Behavioral Foundations for Conditional Markov Models of Aggregate Data

Douglas Miller*
Department of Economics
University of Missouri

September 1, 2007

Abstract

Conditional Markov chain models of observed aggregate share–type data have been used by economic researchers for several years, but the classes of models commonly used in practice are often criticized as being purely ad hoc because they are not derived from micro–behavioral foundations. The primary purpose of this paper is to show that the estimating equations commonly used to estimate these conditional Markov chain models may be derived from the assumed statistical properties of an agent–specific discrete decision process. Thus, any conditional Markov chain model estimated from these estimating equations may be compatible with some underlying agent–specific decision process. The secondary purpose of this paper is to use an information theoretic approach to derive a new class of conditional Markov chain models from this set of estimating equations. The proposed modeling framework is based on the behavioral foundations but does not require specific assumptions about the utility function or other components of the agent–specific discrete decision process. The asymptotic properties of the proposed estimators are developed to facilitate model selection procedures and classical tests of behavioral hypotheses.

Keywords: controlled stochastic process, Fréchet derivative, first–order Markov chain, Cressie–Read power divergence criterion

JEL Classification: C40, C51

*Address for correspondence: 118 Professional Building, University of Missouri, Columbia, MO 65211

1 Introduction

In this paper, we consider first–order Markov chain models of events with a finite number of outcomes measured at discrete time intervals. From the agent–specific or micro prespective, the decision outcomes for agent $i = 1, 2, \ldots, n$ are denoted $Y(i, k, t)$ with finite states $k = 1, 2, \ldots, K$ at time $t = 0, 1, 2, \ldots, T$. For example, count–based outcomes from the decision process are

$$Y(i,k,t) = \begin{cases} 1 & \text{if agent } i \text{ selects state } k \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

If the decision outcomes exhibit first–order Markov character, the dynamic behavior of the agents may be represented by conditional transition probabilities $\pi(j, k, t)$, which represent the probability that agent i moves from state $j = 1, 2, \ldots, K$ to state k at time t (i.e., the transition probabilities only depend on the previous state). Given observations on the micro behavior $Y(i, k, t)$, a number of researchers have used the discrete Markov decision process framework to model the agent–specific dynamic economic behavior. Excellent discussions of the micro–based approach are provided by Rust (1988) and by Hotz and Miller (1993).

In many cases encountered in practice, the observed information may be limited to the aggregated outcomes, $Y(k, t) = n^{-1} \sum_{i=1}^{n} Y(i, k, t)$ for each k and t. These aggregate or macro–data represent the frequency or share of agent–specific outcomes observed in each state k at time t. The controlled stochastic process approach has also been used to model economic decisions given aggregate or macro data, and a prominent example is the article by Berry, Levinsohn, and Pakes (1995). In this case, the authors consider consumer demand for durable goods (i.e., automobiles) and require specific assumptions about the functional form of the utility function, the producer cost function, and the conditional distributions of the state variables. However, the traditional approach to estimating the Markov transition probabilities is based on the estimating equations

$$Y(k,t) = \sum_{j=1}^{K} Y(j, t-1)\pi(j,k,t) \tag{1}$$

that link the observed aggregate data to the transition probabilities. The transition probabilities may be directly estimated from (1) if the Markov process is unconditional or stationary such that $\pi(j, k, t) = \pi(j, k, t + s)$ for all integers s. In this case, the estimation problem reduces to a set of $(K-1)\times(T-1)$ estimating equations with $(K-1)\times(K-1)$ unknown Markov transition probabilities, $\pi(j, k)$, after incorporating the row–sum conditions, $\sum_{k=1}^{K} \pi(j, k, t) = 1$. Following the discussion in Lee, Judge, and Zellner (1977), we may use least squares, quadratic programming, or other established estimation procedures to directly compute estimates of the unconditional Markov transition probabilities if $T \geq K$.

Before continuing, we should note that previous contributors to this literature typically referred to Markov chain models in which $\pi(j,k,t)$ varies with t as non–stationary Markov chains. However, to distinguish this form of non–stationarity from the more widely studied forms of explosive stochastic processes (e.g., random walks), many authors now refer to non–stationary Markov chains as conditional Markov chains. In this case, we have $T \times (K-1) \times (K-1)$ unknown transition probabilities in the $(K-1) \times (T-1)$ estimating equations, and the estimation problem is ill–posed. To overcome the ill–posed character of the problem, many authors have specified parametric functional forms for the Markov transition probabilities. For example, MacRae (1977) follows Theil (1969) and recommends the logistic functional form

$$\pi(j,k,t) = \frac{\exp\left(\boldsymbol{\beta}'_{jk}\boldsymbol{z}_{t-1}\right)}{\sum_{k=1}^{K}\exp\left(\boldsymbol{\beta}'_{jk}\boldsymbol{z}_{t-1}\right)} \tag{2}$$

The logistic specification has some advantages for the purposes of estimation and interpretation (i.e., the fitted transition probabilities satisfy $\pi(j,k,t) > 0$ and the row–sum conditions), and alternative models have been proposed by other authors, including Lee, Judge, and Zellner (1977) and Davis, Heathcote, and O'Neill (2002).

The key criticism of the Markov models based on the estimating equations (1) and specifications like (2) are that the models are convenient but purely ad hoc (MacRae notes this in her own paper). However, we show in the next section that models based on these estimating equations are compatible with the microfoundations of a Markov decision process and are not necessarily ad hoc. In Section 3, we use the properties of a sample analog of the Markov decision process outcomes derived in Section 2 to derive a new class of models for conditional Markov transition probabilities. The asymptotic properties of the proposed parameter estimators are developed in section 4, and concluding remarks are provided in the final section. Endnotes and proofs of the supporting results are provided at the end of the paper.

2 Decisions under a Discrete Markov Decision Process

Following Rust (1994) and White and White (1989), the agent–specific decisions $Y(i,k,t)$ may be viewed as outcomes from a controlled stochastic process. An individual firm or agent selects an optimal sequence of actions from a finite choice set, $\delta_t \in \{1,2,\ldots,K\}$ (i.e., $Y(i,k,t)=1$ if $\delta_t = k$). The problem for individual i is to solve

$$\max_{\boldsymbol{\delta}} \sum_{t=0}^{\infty} \tau^t \; \mathrm{E}\left[u\left(\mathbf{x}_t,\delta_t\right) + \varepsilon_t\left(\delta_t\right)\right] \tag{3}$$

where $\tau \in (0,1)$ is the intertemporal discount factor. In words, the agent selects a set of actions over time in order to maximize the discounted sum of expected agent–specific utilities. The common component of agent utility $u(\mathbf{x}_t, \delta_t)$ is a function of the decision (δ_t) and of the H-vector of observable state variables $(\boldsymbol{x}_t)$. The agent–specific or idiosyncratic component of the utility function is represented by an unobservable state variable (ε_t). The state variables $\mathbf{x}_t$ and ε_t are viewed as random variables with joint distributions that reflect the cross–sectional variation in the these outcomes across individuals $i = 1, 2, \ldots, n$. Accordingly, the control variable δ_t is a random variable resulting from the agent's optimal choice conditional on the outcomes of the state variables.

2.1 Properties of the Markov Decision Process

Next, we can use the assumed properties of the Markov decision process to characterize the Markov transition probabilities for the controlled decision outcomes, $\pi(\delta_t | \mathbf{x}_t, \delta_{t-1})$. To begin, we follow Rust (1994) and define the social surplus function

$$G(v(\boldsymbol{x}_t) | \boldsymbol{x}_t) = \int \left[\max_{\delta_t} \{ u(\boldsymbol{x}_t, \delta_t) + \varepsilon_t(\delta_t) \} \right] q(\varepsilon_t | \boldsymbol{x}_t) \, d\varepsilon_t \tag{4}$$

which represents the sum of maximal utilities across agents. Note that the integral is computed with respect to $q(\varepsilon_t | \boldsymbol{x}_t)$, the conditional distribution of the idiosyncratic or agent–specific state variable ε_t. The social surplus function exists if the conditional distribution $q(\varepsilon_t | \boldsymbol{x}_t)$ has finite first moments (see Rust, Theorem 3.1). The additional conditions required for the existence of a solution to the Markov decision problem are:

- **A1**: The joint distribution of $\mathbf{x}_t$ and ε_t is composed of conditionally independent Markov processes such that[1]

$$p(\boldsymbol{x}_t, \varepsilon_t | \boldsymbol{x}_{t-1}, \varepsilon_{t-1}, \delta_{t-1}) = q(\varepsilon_t | \boldsymbol{x}_t) \ \varphi(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, \delta_{t-1}) \tag{5}$$

- **A2**: The utility function is additively separable in the common and idiosyncratic components, $u(\mathbf{x}_t, \delta_t) + \varepsilon_t(\delta_t)$.

- **A3**: The common component of the utility function $u(\boldsymbol{x}_t, \delta_t)$ is upper semicontinuous in $\boldsymbol{x}_t$ with bounded expected value.

- **A4**: The conditional distribution $\varphi(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, \delta_{t-1})$ is weakly continuous such that for each continuous real–valued function $h(\boldsymbol{x}_t)$

$$\int h(\boldsymbol{x}_t) \varphi(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, \delta_{t-1}) \, d\boldsymbol{x}_t$$

  is continuous in $\boldsymbol{x}_{t-1}$ for each δ_{t-1}.

- **A5**: The common component of the utility function $u(\boldsymbol{x}_t, \delta_t)$ is a bounded Borel–measurable function and the expected value of the social surplus function

$$\int G\left(h\left(\boldsymbol{x}_t\right) \left|\boldsymbol{x}_t\right.\right) \; \varphi\left(\boldsymbol{x}_t \left|\boldsymbol{x}_{t-1}, \delta_{t-1}\right.\right) \; d\boldsymbol{x}_t$$

is a bounded Borel–measurable function for all bounded Borel–measurable functions $h(\boldsymbol{x}_t)$.

For present purposes, we use condition A2 to show that the state variables $\mathbf{x}_t$ and ε_t are orthogonal in expectation:

- **Lemma 1**: Under A2, $\mathrm{E}\left[\varepsilon_t \left|\boldsymbol{x}_t\right.\right] = 0$ and $\mathrm{E}\left[\boldsymbol{x}_t' \varepsilon_t\right] = \mathbf{0}$ for all t.

Under conditions A1–A5, Rust shows (Theorem 3.2) that the optimal value function v is a fixed–point satisfying Bellman's equation

$$\Psi(v) = u\left(\boldsymbol{x}_t, \delta_t\right) + \tau \int G\left(v\left(\boldsymbol{x}_t\right) \left|\boldsymbol{x}_t\right.\right) \; \varphi\left(\boldsymbol{x}_t \left|\boldsymbol{x}_{t-1}, \delta_{t-1}\right.\right) \; d\boldsymbol{x}_t \tag{6}$$

where Ψ is a contraction mapping[2] and the integrand G is the social surplus function. The controlled process $(\boldsymbol{x}_t, \delta_t)$ has a first–order Markov representation

$$P\left(\boldsymbol{x}_t, \delta_t \left|\boldsymbol{x}_{t-1}, \delta_{t-1}\right.\right) = \phi\left(\delta_t \left|\boldsymbol{x}_t\right.\right) \; \varphi\left(\boldsymbol{x}_t \left|\boldsymbol{x}_{t-1}, \delta_{t-1}\right.\right) \tag{7}$$

(see Rust, Theorem 3.3) where $\phi\left(\delta_t \left|\boldsymbol{x}_t\right.\right)$ on the righthand–side is the conditional choice probability for action $\delta_t = k$ given state $\boldsymbol{x}_t$. The conditional choice probability is also the Fréchet derivative of G with respect to the value function[3]

$$\phi\left(\delta_t \left|\boldsymbol{x}_t\right.\right) = \frac{\partial G}{\partial v} = \int_{\Lambda_k} q\left(\varepsilon_t \left|\boldsymbol{x}_t\right.\right) \; d\varepsilon_t \tag{8}$$

where Λ_k is the subset of the sample space for ε_t associated with state $\delta_t = k$. In words, the conditional probability that $\delta_t = k$ at time t is equal to the share of agents for which this state is the optimal decision.

The joint Markov transition probabilities (7) are used to form the temporal relationship among the unconditional (joint) distribution of x_t and δ_t

$$\omega\left(\boldsymbol{x}_t, \delta_t\right) = \int\int P\left(\boldsymbol{x}_t, \delta_t \left|\boldsymbol{x}_{t-1}, \delta_{t-1}\right.\right) \; \omega\left(\boldsymbol{x}_{t-1}, \delta_{t-1}\right) \; d\boldsymbol{x}_{t-1} \; d\delta_{t-1} \tag{9}$$

We can factor the joint distribution as $\omega\left(\boldsymbol{x}_t, \delta_t\right) = \phi\left(\delta_t \left|\boldsymbol{x}_t\right.\right) \varphi\left(\boldsymbol{x}_t\right)$ where $\varphi\left(\boldsymbol{x}_t\right)$ is the unconditional distribution of the vector of observable state variables, and we can factor the Markov transition probability as $P\left(\boldsymbol{x}_t, \delta_t \left|\boldsymbol{x}_{t-1}, \delta_{t-1}\right.\right) = \pi\left(\delta_t \left|\mathbf{x}_t, \delta_{t-1}\right.\right) \; \varphi\left(\boldsymbol{x}_t \left|\boldsymbol{x}_{t-1}\right.\right)$. By substitution of these components into (9), we have

$$\phi\left(\delta_t \left|\boldsymbol{x}_t\right.\right) \varphi\left(\boldsymbol{x}_t\right) = \int\int \pi\left(\delta_t \left|\mathbf{x}_t, \delta_{t-1}\right.\right) \varphi\left(\boldsymbol{x}_t \left|\boldsymbol{x}_{t-1}\right.\right) \phi\left(\delta_{t-1} \left|\boldsymbol{x}_{t-1}\right.\right) \varphi\left(\boldsymbol{x}_{t-1}\right) \; d\boldsymbol{x}_{t-1} \; d\delta_{t-1} \tag{10}$$

After completing the integral with respect to $\boldsymbol{x}_{t-1}$ in (10), we form

$$\phi\left(\delta_t \left| \boldsymbol{x}_t \right.\right) = \int \pi\left(\delta_t \left| \mathbf{x}_t, \delta_{t-1} \right.\right) \phi\left(\delta_{t-1}\right) \; d\delta_{t-1} \tag{11}$$

The resulting set of equations that link the conditional choice probabilities for each time period through the Markov transition probabilities. To simplify notation, we denote the conditional probability that outcomes occur in state k at time t as $\phi\left(\delta_t = k \left| \boldsymbol{x}_t \right.\right) \equiv \phi(k, t)$ and the conditional Markov probability as $\pi\left(\delta_t = k \left| \mathbf{x}_t, \delta_{t-1} = j \right.\right) \equiv \pi(j, k, t)$. Further, given that we only consider a finite number of states, (11) may be restated in more familiar form as

$$\phi(k, t) = \sum_{j=1}^{K} \phi(j, t-1)\pi(j, k, t) \tag{12}$$

which is the set of characteristic equations that link the conditional first–order Markov transition probabilities to the conditional choice probabilities. Although the conditional choice probabilities $\phi(k, t)$ are not observable, we show in the following subsection that we can use the aggregate shares as estimates of these probabilities.

Finally, we can further refine (12) to derive the set of characteristic equations for time–invariant (stationary) Markov chain models. First, multiply both sides of (11) by $\varphi\left(\boldsymbol{x}_t\right)$ and then integrate with respect to $\boldsymbol{x}_t$. The resulting set of equations

$$\phi\left(\delta_t\right) = \int \pi\left(\delta_t \left| \delta_{t-1} \right.\right) \phi\left(\delta_{t-1}\right) \; d\delta_{t-1} \tag{13}$$

link the time–invariant Markov transition probabilities $\pi\left(\delta_t \left| \delta_{t-1} \right.\right)$ to the unconditional choice probabilities. Using the notation defined in the preceding paragraph, we can restate (13) for a finite Markov chain in more familiar form as

$$\phi(k, t) = \sum_{j=1}^{K} \phi(j, t-1)\pi(j, k) \tag{14}$$

where $\pi\left(\delta_t = k \left| \delta_{t-1} = j \right.\right) \equiv \pi(j, k)$.

2.2 Sample Analogs to the Markov Decision Process

The objective of this subsection is to derive a sample analog of the conditional choice probability in (8) and the characteristic equations in (12) and (14) based on the observed aggregate data. The key step in linking the Markov decision process to the observable data is to note that the available data may represent partial or incomplete observations. Although the conditional choice probabilities in (8) are defined with respect to decision outcomes for all agents, observed aggregate data typically represent an incomplete enumeration of agents.

By restating the results from the previous section as sample analogs, we can form estimating equations that link the Markov transition probabilities to the observed aggregate data.

To begin, suppose the aggregate data for a particular set of observations are gathered from a sample of n agents. Let $q_n(\varepsilon_t | \boldsymbol{x}_t)$ represent the empirical conditional distribution of idiosyncratic factors in the observed sample. We assume the empirical conditional distribution converges to the population distribution as $\rho(q_n, q) \to 0$ under some measure ρ as the number of agents in the sample increases ($n \to \infty$). Accordingly, the empirical social surplus function analogous to (4) is

$$G_n(v(\boldsymbol{x}_t) | \boldsymbol{x}_t) = \int \left[\max_{\delta_t} \{u(\boldsymbol{x}_t, \delta_t) + \varepsilon_t(\delta_t)\} \right] q_n(\varepsilon_t | \boldsymbol{x}_t) \; d\varepsilon_t \tag{15}$$

under the empirical distribution q_n. If G is ρ–Fréchet differentiable at q, then we can express the empirical social surplus function as $G_n(\boldsymbol{x}_t, \delta_t) = G(\boldsymbol{x}_t, \delta_t) + \eta_n(q_n - q)$, where η_n is a linear functional that represents the Fréchet differential of G_n. In words, the empirical social surplus function G_n converges to the actual social surplus function G as $n \to \infty$.

Following our discussion of (8), we can interpret the observed share of outcomes in state k at time t as the ρ–Fréchet derivative of the empirical social surplus function, $Y(k,t) = \partial G_n / \partial v$. Then, we can use our differential expression for G_n to derive

$$Y(k,t) = \frac{\partial G_n}{\partial v} = \frac{\partial G}{\partial v} + \frac{\partial \eta_n}{\partial v} = \phi(k,t) + \xi_n(k,t) \tag{16}$$

where ξ_n is the Fréchet derivative of η_n at v. We interpret $\xi_n(k,t)$ as the sampling error in the observed aggregate share relative to the conditional probability $\phi(k,t)$. Given $\mathrm{E}[\varepsilon_t | \boldsymbol{x}_t] = 0$ from Lemma 1, we can prove:

- **Lemma 2**: The conditional moment $\mathrm{E}[\xi_n(k,t) | \boldsymbol{x}_t] = 0$ holds such that

$$\mathrm{E}\left[\boldsymbol{x}_t' \xi_n(k,t)\right] = \boldsymbol{0} \tag{17}$$

given the condition that integration and differentiation are exchangeable operations under $q(\varepsilon_t | \boldsymbol{x}_t)$ and $q_n(\varepsilon_t | \boldsymbol{x}_t)$.

It is important to note that Lemma 2 will not hold if the difference between q_n and q represents selectivity biases (e.g., small firms are under–represented in the observed sample) or other characteristics such that the sampling errors $\xi_n(k,t)$ are correlated with the state variables $\boldsymbol{x}_t$. In this case, we can extend Lemma 2 if there exists a set of H instrumental variables $\boldsymbol{z}_t$ such that

$$\mathrm{E}\left[\boldsymbol{z}_t' \xi_n(k,t)\right] = 0 \tag{18}$$

We can now complete the sample analog to (12) by substituting $Y(k,t)-\xi_n(k,t) = \phi(k,t)$ from (16) into (12) as

$$Y(k,t) - \xi_n(k,t) = \sum_{j=1}^{K} \left(Y(j,t-1) - \xi_n(j,t-1)\right) \pi(j,k,t) \tag{19}$$

which may be rearranged to form

$$Y(k,t) = \sum_{j=1}^{K} Y(j,t-1)\pi(j,k,t) + e\,(k,t) \tag{20}$$

where the composite error term is

$$e\,(k,t) \equiv \xi_n(k,t) - \sum_{j=1}^{K} \xi_n(j,t-1)\pi(j,k,t-1) \tag{21}$$

Note that (20) is the sample analog of (12) and is a noisy version of the traditional estimating equation (1). Accordingly, the sample analog of (14) may be stated as

$$Y(k,t) = \sum_{j=1}^{K} Y(j,t-1)\pi(j,k) + e\,(k,t) \tag{22}$$

for the time–invariant (stationary) case. Thus, any first–order Markov chain model for aggregate data based on (20) or (22) is compatible with agent–specific behavior under this Markov decision process. For empirical purposes, the remaining task is to choose a feasible specification of the Markov transition probabilities.

3 Modeling the Conditional Transition Probabilities

The purpose of this section is to derive a new class of conditional Markov chain models that are compatible with the Markov decision process approach and have favorable statistical properties. The proposed modeling approach is based on a set of estimating equations derived from (20). In particular, we can show that

$$\mathrm{E}\left[\boldsymbol{z}_t' e(k,t)\right] = 0 \tag{23}$$

where $\boldsymbol{z}_t$ is an appropriate set of instrumental variables (perhaps equal to the set of state variables $\boldsymbol{x}_t$ if (17) holds). Then, we can use a goodness–of–fit criterion to derive a conditional Markov model from the sample analog of (23). In this section, we propose using the Cressie–Read power divergence criterion as a goodness–of-fit measure for model selection purposes.

3.1 Cressie–Read Power Divergence (PD) Criterion

The Cressie–Read power divergence (PD) statistic (Cressie and Read, 1984; Read and Cressie, 1988; Baggerly, 1998) is defined for a set of first–order finite and discrete conditional Markov probabilities $\boldsymbol{\pi}$ as

$$I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha) = \frac{2}{\alpha(1+\alpha)} \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \pi(j,k,t) \left[\left(\frac{\pi(j,k,t)}{\overline{\pi}(j,k,t)} \right)^{\alpha} - 1 \right] \tag{24}$$

Here, the PD statistic measures the pseudo–distance between $\boldsymbol{\pi}$ and a set of reference transition probabilities $\overline{\boldsymbol{\pi}}$. The reference distributions may be based on pre–sample or prior information that the researcher may want to add to the Markov estimation problem. Read and Cressie note that (24) encompasses a family of estimation objective functions indexed by α for discrete probability distributions (see Appendix section A.2). For our purposes, the PD statistic is a useful estimation criterion due to its information theoretic properties and because it is strictly convex in $\boldsymbol{\pi}$.

3.2 Minimum Power Divergence (MPD) Models

Given reference weights $\overline{\boldsymbol{\pi}}$ and the orthogonality conditions (23), our proposed method for deriving a model of the condtional Markov transition probabilities is to choose $\boldsymbol{\pi}$ that satisfies the estimating equations and is closest to $\overline{\boldsymbol{\pi}}$ under the PD criterion. The resulting minimum power divergence (MPD) transition probabilities satisfy the behavioral conditions of the Markov decision process while remaining 'least–informative' relative to the set of reference weights. Formally, the MPD problem may be solved by choosing $\boldsymbol{\pi}$ to minimize $I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha)$ (for some α) subject to the sample analog of (23)

$$\sum_{t=1}^{T} \mathbf{z}_t' \left(y(k,t) - \sum_{j=1}^{K} y(j,t-1)\pi(j,k,t) \right) = \mathbf{0} \tag{25}$$

for each $j = 2, \ldots, K$ (after normalization) and the row–sum constraint

$$\sum_{k=1}^{K} \pi(j,k,t) = 1 \tag{26}$$

for all j and t.

For demonstration purposes, we consider two prominent cases of the MPD family. First, the Kullback–Leibler cross–entropy functional is

$$I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha \to 0) = \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \pi(j,k,t) \ln \left[\frac{\pi(j,k,t)}{\overline{\pi}(j,k,t)} \right] \tag{27}$$

Under the cross–entropy criterion, the MPD model of the conditional Markov transition probabilities is

$$\widetilde{\pi}(j,k,t) = \frac{\overline{\pi}(j,k,t)\exp\left(Y(j,t-1)\boldsymbol{z}_t'\widetilde{\boldsymbol{\lambda}}_k\right)}{\sum_{k=1}^{K}\overline{\pi}(j,k,t)\exp\left(Y(j,t-1)\boldsymbol{z}_t'\widetilde{\boldsymbol{\lambda}}_k\right)} \tag{28}$$

The vector $\widetilde{\boldsymbol{\lambda}}_k$ is the set of optimal Lagrange multipliers on (25). In general, there is no explicit solution for the Lagrange multipliers, and the optimal values must be numerically determined. If the reference distribution $\overline{\boldsymbol{\pi}}$ satisfies (25), $\widetilde{\boldsymbol{\lambda}}_k = \mathbf{0}$ for each k and $\widetilde{\pi}(j,k,t) = \overline{\pi}(j,k,t)$ for all j, k, and t.

If the reference weights are discrete uniform, the Kullback–Leibler cross–entropy functional is negatively proportional to Shannon's entropy functional

$$-I\left(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha \to 0\right) \propto -\sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K}\pi(j,k,t)\ln\left(\pi(j,k,t)\right) \tag{29}$$

In this case, the MPD model produces Markov transition probabilities that are equivalent to models derived under Jaynes' maximum entropy criterion. The maximum entropy transition probabilities take the form

$$\widehat{\pi}(j,k,t) = \frac{\exp\left(-Y(j,t-1)\boldsymbol{z}_t'\widehat{\boldsymbol{\lambda}}_k\right)}{\sum_{k=1}^{K}\exp\left(-Y(j,t-1)\boldsymbol{z}_t'\widehat{\boldsymbol{\lambda}}_k\right)} \tag{30}$$

The H-vector $\widehat{\boldsymbol{\lambda}}_k$ is the set of optimal Lagrange multipliers for the constraints (25). As in the minimum cross–entropy case, a closed–form solution for $\widehat{\boldsymbol{\lambda}}_k$ is not known to exist, and the maximum entropy estimates of $\boldsymbol{\pi}$ must be numerically determined. We can also see that the maximum entropy model is a special case of the cross–entropy model in which $\overline{\boldsymbol{\pi}}$ is discrete uniform and $\widetilde{\boldsymbol{\lambda}}_k = -\widehat{\boldsymbol{\lambda}}_k$ for each k. In either case, the entropy–based approach yields a model of the Markov transition probabilities in terms of the observed explanatory variables and a finite number of estimable parameters.

4 Sampling Properties of the MPD Estimators

To demonstrate the large–sample properties of the MPD parameter estimators, we consider the special case of the Kullback–Leibler cross–entropy functional (i.e., $\alpha \to 0$). First, the constrained MPD problem may be reduced to an unconstrained form by concentrating the Lagrangian objective function. If we view the implicitly defined Markov transition probabilities (28) as an intermediate solution and substitute these forms back into the Lagrangian expression, the MPD objective function reduces to

$$m\left(\boldsymbol{\lambda}\right) = \sum_{t=1}^{T}\sum_{k=1}^{K} y(k,t)\boldsymbol{z}_t'\boldsymbol{\lambda}_k - \sum_{t=1}^{T}\sum_{j=1}^{K}\ln\left[\overline{\pi}(j,1,t) + \sum_{k=2}^{K}\overline{\pi}(j,k,t)\exp\left(y(j,t-1)\boldsymbol{z}_t'\boldsymbol{\lambda}_k\right)\right] \tag{31}$$

By the saddle–point property of the constrained minimization problem, (31) is strictly concave in $\boldsymbol{\lambda}$ such that $m(\boldsymbol{\lambda}) < m\left(\widetilde{\boldsymbol{\lambda}}\right)$ $\forall\ \boldsymbol{\lambda} \neq \widetilde{\boldsymbol{\lambda}}$. Accordingly, the optimal values of $\boldsymbol{\lambda}$ can be computed by maximizing $m(\boldsymbol{\lambda})$ by choice of $\boldsymbol{\lambda}$. The gradient vector of $m(\boldsymbol{\lambda})$ is simply the set of sample analog estimating equations (25), and Newton–Raphson or related optimizations algorithms may be used to compute the optimal Lagrange multipliers.

The large–sample properties of the minimum cross–entropy estimator (including the maximum entropy estimator as a special case) may be derived under the following regularity conditions:

- **B1**: There exists $\boldsymbol{\lambda}^0$ such that $\pi\left(\boldsymbol{\lambda}^0; j, k, t\right) = \pi^0(j, k, t)$ for all j, k, and t.
- **B2**: The sample analog of (23) is $\sqrt{T}$-consistent such that

$$T^{-1} \sum_{t=1}^{T} \boldsymbol{z}_t' \left(Y(k,t) - \sum_{j=1}^{K} Y(j, t-1) \pi^0(j,k,t) \right) \xrightarrow{p} \mathbf{0} \tag{32}$$

- **B3**: The moment condition (32) is asymptotically normal as

$$T^{-1/2} \sum_{t=1}^{T} \boldsymbol{z}_t' \left(Y(k,t) - \sum_{j=1}^{K} Y(j, t-1) \pi^0(j,k,t) \right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Delta}\right) \tag{33}$$

Note that we should expect the asymptotic covariance matrix to exhibit heteroskedastic and autocorrelated (HAC) character due to the properties of the composite model errors in (21). In particular, $e(k,t)$ is serially correlated because it includes elements of $\xi_n(k,t)$ at time t and each $\xi_n(j, t-1)$ at time $t-1$. Further, even if the sampling errors $\xi(k,t)$ are homoskedastic, the variance of $e(k,t)$ changes over time if the Markov transition probabilities $\pi\,(j,k,t)$ are conditional and vary with t.

Given these conditions, the consistency and asymptotic normality of the MPD estimator follows under:

- **Proposition 1**: The MPD estimator is $\sqrt{T}$-consistent such that $\widetilde{\boldsymbol{\lambda}} \xrightarrow{p} \boldsymbol{\lambda}_0$ under Assumptions B1 and B2 plus:

  1. there exists function $m_0(\boldsymbol{\lambda})$ that is uniquely maximized at $\boldsymbol{\lambda}^0$
  2. $m(\boldsymbol{\lambda})$ is twice continuously differentiable and concave
  3. $T^{-1} m(\boldsymbol{\lambda}) \xrightarrow{p} m_0(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda}$

- **Proposition 2**: The MPD estimator is asymptotically normal as

$$\sqrt{T}\left(\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Gamma}_0^{-1} \boldsymbol{\Delta} \boldsymbol{\Gamma}_0^{-1}\right) \tag{34}$$

  under the conditions of Proposition 1 plus Assumption B3 and

1. there exists $\boldsymbol{\Gamma}(\boldsymbol{\lambda})$ continuous in $\boldsymbol{\lambda}$ such that

$$\sup_{\boldsymbol{\lambda}} \left\| T^{-1} \frac{\partial^2 m(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} - \boldsymbol{\Gamma}(\boldsymbol{\lambda}) \right\| \xrightarrow{p} 0 \tag{35}$$

2. $\boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}(\boldsymbol{\lambda}_0)$ is nonsingular

In particular, the stated regularity conditions may be used to prove Propositions 1 and 2 under Theorems 2.7 and 3.1 provided by Newey and McFadden (1994).

Following the test results developed for the frequency–based data cases by Kelton and Kelton (1984), we can use the large–sample results presented above to conduct classical hypothesis tests under the fitted MPD Markov models. In particular, we can test sets of linear restrictions under the general null hypothesis $H_0 : \mathbf{c}\boldsymbol{\lambda} = \mathbf{r}$ with the Wald statistic

$$W = T\left(\mathbf{c}\widetilde{\boldsymbol{\lambda}} - \mathbf{r}\right)' \left(\widetilde{\boldsymbol{\Gamma}}^{-1} \widetilde{\boldsymbol{\Delta}} \widetilde{\boldsymbol{\Gamma}}^{-1}\right)^{-1} \left(\mathbf{c}\widetilde{\boldsymbol{\lambda}} - \mathbf{r}\right) \overset{a}{\sim} \chi^2_Q \tag{36}$$

where matrix $\mathbf{c}$ has full row rank of Q. For example, we can conduct a Wald test for time–invariance (i.e., stationarity) of the Markov process under the null hypothesis $H_0 : \boldsymbol{\lambda}^* = \mathbf{0}$ where $\boldsymbol{\lambda}^*$ is the subset of the Lagrange multipliers associated with non–constant elements of $\mathbf{z}_t$ and the reference weights $\overline{\pi}(j, k, t)$ do not vary with t. Following the discussion above, the Wald statistic should incorporate an HAC–consistent covariance estimator (e.g., Newey and West, 1987) in order to accommodate the HAC structure of the underlying error process in (21).

At this point, it is important to note that the proposed estimators are also asymptotically efficient if the specified Markov transition model is correctly specified. Further, the efficiency property is lost if the model specification is incorrect, but versions of Propositions 1 and 2 may hold if Assumption B1 is untrue. In this case, the specified Markov transition model may be viewed as 'pseudo–true', and its properties may be developed from the known results on misspecified models. For example, although $\boldsymbol{\lambda}_0$ is not the vector of 'true' model parameters in this case, these model parameters represent the limiting MPD model based on constraints (25) and (26). Thus, even if the model is not correctly specified, the MPD parameter estimator may yet converge in probability to $\boldsymbol{\lambda}_0$ and exhibit asymptotically normal character. The key implication of this outcome is that the sampling properties may be used to test some relevant hypotheses that do not explicitly require the model specification to be correct. For example, we can still use an incorrectly specified Markov model that does not satisfy Assumption B1 to test the time–invariance hypothesis, $H_0 : \boldsymbol{\lambda}_n = \mathbf{0}$, where $\boldsymbol{\lambda}_n$ is the subset vector of Lagrange multipliers associated with the non–constant instruments in $\mathbf{z}_t$.

5 Concluding Remarks

The framework developed in this paper links an agent–specific controlled decision process to the estimating equations commonly used to estimate conditional Markov models based on aggregate data. Accordingly, these findings weaken the standard criticism that such models are ad hoc and do not reflect behavioral foundations. Based on the findings reported in this paper, there are two key areas for future research. First, given that this modeling framework does not rely on particular specifications for the agents' utility functions or other components of the Markov decision process, the associated class of Markov chain models that are compatible with these microfoundations may be relatively large. The conditional Markov models proposed in this paper belong to this class and have relatively favorable statistical properties, but future studies should evaluate the robustness of the models in cases where they are only pseudo–true. Second, we may be able to further exploit the model framework to develop additional linkages between the behavioral model and the econometric model. For example, it may be possible to develop estimators of the social surplus function that could be used to assess the impact of changes in the state variables (e.g., public policies) that affect the agents' welfare.

References

Anderson, T. W. and L. Goodman, "Statistical Inference about Markov Chains," *Annals of Mathematical Statistics*, 1957, *28*, 89–110.

Baggerly, K., "Empirical Likelihood as a Goodness of Fit Measure," *Biometrika*, 1998, *85*, 535–47.

Bedall, F., "Test Statistics for Simple Markov Chains: A Monte Carlo Study," *Biometrical Journal*, 1978, *20*, 41–9.

Berry, S., J. Levinsohn, and A. Pakes, "Automobile Prices in Equilibrium," *Econometrica*, 1995, *63*, 841–90.

Corcoran, S., "Empirical Exponential Family Likelihood using Several Moment Conditions," *Statistica Sinica*, 2000, *10*, 545–557.

Cressie, N. and T. Read, "Multinomial Goodness–of–Fit Tests," *Journal of the Royal Statistical Society, Series B*, 1984, *46*, 440–64.

Csiszár, I., "*I*–divergence Geometry of Probability Distributions and Minimization Problems," *Annals of Probability*, 1975, *3*, 146–58.

——, "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems," *Annals of Statistics*, 1991, *19*, 2032–66.

Davis, B., C. Heathcote, and T. O'Neill, "Estimating and Interpolating a Markov Chain from Aggregate Data," *Biometrika*, 2002, *89*, 95–110.

Di Cicco, T. and J. Romano, "Nonparametric Confidence Limits by Resampling Methods and Least Favorable Families," *International Statistics Review*, 1990, *58*, 59–76.

Gokhale, D. and S. Kullback, *The Information in Contingency Tables*, New York: Marcel Dekker, 1978.

Golan, A., G. Judge, and D. Miller, *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, New York: John Wiley and Sons, 1996.

Good, I. J., "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," *Annals of Mathematical Statistics*, 1963, *34*, 911–34.

Goodman, L., "A Further Note on Miller's 'Finite Markov Processes in Psychology'," *Psychometrika*, 1953, *18*, 245–8.

Hotz, V. J. and R. A. Miller, "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 1993, *60*, 397–429.

Ireland, C. and S. Kullback, "Minimum Discriminant Information Estimation," *Biometrics*, 1968, *24*, 707–13.

Isaacson, D. and R. Madsen, *Markov Chains: Theory and Applications*, New York: John Wiley and Sons, 1976.

Jaynes, E. T., "Information Theory and Statistical Mechanics, I," *Physical Review*, 1957, *106*, 620–30.

——, "Information Theory and Statistical Mechanics, II," *Physical Review*, 1957, *108*, 171–90.

Kao, R., "Note on Miller's 'Finite Markov Process in Psychology'," *Psychometrika*, 1953, *18*, 241–3.

Kelton, C., "Estimation of Time–Independent Markov Processes with Aggregate Data: A Comparison of Techniques," *Econometrica*, 1981, *49*, 517–8.

—— **and D. W. Kelton**, "Advertising and Intraindustry Brand Shift in the U.S. Brewing Industry," *Journal of Industrial Economics*, 1982, *30*, 293–303.

Kelton, W. D. and C. Kelton, "Hypothesis Tests for Markov Process Models from Aggregate Frequency Data," *Journal of the American Statistical Association*, 1984, *79*, 922–8.

Kullback, S., *Information Theory and Statistics*, New York: John Wiley and Sons, 1959.

——, "Marginal Homogeneity of Multidimensional Contingency Tables," *Annals of Mathematical Statistics*, 1971, *42*, 594–606.

—— **and R. Leibler**, "On Information and Sufficiency," *Annals of Mathematical Statistics*, 1951, *22*, 79–86.

Lee, T. C. and G. Judge, "Entropy and Cross–Entropy Procedures for Estimating Transition Probabilities from Aggregate Data," *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, 1996. John Wiley and Sons, New York.

——, ——, and A. Zellner, *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data*, 2nd ed., North–Holland, 1977.

MacRae, E. C., "Estimation of Time–Varying Markov Processes with Aggregate Data," *Econometrica*, 1977, *45*, 183–98.

Miller, G., "Finite Markov Processes in Psychology," *Psychometrika*, 1952, *17*, 149–67.

Neyman, J., "Contributions to the Theory of χ^2 Testing," *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, 1949, pp. 239–73.

Owen, A., "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 1988, *75*, 237–49.

——, "Empirical Likelihood Ratio Confidence Regions," *Annals of Statistics*, 1990, *18*, 90–120.

——, *Empirical Likelihoood*, Boca Raton, FL: Chapman and Hall, 2001.

Pearson, K., "On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine, 5th Series*, 1900, *50*, 157–75.

Read, T. and N. Cressie, *Goodness–of–Fit Statistics for Discrete Multivariate Data*, New York: Springer–Verlag, 1988.

Rust, J., "Maximum Likelihood Estimation of Discrete Control Processes," *SIAM Journal of Control and Optimization*, 1988, *26*, 1006–24.

——, "Structural Estimation of Markov Decision Processes," in R. Engle and D. McFadden, eds., *Handbook of Econometrics*, Vol. 4, Elsevier, 1994.

Shannon, C., "A Mathematical Theory of Communication," *Bell System Technical Journal*, 1948, *27*, 379–423.

Telser, L., "Advertising and Cigarettes," *Journal of Political Economy*, 1962, *70*, 471–99.

——, "The Demand for Branded Goods as Estimated from Consumer Panel Data," *Review of Economics and Statistics*, 1962, *44*, 300–24.

Theil, H., "A Multinomial Extension of the Linear Logit Model," *International Economic Review*, 1969, *251–9.*

White, C. C. and D. J. White, "Markov Decision Processes," *European Journal of Operational Research*, 1989, *39*, 1–16.

6 End Notes

1. The observable state variables $\mathbf{x}_t$ are also assumed to be sufficient statistics for ε_t such that $q\left(\varepsilon_t \left| \mathbf{x}_t\right.\right)$ is not conditional on ε_{t-1}. As Rust notes, this implies that any serial dependence among the idiosyncratic outcomes of ε_t is fully represented by the Markovian character of $\mathbf{x}_t$.

2. The contraction mapping $\Psi(v)$ also establishes the implicit linkage between the optimal value function v and the intertemporal discount factor τ plus any relevant features of the utility function (e.g., risk preferences).

3. The Fréchet derivative is one member of the class of function space generalizations of the standard derivative used in multivariate calculus. The generalized derivatives are used in this case because the arguments of interest in the social surplus function $G\left(v(\boldsymbol{x}_t) \left| \boldsymbol{x}_t\right.\right)$ are functions (i.e., $v(\boldsymbol{x}_t)$ and $q\left(\varepsilon_t \left| \boldsymbol{x}_t\right.\right)$) rather than real–valued parameters. Following Definition 2.3 in Shao and Tu (1995), a functional G is said to be ρ–Fréchet differentiable at q if

$$\lim_{n\to\infty} \frac{G(q_n) - G(q) - \eta(q_n - q)}{\rho(q_n, q)} = 0$$

where q_n is a local alternative to q such that $\rho\left(q_n, q\right) \to 0$ as $n \to \infty$ under measure ρ (e.g., the sup–norm measure) and η is a linear functional that represents the Fréchet differential.

Appendix

A.1 Proofs of the Lemmas

- **Lemma 1**: If the idiosyncratic state variable has a non–null component to its mean such that $\mathrm{E}\left[\varepsilon_t \,|\boldsymbol{x}_t\right] \neq 0$ for some t, the common component of the utility function could be written as

$$u^*\left(\boldsymbol{x}_t, \delta_t\right) = u\left(\boldsymbol{x}_t, \delta_t\right) + \mathrm{E}\left[\varepsilon_t \,|\boldsymbol{x}_t\right]$$

  Thus, it must be true that $\mathrm{E}\left[\varepsilon_t \,|\boldsymbol{x}_t\right] = 0 \;\; \forall\, t$, and it necessarily follows that $\boldsymbol{x}_t$ and ε_t are orthogonal in expectation (i.e., uncorrelated) such that $\mathrm{E}\left[\boldsymbol{x}_t' \varepsilon_t\right] = \mathbf{0} \;\; \forall\, t$ by the iterated expectations theorem. □

- **Lemma 2**: Given that the conditional distributions of the idiosyncratic state variable ε_t (given x_t) must integrate to one, we know that

$$\int q_n\left(\varepsilon_t \,|\boldsymbol{x}_t\right)\; d\varepsilon_t - \int q\left(\varepsilon_t \,|\boldsymbol{x}_t\right)\; d\varepsilon_t = 0$$

  for all t and n. As noted above and in endnote 2, the Fréchet differential η_n is linear in $q_n - q$, and it must be true that

$$\mathrm{E}\left[\eta_n \,|\boldsymbol{x}_t\right] = \int h\left(q_n\left(\varepsilon_t \,|\boldsymbol{x}_t\right) - q\left(\varepsilon_t \,|\boldsymbol{x}_t\right)\right)\; d\varepsilon_t = \kappa$$

  for some linear functional h, constant κ, and all t and n. Then, we know

$$\mathrm{E}\left[\xi_n(k,t) \,|\boldsymbol{x}_t\right] = \mathrm{E}\left[\left.\frac{\partial \eta_n}{\partial v}\right| \boldsymbol{x}_t\right] = \frac{\partial}{\partial v}\mathrm{E}\left[\eta_n \,|\boldsymbol{x}_t\right] = \frac{\partial \kappa}{\partial v} = 0$$

  provided we can exchange the integration and differentiation operations. Finally, we can use the iterated expectations theorem to show that $\xi_n(k,t)$ is orthogonal in expectation to $\boldsymbol{x}_t$ such that (14) holds. □

A.2 Cressie–Read Power Divergence Family

Read and Cressie note that the PD statistic is strictly convex in its arguments and may be used as a criterion function for minimum distance estimation. The power divergence family of pseudo–distance functionals includes

1. Kullback–Leibler directed divergence or discrimination information statistic (Kullback, 1959; Gokhale and Kullback, 1978)

$$I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha \to 0) \propto \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \pi(j,k,t) \ln\left(\frac{\pi(j,k,t)}{\overline{\pi}(j,k,t)}\right) \tag{37}$$

   and

$$I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha \to -1) \propto \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \overline{\pi}(j,k,t) \ln\left(\frac{\overline{\pi}(j,k,t)}{\pi(j,k,t)}\right) \tag{38}$$

   Note that $I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha \to 0) + I(\overline{\boldsymbol{\pi}}, \boldsymbol{\pi}, \alpha \to -1)$ is a symmetric distance function.

2. Pearson's chi–square statistic (Pearson, 1900)

$$I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha = 1) = \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \frac{(\pi(j,k,t) - \overline{\pi}(j,k,t))^2}{\overline{\pi}(j,k,t)} \tag{39}$$

3. Modified chi–square statistic (Neyman, 1949)

$$I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha = -2) = \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \frac{(\overline{\pi}(j,k,t) - \pi(j,k,t))^2}{\pi(j,k,t)} \tag{40}$$

4. Squared Matusita or Hellinger distance

$$I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha = -1/2) \propto \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \left(\sqrt{\pi(j,k,t)} - \sqrt{\overline{\pi}(j,k,t)}\right)^2 \tag{41}$$

Given uniform reference weights $\overline{\pi}(j,k,t) = K^{-1}$ for all j and t, the negative of (24) also encompasses other prominent distance statistics

1. Empirical likelihood statistic (Owen, 2001)

$$-I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha \to -1) \propto \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \ln\left(\pi(j,k,t)\right) \tag{42}$$

2. Shannon's entropy (Shannon, 1948) or exponential empirical likelihood (Di Cicco and Romano, 1999; Corcoran, 2000) statistic

$$-I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha \to 0) \propto -\sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \pi(j,k,t) \ln\left(\pi(j,k,t)\right) \tag{43}$$

3. Simpson or Gini statistic (Read and Cressie, 1988) or Euclidean likelihood statistic (Owen, 2001)

$$-I(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}, \alpha = 1) = 1 - \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \pi(j, k, t)^2 \tag{44}$$

Finally, it is important to note that our identification of some members of the PD family as likelihood statistics is purely analogical. We do not use these criterion functions as empirical likelihood functions in the context of our Markov chain models. The Markov transition probabilities assign mass to the states $k = 1, \ldots, K$ for each j and t and do not form empirical distributions across the set of observed outcomes as in empirical likelihood analysis.