Heterogenous Markov Chain Models

Bryan S. Graham

March 26, 2021

Many economic behaviors persist over time. For example, the receipt of social assistance in a given month positively correlates with enrollment prior months (e.g., Card & Hyslop, 2005). Similarly, the probability of labor force participation is generally increasing in past participation (e.g., Card & Sullivan, 1988; Magnac, 2001). Heckman (1981b, p. 91) articulates two competing explanations for empirical regularities of this type. First, past participation may induce changes in "preferences, prices or constraints relevant to future choices". Alternatively individuals may simply be heterogeneous in unmeasured attributes (that directly influence choice). Unconditional on these attributes, past choices may help to predict current choices solely because they covary with the omitted attributes. Following Heckman (1981b), Chamberlain (1985) and others I will call the first phenomena *state dependence* and the second *heterogeneity*.

In each of periods $t = 0, \ldots, T$ an agent makes the binary choice $Y_t \in \{0, 1\}$. The econometrician observes the $T+1$ choice sequence, she does not observe the vector of unobserved agent attributes A. I will refer to A as an agent's (unobserved) *type*. Conditional on A choice follows the stationary second order Markov chain:[1]

$$\Pr\left(Y_t = y \middle| Y_0^{t-1}, A\right) \quad = \quad \Pr\left(Y_t = y \middle| Y_{t-1}, Y_{t-2}, A\right),$$

where $Y_0^{t-1} = (Y_{t-1}, Y_{t-2}, \ldots, Y_0)'$ is the $t \times 1$ vector of past choices.

The model enumerates four different transitions: (i) the probability of, say, employment given non-employment in the prior two periods, (ii) the probability of employment given non-employment last period, but employment two periods ago, (iii) the probability of employment given employment last period, but not two periods ago, and (iv) the probability of employment given employment last period as well as two periods ago. Conditional on A

[1] Amemiya (1985, Chapter 11)provides a textbook introduction to Markov chain models.

these four transition probabilities equal:

$$\begin{aligned}\pi_{00}(A) &= \Pr(Y_t = 1 | Y_{t-1} = 0, Y_{t-2} = 0, A)\\ \pi_{01}(A) &= \Pr(Y_t = 1 | Y_{t-1} = 0, Y_{t-2} = 1, A)\\ \pi_{10}(A) &= \Pr(Y_t = 1 | Y_{t-1} = 1, Y_{t-2} = 0, A)\\ \pi_{11}(A) &= \Pr(Y_t = 1 | Y_{t-1} = 1, Y_{t-2} = 1, A).\end{aligned}$$

It is convenient to analyze this second order Markov chain as a first order chain for the two-dimensional outcome (Y_t, Y_{t-1}). In this formulation there are four states – $\{(0,0),(0,1),(1,0),(1,1)\}$ – with a transition matrix of

$$\Pi(A) = \begin{pmatrix} 1-\pi_{00}(A) & 0 & \pi_{00}(A) & 0 \\ 1-\pi_{01}(A) & 0 & \pi_{01}(A) & 0 \\ 0 & 1-\pi_{10}(A) & 0 & \pi_{10}(A) \\ 0 & 1-\pi_{11}(A) & 0 & \pi_{11}(A) \end{pmatrix}. \tag{1}$$

The transition matrix has several structural zeros. It is, for example, impossible to transition from the $(0,0)$ to the $(1,1)$ state in a single period. The stationary distribution of the chain coincides with the solution to

$$p(A) = p(A)\Pi(A).$$

Hence $p(A)$ is a normalized left eigenvector of $\Pi(A)$. It is easy to verify that the stationary distribution associated with (1) is proportional to

$$p(A) \propto \begin{pmatrix} (1-\pi_{01}(A))(1-\pi_{11}(A)) \\ \pi_{00}(A)(1-\pi_{11}(A)) \\ \pi_{00}(A)(1-\pi_{11}(A)) \\ \pi_{00}(A)\pi_{10}(A) \end{pmatrix}'. \tag{2}$$

Let $p^{\infty}(y_t | y_{t-1}, A) = \lim_{t\to\infty} \Pr(Y_t = y_t | Y_{t-1} = y_{t-1}, A)$. It follows from (2) that the *stationary transition* process from Y_{t-1} to Y_t is given by

$$\begin{pmatrix} p^{\infty}(0|0,A) & p^{\infty}(1|0,A) \\ p^{\infty}(0|1,A) & p^{\infty}(1|1,A) \end{pmatrix} \propto \begin{pmatrix} (1-\pi_{01}(A))(1-\pi_{11}(A)) & \pi_{00}(A)(1-\pi_{11}(A)) \\ \pi_{00}(A)(1-\pi_{11}(A)) & \pi_{00}(A)\pi_{10}(A) \end{pmatrix}, \tag{3}$$

where the constant of proportionality, $1/\Delta(A)$, is given by

$$\Delta(A) = (1-\pi_{01}(A))(1-\pi_{11}(A)) + 2\pi_{00}(A)(1-\pi_{11}(A)) + \pi_{00}(A)\pi_{10}(A).$$

One estimand of interest is the relative *limiting* frequency of employment, given employment vs. non-employment in the prior period:

$$\lambda(A) = p^{\infty}(1|1,A) - p^{\infty}(1|0,A). \tag{4}$$

This can be thought of as a type-specific measure of (limiting) state dependence. In the current set-up, manipulation of (3) yields:

$$\lambda(A) = \frac{\pi_{00}(A)\pi_{10}(A) - \pi_{00}(A)(1-\pi_{11}(A))}{(1-\pi_{01}(A))(1-\pi_{11}(A)) + 2\pi_{00}(A)(1-\pi_{11}(A)) + \pi_{00}(A)\pi_{10}(A)}.$$

Since A is unobserved, the identification of $\lambda(A)$ is not possible. Instead I will explore how the econometrician might instead learn about is average:

$$\lambda = \mathbb{E}_A[\lambda(A)]. \tag{5}$$

In addition to the average (limiting) state dependence measure, λ, another object of interest is the average survivor function

$$\Lambda(y_1, y_0, s) = \mathbb{E}_A[\Pr(Y_{s+2} = Y_{s+1} = Y_s = Y_{s-1} = \cdots = Y_2 = 1|Y_1 = y_1, Y_0 = y_0, A)]. \tag{6}$$

Equation (6) gives the probability of s periods of continuous employment (from $t = 2$ to $t = s+2$) conditional on the initial conditio (y_1, y_0). The average is across the unconditional distribution of types. Hence, in general, (6) would not coincide with the empirical survival function

$$\Pr(Y_{s+2} = Y_{s+1} = Y_s = Y_{s-1} = \cdots = Y_2 = 1|Y_1 = y_1, Y_0 = y_0),$$

since, in general, we would not expect $\Pr(A \leq a|Y_1 = y_1, Y_0 = y_0) = \Pr(A \leq a)$.

Discrete heterogeneity

I assume a discrete heterogeneity structure. Specifically, the unobserved attribute vector may take one of K configurations:

$$A \in \mathbb{A} = \{a_1, \ldots, a_K\}.$$

Browning & Carro (2013) study identification of a first order heterogenous Markov chain under a K types discrete heterogeneity assumption. They develop results connecting the

length of the panel (T) and the maximum number of identifiably discrete types (K).

Let $\underline{\rho} = (\rho_1, \ldots, \rho_K)'$ denote the population frequency of each type of agent. Let

$$\Pr\left(Y_{i1} = y_1, Y_{i0} = y_0 \middle| A_i = a_k\right) = \gamma_{y_1 y_{0,k}}$$

for $(y_1, y_0) \in \{0,1\}^2$, $a_k \in \mathbb{A}$ and $k = 1, \ldots, K$ parameterize the initial condition of the process for each type of agent. Similarly let $\pi_{00}(a_k) = \pi_{00,k}$ for $k = 1, \ldots, K$ be the probability of choice $Y_t = 1$ given that $Y_{t-1} = Y_{t-2} = 0$ for each type of agent. Define $\pi_{01,k}$, $\pi_{10,k}$ and $\pi_{11,k}$ similarly.

The model implies the following adding-up conditions: (i) $\sum_{k=1}^K \rho_k = 1$ and (ii) $\gamma_{00,k} + \gamma_{01,k} + \gamma_{10,k} + \gamma_{11,k} = 1$ for $k = 1, \ldots, K$.

There are a total of $\dim\left(\underline{\rho}\right) - 1 = K - 1$ non-redundant marginal type probabilities, $\dim\left(\underline{\gamma}\right) - K = 3K$ initial conditions and $\dim\left(\underline{\pi}\right) = 4K$ transition matrix probabilities. The dimension of the entire parameter space is therefore $8K - 1$.

Since there are 2^{T+1} choice sequences a necessary condition for identification is that $2^{T+1} \geq 8K - 1$.

Under the discrete heterogeneity structure (5) and (6) coincide with

$$\lambda = \sum_{k=1}^K \rho_k \frac{\pi_{00,k}\pi_{10,k} - \pi_{00,k}\left(1 - \pi_{11,k}\right)}{\left(1 - \pi_{01,k}\right)\left(1 - \pi_{11,k}\right) + 2\pi_{00,k}\left(1 - \pi_{11,k}\right) + \pi_{00,k}\pi_{10,k}} \tag{7}$$

$$\Lambda(y_1, y_0, s) = \sum_{k=1}^K \rho_k \left[\pi_{y_1 y_0,k}\pi_{1 y_1,k}\pi_{11,k}^{s-2}\right]. \tag{8}$$

A plot of $\Lambda(1,1,s) - \Lambda(0,0,s)$ with s on the horizontal axis provides a simple summary of the structural relationship between initial conditions and the length of employment spells. A comparison of this plot with one of the difference

$$\begin{aligned} &\Pr\left(Y_{s+2} = Y_{s+1} = Y_s = Y_{s-1} = \cdots = Y_2 = 1 \middle| Y_1 = 1, Y_0 = 1\right) \\ &- \Pr\left(Y_{s+2} = Y_{s+1} = Y_s = Y_{s-1} = \cdots = Y_2 = 1 \middle| Y_1 = 0, Y_0 = 0\right), \end{aligned}$$

provides a simple assessment of the magnitude of any heterogeneity bias.

Likelihood

I begin by considering the complete data likelihood. The econometrician observes choice in periods $t = 0, \ldots, T$ for each of $i = 1, \ldots, N$ randomly sampled agents. Let $\theta = \left(\underline{\pi}', \underline{\gamma}', \underline{\rho}'\right)'$

be the full "common" parameter; the i^{th} agent's contribution to the complete data likelihood is

$$\begin{aligned} L_i^C(\theta, A_i) &= \left[\prod_{k=1}^{K}\left(\Pr(Y_{i1}, Y_{i0}|\, a_k;\theta)\prod_{t=2}^{T}\Pr(Y_{it}|\, Y_{it-1}, Y_{it-1}, a_k;\theta)\right)^{\mathbf{1}(A_i=a_k)}\right] \\ &\quad\times\left[\prod_{k=1}^{K}\Pr(A_i=a_k)^{\mathbf{1}(A_i=a_k)}\right] \end{aligned} \tag{9}$$

with, for $k=1,\ldots,K$,

$$\Pr(Y_{i1}, Y_{i0}|\, a_k;\theta) = \gamma_{00,k}^{(1-Y_{i1})(1-Y_{i0})}\gamma_{01,k}^{(1-Y_{i1})Y_{i0}}\gamma_{10,k}^{Y_{i1}(1-Y_{i0})}\gamma_{11,k}^{Y_{i0}Y_{i1}} \tag{10}$$

$$\begin{aligned} \Pr(Y_{it}|\, Y_{it-1}, Y_{it-1}, a_k;\theta) &= \pi_{00,k}{}^{Y_{it}(1-Y_{it-1})(1-Y_{it-2})}\left(1-\pi_{00,k}\right)^{(1-Y_{it})(1-Y_{it-1})(1-Y_{it-2})} \\ &\quad\times\pi_{01,k}{}^{Y_{it}(1-Y_{it-1})Y_{it-2}}\left(1-\pi_{01,k}\right)^{(1-Y_{it})(1-Y_{it-1})Y_{it-2}} \\ &\quad\times\pi_{10,k}{}^{Y_{it}Y_{it-1}(1-Y_{it-2})}\left(1-\pi_{10,k}\right)^{(1-Y_{it})Y_{it-1}(1-Y_{it-2})} \\ &\quad\times\pi_{11,k}{}^{Y_{it}Y_{it-1}Y_{it-2}}\left(1-\pi_{11,k}\right)^{(1-Y_{it})Y_{it-1}Y_{it-2}}\Big\}. \end{aligned} \tag{11}$$

$$\Pr(A_i=a_k;\theta) = \rho_k. \tag{12}$$

The three components of (9) are the initial condition (10), the sequence probability (11) and the type probability (12). We could reduce the dimensionality of the parameter space by assuming that the initial conditions correspond to draws from the stationary distribution of the Markov chain:

$$\begin{aligned} \gamma_{00,k} &= \frac{1}{\triangle_k}\left(1-\pi_{01,k}\right)\left(1-\pi_{11,k}\right) \\ \gamma_{01,k} &= \frac{1}{\triangle_k}\pi_{00,k}\left(1-\pi_{11,k}\right) \\ \gamma_{10,k} &= \frac{1}{\triangle_k}\pi_{00,k}\left(1-\pi_{11,k}\right) \\ \gamma_{11,k} &= \frac{1}{\triangle_k}\pi_{00,k}\pi_{10,k} \end{aligned}$$

for $k=1,\ldots,K$ and $\triangle_k=\triangle(a_k)$. Such restrictions are sometimes imposed in empirical work (e.g., Card & Sullivan, 1988); however they are also often rejected by the data. I will develop results for the case that leaves the specification of the initial condition unrestricted. Taking logarithms and summing across agents $i=1,\ldots,N$ yields the sample log-likelihood

$$l_N^C(\theta, \mathbf{A}) = \sum_{i=1}^{N}\ln L_i^C(\theta, A_i). \tag{13}$$

If $\mathbf{A} = (A_1, \ldots, A_N)'$ were observed by the econometrician, maximization of (13) would be straightforward; as would be the analysis of the asymptotic sampling properties of the resultant complete data estimate, say, $\hat{\theta}_{\text{CD}}$. In practice, of course, $\mathbf{A}$ is unobserved. There are a variety of approaches to dealing with this fact. Each requires different assumptions and, consequently, is associated with different advantages and disadvantages.

Joint fixed effects estimation

One approach treats $\mathbf{A}$ as an additional parameter to be estimated along with $\mathbf{A}$. This makes the problem non-standard, since the dimension of $\mathbf{A}$ grows with the sample size N. Following Neyman & Scott (1948) I will call $\{A_i\}_{i=1}^N$ *incidental parameter*s.[2] A procedure which estimates $\mathbf{A}$ jointly with θ, the common parameter, is a *joint fixed effects* estimator. The reasoning behind this terminology will become clearer later. Without loss of generality we can let $\mathbb{A}$ coincide with the vertices of the $K-1$ simplex. This conceptualizes A as a type assignment vector. Specifically A is a $K \times 1$ vector with a one for its k^{th} element if an agent is of type k and zeros elsewhere. Now consider the following iterative procedure:

1. Choose an initial (random) assignment of agents to types: $\hat{\mathbf{A}}^{(s)}$ for $s = 0$.
2. Maximize $l_N^C\left(\theta, \hat{\mathbf{A}}^{(s)}\right)$ with respect to θ, generating $\hat{\theta}^{(s+1)}$.
3. For each $i = 1, \ldots, N$ choose $\hat{A}_i^{(s+1)} \in \mathbb{A}$ to maximize $L_i^C\left(\hat{\theta}^{(s+1)}, A_i\right)$ as defined in (9).
4. Repeat steps 2 and 3 until the change in (13) is small and/or until $\hat{\theta}^{(s+1)} \approx \hat{\theta}^{(s)}$.

This procedure is closely related to the so-called k-means algorithm from cluster analysis (e.g., Murphy, 2012, Chapter 25). It is very easy to implement in the present context. For

[2]Jerzy Neyman was the founding chair of the Statistics Department at the University of California - Berkeley. Elizabeth Scott was a Professor of Statistics and an astronomer, also at Berkeley. The 1948 paper was Scott's first in statistics (completed as a graduate student), although she had already published widely in the field of astronomy. Neyman's influence on the development of the field of post WW II statistics is well-known. Scott was also a highly accomplished statistician and served as president of the Institute of Mathematical Statistics in the early 1970s. In addition to her considerable scientific work, Scott was strong advocate for gender equality in the academy and played a leading role in documenting discrimination against women graduate students and faculty. This paper has proved to be a landmark, especially in the field of econometrics. Interestingly it was published in *Econometrica*, not a statistics journal. This was apparently the only occasion in which Neyman published in an economics journal (Lancaster, 2000).

example, $\hat{\theta}^{(s+1)}$ has the closed form solution (Step 2)

$$\begin{aligned}
\hat{\gamma}^{(s+1)}_{y_1 y_0, k} &= \frac{\sum_{i=1}^N \mathbf{1}\left(\hat{A}_i^{(s)} = e_k\right) \mathbf{1}\left(Y_{i1} = y_1\right) \mathbf{1}\left(Y_{i0} = y_0\right)}{\sum_{i=1}^N \mathbf{1}\left(\hat{A}_i^{(s)} = e_k\right)} \\
\hat{\pi}^{(s+1)}_{y_{t-1} y_{t-2}, k} &= \frac{\sum_{i=1}^N \mathbf{1}\left(\hat{A}_i^{(s)} = e_k\right) \left[\sum_{t=2}^T \mathbf{1}\left(Y_{it} = 1\right) \mathbf{1}\left(Y_{it-1} = y_{t-1}\right) \mathbf{1}\left(Y_{it-2} = y_{t-2}\right)\right]}{\sum_{i=1}^N \mathbf{1}\left(\hat{A}_i^{(s)} = e_k\right) \left[\sum_{t=2}^T \mathbf{1}\left(Y_{it-1} = y_{t-1}\right) \mathbf{1}\left(Y_{it-2} = y_{t-2}\right)\right]} \\
\hat{\rho}_k^{(s+1)} &= \frac{1}{N} \sum_{i=1}^N \mathbf{1}\left(\hat{A}_i^{(s)} = e_k\right)
\end{aligned}$$

for $(y_1, y_0) \in \{0,1\}^2$, $(y_{t-1}, y_{t-2}) \in \{0,1\}^2$ and $k = 1, \ldots, K$. Here e_k is a $K \times 1$ vector with a 1 in its k^{th} row and zeros elsewhere. Step 3 is a maximization over a finite set

$$\hat{A}_i^{(s+1)} = \max_{k \in \{1,\ldots,K\}} L_i^C\left(\hat{\theta}^{(s)}, a_k\right).$$

Unfortunately, the algorithm is sensitive to the starting value $\hat{\mathbf{A}}^{(0)}$. In practice it should be repeated with several different starting values to ensure convergence to a global maximum. Under large N, fixed T asymptotics Hahn & Moon (2010) show that the resulting estimate, $\hat{\theta}_{\text{JFE}}$, is not consistent. However, if T grows along with N, they show that the procedure is consistent. The required rate of growth is slow: $T = O\left(\ln\left(N\right)\right)$. Bonhomme & Manresa (2015) develop additional related results.

We will not develop the theoretical properties of $\hat{\theta}_{\text{JFE}}$ in detail. However the formulation and analysis of fixed effects estimators under discrete heterogeneity is an interesting area of research.

Random effects estimation

A random effects estimator is based on the "integrated" or *observed log-likelihood*

$$l_N^I\left(\theta\right) = \sum_{i=1}^N \ln\left(\sum_{k=1}^K L_i^C\left(\theta, a_k\right)\right), \tag{14}$$

which is a function of the common parameter, θ, alone. Direct maximization of (14) is feasible; evaluating each agent's contribution to the log-likelihood involves computing a sum of K elements which is tractable. However, it turns out that (14) is a good candidate for maximization via the EM algorithm. This procedure is very convenient for the model

outlined here.

EM algorithm

Let $q(a)$ be some assignment of probability mass to the K possible types such that $q(a_k) > 0$ for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} q(a_k) = 1$. We can show that the i^{th} unit's contribution to the observed log likelihood is bounded below by

$$\begin{aligned} \ln\left(\sum_{l=1}^{K} L_i^C(\theta, a_l)\right) &= \ln\left(\sum_{l=1}^{K} q(a_l) \frac{L_i^C(\theta, a_l)}{q(a_l)}\right) \\ &\geq \sum_{l=1}^{K} q(a_l) \ln\left(\frac{L_i^C(\theta, a_l)}{q(a_l)}\right) \\ &= Q_i^*(\theta, q) \end{aligned} \tag{15}$$

where the middle line follows from Jensen's inequality: $g(\mathbb{E}[Y]) \geq \mathbb{E}[g(Y)]$ for $g(\cdot)$ concave. Here $\ln(\cdot)$ is concave and expectations are with respect to $q(a)$. The last line defines $Q_i^*(\theta, q)$ Equation (15) gives

$$l_N^I(\theta) \geq \sum_{i=1}^{N} Q_i^*(\theta, q_i)$$

for any set of valid distribution functions $\{q_i\}_{i=1}^N$ that assign positive probability to each $\{a_k\}_{k=1}^K$.

Bayes' Theorem, and the form of the complete data likelihood (9), yields the conditional type probabilities

$$\Pr\left(A = a_k \middle| Y_0^T; \theta\right) \overset{def}{=} \tilde{\rho}_{ki}(\theta) = \frac{L_i^C(\theta, a_k)}{\sum_{l=1}^{K} L_i^C(\theta, a_l)}, \tag{16}$$

for $k = 1, \ldots, K$. In machine learning literature on classification (16) is called the "responsibility" of cluster k for unit i. We can use (16) to factor the i^{th} unit's contribution to the complete data likelihood as

$$L_i^C(\theta; a_k) = \tilde{\rho}_{ki}(\theta) \left[\sum_{l=1}^{K} L_i^C(\theta, a_l)\right].$$

This gives a re-arrangement of the lower bound (15) equal to

$$\begin{aligned} Q_i^*(\theta, q_i) &= \sum_{l=1}^{K} q_i(a_l) \ln\left(\frac{L_i^C(\theta, a_l)}{q_i(a_l)}\right) \\ &= \sum_{l=1}^{K} q_i(a_l) \ln\left(\frac{\tilde{\rho}_{li}(\theta)\left[\sum_{m=1}^{K} L_i^C(\theta, a_m)\right]}{q_i(a_l)}\right) \\ &= -D_{\mathrm{KL}}(q_i \| \tilde{\rho}_i) + \left[\sum_{l=1}^{K} q_i(a_l)\right] \ln\left(\sum_{m=1}^{K} L_i^C(\theta, a_m)\right) \\ &= -D_{\mathrm{KL}}(q_i \| \tilde{\rho}_i) + \ln\left(\sum_{m=1}^{K} L_i^C(\theta, a_m)\right), \end{aligned} \tag{17}$$

where $D_{\mathrm{KL}}(q_i \| \tilde{\rho}_i) = \sum_{l=1}^{K} q_i(a_l) \ln\left(\frac{q_i(a_l)}{\tilde{\rho}_{li}(\theta)}\right)$ is the Kullback-Leibler divergence of $\tilde{\rho}_i$ from q_i. Now consider the following procedure:

1. Let $\hat{\theta}^{(s)}$ for $s = 0$ be an initial value for θ.
2. **E-Step:** Set $q(a_k) = \tilde{\rho}_{ki}\left(\hat{\theta}^{(s)}\right)$ for $k = 1, \ldots, K$ and form the observed log-likelihood lower bound

$$\begin{aligned} Q\left(\theta, \hat{\theta}^{(s)}\right) &= \sum_{i=1}^{N} Q_i^*\left(\theta, \tilde{\rho}_i\left(\hat{\theta}^{(s)}\right)\right) \\ &= \sum_{i=1}^{N} \left\{ \sum_{l=1}^{K} \tilde{\rho}_{li}\left(\hat{\theta}^{(s)}\right) \ln\left(L_i^C(\theta, a_l)\right) + \mathbf{S}\left(\tilde{\rho}_i\left(\hat{\theta}^{(s)}\right)\right) \right\} \\ &= \sum_{i=1}^{N} \left\{ \mathbb{E}\left[\ln\left(L_i^C(\theta, A)\right) \middle| Y_{i0}^T; \hat{\theta}^{(s)} \right] + \mathbf{S}\left(\tilde{\rho}_i\left(\hat{\theta}^{(s)}\right)\right) \right\} \end{aligned} \tag{18}$$

   where $\mathbb{E}\left[\ln\left(L_i^C(\theta, A)\right) \middle| Y_{i0}^T; \hat{\theta}^{(s)} \right]$ is the expected value of the i^{th} unit's contribution to the complete data log-likelihood (given her observed choice sequence and the current parameter value $\hat{\theta}^{(s)}$) and $\mathbf{S}(q) = -\sum_l q_l \ln q_l$ is the entropy of q.
3. **M-Step**: Choose $\hat{\theta}^{(s+1)}$ to maximize $Q\left(\theta, \hat{\theta}^{(s)}\right)$ with respect to θ. Note that since $\mathbf{S}\left(\tilde{\rho}_i\left(\hat{\theta}^{(s)}\right)\right)$ is constant in θ this term is often omitted from the "Q-function" in practice.
4. Repeat steps 2 and 3 until $Q\left(\hat{\theta}^{(s+1)}, \hat{\theta}^{(s)}\right) \approx Q\left(\hat{\theta}^{(s)}, \hat{\theta}^{(s-1)}\right)$ is small and/or until $\hat{\theta}^{(s+1)} \approx \hat{\theta}^{(s)}$.

Note that $Q(\theta,\theta) = l_N^I(\theta)$: after the E-Step the "Q-function" coincides with the observed log-likelihood (the Kullback-Leibler term is zero at $q_i = \tilde{\rho}_i(\theta)$). We also have that the M-Step weakly increases the "Q-function". Putting things together we have

$$l_N^I\left(\hat{\theta}^{(s+1)}\right) \geq Q\left(\hat{\theta}^{(s+1)},\hat{\theta}^{(s)}\right) \geq Q\left(\hat{\theta}^{(s)},\hat{\theta}^{(s)}\right) = l_N^I\left(\hat{\theta}^{(s)}\right). \tag{19}$$

From left-to-right the first inequality follows from (15), the second from the definition of maximization, and the third from (17) evaluated at $q_i = \tilde{\rho}_i(\theta)$. By virtue of (19) the observed log-likelihood $Q\left(\hat{\theta}^{(s)},\hat{\theta}^{(s)}\right) = l_N^I\left(\hat{\theta}^{(s)}\right)$ is monotonically increasing in s. The EM algorithm will therefore find a local maximum (or saddle point) of the *observed log-likelihood* (14). As with the joint fixed effects procedure outlined above, running the algorithm from a variety of starting points is advised.

For our heterogenous Markov chain, both the E- and M-Steps are straightforward. Calculation of $Q\left(\theta,\hat{\theta}^{(s)}\right)$ follows from (9), (16) and (18), while $\hat{\theta}^{(s+1)}$ has the closed form solution

$$\begin{aligned}
\hat{\gamma}_{y_1y_0,k}^{(s+1)} &= \frac{\sum_{i=1}^N \tilde{\rho}_{ki}\left(\hat{\theta}^{(s)}\right)\mathbf{1}(Y_{i1}=y_1)\mathbf{1}(Y_{i0}=y_0)}{\sum_{i=1}^N \tilde{\rho}_{ki}\left(\hat{\theta}^{(s)}\right)} \\
\hat{\pi}_{y_{t-1}y_{t-2},k}^{(s+1)} &= \frac{\sum_{i=1}^N \tilde{\rho}_{ki}\left(\hat{\theta}^{(s)}\right)\left[\sum_{t=2}^T \mathbf{1}(Y_{it}=1)\mathbf{1}(Y_{it-1}=y_{t-1})\mathbf{1}(Y_{it-2}=y_{t-2})\right]}{\sum_{i=1}^N \tilde{\rho}_{ki}\left(\hat{\theta}^{(s)}\right)\left[\sum_{t=2}^T \mathbf{1}(Y_{it-1}=y_{t-1})\mathbf{1}(Y_{it-2}=y_{t-2})\right]} \\
\hat{\rho}_k^{(s+1)} &= \frac{1}{N}\sum_{i=1}^N \tilde{\rho}_{ki}\left(\hat{\theta}^{(s)}\right)
\end{aligned}$$

for $(y_1,y_0) \in \{0,1\}^2$, $(y_{t-1},y_{t-2}) \in \{0,1\}^2$ and $k = 1,\ldots,K$. Note that these estimators are smoothed versions of the fixed effects ones defined above.

Extensions and further reading

Murphy (2012, Chapter 12) provides a elementary introduction to the EM algorithm from a machine learning perspective. Gupta & Chen (2010) provide a survey with signal processing applications. Ruud (1991) provides a nice theoretical discussion with applications to discrete choice models common in econometrics. An basic introduction to Markov chain models from an econometrics perspective is Chapter 11 of Amemiya (1985). Chapter 17 of Murphy (2012) surveys applications in Machine Learning.

The use of discrete heterogeneity distributions has a long history in various branches of

applied statistics. Langeheine & Van de Pol (1990) is a common reference in sociology with useful historical background. A common early reference in econometrics is Heckman & Singer (1984). Prominent applications in the context of dynamic discrete choice analysis are provided by Card & Sullivan (1988), Cameron and Heckman (2001) and Card & Hyslop (2005). This last paper won the Frisch Medal of the Econometric Society. My read of this literature is that estimation is generally done by some variant of the joint fixed effects procedure outlined above, but this is not always clear from the printed discussion. Similarly methods of inference are not always clearly articulated.

In a series of recent papers Martin Browning and Jesus Carro have advocated for an approach to dynamic panel data analysis with "rich" forms of unobserved heterogeneity. Browning & Carro (2013) study heterogenous first order Markov chains. The treatment of heterogenous second order Markov chains given here is very much inspired by their analysis. Chamberlain (1985) is an important early reference on heterogenous Markov chain analysis in econometrics. His discussion provides motivation for the focus on second order chains.

In microeconometric analyses large portions of the sample may consist of "stayers": individuals who undergo no transitions during the observed time periods (e.g., continuously employed or unemployed individuals). A similar phenomenon inspired the development of Mover-Stayer models in sociology. In such settings it can be advantageous to fix the elements of $\Pi(a_k)$ for two types a priori with zeros and ones, parameterizing "always working" and "never working" latent classes.

The model developed in this note is isomorphic to a logit model of

$$\Pr(Y_t = y | Y_{t-1}, Y_{t-2}, A) = \frac{\exp(A_0 + A_1 Y_{t-1} + A_2 Y_{t-2} + A_3 Y_{t-1} Y_{t-2})}{1 + \exp(A_0 + A_1 Y_{t-1} + A_2 Y_{t-2} + A_3 Y_{t-1} Y_{t-2})}.$$

Chamberlain (1985) studies this model with $A_2 = \alpha_2$ (i.e., homogenous across agents) and $A_3 = 0$. His analysis does not require a discrete heterogeneity assumption on (A_0, A_1). Card & Hyslop (2005) study this model with an additional single factor restriction of

$$A_j = \alpha_j + B,\ j = 0, 1, 2, 3,$$

and B either discrete or belonging to a parametric continuous distribution. Card & Hyslop (2005) find that their model fits the data well, but the single factor assumption does have restrictive implications for behavior. Consider a simple heterogenous first order Markov chain. It is possible that some types of agents may have a low A_0 and a high A_1. Such agents exit employment with low probability, but may also exit unemployment more slowly (perhaps they are "picky"). A single factor model implies that those individuals most likely

to retain a job will also exit unemployment more quickly after a job loss.

One approach to empirical analysis would be to begin with a flexible model of the type discussed above and then impose additional restrictions using minimum distance procedures (e.g., Chamberlain, 1984).

Given the discrete heterogeneity assumption, the treatment of the initial condition outlined here is unrestrictive. In other settings (of dynamic discrete choice) the modelling of the initial condition can be challenging. Heckman (1981a) provides a classic and still useful discussion of this problem. (Honoré & Tamer, 2006) represents a modern, "set identification", treatment. Wooldridge (2005) is a widely-cited pedagogical overview.

We will return to more structured forms of dynamic discrete choice analysis in a few lectures (e.g., Honoré & Kyriazidou, 2000).

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Bonhomme, S. & Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3), 1147 – 1184.

Browning, M. & Carro, J. M. (2013). Dynamic binary outcome models with maximal heterogeneity. *Journal of Econometrics*, 178(2), 805 – 823.

Card, D. & Hyslop, D. R. (2005). Estimating the effect of a time-limited subsidy for welfare-leavers. *Econometrica*, 73(6), 1723 – 1770.

Card, D. & Sullivan, D. (1988). Measuring the effect of a subsidized training program on movements in and out of employment. *Econometrica*, 56(3), 497 – 530.

Chamberlain, G. (1984). *Handbook of Econometrics*, volume 2, chapter Panel Data, (pp. 1247 – 1318). North-Holland: Amsterdam.

Chamberlain, G. (1985). *Longitudinal Analysis of Labor Market Data*, chapter Heterogeneity, omitted variable bias, and duration dependence, (pp. 3 – 38). Cambridge University Press: Cambridge.

Gupta, M. R. & Chen, Y. (2010). Theory and use of the em algorithm. *Foundations and Trends in Signal Processing*, 4(3), 223 – 296.

Hahn, J. & Moon, H. R. (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory 26*, 26(3), 863 – 881.

Heckman, J. J. (1981a). *Structural Analysis of Discrete Data and Econometric Applications*, chapter The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process, (pp. 179 – 195). The MIT Press: Cambridge, MA.

Heckman, J. J. (1981b). *Studies in Labor Markets*, chapter Heterogeneity and state dependence, (pp. 91 – 139). Chicago University Press.: Chicago.

Heckman, J. J. & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52(2), 271 – 320.

Honoré, B. E. & Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68(4), 839 – 874.

Honoré, B. E. & Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, 73(3), 611 – 629.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2), 391 – 413.

Langeheine, R. & Van de Pol, F. (1990). A unifying framework for markov modeling in discrete space and discrete time. *Sociological Methods and Research*, 18(4), 416 – 441.

Magnac, T. (2001). Subsidised training and youth employment: distinguishing unobserved heterogeneity from state dependence in labor market histories. *Economic Journal*, 110(466), 805 – 837.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: The MIT Press.

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1), 1 – 32.

Ruud, P. A. (1991). Extensions of estimation methods using the em algorithm. *Journal of Econometrics*, 49(3), 305 – 341.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 20(1), 39 – 54.

