Institute for Food and Resource Economics
University of Bonn

Discussion Paper 2011:2

Bayesian estimation of non-stationary Markov models combining micro and macro data

Hugo Storm
University of Bonn, Institute for Food and Resource Economics, Bonn, Germany
hugo.storm@ilr.uni-bonn.de

Thomas Heckelei
University of Bonn, Institute for Food and Resource Economics, Bonn, Germany
thomas.heckelei@ilr.uni-bonn.de

Ron C. Mittelhammer
School of Economic Sciences, Washington State University, Pullman
mittelha@wsu.edu

Editor: Thomas Heckelei
Institute for Food and Resource Economics
University of Bonn
Nußallee 21
53115 Bonn, Germany

Phone: +49-228-732332
Fax: +49-228-734693
E-mail: thomas.heckelei@ilr.uni-bonn.de

Bayesian estimation of non-stationary Markov models combining micro and macro data

Hugo Storm, Thomas Heckelei, Ron C. Mittelhammer

Abstract

We develop a Bayesian estimation framework for non-stationary Markov models for situations where both sample data on observed transitions between states (micro data) and population data, where only the proportion of individuals in each state is observed (macro data), are available. Posterior distributions on transition probabilities are derived from a micro-based prior and a macro-based likelihood, thereby providing a new method that combines micro and macro information in a logically consistent manner and merges previously disparate approaches for inferring transition probabilities. Monte Carlo simulations for ordered and unordered states show how observed micro transitions improve the precision of posterior knowledge.

Keywords: Bayesian estimation, Markov transitions, prior information, multinomial logit, ordered multinomial logit
JEL-classification: C11, C81

1 Introduction

In this paper a new Bayesian estimation framework for inferring the transition probabilities of non-stationary Markov models is developed. Non-stationary Markov models facilitate analysis of factors influencing the probability that an individual will transition between predefined states. Data used for estimating Markov models can either be panel data, where the specific movement of an individual between states is observed over time, or aggregated data, providing only the number of individuals residing in each state over time. Following Markov terminology, we refer to such panel data and aggregated data as *micro* and *macro* data, respectively. The overall objective of our approach is to combine micro and macro information into a unified and consistent estimation methodology

Examples of empirical problems for which the preceding types of macro and micro data are relevant are readily available in the literature. One example is the analysis of EU farm structural change, where structural change is defined as farm size or production specialization change over time (see ZIMMERMANN et al., 2009 for a review of that strand of literature). In that instance population data on the number of farms in specific size or specialization states is available from the Farm Structure Survey. Micro data, offering observed transitions of individual farms between the states, is available in the Farm Accountancy Data Network, albeit for a relatively small sample of farms. Another example is the analysis of voter transitions in political science. Here, macro data on the shares of candidates is usually available from official statistics, whereas micro data can be obtained from voter (transition) surveys. Additional examples of similar data situations can be found in the context of Ecological inference problems, which are closely related to Markov processes (WAKEFIELD, 2004, LANCASTER et al., 2006).

This paper utilizes a Bayesian framework that allows prior information to be incorporated into the estimation of non-stationary Markov models within an established and consistent probabilistic framework. Moreover, it defines a rigorous statistical method for combining previously distinct micro and macro data-based classical estimators. Specifically, we offer a method for utilizing a sample of micro observations as prior information weighting a macro data-based likelihood function. The approach is presented for both ordered and unordered Markov states. Monte Carlo simulations are used to assess how the inclusion of prior information affects the posterior as well as the numerical stability of the sampling algorithm, and the degree to which estimator performance is improved under different micro sample sizes for both specifications. The combination of micro and macro data was considered previously in the context of a medical application by HAWKINS and HAN (2000). They analyzed macro data obtained in repeated independent cross sectional surveys within a city district together with limited micro data obtained from respondents who where 'coincidently' interviewed in two consecutive cross sectional surveys. The behavior under study was the benefits of an intervention program attempting to modify drug use-related behavior. They defined a linear model that jointly explained the marginal probabilities of being in a use state in a certain time period (based on "standard observed proportion estimates" from aggregate data) and bivariate transition probabilities relating to state transitions (from the micro data), linking the two through an appropriate asymptotic covariance structure and constraints imposed by the sampling design. The Bayesian approach that we provide offers a more general full posterior information approach for combining micro and macro data-based information on transition probabilities and allows the estimation of functional relationships that link transition probabilities with their determinants, in addition to not relying on asymptotic properties.

The paper is organized as follows: First, a Bayesian framework for non-stationary Markov models is developed in section 2. Two different specifications of the transition probabilities are discussed, an appropriate likelihood function and prior density are defined, and issues relating to computational implementation are identified. The design and results of a Monte Carlo simulation experiment are presented in section 3, where the impact of the prior on the posterior distribution and estimator performance are analyzed. Section 4 concludes and discusses areas for further research.

2 Bayesian Approach for non-stationary Markov models

A Markov process provides a conceptual model for the movement of individuals between a finite number of predefined states, $i = 1,...,k$, within the context of a stochastic process. The k states are mutually exclusive and exhaustive. A Markov process is characterized by a $(k \times k)$ transition probability (TP) matrix[1] $\mathbf{P}_t$. The elements P_{ijt} of $\mathbf{P}_t$ represent the probability that an individual moves from state i in time $t-1$ to j in time t. The $(k \times 1)$-vector $\mathbf{n}_t$ denotes the number of individuals in each state i at time t and evolves over time according to a (first order) Markov process

$$\mathbf{n}_t = \mathbf{P}_t' \mathbf{n}_{t-1} . \tag{1}$$

In a non-stationary Markov process, the TPs change over time[2] $t = 1,...,T$. Data used for estimating a non-stationary Markov process can either be macro or micro level. In the case of macro data, only the aggregate numbers of individuals in the states, $\mathbf{n}_t$, is observed at each time period. For micro data, the movement of each individual between states is also observed over time. Thus, the $(k \times k)$-matrix $\mathbf{N}_t$ with elements n_{ijt} representing the number of individuals that transition from state i at $t-1$ to j in t, is directly observed.

The specification of the TP matrix $\mathbf{P}_t$ depends on the underlying behavioral model. In the following subsection we discuss specifications corresponding to ordered as well as unordered Markov states. Afterwards the posterior density consisting of a data likelihood function $L(\mathbf{n}_1,...,\mathbf{n}_T | \boldsymbol{\beta})$, representing the macro data, and a prior density $p(\boldsymbol{\beta})$, representing the micro data is derived.

2.1 Specification of the Transition Probability Matrix

For appropriate specification of the TPs, the nature of the relationship between Markov states need to be considered, and we discuss two different behavioral models that differentiate between ordered and unordered Markov states. We argue that for ordered Markov states the ordered logit model is superior to the more common multinomial logit model with respect to both model assumptions and from a computational point of view.

[1] Bold letters are used for vectors or matrices.

[2] Depending on the problem context, one could also consider only two time periods observed over various regions, or a combination of multiple time and regional observations.

2.1.1 Multinomial logit model

In cases where the states of the Markov process are unordered, the multinomial logit model is a suitable specification for the TPs[3]. In order to define notation, and to establish a consistent modeling context for use throughout the rest of the paper, we provide a brief review of the logit model derivation.[4] The specification based on the multinomial logit model assumes that the transition of individuals between different states can be represented by a random utility model. The utility that would accrue to individual l upon moving from state i in $t-1$ to j in t is denoted as U_{ijtl}:

$$U_{ijtl}=V_{ijt}+\varepsilon_{ijtl}, \tag{2}$$

where the deterministic component of utility is specified as $V_{ijt}=\mathbf{z}'_{t-1}\mathbf{b}_{ij}$, the $\left(n_z \times 1\right)$ vector $\mathbf{z}_{t-1}$ represents observations on lagged exogenous variables, and $\mathbf{b}_{ij}$ is a $\left(n_z \times 1\right)$ vector of unknown parameters. Note that the deterministic part varies only over time and not over individuals because aggregated data is considered. Consequently, the deterministic component of utility reflects exogenous variables that affect the utility of all individuals alike. The random error ε_{ijtl} varies over time and individuals. An individual chooses a transition that maximizes its utility, so movement from state i in $t-1$ to state j in t occurs if $U_{ijtl}=Max\left(U_{i1tl},U_{i2tl},...,U_{iktl}\right)$. The probability that an individual chooses the transition from state i in $t-1$ to state j in t is

$$\begin{aligned} P_{ijtl} &= \Pr\left(U_{ijtl} > U_{iftl}, \forall\ f \neq j\right) \\ &= \Pr\left(V_{ijt}+\varepsilon_{ijtl} > V_{ift}+\varepsilon_{iftl}, \forall\ f \neq j\right) \\ &= \Pr\left(\varepsilon_{iftl}-\varepsilon_{ijtl} < V_{ijt}-V_{ift}, \forall\ f \neq j\right), \end{aligned} \tag{3}$$

which can be rewritten as the value of the cumulative distribution of $\boldsymbol{\varepsilon}_{tl}^{j}$ evaluated at the argument $V_{ijt}-V_{ift},\ for\ f \neq j$, where $\boldsymbol{\varepsilon}_{tl}^{j}$ denotes a vector whose elements are given respectively by $\varepsilon_{iftl}-\varepsilon_{ijtl}$ for $f \neq j$. Letting $f\left(\boldsymbol{\varepsilon}_{tl}^{j}\right)$ denote the probability density of $\boldsymbol{\varepsilon}_{tl}^{j}$, the appropriate cumulative distribution value can be expressed as

[3] A multinomial probit model could be an appropriate alternative and provides flexibility in the error structure specification, but is left to future work because of the additional computational complexities involved.

[4] Textbook expositions of the standard multinomial logit model, which are the foundation of the derived model, can be found in TRAIN (2009) or MITTELHAMMER et al. (2000).

$$\begin{aligned} P_{ijtl} &= \Pr\left(\varepsilon_{iftl} - \varepsilon_{ijtl} < V_{ijt} - V_{ift}, \forall\ f \neq j\right) \\ &= \int \left[\prod_{f \neq j} I_{\left(-\infty, V_{ijt} - V_{ift}\right)} \left(\varepsilon_{iftl} - \varepsilon_{ijtl}\right) \right] f\left(\boldsymbol{\varepsilon}_{tl}^{j}\right) d\boldsymbol{\varepsilon}_{tl}^{j}, \end{aligned} \tag{4}$$

where the indicator function $I_{(a,b)}(x)$ takes the values 1 if $a < x < b$ and equals 0 otherwise. The logit model assumes that the individual ε_{ijtl} are *iid* random draws from a Gumbel distribution. The random vector $\boldsymbol{\varepsilon}_{tl}^{j}$ then follows a logistic distribution for which a closed form expression for the integral in (4) can be expressed as

$$P_{ijtl} = \frac{e^{V_{ijt}}}{\sum_f e^{V_{ift}}} = \frac{e^{\mathbf{z}'_{t-1}\mathbf{b}_{ij}}}{\sum_f e^{\mathbf{z}'_{t-1}\mathbf{b}_{if}}} = P_{ijt}, \tag{5}$$

where the last equality follows because the deterministic component of utility does not vary between individuals (recall (2)). In order to identify the parameters in (5) normalization is required because only the difference in utility matters, and not their absolute value. Normalization is achieved by using the last state as a reference case and transforming equation (5) to

$$P_{ijt} = \frac{e^{\mathbf{z}'_{t-1}\left(\mathbf{b}_{ij} - \mathbf{b}_{ik}\right)}}{\sum_f e^{\mathbf{z}'_{t-1}\left(\mathbf{b}_{if} - \mathbf{b}_{ik}\right)}} = \frac{e^{\mathbf{z}'_{t-1}\boldsymbol{\beta}_{ij}}}{1 + \sum_{f=1}^{k-1} e^{\mathbf{z}'_{t-1}\boldsymbol{\beta}_{if}}}, \tag{6}$$

where $\boldsymbol{\beta}_{ij} \equiv \mathbf{b}_{ij} - \mathbf{b}_{ik}$, and thus $\boldsymbol{\beta}_{ik} = 0$. The development of the model implies that for each row $i = 1, \ldots, k$ of the transition matrix, $\mathbf{P}_t$, there is one multinomial model analogous to (6) specified across states $j = 1, \ldots, k$.

2.1.2 Ordered logit model

If the Markov states are ordered, an ordered choice model is an appropriate specification for the underlying behavioral model. In this case it is assumed that there exists an unobserved continuous latent variable Y_{itl}^{*} for each individual l that determines the value of the observed variable Y_{itl} according to

$$\begin{aligned} &Y_{itl} = 1 \quad \text{if} \quad Y_{itl}^{*} \leq c_1 \\ &Y_{itl} = j \quad \text{if} \quad c_{j-1} < Y_{itl}^{*} \leq c_j \quad \forall\ j = 2, \ldots, k-1 \\ &Y_{itl} = k \quad \text{if} \quad c_{k-1} < Y_{itl}^{*} \end{aligned} \tag{7}$$

for $i = 1, \ldots, k$ where the c_j 's are the thresholds for each Markov state and the index i indicates the an individual was in state i at $t-1$. The unobserved latent variable Y_{itl}^{*} consists of a deterministic part $\mathbf{z}'_{t-1}\boldsymbol{\beta}_i$ and a random part ε_{itl}^{*}, and is defined by

$$Y_{itl}^{*} = \mathbf{z}'_{t-1}\boldsymbol{\beta}_i + \varepsilon_{itl}^{*} \quad \forall\ i = 1,...,k\ . \tag{8}$$

For the deterministic part the $(n_z \times 1)$ vector of unknown parameters $\boldsymbol{\beta}_i$ are allowed to differ between the k different states in $t-1$. As in the preceding multinomial logit model, the deterministic part varies over time but not over individuals. Letting $c_o \equiv -\infty$ *and* $c_k \equiv \infty$, the probability of an individual being in state j at t , given that it is in state i at $t-1$, is then given by

$$\Pr\left(Y_{itl} = j\right) = \Pr\left(c_{j-1} < Y_{itl}^{*} \le c_j\right) = P_{ijtl} = P_{ijt} \quad \forall\ j = 1,...,k \tag{9}$$

and similarly for $j=1$ and $j=k$. The last equality follows from the fact that the exogenous variables do not vary over individuals and the errors ε_{itl}^{*} are *iid* over individuals. If it is assumed that the errors ε_{it}^{*} are *iid* random draws from a logistic distribution the model results in an order logit model[5] and the TPs in (9) can be expressed in closed form as

$$\begin{aligned} P_{ijt} &= \Pr\left(c_{j-1} < Y_{it}^{*} \le c_j\right) \\ &= \Pr\left(c_{j-1} < \mathbf{z}'_{t-1}\boldsymbol{\beta}_i + \varepsilon_{it}^{*} \le c_j\right) \\ &= \Pr\left(\varepsilon_{it}^{*} \le c_j - \mathbf{z}'_{t-1}\boldsymbol{\beta}_i\right) - \Pr\left(\varepsilon_{it}^{*} < c_{j-1} - \mathbf{z}'_{t-1}\boldsymbol{\beta}_i\right) \\ &= \frac{e^{c_j - \mathbf{z}'_{t-1}\boldsymbol{\beta}_i}}{1 + e^{c_j - \mathbf{z}'_{t-1}\boldsymbol{\beta}_i}} - \frac{e^{c_{j-1} - \mathbf{z}'_{t-1}\boldsymbol{\beta}_i}}{1 + e^{c_{j-1} - \mathbf{z}'_{t-1}\boldsymbol{\beta}_i}} \qquad \forall\ j = 1,...,k. \end{aligned} \tag{10}$$

$j=k$. The specification, which consists of one ordered choice model for each of the $i = 1,...,k$ Markov states, allows interpreting the probabilities in (10) as one row of $\mathbf{P}_t$.

One important difference between the ordered logit and the multinomial logit model is that only one error term, instead of one error term for each alternative, is considered for each individual. This implies that the assumption of "Independence of Irrelevant Alternatives" (IIA) does not apply to the ordered logit model. This is more appropriate whenever the alternatives are ordered since in this case it can be expected that the error associated with one state is more similar to the error of an alternative close to it than to an alternative further away (TRAIN, 2009). Also from a computational point of view, the ordered logit specification is often preferable since only $k\,n_z$ (or $k\,n_z + (k-1)n_z$ if thresholds c_j are estimated) parameters

[5] Assuming that the ε_{it}^{*} are random draws from a normal distribution would result in a probit (see footnote 3).

need to be estimated, as compared to $k(k-1)n_z$ parameters for the multinomial logit model.

A further advantage of the ordered choice model is that the interpretation of the latent variable is often straightforward. For example, in the case of farm structural change where Markov states refer to size classes, the latent variable can be interpreted as farm size or in the medical context where classes refer to different stages of illness, the latent variable can be interpreted as the degree of illness. The decision between an ordered and unordered choice model is, however, not always straightforward and can depend on the problem context and decision makers' behavioral characteristics. In the voter transition example, one could regard the candidates as unordered choices, but alternatively one could also argue that they are ordered according to a one-dimensional political spectrum ("right" to "left"). In that case both models have their justification and the choice between the two must be guided by theoretical and/or substantive behavioral arguments.

2.2 Data likelihood

In order to implement a Bayesian framework of analysis, a likelihood function needs to be defined. The foundation for this likelihood specification is provided by the first-order non-stationary Markov process proposed by MACRAE (1977). For the specification of a macro data based likelihood function MACRAE (1977) points out that the type of available observations needs to be considered. She distinguishes the case of perfect observations, where the state proportions, $\mathbf{x}_t$, are observed over time for the entire population of size N, from imperfect observations where only the state proportions, $\mathbf{y}_t$, of a random sample of size $M_t < N$ is drawn and observed at each time period. In the case of perfect observations the distribution of $\mathbf{x}_t$ is fully characterized by $\mathbf{x}_{t-1}$, which is not the case for imperfect observation where the distribution of $\mathbf{y}_t$ also depends on earlier observations, $y_{t-2},...,y_0$, that provide additional information on y_t. For the latter case MACRAE (1977) proposed a limited information likelihood concept which is appropriate whenever macro data is available for a sample of the population. In the following, however, we restrict ourselves to the case of perfect observations, i.e., a census type of data set.

MACRAE (1977) shows that in the case of perfect observations, the state proportions are distributed as a weighted sum of independent multinomial random variables with probabilities equal to the corresponding rows in $\mathbf{P}_t$ and weights equal to the state proportions in $t-1$. The resulting likelihood function is given by

$$L\left(\boldsymbol{\beta}\middle|\mathbf{n}_1,...,\mathbf{n}_T\right)=\prod_{t=1}^{T}\sum_{\mathbf{H}_t\in\mathbb{H}_t}\prod_{i=1}^{k}\left(n_{i,t-1}!\right)\left(\prod_{j=1}^{k}P_{ijt}^{\eta_{ijt}}/\eta_{ijt}!\right), \tag{11}$$

where n_{it} 's are the elements of the data vector $\mathbf{n}_t$. The summation in the likelihood expression (11) is over the set $\mathbb{H}_t$ of all matrices $\mathbf{H}_t$ having rows sum to corresponding elements in $\mathbf{n}_{t-1}$ and columns sum to the corresponding entries in $\mathbf{n}_t$, so that

$$\mathbb{H}_t = \left\{ \mathbf{H}_t \left| \sum_h \eta_{iht} = n_{i,t-1}, \ \sum_h \eta_{hjt} = n_{jt} \quad \forall i, j \right. \right\}, \tag{12}$$

where η_{ijt} denote the (unobserved) number of individuals transitioning from state i at time $t-1$ to state j at time t, and $\mathbf{H}_t$ is a matrix whose $(i,j)^{th}$ element is η_{ijt}.

The set of matrices represented by $\mathbb{H}_t$ is the collection of all conceptually possible outcomes of between-states transition numbers when moving from observed state distribution $\mathbf{n}_{t-1}$ in time $t-1$ to the observed state distribution $\mathbf{n}_t$ in time t. The number of elements in set $\mathbb{H}_t$ increases exponentially with the number of states, making the implementation of expression (11) for larger samples challenging (or impossible) from a computational point of view (for example, in the case of only 3 states and 200 observations, there are over 2.5 million combinations of (3×3)-matrices possible if approximately the same number of individuals reside in each of the three states). To mitigate this dimensionality problem, a large sample approximation that avoids the computation of the set $\mathbb{H}_t$ is employed (see HAWKES, 1969 and BROWN and PAYNE, 1986). In particular, letting $\mathbf{n}_t^*$ represent $\mathbf{n}_t$ without the last row and $\mathbf{P}_t^*$ represent $\mathbf{P}_t$ without the last column, one can assume, in large samples, that $\mathbf{n}_t^*$ is distributed as a $(k-1)$-variate normal vector with mean vector $\mathbf{P}_t^{*\prime}\mathbf{n}_{t-1}$ and covariance matrix

$$\operatorname{cov}\left(\mathbf{n}_t^*\right) = diag\left(\mathbf{P}_t^{*\prime}\mathbf{n}_{t-1}\right) - \mathbf{P}_t^{*\prime} diag\left(\mathbf{n}_{t-1}\right)\mathbf{P}_t^* = \mathbf{\Gamma}_t, \tag{13}$$

where $diag(\cdot)$ is a square matrix with the argument vector as the main diagonal and zero off-diagonal elements. The large sample log-likelihood, L_{la}, can then be written as

$$L_{la}\left(\boldsymbol{\beta} \middle| \mathbf{n}_1, ..., \mathbf{n}_T\right) = \\ \sum_{t=1}^{T} -0.5\left(\log\left|\mathbf{\Gamma}_t\right| + \left(\mathbf{n}_t^* - \mathbf{P}_t^{*\prime}\mathbf{n}_{t-1}\right)' \left(\mathbf{\Gamma}_t\right)^{-1} \left(\mathbf{n}_t^* - \mathbf{P}_t^{*\prime}\mathbf{n}_{t-1}\right) \right). \tag{14}$$

2.3 Prior information

As noted in the introduction, the intent of the Bayesian framework is to combine a macro-data likelihood function, as derived in the previous section, with a prior density representing information derived from a sample of micro observations on state transitions. To specify an appropriate prior density $p(\boldsymbol{\beta})$, consider the underlying sampling distribution of the micro observations. Let n_{it} be the number of individuals that were in state i at time t, let $\mathbf{X}_t^i$ be the vector of shares across states in *t* of individuals who were in state i in $t-1$, and let $\mathbf{P}_{it}$ be the i-th row of $\mathbf{P}_t$. The propensity of each individual in the micro sample to transition between states is in accordance with the appropriate elements of $\mathbf{P}_t$. Analogous to the case of macro data, the distribution across states in *t* of individuals who were in state i in $t-1$ is multinomial around mean $\mathbf{P}_{it}$ with size n_{it}. The observed number of individuals in each of the *k* states in *t*, n_{it}, $i=1,...,k$, is then the corresponding weighted sum of vectors $\mathbf{X}_t^i$, $i=1,...,k.$ Therefore, the prior density can be represented as a likelihood similar to (11), except that now information about the individual transitions n_{ijt} are available, making the summation over the set $\mathbb{H}_t$ unnecessary because the actual transitions are observed. Hence the likelihood simplifies to

$$p(\boldsymbol{\beta}) = L(\boldsymbol{\beta}|\mathbf{N}_1,...,\mathbf{N}_T) = \prod_{t=1}^{T}\prod_{i=1}^{k}(n_{i,t-1}!)\left(\prod_{j=1}^{k}\mathbf{P}_{ijt}^{\;n_{ijt}} / n_{ijt}!\right), \tag{15}$$

where the $(k \times k)$-matrix $\mathbf{N}_t$ has elements n_{ijt} representing the number of individuals that transition from state i at $t-1$ to j in t. We emphasize that for the case of aggregated data discussed above, the distribution of $\mathbf{n}_t$ differs between imperfect and perfect observations, while for micro observations, this distinction does not apply. In the latter case, the distribution of $\mathbf{x}_t$ is fully characterized by $\mathbf{x}_{t-1}$ regardless of whether a sample or the entire population is observed. The fundamental difference is that in the case of micro observations, individuals in the sample in time period t are the same as in $t-1$ which is usually not the case for macro data. Consequently, information earlier than $\mathbf{x}_{t-1}$ contains no additional information.

2.4 Computational Implementation

In order to conduct Bayesian inference in the model depicted above, integrating and/or taking expectations with respect to the posterior density $h(\boldsymbol{\beta}|\mathbf{d})$ is required. An analytical approach to such computations is intractable, and therefore sampling from the posterior density to implement Monte Carlo integration is pursued in this section. The sampling is accomplished via a Markov Chain Monte Carlo (MCMC) method, namely, the Metropolis Hastings (MH) algorithm. The MH sampler is capable of generating a (pseudo-) random sample from almost any target distribution that is known up to a normalizing constant (see CHIB and GREENBERG, 1995). For our purposes, we approximate the posterior mean, which is the optimal Bayesian estimator under squared error loss, by calculating the mean of an *iid* sample from $h(\boldsymbol{\beta}|\mathbf{d})$ for sufficiently large sample sizes.

Specifically, a simple random walk MH algorithm is employed to obtain a sample of R outcomes, $\boldsymbol{\beta}^{(1)},...,\boldsymbol{\beta}^{(R)}$, from the posterior density. For each iteration $r=1,...,R$ of the MH algorithm u is drawn from $\mathcal{U}\left(0,1\right)$ and a candidate $\boldsymbol{\beta}^{can}$ is drawn from the multivariate normal generating density, $\mathcal{N}\left(\boldsymbol{\beta}^{(r)},\sigma^2\mathbf{I}\right)$. The tuning parameter σ can be used to control the acceptance rate of the algorithm. The candidate is accepted, $\boldsymbol{\beta}^{(r+1)}=\boldsymbol{\beta}^{can}$, if $u\leq\alpha\left(\boldsymbol{\beta}^{(r)},\boldsymbol{\beta}^{can}\right)$, otherwise the chain remains in its current state, $\boldsymbol{\beta}^{(r+1)}=\boldsymbol{\beta}^{(r)}$, where following CHIB and GREENBERG (1995),

$$\begin{aligned}\alpha\left(\boldsymbol{\beta}^{(r)},\boldsymbol{\beta}^{can}\right)&=\min\left[\frac{h\left(\boldsymbol{\beta}^{can}\middle|\mathbf{d}\right)}{h\left(\boldsymbol{\beta}^{(r)}\middle|\mathbf{d}\right)}\frac{f\left(\boldsymbol{\beta}^{(r)};\boldsymbol{\beta}^{can},\sigma^2\mathbf{I}\right)}{f\left(\boldsymbol{\beta}^{can};\boldsymbol{\beta}^{(r)},\sigma^2\mathbf{I}\right)},1\right]\\&=\min\left[\frac{h\left(\boldsymbol{\beta}^{can}\middle|\mathbf{d}\right)}{h\left(\boldsymbol{\beta}^{(r)}\middle|\mathbf{d}\right)},1\right]\end{aligned}, \tag{16}$$

with the last equality following from symmetry of the multivariate normal probability density function $f\left(\boldsymbol{\beta}^{can};\boldsymbol{\beta}^{(r)},\sigma^2\mathbf{I}\right)$. For reasons of computational stability, (16) is transformed to

$$\alpha\left(\boldsymbol{\beta}^{(r)},\boldsymbol{\beta}^{can}\right)=\min\left[\exp\left(\ln h\left(\boldsymbol{\beta}^{can}\middle|\mathbf{d}\right)-\ln h\left(\boldsymbol{\beta}^{(r)}\middle|\mathbf{d}\right)\right),1\right], \tag{17}$$

and $\ln h\left(\boldsymbol{\beta}\middle|\mathbf{d}\right)\propto\ln L(\mathbf{d}|\boldsymbol{\beta})+\ln p(\boldsymbol{\beta})$, which mitigates computer overflow problems. In cases where the number of parameters to be estimated is large, a "Block-at-a-Time" algorithm proposed by CHIB and GREENBERG (1995) is employed in which the parameters to be estimated are divided into two blocks.

The sampling algorithm is used to obtain an *iid* sample from the posterior of size n_{sample} after a burn-in period of n_{burn} iterations. The tuning parameter σ of the proposal density is chosen such that an acceptance rate in the interval $\left[.2,\ .3\right]$is obtained.

3 Monte Carlo simulations on the effects of prior information

In this section we analyze the influence of prior information, in the form of a sample of micro observations, on the posterior distribution and associated estimators' performance as well as on the behavior of the sampling algorithm via Monte Carlo simulations. Based on an underlying population of $n_{ind}=10,000$ individuals, four different scenarios are considered regarding the availability of prior information, including a case of no micro observations, and micro samples of *n* = 100, 500, and 1000. The scenarios are further distinguished by the number of Markov states ($k=3,4,5$). Data is generated for $T=100$time periods and

$n_z = 6$ explanatory variables including a constant. All simulations are undertaken for a Markov model based on either the multinomial logit specification or the ordered logit specification discussed above, and are performed using Aptech's GAUSS$^{\text{TM}}$ 11.

3.5 Data generating process

The data generating process distinguishes between the two different behavioral models, based on the multinomial logit and ordered logit specification discussed in section 2.1. In both cases the parameterization is chosen in such a way that the deterministic part constitutes roughly one third of the model's total variation. Further, in both cases n_{ind} individuals are considered that transition over time between the k states in accordance with the underlying behavioral model. The initial state of each individual in $t = 1$ is randomly chosen with probability equal to $u_i \; \forall \, i = 1,...,k$, where the probability is the same for all individuals and given by $u_i = \tilde{u}_i \Big/ \sum_{h=1}^{k} \tilde{u}_h$ with $\tilde{u}_i \sim iid \; \mathcal{U}(0,1)$.

In the multinomial logit model each individual l chooses the state of the next period based on the utility, U_{ijtl}, associated with a specific transition from state i in $t-1$ to j in t. The utility $U_{ijtl} = V_{ijt} + \varepsilon_{ijtl}$ consists of a deterministic part $V_{ijt} = \mathbf{z}'_{t-1}\mathbf{b}_{ij}$ and an individual random part ε_{ijtl} (see equation (2)) and is generated by drawing the elements of the (lagged) exogenous variables $\mathbf{z}_{t-1}$ from $\mathcal{N}(1,4)$ and the elements of the $(n_z \times 1)$ "true" parameter vectors $\mathbf{b}_{ij}$ from $\mathcal{U}(-1,1)$. Since only differences in utilities are relevant, the parameters of the last alternative are set to zero, $\mathbf{b}_{ik} = \mathbf{0} \; \forall \, i = 1,...,k$, in order to identify the model. To obtain a logit model, the ε_{ijtl} are drawn from a Gumbel (type I extreme value) distribution, specified by $F_g(\varepsilon_{ijtl}; 0,3) = e^{-e^{-\varepsilon_{ijtl}/3}}$. In each time period an individual chooses the transition that maximizes utility, moving from state i in $t-1$ to state j in t if $U_{ijtl} = Max(U_{i1tl}, U_{i2tl}, ..., U_{iktl})$.

For the ordered logit model, the transition between states is based on a latent index value $Y^*_{itl} = \mathbf{z}'_{t-1}\boldsymbol{\beta}_i + \varepsilon^*_{itl}$ consisting of a deterministic part $\mathbf{z}'_{t-1}\boldsymbol{\beta}_i$ and a random part ε^*_{itl} (see equation (8)). The index value is generated by drawing the elements of the (lagged) exogenous variables $\mathbf{z}_{t-1}$ from $\mathcal{N}(1,4)$ and the elements of the $(n_z \times 1)$ *true* parameter vectors $\boldsymbol{\beta}_i$ from $\mathcal{U}(-1,1)$. The random errors ε^*_{itl} are *iid* random draws from a logistic distribution, specified by $F_l(\varepsilon^*_{itl}; 0, 2.3) = \left(1 + e^{-\varepsilon^*_{itl}/2.3}\right)^{-1}$. The latent index value determines the outcome of Y_{itl} for each individual in each time period according to (7).

With this sampling design a micro dataset for n_{ind} individuals and T time periods is obtained for both the multinomial logit and the ordered logit specification, and represents the full population of individuals under study. For the specification of the prior density, random samples of size 100, 500, and 1000 are drawn without replacement from these micro datasets. The population is transformed into macro datasets by simply counting the number of individuals in each state in each time period.

In order to avoid dependency of the results on a specific set of parameters, $n_{true} = 10$ *true* models are generated using the data generating process. For each of the n_{true} true models the process is repeated $n_{rep} = 20$ times with the same parameters, but with new draws of the random errors ε_{ijtl} or ε^{*}_{itl} in each repetition.

3.6 Performance measures

The influence of prior information is assessed by a comparison of measures characterizing features of the posterior density, the performance of the associated estimator, i.e., the mean of the posterior density, and the numerical stability of the sampling algorithm. Regarding performance of the estimator, for each of the $n_{true}n_{rep} = 200$ simulation outcomes the squared error, i.e., the squared deviation of the estimates from the true value, is calculated. To obtain one scalar value measure for each simulation outcome the squared errors are summed over all n_z parameters.

For the Monte Carlo simulation a fixed burn-in period and a fixed sample size was employed for the MH sampler. Even though burn-in periods and the sample sizes were assessed using graphical measures in trial runs for each scenario and resulted in substantially large burn-in periods, it still cannot be guaranteed that the MH sample converged correctly for every simulation run. Therefore, Box-Whisker-Plots are employed to detect outliers among the sum of squared errors of the $n_{true}n_{rep}$ simulations as an indication that the MH sample had not converged appropriately. Measures characterizing the posterior density and performance measures relating to the estimator are then calculated based on only those runs that were not detected to be outliers.

The effect of prior information on the spread of the posterior was assessed based on posterior variances, and was calculated on the basis of the posterior sample outcomes. The total variance of the posterior density was calculated by summing over the posterior variances of all n_z parameters in each run and the mean over all $n_{true}n_{rep}$ simulation runs, except for any outliers, was then calculated to obtain one scalar value measure of the total variance.

The analysis of the influence of prior information on the Bayes estimator is based on the mean square error (MSE) criterion, calculated as the mean of the summed squared errors between estimates and true parameter values, where the mean is calculated over all of the $n_{true}n_{rep}$ simulation runs not detected as outliers. Further, the MSE is decomposed into variance and bias components, where the squared bias is again summed over all parameters. Both the distribution of the sum of squared errors and the number of outliers detected for each scenario provides an assessment of the numerical stability of the MH sampler, and the effects of prior information on that numerical stability.

3.7 Results of the Monte Carlo Simulation

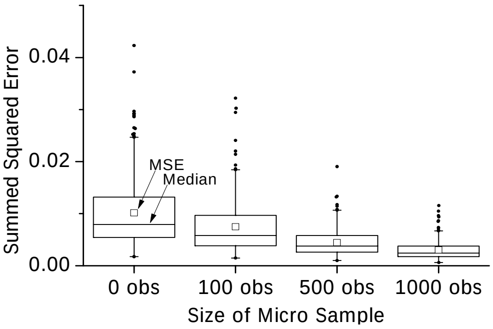
The results of the Monte Carlo Simulations for the multinomial logit model are presented in Figure 1. Results show that considering prior information in the form of a micro sample decreases the total variance of the posterior density, and more so the larger the micro sample. The effect of prior information becomes even more pronounced the more Markov states are considered. Similarly, prior information decreases the MSE of the estimator, and more so the more Markov states that are being considered. Decomposing the MSE into bias and variance suggests that the MSE is primarily determined by the variance of the estimator. In all scenarios the share of the squared bias is only 4 to 9 % of total MSE.

The distribution of the summed squared errors, as depicted in the Box-Whisker-Plots in Figure 1, provides information about the numerical performance of the MH sampling algorithm. Results show that more simulation runs are detected as outlier in the no prior information scenario (i.e. micro sample with 0 obs.), especially when considering $k=4$ or $k=5$ Markov states. This observation indicates problems relating to the numerical stability of the MH sampler, in the sense that the algorithm does not converge correctly for some simulation runs. When considering a micro sample as prior information, substantially fewer simulation runs are detected as outliers, indicating that the use of prior information improves the numerical stability of MH sampler.

Similar results are obtained for the ordered logit model as depicted in Figure 2. As in the multinomial logit simulation, results indicate that prior information reduces the variance of the posterior density, and more so the larger the micro sample considered. The same can be observed for the MSE, which decreases with increasing micro sample size. MSE is mainly determined by the variance of the estimator and the share of the squared bias is only 3 to 9 % of total MSE in all scenarios except for the no prior information scenario (i.e., no micro sample for $k=4$ and $k=5$ Markov states, for which the bias share is substantially larger with 44 and 42 %, respectively.

Figure 1: Results for the multinomial logit model of a Monte Carlo simulation to analyse the influence of prior information, in the form of a micro sample, on the posterior and the posterior mean estimator

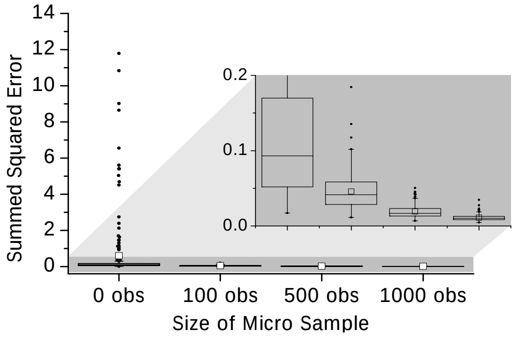
Number of Markov states: k=3



	Size of micro sample			
Measures	0	100	500	1000
MSE[a], Estimator	0.00892	0.00672	0.00414	0.00275
Sq. Bias[a], Estimator	0.00042	0.00041	0.00024	0.00014
Variance[a], Posterior	0.00592	0.00496	0.00312	0.00219
Outlier	12	9	7	9

Sample: 50,000; Burn-In: 100,000; Blocks: 1;
σ: 1/800; num. o coef.: 36

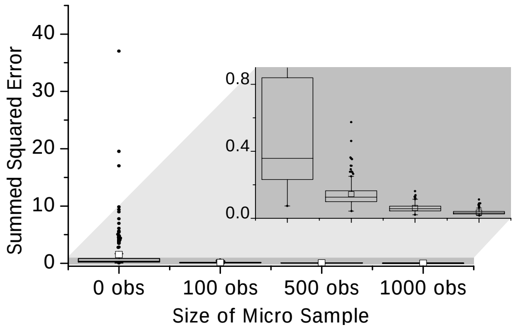
k=4



	Sizeof micro sample			
Measures		100	500	1000
MSE[a], Estimator	0.09394	0.04425	0.01802	0.01036
Sq. Bias[a], Estimator	0.00808	0.00217	0.00108	0.00045
Variance[a], Posterior	0.03585	0.02433	0.01340	0.00891
Outlier	34		9	11

Sample: 100,000; Burn-In: 200,000; Blocks: 1;
σ: 1/870; num. of oef.: 72

k=5



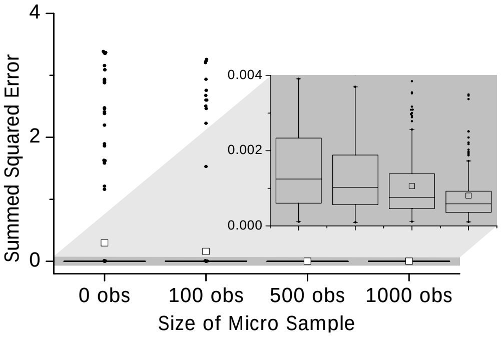
	Size of micro sample			
Measures	0	100	500	1000
MSE[a], Estimator	0.39920	0.13130	0.0586	0.03296
Sq. Bias[a], Estimator	0.03678	0.00758	0.00336	0.00179
Variance[a], Posterior	0.18702	0.10570	0.04839	0.02992
Outlier	3	1	7	9

Sample: 250,000; Burn-In: 500,000; Blocks: 2;
σ: 1/580; num. of coef.: 12

[a] Calculated without simulation runs detected as outliers.

Figure 2: Results for the ordered logit model of a Monte Carlo simulation to analyse the influence of prior information, in the form of a micro sample, on the posterior and the posterior mean estimator

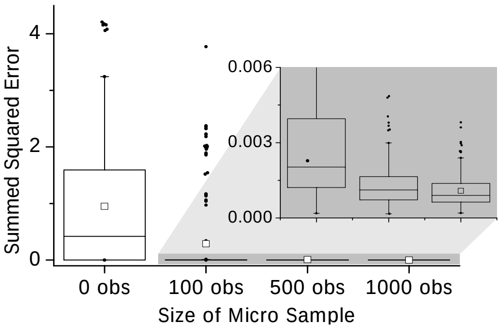
Number of Markov states: k=3



	Size of micro sample			
Measures	0	100	500	1000
MSE[a], Estimator	0.00124	0.00115	0.00088	0.00064
Sq. Bias[a], Estimator	0.00006	0.00005	0.00003	0.00004
Variance[a], Posterior	0.00108	0.00103	0.00076	0.00061
Outlier	29	18	13	15

Sample: 20,000; Burn-In: 50,000; Blocks: 1;
σ: 1/750; num. ofcoef.: 15

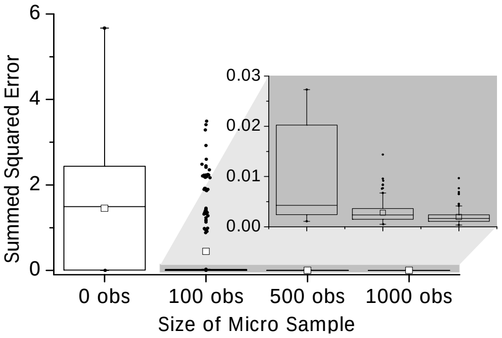
k=4



	Size of micro sample			
Meures	0	10	500	000
MSE[a], Estimator	0.85311	0.00205	0.00118	0.00100
Sq. Bias[a], Estimator	0.37768	0.00015	0.00006	0.00004
Variance[a] Posterior	0.00219	0.00157	0.00111	0.00082
Outlier	6	35	8	8

Sample: 20,000; Burn-In: 60,000; Blocks: 1;
σ: 1/750 num. of coef.: 20

k=5



	Size of micro sample			
Measures	0	10	500	1000
MSE[a], Estimator	1.45498	0.00420	0.0255	0.00176
Sq. Bias[a], Estimator	0.61503	0.00030	0.00021	0.00016
Variance[a], Posterior	0.00576	0.00360	0.00232	0.00169
Outlier	0	48	7	7

Sample: 50,000; Burn-In: 100,000; Blocks: 1;
σ 1/800; num. o coef.: 25

[a] Calculated without simulation runs detected as outliers.

The number of outliers detected by the Box-Whisker-Plots are used again to assess the numerical stability of the MH sampler. For $k = 4$ and $k = 5$ Markov states, however, no outlier are detected and the entire shape of the distribution of the summed squared errors needs to be taken into account to obtain a more complete assessment of the performance of the MH sampler. The fact that no outliers are detected should not be interpreted as indicating the MH sampler converges correctly for every run. On the contrary, when considering the distribution of the summed squared errors in these two scenarios it could be that - due to a poor performance of the MH sampler and a large number of outliers - the Box-Whisker-Plots failed to distinguish between correct and failed convergence of the MH sampler. This is important with respect to the comparison of these two scenarios to other scenarios because it implies that the calculated MSE and posterior variance might not be accurate. It also can explain the substantially larger share of the bias of the MSE in these two scenarios as mentioned above.

With respect to an assessment of the performance of the MH sampler, the results are consistent with the findings in the multinomial logit case, where performance of the MH sampler improves the larger the micro sample size considered as prior information. It is worth noting that the numerical problems in cases without prior information persist (and seem to be more severe) in the ordered logit model compared to the multinomial logit model even though substantially fewer coefficients need to be estimated (e.g. 25 compared to 120 for $k = 5$).

Overall the results suggest that without prior information, alternative individualized sampling strategies or extensions of the simple MH sampler (e.g. Parallel Tempering (LIU, 2008) or Multiple Try Method (LIU et al., 2000)) should be considered for successful sampling from the posterior, which could not be automated for the Monte Carlo simulations. This suggests that through prior information, the computational demands with respect to the sampling algorithm are reduced and that precise estimation in terms of the MSE can be achieved with the simple MH sampler in both the multinomial and the ordered logit model with a moderately sized micro sample.

4 Conclusion

In this paper a Bayesian estimation framework for non-stationary Markov models is derived that allows micro and macro data to be combined in estimation to provide more precise inference regarding model parameters. Specifically, it is shown how a sample of observed transitions between states at the individual level can be implemented as prior information within an otherwise macro data Bayesian estimation framework. Moreover, the paper proposes two different models for the specification of the transition probabilities depending on whether the Markov states are unordered or ordered, using a multinomial logit and an ordered logit model, respectively. In contrast earlier approaches for combining micro and macro data offered in the literature, the Bayesian framework considered here offers a more general full posterior information approach for combining micro and macro data-based information on transition probabilities and allows the estimation of functional relationships that link transition probabilities with their determinants, in addition to not relying on asymptotic properties. Monte Carlo simulations are used to analyze the influence of prior information on the posterior distribution and the performance of the posterior mean, which is the widely used Bayesian estimator. Results indicate that prior information, in the form of a micro sample of data, improves the performance of the posterior mean estimator and reduces the total variance of the posterior distribution substantially. This reduction becomes more important, the more Markov states are considered. The results of the Monte Carlo Simulation also indicate that the numerical implementation of the employed MH algorithm improves the larger the size of the micro sample. Thus prior information in the form of a sample of micro transitions can improve estimation in at least two ways: with respect to the accuracy of the posterior information on the parameters of interest as well as the numerical stability of the estimation approach.

These findings and the proposed approach are subject to some limitations. First of all, the considered likelihood specification is only applicable to the case of perfect observations, i.e. if aggregated data is observed for the entire population over time. For situations where aggregated data is only available from a sample of the entire population, there are other likelihood specifications that can be considered for use in the proposed Bayesian framework, such as MACRAE (1977) limited information likelihood specification. Secondly, the number of parameters that need to be estimated increases with the number of Markov states, often limiting the number of Markov states that can be feasibly considered in empirical applications. The proposed ordered logit approach addresses this problem since substantially fewer parameters need to be estimated compared to the commonly applied multinomial logit model. Other model specifications based on continuous Markov chains could be developed in which the number of model parameters is independent from the number of Markov states. First attempts in this respect are undertaken by PIET (2010).

Overall, this paper contributes to the existing literature by providing an estimation framework that allows for combining micro and macro data information relating to non-stationary Markov models in a way that is consistent with well-established probability calculus and leads to a minimum loss estimator that is based on full posterior information. The approach is relevant for a broad range of empirical applications in which macro and micro data are available and one is interested in quantifying the effect of factors that cause individuals to switch between predefined states.

Acknowledgements

This research was funded by the German Research Foundation (DFG), grant No. 585064. The authors are grateful to Andrea Zimmermann for countless helpful comments and Markov discussions.

References

BROWN, P.J., PAYNE, C.D., 1986. Aggregate data, ecological regression, and voting transitions. Journal of the American Statistical Association. 81, 452–460.

CHIB, S., GREENBERG, E., 1995. Understanding the Metropolis-Hastings Algorithm. The American Statistician. 49, 327–335.

HAWKES, A.G., 1969. An Approach to the Analysis of Electoral Swing. Journal of the Royal Statistical Society: Series A (Statistics in Society). 132, 68–79.

HAWKINS, D.L., HAN, C.-P., 2000. Estimating Transition Probabilities from Aggregate Samples Plus Partial Transition Data. Biometrics. 56, 848–854.

LANCASTER, G.A., GREEN, M., LANE, S., 2006. Reducing bias in ecological studies: an evaluation of different methodologies. Journal of the Royal Statistical Society: Series A (Statistics in Society). 169, 681–700.

LIU, J.S., 2008. Monte Carlo strategies in scientific computing. Springer, New York, NY.

LIU, J.S., LIANG, F., WONG, W.H., 2000. The Multiple-Try Method and Local Optimization in Metropolis Sampling. Journal of the American Statistical Association. 95, 121–134.

MACRAE, E.C., 1977. Estimation of Time-Varying Markov Processes with Aggregate Data. Econometrica. 45, 183–198.

MITTELHAMMER, R.C., JUDGE, G.G., MILLER, D.J., 2000. Econometric foundations. Cambridge Univ. Press, Cambridge.

PIET, L., 2010. A structural approach to the Markov chain model with an application to the commercial French farms. Paper presented: 4èmes journées de recherches en sciences sociales, 9-10.12.2010, Rennes, France.

TRAIN, K.E., 2009. Discrete choice methods with simulation, 2nd. Cambridge University Press, New York.

WAKEFIELD, J., 2004. Ecological inference for 2×2 tables. Journal of the Royal Statistical Society: Series A (Statistics in Society). 167, 385–445.

ZIMMERMANN, A., HECKELEI, T., DOMÍNGUEZ, I.P., 2009. Modelling farm structural change for integrated ex-ante assessment: review of methods and determinants. Environmental Science & Policy. 12, 601–618.