_{xy}^1 Q(y) + \mu_k \right].$$

Hence, we have

$$\frac{\partial f_x}{\partial Q(x)} = 1 - \sum_{k=1}^{m} \omega_k \sigma_k \left[\frac{\partial A_{xk}}{\partial Q(x)} + \frac{\partial E1(e^{-a_{xk}})}{\partial Q(x)} \right],$$

$$\frac{\partial A_{xk}}{\partial Q(x)} = \frac{\beta}{\sigma_k} G_{xx}^1,$$

$$\frac{\partial E1(e^{-a_{xk}})}{\partial Q(x)} = \exp\big[-e^{-a_{xk}} \big] \frac{\partial a_{xk}}{\partial Q(x)} = \exp\big[-e^{-a_{xk}} \big] \frac{\beta}{\sigma_k} (G_{xx}^0 - G_{xx}^1).$$

$$\implies \frac{\partial f_x}{\partial Q(x)} = 1 - \beta G_{xx}^2 - \beta(G_{xx}^1 - G_{xx}^2) \sum_{k=1}^{m} \omega_k \exp\big[-e^{-a_{xk}} \big].$$

It is easy to show that $\frac{\partial f_x}{\partial \omega_k} = -\sigma_k \big[A_{xk} + E1(e^{-a_{xk}}) \big] + \sigma_m \big[A_{xm} + E1(e^{-a_{xm}}) \big]$ and $\frac{\partial f_x}{\partial \mu_k} = -\omega_k \sigma_k \left[\frac{\partial A_{xk}}{\partial \mu_k} + \frac{\partial E1(e^{-a_{xk}})}{\partial \mu_k} \right] = -\omega_k \big[1 - \exp(-e^{-a_x k}) \big]$.

Finally,

$$\frac{\partial f_x}{\partial \sigma_k} = -\omega_k \big[A_{xk} + E1(e^{-a_{xk}}) \big] - \omega_k \sigma_k \left[\frac{\partial A_{xk}}{\partial \sigma_k} + \frac{\partial E1(e^{-a_{xk}})}{\partial \sigma_k} \right],$$

$$\frac{\partial A_{xk}}{\partial \sigma_k} = \frac{1}{\sigma_k^2} \big[v(x,1) + \mu_k \big] = \frac{1}{\sigma_k} A_{xk},$$

$$\frac{\partial E1(e^{-a_{xk}})}{\partial Q(x)} = \exp\big[-e^{-a_{xk}} \big] \frac{\partial a_{xk}}{\partial \sigma_k} = \exp\big[-e^{-a_{xk}} \big] \left[-\frac{1}{\sigma_k^2} (v(x,0) - v(x,1) - \mu_k) \right]$$

$$= \exp\big[-e^{-a_{xk}}\big]\frac{1}{\sigma_k}\big[\gamma - a_{xk}\big],$$

$$\implies \frac{\partial f_x}{\partial \sigma_k} = -\omega_k\big[E1(e^{-a_{xk}}) + \exp(-e^{-a_{xk}})(\gamma - a_{xk})\big].$$

Q.E.D.

APPENDIX E: SUPPLEMENT. IMPLEMENTATION DETAILS FOR RUSTS'S MODEL

E.1. *Norets and Tang (2013)*

Norets and Tang (2013) showed that if we assume that the distribution is unknown, then the preference-parameter $\theta = (\theta_0, \theta_1)$ is only set-identified and proposed an algorithm to compute the identified set. In Section 5.1 of their paper, they applied the method to the Rust's model. In their paper, the following setting of the utility function is used:

$$u(x, 0) = \theta_0 + \theta_1 x + \Delta\epsilon,$$
$$u(x, 1) = 0,$$

where θ_0 and θ_1 are the data generating values of preference parameters. The term $\Delta\epsilon = \epsilon_1 - \epsilon_2$ follows some unknown distribution F, which is assumed to have the same location-scale normalization as the logistic distribution. Specifically, $\int z \, dF(z) = 0$ and $\int_{M_F}^{\infty} z \, dF(z) = \log 2$, where M_F is the median of F or $F(M_F) = 0.5$. Figures 1 and 2 show the identified set of preference-parameters computed by their algorithm. For each θ inside of the black lines, there is a corresponding unknown distribution F such that the pair (θ, F) implies the true vector of CCPs.

E.2. *MCMC convergence*

Let N be the number of samples per state. We generate the data by multiplying N by the true CCP. We run MCMC with variable m with $N \in \{3, 10\}$. After each jump proposal block, we run 10 iterations of the HMC block. We obtained 500,000 draws.

When checking for convergence of the chain, we have to be careful because we are using a mixture model. It is well-known that if the likelihood of a mixture model with m components

has one mode for a fixed labeling of the components, then it can have $m!$ modes because the likelihood is invariant to a re-labeling of the components and there are $m!$ ways to label components. Although it is not possible to empirically detect label-switching(s) of our chain, if there was label-switching(s), it would be misleading, for example, to focus our attention on μ_1 when checking convergence of the chain. Geweke (2007) points out that this is not a problem as long as both the object of interest and the prior distribution are permutation-invariant. Note that we are using exchangeable priors and the object of our interest, the density estimate, is also permutation-invariant. We conduct a convergence diagnosis on the mean of ϵ or $\sum_{k=1}^{m} \omega_k \mu_k$ which is a permutation-invariant object. To check the convergence we perform the mean equality test for the first 10% and the last 50% of the samples. We conclude that the HMC samples for both $N = 3$ and $N = 10$ come from stationary distributions after a burn-in period of 10,000 draws.
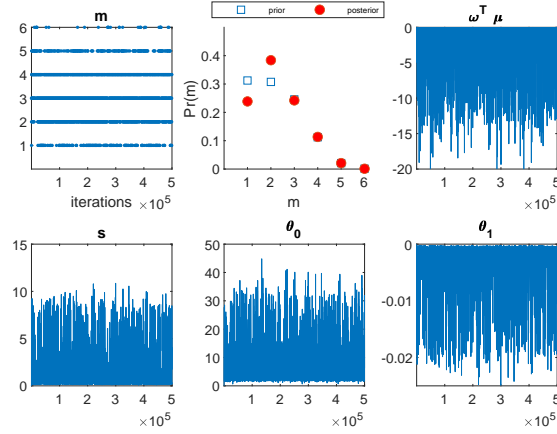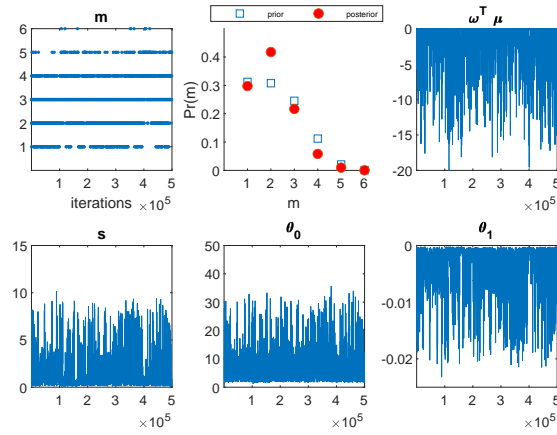
(a) $N = 3$

(b) $N = 10$

Figure 6: Posterior draws from Rust example. Trace plot of m (upper-left), p.m.f. of m (upper-middle), trace plots of $\sum_{k=1}^{m} \omega_k \mu_k$ (upper-right), s (bottom-left), θ_0 (bottom-middle), and θ_1 (bottom-right).

APPENDIX F: SUPPLEMENT. IMPLEMENTATION DETAILS FOR GILLESKIE'S MODEL

F.1. *More on State Transitions*

The individual contracts an illness and moves to the state $x = (1, 0, 0)$ with probability $\pi^S(H) = 1/[1 + \exp(\delta' H)]$. In each illness period $t \in \{1, \dots, T\}$, the individual recovers and returns to the state of being well with probability $\pi^W(x_t, d_t) = \exp(\eta^T E(x_t, d_t))/[1 +$

time	value	# possible values
t=0	$(0,0,0)$	1
t=1	$(1,0,0)$	1
t=2	$(2,v,a)$	$2^2 = 4$ $(v=0,1, a=0,1)$
t=3	$(3,v,a)$	$3^2 = 9$ $(v=0,1,2, a=0,1,2)$
$\vdots$	$\vdots$	$\vdots$
t	(t,v,a)	t^2 $(v=0,1,...t-1, a=0,1,...t-1)$
$\vdots$	$\vdots$	$\vdots$
T	(T,v,a)	T^2 $(v=0,1,...T-1, a=0,1,...T-1)$
Total		$K = 1+1+2^2+3^2+\cdots T^2 = 1+\sum_{j=1}^{T} j^2 = 1 + \frac{T(T+1)(2T+1)}{6}$

Table II: State transitions

$\exp(\eta^T E(x_t, d_t))]$, where $\eta^T E(x_t, d_t) = \eta_0 + \eta_1 v_{t+1} + \eta_2 (v_{t+1})^2 + \eta_3 a_{t+1} + \eta_4 (a_{t+1})^2 + \eta_5 v_{t+1} a_{t+1} + \eta_6 t + \eta_7 t^2 + \eta_8 t^3 + \xi' H$.

As shown in Table II, at the beginning of time t, there are t^2 possible values of the current state.

We summarize the state transitions below.

1. The initial state is $(0,0,0)$.

2. If the state is currently at $(0,0,0)$,

   - wp π^S, $(0,0,0) \rightarrow (1,0,0)$,

   - wp $1 - \pi^S$, $(0,0,0) \rightarrow (0,0,0)$.

3. In general, for $t = 1, ..., T-1$, if currently at (t, v_t, a_t),

   - wp $1 - \pi^W(x_t, d_t)$,

   $$(t, v_t, a_t) \rightarrow \begin{cases} (t+1, v_t, a_t), & \text{if } d_t = 0 \\ (t+1, v_t + 1, a_t), & \text{if } d_t = 1 \\ (t+1, v_t, a_t + 1), & \text{if } d_t = 2 \\ (t+1, v_t + 1, a_t + 1), & \text{if } d_t = 3 \end{cases}$$

   - wp $\pi^W(x_t, d_t)$, $(t, v_t, a_t) \rightarrow (0,0,0)$.

$d = 0$

	000	100	200	201	210	211
000	$1 - \pi^S$	π^S	0	0	0	0
100	$\pi^W(100,0)$	0	$1 - \pi^W(100,0)$	0	0	0
200	1	0	0	0	0	0
201	1	0	0	0	0	0
210	1	0	0	0	0	0
211	1	0	0	0	0	0

$d = 1$

	000	100	200	201	210	211
000	$1 - \pi^S$	π^S	0	0	0	0
100	$\pi^W(100,1)$	0	0	0	$1 - \pi^W(100,1)$	0
200	1	0	0	0	0	0
201	1	0	0	0	0	0
210	1	0	0	0	0	0
211	1	0	0	0	0	0

$d = 2$

	000	100	200	201	210	211
000	$1 - \pi^S$	π^S	0	0	0	0
100	$\pi^W(100,2)$	0	0	$1 - \pi^W(100,2)$	0	0
200	1	0	0	0	0	0
201	1	0	0	0	0	0
210	1	0	0	0	0	0
211	1	0	0	0	0	0

$d = 3$

	000	100	200	201	210	211
000	$1 - \pi^S$	π^S	0	0	0	0
100	$\pi^W(100,3)$	0	0	0	0	$1 - \pi^W(100,3)$
200	1	0	0	0	0	0
201	1	0	0	0	0	0
210	1	0	0	0	0	0
211	1	0	0	0	0	0

Figure 7: Example of transition matrices with $T = 2$.

4. The end of an episode

$$(T, v, a) \to (0,0,0) \quad \forall v, a,$$
$$(t, T-1, a) \to (0,0,0) \quad \forall t, a,$$
$$(t, v, T-1) \to (0,0,0) \quad \forall t, v.$$

Figure 7 shows the state transition matrices when $T = 2$. In this case, the possible states are $x = (t, v, a) \in \{000, 100, 200, 201, 210, 211\}$. The number of states is $K = 6$.

F.2. *More on Utility Function*

The system of utility functions in Gilleskie (1998) can be written as follows

$$u(d_t = 1, x_t, \epsilon_t) = \theta_1 + \theta_4 1(t=0) + \theta_6 C(x_t, 1)1(t>0) + \epsilon_{t1},$$
$$u(d_t = 2, x_t, \epsilon_t) = \theta_2 + \theta_4 1(t=0) + \theta_6 C(x_t, 2)1(t>0) + \epsilon_{t2},$$
$$u(d_t = 3, x_t, \epsilon_t) = \theta_3 + \theta_4 1(t=0) + \theta_6 C(x_t, 3)1(t>0) + \epsilon_{t3},$$
$$u(d_t = 0, x_t, \epsilon_t) = \theta_5 1(t=0) + \theta_6 C(x_t, 0)1(t>0).$$

Since we do not restrict the location and scale of ϵ_{tj}, $j = 1, 2, 3$, the values of the intercepts θ_j, $j = 1, 2, 3$ can be set to arbitrary values. The coefficient θ_4 is set to a large negative value

so that $d_t = 0$ is always chosen when $t = 0$ (i.e., the individual is well) and θ_5 can be set to an arbitrary positive value. Define $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)'$. Then

$$u(d_t = 1, x_t, \epsilon_t) = \Big(1{,}0{,}0{,}1(t = 0){,}0{,} \qquad C(x_t, 1)1(t > 0) \Big) \boldsymbol{\theta} + \epsilon_{t1} = \boldsymbol{Z_1}(x_t)\boldsymbol{\theta} + \epsilon_{t1},$$

$$u(d_t = 2, x_t, \epsilon_t) = \Big(0{,}1{,}0{,}1(t = 0){,}0{,} \qquad C(x_t, 2)1(t > 0) \Big) \boldsymbol{\theta} + \epsilon_{t2} = \boldsymbol{Z_2}(x_t)\boldsymbol{\theta} + \epsilon_{t2},$$

$$u(d_t = 3, x_t, \epsilon_t) = \Big(0{,}0{,}1{,}1(t = 0){,}0{,} \qquad C(x_t, 3)1(t > 0) \Big) \boldsymbol{\theta} + \epsilon_{t3} = \boldsymbol{Z_3}(x_t)\boldsymbol{\theta} + \epsilon_{t3},$$

$$u(d_t = 0, x_t, \epsilon_t) = \Big(0{,}0{,}0{,}0{,} \qquad 1(t = 0){,}C(x_t, 0)1(t > 0) \Big) \boldsymbol{\theta} \qquad = \boldsymbol{Z_0}(x_t)\boldsymbol{\theta},$$

where

$$\boldsymbol{Z_1}(x) = \big(Z_{11}(x), Z_{12}(x), Z_{13}(x), Z_{14}(x), Z_{15}(x), Z_{16}(x) \big) = \Big(1, 0, 0, 1(t = 0), 0, C(x, 1)1(t > 0) \Big),$$

$$\boldsymbol{Z_2}(x) = \big(Z_{21}(x), Z_{22}(x), Z_{23}(x), Z_{24}(x), Z_{25}(x), Z_{26}(x) \big) = \Big(0, 1, 0, 1(t = 0), 0, C(x, 2)1(t > 0) \Big),$$

$$\boldsymbol{Z_3}(x) = \big(Z_{31}(x), Z_{32}(x), Z_{33}(x), Z_{34}(x), Z_{35}(x), Z_{36}(x) \big) = \Big(0, 0, 1, 1(t = 0), 0, C(x, 3)1(t > 0) \Big),$$

$$\boldsymbol{Z_0}(x) = \big(Z_{01}(x), Z_{02}(x), Z_{03}(x), Z_{04}(x), Z_{05}(x), Z_{06}(x) \big) = \Big(0, 0, 0, 0, 1(t = 0), C(x, 0)1(t > 0) \Big).$$

Letting $\boldsymbol{Z}_{j,1:5}(x)$ to be the first 5 columns of $\boldsymbol{Z_j}(x)$, $j = 0, 1, 2, 3$, we can also write

$$u(1, x_t, \epsilon_t) = \boldsymbol{Z}_{1,1:5}(x_t)\hat{\boldsymbol{\theta}}_{1:5} + \theta_6 Z_{16}(x_t) + \epsilon_{t1},$$

$$u(2, x_t, \epsilon_t) = \boldsymbol{Z}_{2,1:5}(x_t)\hat{\boldsymbol{\theta}}_{1:5} + \theta_6 Z_{26}(x_t) + \epsilon_{t2},$$

$$u(3, x_t, \epsilon_t) = \boldsymbol{Z}_{3,1:5}(x_t)\hat{\boldsymbol{\theta}}_{1:5} + \theta_6 Z_{36}(x_t) + \epsilon_{t3},$$

$$u(0, x_t, \epsilon_t) = \boldsymbol{Z}_{0,1:5}(x_t)\hat{\boldsymbol{\theta}}_{1:5} + \theta_6 Z_{06}(x_t),$$

where $\hat{\boldsymbol{\theta}}_{1:5}$ is the vector of pre-specified coefficients and θ_6 is a parameter to be estimated.

F.3. *Data Generating Parameters*

We compute the data generating CCPs based on the following values, mostly based on estimates in Gilleskie (1998) for Type 2 illness with some adjustments so that the expected number of doctor visits and work absences roughly match with Gilleskie's sample. We use the following data-generating parameter values: $T = 8$, $\beta = 0.9$, $\theta_1 = -1.25$, $\theta_2 = -0.83$, $\theta_3 = -2.08$, $\theta_4 = -10000$, $\theta_5 = 1$, $\theta_6 = 0.0469$, $Y = 100$, $PC = 15$, $L = 0.7$, $\phi_1 = 5.6$, and $\phi_2 = -1.75$. For the state transition parameters η, δ, and ξ, we use the estimated values in Gilleskie (1998). We compute the data generating CCPs based on the parameter values mentioned above assuming that the data generating distribution of the utility shock is the following two-component mixture of extreme value distributions: $\sum_{k=1}^{2} \omega_k \phi\left(\epsilon; \mu_k, \sigma_k\right)$ where $\omega_1 = 0.5568, \omega_2 = 0.4432$, $\mu_1 = [-0.4683, 3.4628, -0.0914]'$, $\mu_2 = [0.9798, -2.2437, 1.3496]'$, $\sigma_1 = 3.7045$, and $\sigma_2 = 0.6378$. We then use the CCPs and the state-transition probabilities to sequentially generate 100 illness episodes.

F.4. *MCMC Plots from Gilleskie's Application*
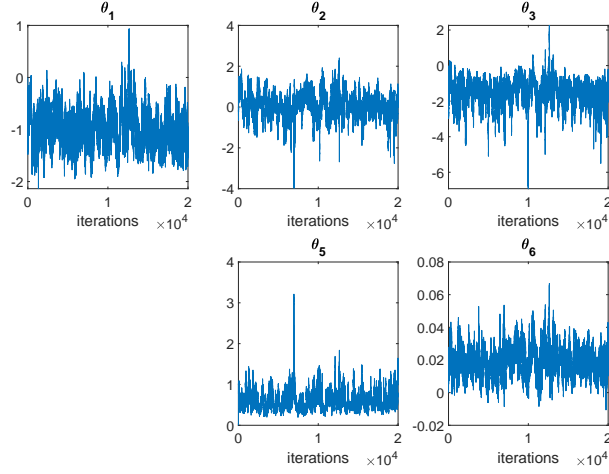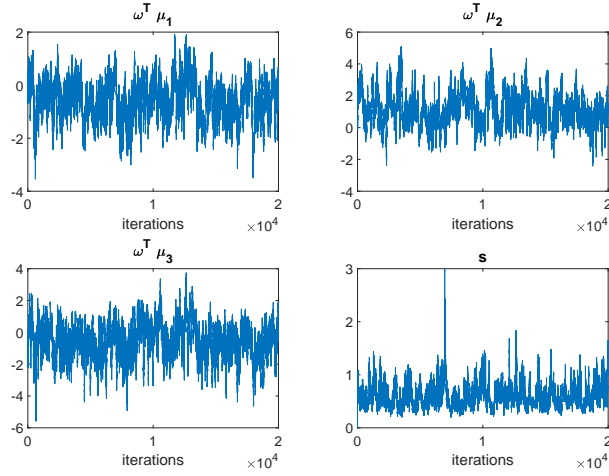
(a) Trace plots of utility parameters

(b) Trace plots of $\sum_{k=1}^{m} \omega_k \mu_{1k}$ (upper-left), $\sum_{k=1}^{m} \omega_k \mu_{2k}$ (upper-right), $\sum_{k=1}^{m} \omega_k \mu_{3k}$ (bottom-left), s (bottom-right), which is defined in Section 4.2.

Figure 8: Trace plots in Gilleskie model

F.5. *Prior sensitivity check for Gilleskie's model*

We consider two additional sets of priors for estimating Gilleskie's model in Section 7 and present results of estimation and counterfactual analysis. Despite slight differences, the overall findings remain the same. That is, the credible intervals for the identified sets introduced

in Section 5.4 include the true counterfactual value of mean doctor visits, and the HPD intervals get very close to it. In contrast, the confidence intervals computed via MLE assuming the dynamic logit misses the true value.

F.5.1. *Prior sensitivity check 1*

First, we consider increasing the prior variances for the component specific scale parameters σ_k's and the preference parameter θ_6. The priors are now specified as follows, $\underline{a} = 10$, $A_m = 0.05$, and $\tau = 5$, $\mu_{jk} \sim N(0, 3^2)$, $\log \sigma_k \sim N(0, 1)$, $\log \sigma \sim N(0, 0.01^2)$, and $\theta_6 \sim N(0, 5^2)$.

		MLE under logit		Semiparametric Bayes		
	True	Est.	95%CI (length)	Est.	95%HPD (length)	$\hat{B}_{0.95}^{I_{E(v)}}$ (length)
$E(v)$	1.49	1.42	[1.26, 1.58] (0.32)	1.44	[1.23, 1.65] (0.41)	[1.10, 1.94] (0.84)
$c.f.E(v)$	2.31	1.71	[1.53, 1.89] (0.36)	1.75	[1.37, 2.40] (1.03)	[1.14, 5.00] (3.86)

Table III: Estimated (Est.) $E(v)$ and $c.f.E(v)$: the MLE with its 95% confidence interval and the posterior mean with the 95% HPD interval and $\hat{B}_{0.95}^{I_{E(v)}}$.

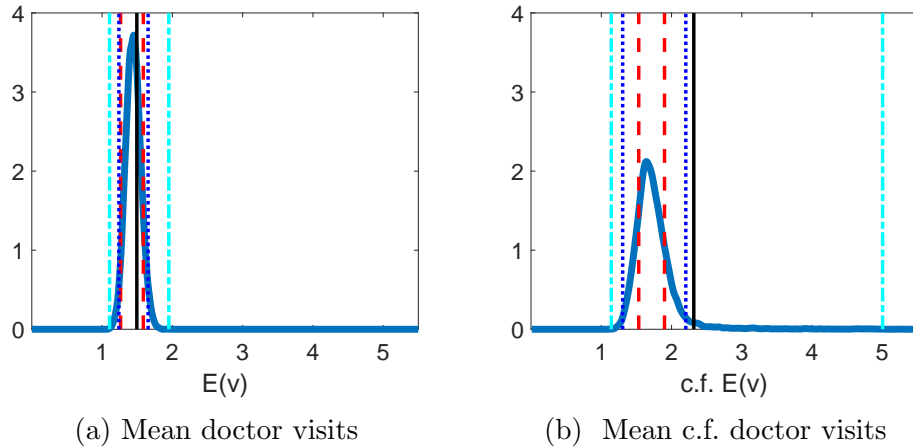(a) Mean doctor visits   (b) Mean c.f. doctor visits

Figure 9: Actual and counterfactual (c.f.) expected number of doctor visits. Posterior density (blue solid), 95% confidence interval (red dashed), 95% HPD interval (blue dotted), $\hat{B}_{0.95}^{I_{E(v)}}$ (light blue dash-dotted), and data generating value (black solid).

F.5.2. *Prior sensitivity check 2*

Second, we consider using a set of priors similar to the one used to estimate Rust model in Section 6. The priors are now specified as follows, $\underline{a} = 10$, $A_m = 0.05$, and $\tau = 5$, $\mu_{jk} \sim 0.5N(2, 2^2) + 0.5N(-3, 2^2)$, $\log \sigma_k \sim 0.4N(0, 1) + 0.6N(-1, 1)$, $\log \sigma \sim N(0, 0.01^2)$, and $\theta_6 \sim N(0, 3^2)$.

		MLE under logit		Semiparametric Bayes		
	True	Est.	95%CI (length)	Est.	95%HPD (length)	$\hat{B}_{0.95}^{I_{E(v)}}$ (length)
$E(v)$	1.49	1.42	[1.26, 1.58] (0.32)	1.44	[1.23, 1.64] (0.41)	[1.03, 1.79] (0.76)
$c.f.E(v)$	2.31	1.71	[1.53, 1.89] (0.36)	1.61	[1.32, 2.28] (0.96)	[1.12, 4.37] (3.25)

Table IV: Estimated (Est.) $E(v)$ and $c.f.E(v)$: the MLE with its 95% confidence interval and the posterior mean with the 95% HPD interval and $\hat{B}_{0.95}^{I_{E(v)}}$.

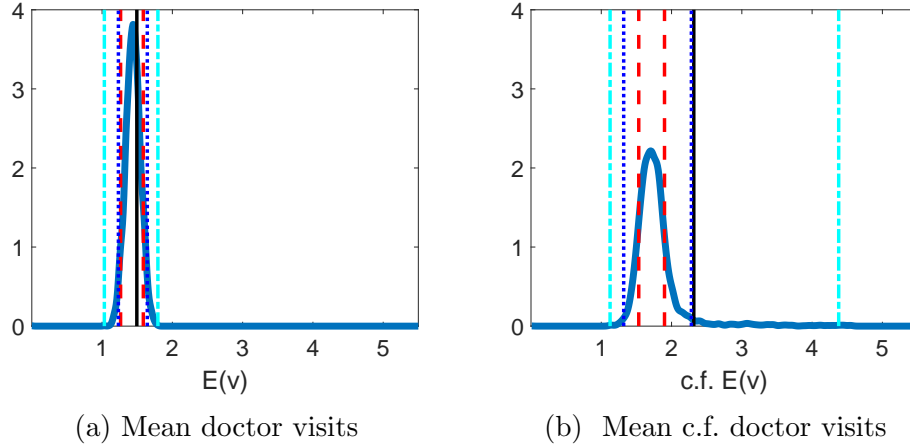(a) Mean doctor visits   (b) Mean c.f. doctor visits

Figure 10: Actual and counterfactual (c.f.) expected number of doctor visits. Posterior density (blue solid), 95% confidence interval (red dashed), 95% HPD interval (blue dotted), $\hat{B}_{0.95}^{I_{E(v)}}$ (light blue dash-dotted), and data generating value (black solid).

F.6. *Results of Gilleskie's model when the data generating process is dynamic logit*

Recall that in Section 7, we let the data generating distribution of the utility shocks to be a two-component mixture of extreme value distributions. Here we show results when it is logistic, and therefore the dynamic logit model is correctly specified. Table V shows estimation results. The overall findings are similar to the case with a mixture data generating distribution: the credible intervals of our semiparametric approach are much wider than the confidence intervals based on the dynamic logit assumption. However, now the confidence interval includes the true counterfactual value of doctor visits $E(v)$ (See Figure 11) while it did not under the mixture data generating distribution in Section 7. This is probably a consequence of the dynamic logit model being correctly specified in this particular example. We conducted some prior sensitivity checks in Section F.6.1 and confirm that the general findings are the same.

		MLE under logit		Semiparametric Bayes		
	True	Est.	95%CI (length)	Est.	95%HPD (length)	$\hat{B}_{0.95}^{I_{E(v)}}$ (length)
$E(v)$	1.53	1.50	[1.39, 1.62] (0.22)	1.52	[1.31, 1.74] (0.43)	[1.13, 1.94] (0.80)
$c.f.E(v)$	1.64	1.71	[1.59, 1.83] (0.24)	1.74	[1.34, 2.16] (0.82)	[1.03, 4.08] (3.04)

Table V: Estimated (Est.) $E(v)$ and $c.f.E(v)$: the MLE with its 95% confidence interval and the posterior mean with the 95% HPD interval and $\hat{B}_{0.95}^{I_{E(v)}}$.

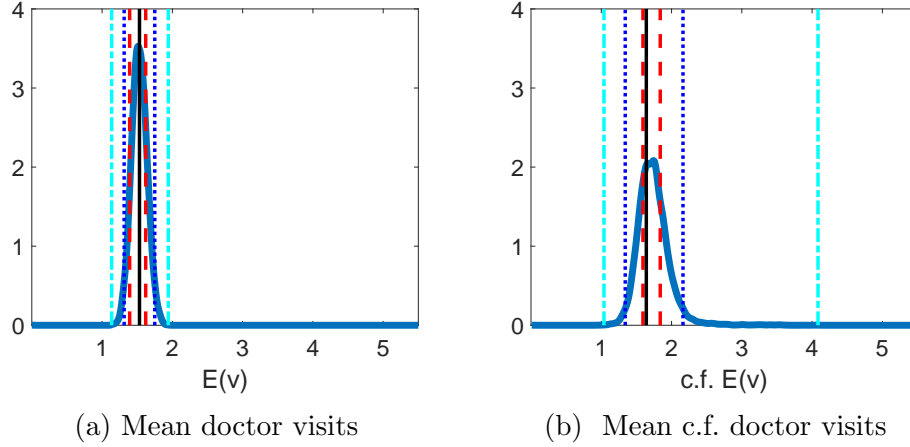(a) Mean doctor visits   (b) Mean c.f. doctor visits

Figure 11: Actual and counterfactual (c.f.) expected number of doctor visits. Posterior density (blue solid), 95% confidence interval (red dashed), 95% HPD interval (blue dotted), $\hat{B}_{0.95}^{I_{E(v)}}$ (light blue dash-dotted), and data generating value (black solid).

F.6.1. *Prior sensitivity checks*

Prior sensitivity check 1

First, we consider increasing the prior variances for the component specific scale parameters σ_k's and the preference parameter θ_6. The priors are now specified as follows, $\underline{a} = 10$, $A_m = 0.05$, and $\tau = 5$, $\mu_{jk} \sim N(0, 3^2)$, $\log \sigma_k \sim N(0, 1)$, $\log \sigma \sim N(0, 0.01^2)$, and $\theta_6 \sim N(0, 5^2)$. The results are presented below.

		MLE under logit			Semiparametric Bayes	
	True	Est.	95%CI (length)	Est.	95%HPD (length)	$\hat{B}_{0.95}^{I_{E(v)}}$ (length)
$E(v)$	1.53	1.50	[1.39, 1.62] (0.22)	1.52	[1.31, 1.74] (0.42)	[1.13, 1.95] (0.81)
$c.f.E(v)$	1.64	1.71	[1.59, 1.83] (0.24)	1.74	[1.32, 2.20] (0.87)	[1.16, 4.53] (3.37)

Table VI: Estimated (Est.) $E(v)$ and $c.f.E(v)$: the MLE with its 95% confidence interval and the posterior mean with the 95% HPD interval and $\hat{B}_{0.95}^{I_{E(v)}}$.

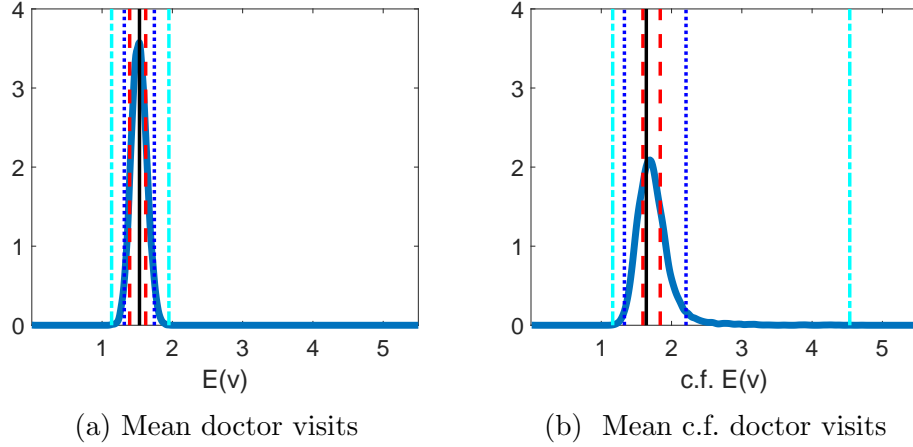(a) Mean doctor visits   (b) Mean c.f. doctor visits

Figure 12: Actual and counterfactual (c.f.) expected number of doctor visits. Posterior density (blue solid), 95% confidence interval (red dashed), 95% HPD interval (blue dotted), $\hat{B}_{0.95}^{I_{E(v)}}$ (light blue dash-dotted), and data generating value (black solid).

Prior sensitivity check 2

Second, we consider using a set of priors similar to the one used to estimate Rust model in Section 6. The priors are now specified as follows, $\underline{a} = 10$, $A_m = 0.05$, and $\tau = 5$, $\mu_{jk} \sim 0.5N(2, 2^2) + 0.5N(-3, 2^2)$, $\log \sigma_k \sim 0.4N(0,1) + 0.6N(-1,1)$, $\log \sigma \sim N(0, 0.01^2)$, and $\theta_6 \sim N(0, 3^2)$. The results are shown below.

		MLE under logit		Semiparametric Bayes		
	True	Est.	95%CI (length)	Est.	95%HPD (length)	$\hat{B}_{0.95}^{I_{E(v)}}$ (length)
$E(v)$	1.53	1.50	[1.39, 1.62] (0.22)	1.53	[1.31, 1.73] (0.42)	[1.11, 1.91] (0.80)
$c.f.E(v)$	1.64	1.71	[1.59, 1.83] (0.24)	1.75	[1.27, 2.33] (1.06)	[1.09, 5.29] (4.20)

Table VII: Estimated (Est.) $E(v)$ and $c.f.E(v)$: the MLE with its 95% confidence interval and the posterior mean with the 95% HPD interval and $\hat{B}_{0.95}^{I_{E(v)}}$.

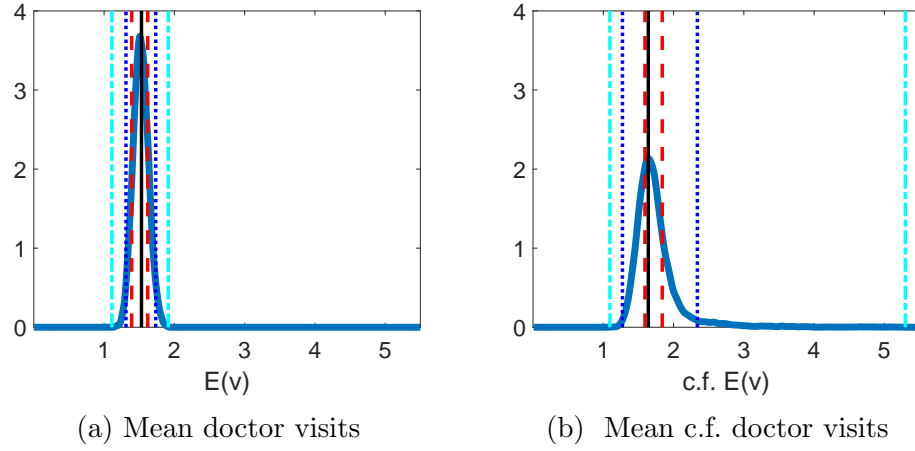(a) Mean doctor visits (b) Mean c.f. doctor visits

Figure 13: Actual and counterfactual (c.f.) expected number of doctor visits. Posterior density (blue solid), 95% confidence interval (red dashed), 95% HPD interval (blue dotted), $\hat{B}_{0.95}^{I_{E(v)}}$ (light blue dash-dotted), and data generating value (black solid).