Bayesian Estimation of Transition Probabilities Without Panel Data: Applications and Limitations

Umberto Barbieri* Giulio Radaelli†

January 2024

Abstract

We follow the approach of Lee, Judge, and Zellner 1968 to estimate transition probabilities with a Bayesian approach when panel data are not available, but we only have state proportions for each period. After discussing the modeling assumptions we find the posterior pdf up to a normalizing constant and, as opposed to Lee, Judge, and Zellner 1968, we do not derive an estimator for the MAP but proceed with a MCMC approximation of the posterior density. This procedure is applied to simulated data, to a Labor Force Survey dataset, and to an electoral dataset. We conclude by showing the limitations of this approach in terms of robustness and computational complexity.

*MSc Candidate in Finance at Bocconi University
†MSc Candidate in Economic and Social Sciences at Bocconi University

Introduction

In this paper, we briefly present the model proposed by Lee, Judge, and Zellner (1968) (hence, LJZ), to estimate transition probabilities of discrete stationary Markov processes of order one, when the only available data are proportions for each period.
First, we re-derive the posterior, following the passages of LJZ. Then, we illustrate the methodology to use an MCMC methodology to find the MAP estimate of the transition matrixes.
We apply it to three datasets. The first one is a simulated dataset, with two states; the second one is from Italian electoral records, and we estimate the transition probabilities of voters from one coalition to another; the third one is from Italian Labor Force Surveys, and we estimate the transition probabilities between employment, unemployment, and out-of-labor-force situation.
In Section 1, we present the model and we re-derive the posterior. In Section 2, we present our empirical methodology, in Section 3 we present two applications, and in Section 4 we conclude.

1 The model

1.1 Setting

We now describe the stochastic process that we are going to assume. In each period $t \in \{0, T\}$, the feature x of the unit of observation can be in a different state i, with $i \in \{1, r\}$. If the feature x is in state i at time t, we write $x_t = s_i$.
This model needs two assumptions: (**1.**) the probability with which the unit moves from one state to another (in the following period) depends only on the current state[1] and (**2.**) it is constant through the periods.
Thus,

Assumption 1. Order 1. Using $\mathbf{x}^t = [x_{1:t}]$ to denote the history of x up to time t, we have:

$$\Pr\{x_t = s_i | \mathbf{x}^{t-1}\} = \Pr\{x_t = s_i | x_{t-1}\} \tag{1}$$

Assumption 2. Time-Homogeneity[2] Building on Assumption 1,

$$\Pr\{x_t = s_j | x_{t-1} = s_i\} = p_{ij} \ \forall t \in \{0, T\}, \ i, j \in \{0, r\} \tag{2}$$

Assumption 3. Independence.[3]

$$\text{cov}[\Pr\{x_t = k | x_{t-1} = i\}, \Pr\{x_t = h | x_{t-1} = j\}] = 0 \ \forall i \neq j \tag{3}$$

The first two assumptions allow us to build the transition matrix T of this process, i.e. the $r \times r$ matrix that collects all p_{ij}:

1. Indeed, the process has memory 1.
2. This assumption is Equation (1) in LJZ.
3. Equivalent to Equation (45) in LJZ.

$$T = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1r} \\ p_{21} & p_{22} & \dots & p_{2r} \\ \dots & \dots & \dots & \dots \\ p_{r1} & p_{r2} & \dots & p_{rr} \end{bmatrix} \tag{4}$$

Notice that each row has to sum to 1 because the row collects the probabilities of all the possible scenarios that can happen in $t+1$ conditional on $x_t = s_i$.

$$\sum_j p_{ij} = 1 \ \forall i \in \{1, r\}$$

The diagonal of the matrix T collects all the probabilities of remaining in the same state. In other words, p_{ii} is the probability of being in state i in $t+1$, conditional on being in the same state already in t.
We discuss the strength of the two assumptions when we show how we applied the estimator to real phenomena.

1.2 Available data

The dataset to which the model by Lee, Judge, and Zellner (1968) can be applied is a dataset in which only aggregated proportions are available. In other terms, we do not observe, for each unit, the path $\mathbf{x}^T$ through each period, but only the proportion, for each state i, of the observations in that state. In other words, we observe, for each time t and for each state i

$$w_i(t) := \frac{\sum_{d=1}^{N(t)} \mathbb{1}[x_{d,t} = s_i]}{N(t)}$$

where d is the subscript that indicates the unit of observation (e.g., in the data examples in this work, the individual in the labor force, the elector[4]), and $N(t)$ is the number of units for each period (e.g., the number of individuals in the labor force or the number of electors).

The resulting dataset D has this structure:

$$D = \begin{bmatrix} w_1(1) & w_1(2) & \dots & w_1(T) \\ w_2(1) & w_2(2) & \dots & w_2(T) \\ \dots & \dots & \dots & \dots \\ w_r(1) & w_r(2) & \dots & w_r(T) \end{bmatrix} \tag{5}$$

As assumed in LJZ, the dataset D is generated from T sets of $N(t)$ independent trials.

1.3 The Bayesian Model

1.3.1 The likelihood function

We show here the derivation that needs to be followed to obtain the likelihood function specified by LJZ. We use similar notation.

4. In the electoral example, this is particularly intuitive, as what we observe, for each time t, and for each party i, is the percentage of people that voted for i.

LJZ have to write the problem using the proportions only since micro-data are not available. They use the following reasoning:

$$\begin{aligned}\Pr\{x_t = s_j\} &= \sum_i \left[\Pr\{x_t = s_j | x_{t-1} = s_i\} \cdot \Pr\{x_{t-1} = s_i\}\right] \\ &= \sum_i [p_{ij} \Pr\{x_{t-1} = x_i\}]\end{aligned}$$

Thus, one can write

$$v_j(t) = \sum_i p_{ij} v_i(t-1) \tag{6}$$

where $v_i(t)$ is the probability that the unit belongs to state i at time t.

Now, the dataset enters the picture. The structure of the data is as in (5).
Call $q_j(t)$ the constant (within each of the T sets) probability $\Pr\{x_t = s_j\}$. One obtains:

$$q_j(t) = \sum_i p_{ij} w_i(t-1) \tag{7}$$

Call $n_j(t)$ the number of units that are in state j at time t:

$$n_j(t) := N(t) w_j(t) \tag{8}$$

The distribution of $n_j(t)$ is binomial, with probability of success $q_j(t)$, and number of independent trials $N(t)$:

$$\begin{aligned}f(n_j(t)) &= \binom{N(t)}{n_j(t)} q_j(t)^{n_j(t)} [1 - q_j(t)]^{N(t) - n_j(t)} \\ &= \frac{N(t)!}{n_j(t)![N(t) - n_j(t)]!} q_j(t)^{n_j(t)} [1 - q_j(t)]^{N(t) - n_j(t)} \ \forall j \in \{1, ..., r\}\end{aligned}$$

On the other end, the joint distribution of $\boldsymbol{n}(t) = (n_j(t))_{j=1}^r$ is multinomial[5]:

$$f(\boldsymbol{n}(t)) = \binom{N(t)}{n_1(t), ..., n_r(t)} \prod_{j=1}^r q_j(t)^{n_j(t)} \tag{9}$$

$$= \frac{N(t)!}{\prod_{j=1}^r n_j(t)!} \prod_{j=1}^r q_j(t)^{n_j(t)} \tag{10}$$

Conditional on probabilities $q_{1:r}(t)$, the elements of $\boldsymbol{n}(t)$ are independent. Thus, LJZ write the distribution of the matrix $\boldsymbol{n} = [\boldsymbol{n}(1), \boldsymbol{n}(2), ..., \boldsymbol{n}(T)]$:

$$f(\boldsymbol{n}) = \prod_{t=1}^T \left[\frac{N(t)!}{\prod_{j=1}^r n_j(t)!} \prod_{j=1}^r q_j(t)^{n_j(t)} \right]$$

Then, they exploit the following facts:

5. Equation (10) in LJZ.

$$\begin{cases} \sum_j q_j(t) = 1 \ \forall t \\ \sum_j n_j(t) = N(t) \ \forall t \end{cases} \Rightarrow \begin{cases} q_r(t) = 1 - \sum_{j=1}^{r-1} q_j(t) \\ n_r(t) = N(t) - \sum_{j=1}^{r-1} n_j(t) \end{cases}$$

and obtain[6]:

$$f(\boldsymbol{n}|T) = \prod_{t=1}^{T} \frac{N(t)!}{\left[\prod_{k=1}^{r-1} n_k(t)!\right] n_r(t)!} \left[\prod_{k=1}^{r-1} q_k(t)^{n_k(t)}\right] q_r(t)^{n_r(t)} \tag{11}$$

$$= \prod_{t=1}^{T} \frac{N(t)!}{\left[\prod_{k=1}^{r-1} n_k(t)!\right] \left[N_t - \sum_{k=1}^{r-1} n_k(t)\right]!} \left[\prod_{k=1}^{r-1} q_k(t)^{n_k(t)}\right] \left(1 - \sum_{k=1}^{r-1} q_k(t)\right)^{N(t) - \sum_{k=1}^{r-1} n_k(t)} \tag{12}$$

Notice that we just obtained the likelihood, since $\boldsymbol{n} = D$. In other words, $f(\boldsymbol{n}|T) = L(D|T)$.

1.3.2 The prior

LJZ put a prior on the elements of each row of the transition matrix. The distribution is a Dirichlet[7] (or multivariate beta) for the vector $\boldsymbol{p}_i = (p_{i1:ir})$ over the simplex Δ_{r-1}[8]:

$$f(\boldsymbol{p}_i, \boldsymbol{\alpha}_i) = \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{r} p_{ij}^{\alpha_{ij}-1} \tag{13}$$

where $\boldsymbol{\alpha}$ is the vector of length r of the r hyper parameters and $B(\boldsymbol{\alpha})$ is the Generalized Beta function[9]. The choice of $\boldsymbol{\alpha}$ can be strategic in the estimation of the model. Indeed, prior information about the process can be stored in these parameters.

Consider the example of the transition probabilities in the labor market, between the states "employed" (e), "unemployed" (u), and "out of labor force" (o). The only state that, by definition, implies the willing of the individual to move from it is the unemployed one[10]. On the other hand, employed and out-of-labor force individuals are less likely to be willing to move to another state. As a result, we can use the hyperparameters to store in the model the information that e and o states tend to be more absorbing then u. Thus, we will have $\alpha_{ee} > \alpha_{eu}$, α_{eo} and $\alpha_{oo} > \alpha_{oe}, \alpha_{ou}$.

6. Equation (12) in LJZ
7. The following is equation (40) in Lee, Judge, and Zellner (1968)
8. The definition of the simplex Δ_{r-1} is:

$$\Delta_{r-1} := \left\{ (a_i)_{i=1}^{r} \in \mathbb{R}^r \text{ s.t. } \sum_{i=1}^{r} a_i = 1 \wedge a_i \in (0,1) \ \forall i \right\}$$

9. The Generalized Beta function is defined as follows:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^{r} \Gamma(a_{ij})}{\Gamma\left(\sum_{j=1}^{r} a_{ij}\right)}$$

10. OECD definition of unemployment: "The unemployed are people of working age who are without work, are available for work, and have taken specific steps to find work."

Notice that if we set $\boldsymbol{\alpha} = (1, ..., 1)$, our prior will be uniform with density $\frac{1}{B(\boldsymbol{\alpha})}$.
We will discuss later how also the sum of the elements of $\boldsymbol{\alpha}$ is informative for the prior.

From (13), LJZ exploit the relation between the Dirichlet and the Beta distributions[11] to obtain the marginal distribution[12] of each transition probability p_{ij}:

$$f(p_{ij}) = \frac{1}{B\left(\alpha_{ij}, \sum_{k \neq j} a_{ik}\right)} p_{ij}^{a_{ij}-1} (1 - p_{ij})^{\sum_{k \neq j} a_{ik} - 1} \tag{14}$$

Assumption 3 allows to write the joint prior for the entire matrix:

$$\begin{aligned} f(T, \boldsymbol{\alpha}) &= \prod_{i=1}^{r} f(\boldsymbol{p}_i, \boldsymbol{\alpha}_i) \\ &= \prod_{i=1}^{r} \left[\frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{r} p_{ij}^{\alpha_{ij}-1} \right] \end{aligned} \tag{15}$$

1.3.3 Posterior

To obtain the posterior distribution[13], LJZ combine the prior in (15) and the likelihood in (12):

$$\begin{aligned} f(T|D, \boldsymbol{\alpha}) &\propto L(T|D) \times f(T, \boldsymbol{\alpha}) \\ &\propto \left[\prod_{k=1}^{r-1} \left(\sum_{i=1}^{r} w_i(t-1) p_{ik} \right)^{N(t) w_k(t)} \right] \times \left(1 - \sum_{k=1}^{r-1} \sum_{i=1}^{r} w_i(t-1) p_{ik} \right)^{N(t) - \sum_{k=1}^{r-1} N(t) w_k(t)} \\ &\times \prod_{i=1}^{r} \left[\frac{1}{B(\boldsymbol{\alpha}_i)} \times \prod_{j=1}^{r} p_{ij}^{\alpha_{ij}-1} \right] \end{aligned} \tag{16}$$

2 Methodology

In this work, we do not use the closed-form MAP estimator proposed by LJZ. Indeed, we aim to investigate how far we can go with the study of the posterior.

In this section, we expose our methodology and we apply it to simulated data. To generate the data, we set $N(t) = N = 50 \ \forall t$, $t = 25$. The states are A and B ($r = 2$). The true transition probabilities are

$$T_{\text{True}} = \begin{bmatrix} p_{AA} = 0.7 & p_{AB} = 0.3 \\ p_{BA} = 0.4 & p_{BB} = 0.6 \end{bmatrix}$$

11. Given a stochastic vector $\boldsymbol{\theta} \in \Delta_{r-1}$ distributed as a Dirichlet with parameters $\boldsymbol{\alpha} \in (\mathbb{R}^+)^r$, the i-th component of $\boldsymbol{\theta}$ behaves as follows:

$$\theta_i \sim \text{Beta}\left(\alpha_i, \sum_{k \neq i} \alpha_k \right)$$

12. Equation (41) in LJZ.
13. Equation (47) in LJZ.

Figure [6] in the Appendix shows the trend of the proportions in the simulated data.

For this example we consider a Dirichlet($\boldsymbol{\alpha}$) as uninformative prior for each row, where $\boldsymbol{\alpha} = k\mathbf{1}$, with $\mathbf{1}$ being a vector of ones with length 2 and k a positive scalar. Changing the value of k is an interesting exercise if we want to understand how the posterior is affected by changes in the prior sample size for each row, i.e. $\sum_{i=1}^{r} \alpha_i$. Figure [2] indeed shows how the posterior pdf becomes more concentrated in space as k increases.

In order to get a numerical approximation of the posterior density we adopt a Markov Chain Monte Carlo (MCMC) methodology. We build a standard Metropolis simulator that samples from a symmetric proposal distribution for T, consisting of an independent Dirichlet($\boldsymbol{\beta}$) for each row. To induce symmetry we set $\beta_i = \beta \ \forall i = 1, ...r$. This is equivalent to having the hyper-parameter $\boldsymbol{\beta} = m\mathbf{1}$, where $\mathbf{1}$ is a vector of ones of length r, and m is a scaling factor. The value of m affects the convergence as well as the mixing of our chain (intuitively, reducing m increases the variance of each component of a Dirichlet distribution, making a larger jump to another region of the space more likely) while retaining symmetry.

In order to estimate the burn-in we resort to the running-mean plot of the Frobenius norm of each sampled transition matrix. The Frobenius norm of a matrix is defined as the square root of the sum of the absolute squares of its elements. As a rule of thumb, we discard as burn-in the first samples before the running-mean plot converges (Figure [1]).

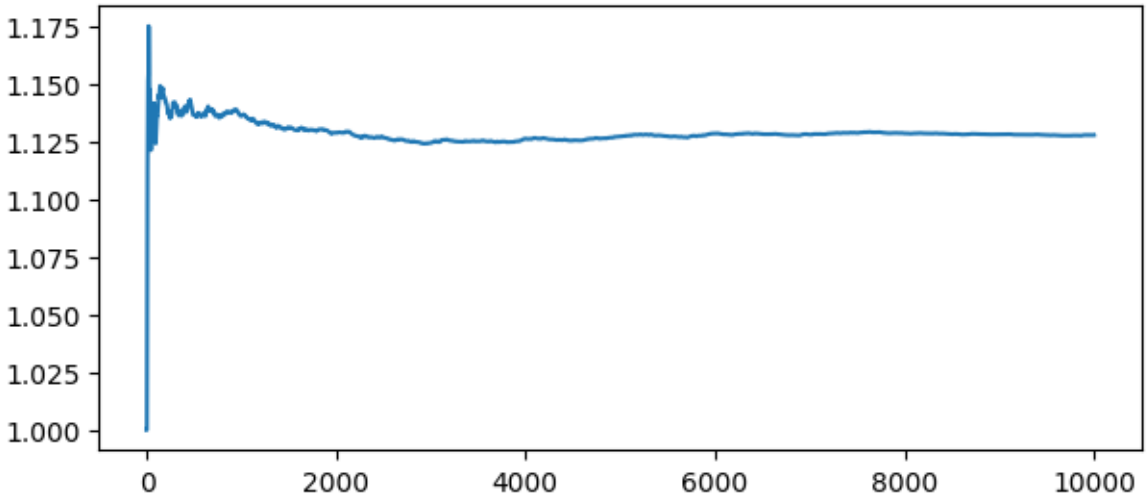


Figure 1: Frobenius Norm for the sampled observations. We discard the first 500 observations.

Once we have our sample, we can assess the quality of our sampler: we compute the generalization of the Effective Sample Size to the multivariate case as proposed by Vats, Flegal, and Jones (2019). Table [1] reports how observations are accounted for in the analysis:
We notice that the ESS is a small proportion of the number of draws. This is a limitation of our analysis and the first (of many) signal that effectively sampling from this prior can be particularly complex with this prior.

Two-State Example	
Number of draws	10,000
After Burn-in	9,500
ESS	2027

Table 1: Number of draws in the two-state example.

A thorough analysis of this metric is beyond the scope of this project, as the authors also suggest stopping rules to make the MCMC sampler more efficient. In the appendix to our paper, we provide details concerning this estimate.

Once we have discarded the burn-in, we consider two estimators for our transition matrix T. Both estimators are computed from the samples that we have just simulated: the sample posterior mean $\hat{T}_{\mathbb{E}}$, and the Maximum A Posteriori (MAP) estimator $\hat{T}_{\text{MAP}}$.

- The sample posterior mean $\hat{T}_{\mathbb{E}}$ is an unbiased and consistent estimator for the posterior expectation of T, i.e. $\mathbb{E}[T|\boldsymbol{\alpha}, D]$.
- The MAP is the value of P that maximizes the posterior pdf. We estimate MAP with the simulated (from MCMC) P that is associated with the highest value of the posterior. In other words,

$$\hat{T}_{\text{MAP}} = \arg \max_{P \in \{\boldsymbol{P}_{\text{MCMC}}\}} f(P|D, \boldsymbol{\alpha})$$

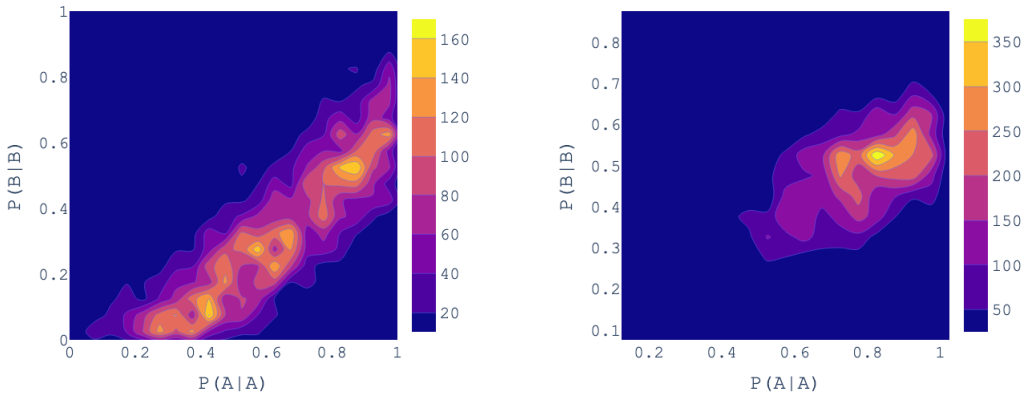
  where $\{\boldsymbol{P}_{MCMC}\}$ is the set of all the matrixes drawn from the MCMC.

Notice that LJZ also provide a way to obtain an exact MAP estimator by solving a constrained maximization problem. However our aim is to perform a broader analysis of the posterior. As a matter of fact, comparing $\hat{T}_{\mathbb{E}}$, $\hat{T}_{\text{MAP}}$, and the scatter density plot, one can grasp the behavior of the posterior distribution.

We provide here the results:

$$\hat{T}_{\mathbb{E}} = \begin{bmatrix} \hat{p}_{AA} = 0.63 & \hat{p}_{AB} = 0.37 \\ \hat{p}_{BA} = 0.66 & \hat{p}_{BB} = 0.34 \end{bmatrix}, \; \hat{T}_{\text{MAP}} = \begin{bmatrix} \hat{p}_{AA} = 0.85 & \hat{p}_{AB} = 0.15 \\ \hat{p}_{BA} = 0.49 & \hat{p}_{BB} = 0.51 \end{bmatrix} \tag{17}$$

Notice that the transition probabilities captured by the two estimators are very different. We could say that, in this case, $\hat{T}_{\text{MAP}}$ grasps better than $\hat{T}_{\mathbb{E}}$ the nature of the phenomena (comparing it with T_{True}).
We think that there are two reasons for such a difference. The first one is linked to the fact that, as caught by the multivariate ESS analysis, our sample generated through MCMC is not particularly informative. The second one, instead, is related to the nature itself of the phenomena we are trying to study. Indeed, the posterior distribution we try to approximate might have some degree of multimodality: this makes the expected value less relevant and the MAP much harder to estimate because of local maxima in the posterior pdf: from Figure [2] one can indeed notice the multimodality.



(a) $k = 1$ (b) $k = 10$

Figure 2: Scatter of the Posterior Density with parameters that differ in the multiplicative factor k.

3 Applications

We now apply this methodology to two real-life problems. In doing so, we are aware of the deep limitations of our estimators: they are going to be amplified by the fact that the phenomena we are studying could show significant persistence in states (this is true in particular for employment/unemployment conditions) or hard to fit in a Markovian framework.

3.1 Labor Force

3.1.1 The Research Question

In this application, we estimate the probabilities with which individuals of working age (15-64) move among the states of employment, unemployment, and out-of-labor-force situation.
It could give a hint about how precarious is on average the employment situation for a specific group of people, or, on the other hand, how easy is to find a job once one is unemployed. Those probabilities are relevant in the context of the Search and Matching model (Mortensen and Pissarides 1994).
Estimating the probability of going from unemployment to employment allows us to provide an estimate of the average duration of unemployment[14]:

$$D_u = \sum_{k=o}^{\infty} p_{uu}^k = \frac{1}{1 - p_{uu}} = \frac{1}{p_{ue} + p_{uo}}$$

where

$$p_{uu} = \Pr\{\text{being unemployed}|\text{being unemployed in the previous period}\}$$

$$p_{ue} = \Pr\{\text{being employed}|\text{being unemployed in the previous period}\}$$

$$p_{uo} = \Pr\{\text{being out of labor force}|\text{being unemployed in the previous period}\}$$

14. All these considerations ca be done assuming $T_t = T \ \forall t$

This research is of particular interest when the probabilities are estimated for different groups of people. For example, one may want to understand whether the employment situation of college graduates is less precarious than the one of lower-educated individuals. Depalo and Lattanzio (2023) estimates the transition probabilities among the employed individuals between different types of contracts (e.g., full-time, part-time, open-ended, fixed term, etc.).

Usually, economists use panel data, and thus they have more information than the one that we use in our dataset. For example, the Current Population Survey is a rich labor market dataset for the US labor force (Madrian and Lefgren 1999), and its main function, according to the Bureau of Labor Statistics Handbook of Methods[15], is indeed to "classify the sample population into three basic economic groups: the employed, the unemployed, and those not in the labor force". However, to test the dataset on this application, we employ only the Labor Force Surveys cross-sectional data[16], and thus only the proportions of people in each state.

3.1.2 The Stochastic Process

The process for which we want to find the transition matrix is the one of a person of working age (15-64, from now on, just individuals) throughout employment, non-employment, and out-of-labor force conditions.
We now discuss how the three assumptions presented in subsection 1.1 specialize in this case, and which are the circumstances under which those may be violated.
Assumption 1: all the individuals x that at time $t-1$ were in situation $i \in \{e, u, o\}$ have the same probability of being in situation j at time t.

We realize that this assumption is quite strong. Indeed, for example, the individual probability of job separation (i.e., p_{eu}) depends on the type of contract (e.g., open-ended *vs* fixed-term), and this influences the probability that an employed person in $t-1$ was employed in $t-2$. For this reason, a more precise analysis could consider different types of contracts, and apply the same assumption to more granular classifications of individuals.

Assumption 2: the probability of becoming unemployed (p_{eu} and p_{oe}), exit from the labor force (p_{uo} and p_{eo}), finding a job (p_{ue} and p_{oe}) do not vary throughout the years.

The strength of this assumption depends on the time horizon one considers. On one side, the longer the time horizon, the less plausible the assumption. However, one should also avoid structural breaks. For example, we select the time horizon of our data to avoid the 2008 Financial Crisis, the Sovereign Debt Crisis, and the COVID Recession.
Assumption 3: the probability that an individual x goes from state i at time $t-1$ to state j at time t is independent from the probability that an individual y goes from state k at time $t-1$ to state z at time t.

This assumption is quite strong and probably often violated in real life. For example, when unemployed people are hired, some of the individuals out-of-the-labor force may begin looking for a job and then become unemployed.

15. Cited in Madrian and Lefgren (1999)
16. The Labor Force Survey available online includes the answer to the question about the employment situation one year before the interview. Thus, we have a panel component in this dataset. In the Appendix, we plot the transition probabilities that we have estimated using all the information available.

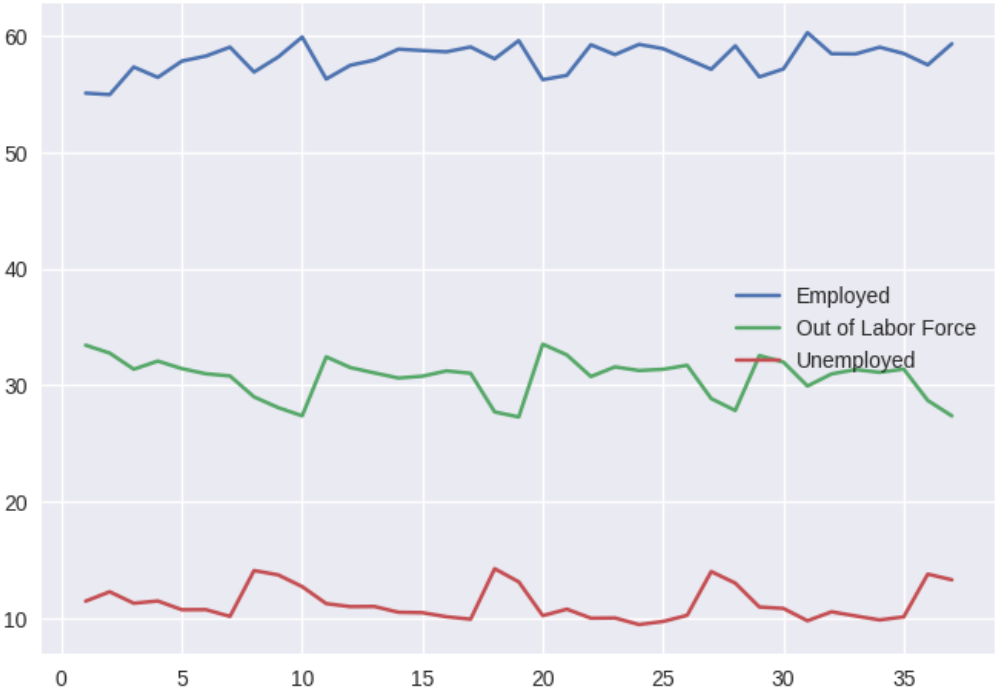


Figure 3: Proportion of Employed, Unemployed, and Out of Labor Force over total sample, by year.

3.1.3 Dataset

We merge ISTAT Labor Force Surveys[17] from 2014 (First-Quarter) to 2020 (First-Quarter). We keep observations related to individuals in the working age.
Notice that we may have absorbing-states situations (e.g., people who are not able to work because invalid, for example), but we are not able to better refine our dataset to account for those situations.

In particular, we observe the question in which respondents are asked how they perceive their employment situation.
Respondents can select only one of the following options: employed, unemployed, retired from labor, student, homemaker, or other.
We drop observations with "other" answers, and we classify individuals that answer "retired from labor", "student", and "homemaker" as "out of labor force".[18]
Figure [3] visualizes the evolution of means (over quarters) proportions of individuals in each of the three states we are considering. These are the only data that we use as input for our estimator.

3.1.4 Methodology

We perform the same analysis as before. In the simulated data, we had $N(t) = 50$ and time-invariant. For the real-world applications of this paper, we allow $N(t)$ to change

17. Data can be downloaded from the ISTAT webpage.
18. Further analysis could check the robustness of this classification. For example, the fact that students usually enter the labor force (and rarely happens the inverse) may be a good reason to drop those observations).

over time but we rescale the original count dataset in such a way that $N(t)$ is of order 10^2. This is necessary to keep the computations of the posterior feasible, both in terms of execution time and precision (the unnormalized posterior pdf can reach extremely small orders of magnitude very quickly). This approach introduces some discretization error in computing the relative proportions $w_i(t)$ when moving from the original to the rescaled dataset. We do not deem this to significantly influence our estimation process.

We are willing to estimate the transition matrix

$$T_{\text{labor}} = \begin{bmatrix} p_{ee} & p_{eo} & p_{eu} \\ p_{oe} & p_{oo} & p_{ou} \\ p_{ue} & p_{uo} & p_{uu} \end{bmatrix}$$

where e stands for employment, u for unemployment, and o for out-of-labor force.
We set the following hyperparameters:

$$\boldsymbol{\alpha} = \begin{bmatrix} 0.90 & 0.05 & 0.05 \\ 0.05 & 0.90 & 0.05 \\ 0.20 & 0.05 & 0.75 \end{bmatrix}$$

Notice that this prior is quite informative, as it considers all three states quite persistent (i.e., high p_{ii} for each state). In particular, employment and out-of-labor force are more persistent. This is consistent with the definition of the three states, and with the panel-data-estimated transition probabilities (e.g., see Figure[7])[19]
Notice also that $\sum_{j=1}^{r} \alpha_{ij} = 1 \ \forall i$: the degree at which prior information matters in the posterior is at the same level as in Figure [2a] for simulated data.

We generate 10.000 samples through our MCMC sampler and we discard the first 500 (we plot the Frobenius Norm in Figure [8] in the Appendix).
The multivariate Effective Sample Size is particularly low in this case (218).

Labor Economics Application	
Number of draws	10,000
After Burn-in	9,500
ESS	218

Table 2: Number of draws in the labor example.

3.1.5 Results

The two estimators give the following results:

$$\hat{T}_{\mathbb{E}} = \begin{bmatrix} 0.64 & 0.27 & 0.09 \\ 0.46 & 0.33 & 0.21 \\ 0.36 & 0.25 & 0.39 \end{bmatrix}, \ \hat{T}_{\text{MAP}} = \begin{bmatrix} 0.75 & 0.19 & 0.06 \\ 0.51 & 0.41 & 0.08 \\ 0.03 & 0.01 & 0.96 \end{bmatrix} \tag{18}$$

19. Notice that we are not using the panel dimension of the Labor Force Survey dataset to build our prior (since we are ignoring the panel dimension): we just use it to show that our prior choice is reasonable).

Very high-levels economic conclusions may be derived from these estimates. Methodological limitations presented for the two-state case are exacerbated by the high degree of stickiness of this phenomenon.

3.2 Electoral Flows

3.2.1 The Research Question

In this application, we estimate the probabilities with which voters change their voting preferences from one political area to another. For example, the probability with which a voter votes for the right parties if she voted for the left parties in the previous election. This question would be of interest to a political economist and to a political scientist. We estimate this transition matrix for the Italian voters from 1999 to 2022.

These probabilities can help the understanding of the degree of stability of voting choices. As in the labor economics application, transition probabilities of different groups can be compared (e.g., populist supporters Voogd and Dassonneville (2018)). Also, one can investigate whether the transition probabilities are correlated to economic indicators (e.g., a similar research question has been addressed by Dassoneville and Hooghe (2017)).

3.2.2 The Stochastic Process

The process for which we want to find the transition matrix is the one of a voter choosing which party to vote for in each election.
We now discuss how the three assumptions presented in subsection 1.1 specialize in this case, and which are the circumstances under which those may be violated.

Assumption 1: all the voters x that at time $t-1$ voted for party i has the same probability of voting for party j at time t, notwithstanding previous choices, $\forall i, j \in (1, r)$.

A reasonable violation may be the positive relationship of p_i (i.e., the probability of voting for party i) wuth the number of times one has voted for i.

Assumption 2: the probability that the elector x switch from party i (voted in $t-1$) to party j voted in t is constant in each election.

This assumption is particularly strong as it is reasonable to think that there are periods of political stagnation when the probabilities of non-switching (i.e., p_{ii}) increase, and periods of higher volatility when the p_{ii}s decrease. It is for this reason that our analysis should be applied to periods without structural breaks in the political scenario[20]

Assumption 3: the probability with which a voter x switches from party i to party j is independent from the probability with which a voter y switches from party z to party k.

This assumption may be violated if there are parties that attract voters from all coalitions.

While we understand that the assumptions are quite strong, we are still convinced that estimating a transition matrix can be a good way to grasp the mechanism with which an average voter switches across parties.

20. It is for this reason that our dataset includes only elections after the 1991 Tangentopoli scandal. However, we do not exclude that other occurrences (e.g., the creation of the M5S party) may have structurally altered the voting choice mechanism.

3.2.3 The Dataset

We use results of Italian political elections[21]. During normal times, Italian citizens are supposed to participate in four different types of elections: the elections for the (a) upper ("Senato della Repubblica") and (b) lower house ("Camera dei Deputati"), (c) European Parliament, and (d) municipalities' and regional administrations. (a) and (b) always occur on the same day.
However, in this work, we only consider (b) and (c). Indeed, they are the only elections in which all electors can cast their votes. As a matter of fact, "Senato" elections involve people older than 25, and regional and local elections do not involve all Italians.
Also, we consider elections after 1999. This choice is because the Italian political system witnessed a deep reassessment in 1992. Thus, we did not want to include such structural change in our observations. As a result, our dataset includes the outcomes of "Camera" elections in 2001, 2006, 2008, 2013, 2018, 2022 and the outcomes of European Elections in 1999, 2004, 2009, 2019.
One can easily notice that the time intervals are not uniform between elections. Both European and "Camera" elections occur every 5 years, but "Camera" elections can be called before the end of the legislature (e.g., Italy had "Camera" elections in 2006 and 2008).

First, we need to organize parties in political areas, since the set of existing parties is not constant. We group parties in three areas: left, right, and other. The left area includes parties that are usually identified with left and center-left positions (e.g., Rifondazione Comunista, Partito Democratico, La Margherita, etc.); the same goes for the right area with right and center-right positions (e.g., Lega, Fratelli d'Italia, Alleanza Nazionale, Forza Italia).
The "other" category includes the parties that cannot be categorized either in "Left" or in "Right". This category is probably too wide, as, for example, it includes "Scelta Civica", "Italia Viva", and "Movimento 5 Stelle", which are parties whose constituencies are far from each other. To address this limitation one could refer to the Chapel Hill Expert Survey (see Jolly et al. 2022), which categorizes the parties' positions for many countries (including Italy).

Figure [4] visualizes the trend of proportions for each election year.

Notice that, while there exist panel surveys that allow the researchers to estimate transition probabilities, what is interesting in this application is the large availability of data. Indeed, what one needs to perform this analysis is just the electoral results (that are easily available for each election, for each country).

3.2.4 Methodology

We perform the same analysis as before. The same considerations about $N(t)$ are valid here. However, notice that the rescaling in this case is 100 times bigger since the number of units (i.e., voters) in this case is 10^6.

We are willing to estimate the transition matrix

21. Data are from Eligendo, the platform of the Interior Ministry which collects the results of the elections.

Figure 4: Proportion of Voters for Right, Left and other parties, over total sample, by year.

$$T_{\text{politics}} = \begin{bmatrix} p_{rr} & p_{ro} & p_{rl} \\ p_{or} & p_{oo} & p_{ol} \\ p_{lr} & p_{lo} & p_{ll} \end{bmatrix}$$

where r stands for right, o for other, and l for left. We set the following hyperparameters:

$$\boldsymbol{\alpha} = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

As for the labor case, we use an informative prior. We assume that the probability of persistence is higher for right and left parties. This is reasonable since the "other" category includes several parties, and its composition evolves throughout the time series.

We generate 50.000 samples through our MCMC sampler and we discard the first 3000 (we plot the Frobenius Norm in Figure [9] in the Appendix). The multivariate ESS is 1988. Table [3] reports how the observations are accounted for.

3.2.5 Results

The two estimators give the following results:

$$\hat{T}_{\mathbb{E}} = \begin{bmatrix} 0.34 & 0.33 & 0.33 \\ 0.33 & 0.34 & 0.33 \\ 0.36 & 0.32 & 0.32 \end{bmatrix}, \quad \hat{T}_{\text{MAP}} = \begin{bmatrix} 0.32 & 0.04 & 0.64 \\ 0.23 & 0.67 & 0.10 \\ 0 & 0.64 & 0.36 \end{bmatrix} \tag{19}$$

Political Economics Application	
Number of draws	50,000
After Burn-in	47,000
ESS	1,988

Table 3: Number of draws in the politics example.

Again, $\hat{T}_{\mathbb{E}}$ and $\hat{T}_{\text{MAP}}$ differ a lot, but $\hat{T}$MAP is more informative, since $\hat{T}_{\mathbb{E}}$ is basically uniform.
High-level considerations may be drawn. For example, one can observe that the higher persistence is found for the Right coalitions. Also, the higher probability $p_{\text{Left}\to\text{Other}}$ may be explained by the voters going from the Left parties to the Five Star Movement (categorized as other).

4 Conclusion

In this paper, we deal with the estimation transition matrices when only aggregate counts for each state are available, without knowing individual transitions or having data in panel form. After a brief review of the relevant theory of Markov chain we highlight the main assumptions on the building blocks of our Bayesian model. We specify a likelihood that embodies the Markovian time dependency of state proportions and pair it with a prior distribution. We follow LJZ by naturally modeling each row of the transition matrix with a prior Dirichlet distribution, where the hyperparameter allows us to incorporate prior beliefs as well as the prior sample size. The posterior pdf is derived up to a normalizing constant by employing the Bayes theorem in proportional form.
Our estimation process revolves around an MCMC approximation of the posterior via Metropolis algorithm. MCMC diagnostic is carried out through running-mean plots and the multivariate ESS. Point estimates of the transition probabilities are computed as the sample mean and MAP of the simulated chain.
We apply this procedure to simulated data first, and then to labor and electoral flows data. The limitations of this approach become apparent when checking the robustness of our estimators. The MCMC approximation is in general poor (as signaled by the low multivariate ESS) and the computational complexity of the sampling process increases non-linearly with the number of states. Moreover, sample estimators can be influenced by setting-specific features, such as multimodality in the posterior distribution. A potential way out of this scenario is the MAP estimator that LJZ find as a solution to a constrained optimization problem, even if this comes at the cost of very limited information about the posterior distribution.

References

Dassoneville, R., and M. Hooghe. 2017. "Economic indicators and electoral volatility: economic effects on electoral volatility in Western Europe, 1950-2013." *Comparative European Politics* (15): 919–943. https://doi.org/https://doi.org/10.1057/cep.2015.3.

Depalo, D., and S. Lattanzio. 2023. "The increase in earnings inequality and volatility in Italy: the role and persistence of atypical constracts." *Questioni di Economia e Finanza* (801).

Flegal, J. M., and G. L. Jones. 2010. "Batch Means and Spectral Variance Estimators in Markov Chain Monte Carlo." *The Annals of Statistics* 38 (2). https://doi.org/https://doi.org/10.1214/09-AOS735.

Jolly, S., R. Bakker, L. Hooghe, G. Marks, J. Polf, J. Rovny, M. Steenbergen, and M. A. Vachudova. 2022. "Chapel Hill Expert Survey trend file, 1999-2019." *Electoral Studies* (75). https://doi.org/https://doi.org/10.1016/j.elecstud.2021.102420.

Lee, T. C., G. G. Judge, and A. Zellner. 1968. "Maximum Likelihood and Bayesian Estimation of Transition Probabilities." *Journal of the American Statistical Association* 63 (324). ISSN: 1537274X. https://doi.org/10.1080/01621459.1968.10480918.

Madrian, B. C., and L. J. Lefgren. 1999. "A note on longitudinally matching current population survey (CPS) respondents." *NBER Technical Working Paper* (247).

Morita, S., P. F. Thall, and P. Muller. 2008. "Determining the Effective Sample Size of a Parametric Prior." *Biometrics* (64). https://doi.org/https://doi.org/10.1111/j.1541-0420.2007.00888.x.

Mortensen, D. T., and C. A. Pissarides. 1994. "Job Creation and Job Destruction in the Theory of Unemployment." *The Review of Economic Studies* 61 (3): 397–415. https://doi.org/https://doi.org/10.2307/2297896.

Vats, D., J. M. Flegal, and G. L. Jones. 2019. "Multivariate output analysis for Markov chain Monte Carlo." *Biometrika* 106 (2). https://doi.org/https://doi.org/10.1093/biomet/asz002.

Voogd, R., and R. Dassonneville. 2018. "Are the supporters of populist parties loyal voters? Dissatisfaction and stable voting for populist parties." *Government & Opposition* (55): 349–370. https://doi.org/https://doi.org/10.1017/gov.2018.24.

Appendix

Multivariate Effective Sample Size

The Effective Sample Size is a metric that allows us to understand the amount of information contained in the sample that we have drawn from the MCMC sampler (Morita, Thall, and Muller 2008). Indeed, it is the number of independent draws that has the same amount of information as our sample.

In this work, we need an expression for the ESS in the multivariate case: the following discussion about the Multivariate Effective Sample Size closely follows Vats, Flegal, and Jones (2019) (from now on, VFJ).
The multivariate case applies to a vector of parameters. In this case, however, we have a $r \times r$ matrix T of parameters, where r is the number of states.
Moreover, since for each row i of the matrix T we have $\sum_j p_{ij} = 1$, we need to estimate $r - 1$ parameters[22].
Thus, we define a vector $\bar{T}$ of length $s = r(r-1)$ that is built as follows:

$$\bar{T} = [p_{11}\ p_{12}\ \dots\ p_{1(r-1)}\ p_{21}\ p_{22}\ \dots\ p_{2(r-1)}\ \dots\ p_{r(r-1)}]^T \tag{20}$$

In other words, we cut the last column of the matrix, and we open out the other elements.

To estimate ESS we first need to group the samples into b_n batches of batch size a_n (chosen such that $b_n \times a_n = n$). For example, the first batch contains the first a_n samples after burn-in.
Batches are needed to compute the multivariate batch means estimator (mBM) defined as:

$$\hat{\Sigma} = \Sigma_n = \frac{b_n}{a_n - 1} \sum_{k=1}^{a_n} \left(\bar{Y}_k - \theta_n\right)\left(\bar{Y}_k - \theta_n\right)^T$$

where θ_n is the sample mean over the entire set of samples. Under suitable conditions, Σ_n is a strongly consistent estimator for Σ, which represents the asymptotic covariance matrix in the Markov chain Central Limit Theorem as $n \to \infty$:

$$\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N_p(0, \Sigma).$$

VFJ propose the following expression for the multivariate ESS if Λ and Σ are full-rank:

$$\text{ESS} = n \left[\frac{\det(\Lambda)}{\det(\Sigma)}\right]^{\frac{1}{p}} \tag{21}$$

where Λ is estimated with the sample covariance matrix of the sample generated through the MCMC and Σ with Σ_n. Notice that if $\Sigma = \Lambda$ (i.e., uncorrelated samples), ESS $= n$.

A non-trivial problem underlying the multivariate ESS concerns the "optimal" batch size for the computation of the multivariate batch means estimator. If b_n is too high, the

22. We proceed with the vector $\bar{T}$ because considering the "full" vec(T) would obviously lead to singular matrices and to an undefined Multivariate ESS estimator.

number of elements a_n of each batch will not allow estimating a"variance-covariance"[23] matrix precisely. On the other hand, if b_n is too low, the arithmetic mean will be over few elements.

As a rule of thumb, we set the batch size to $b = \lfloor n^{\frac{1}{2}} \rfloor$ (Flegal and Jones 2010), where n is the number of MCMC samples (post-burn-in, in our case). Then we proceed to perform some sensitivity analysis around this value (Figure [5]): if we want to adopt a conservative approach we should consider the lowest possible value of multivariate ESS.

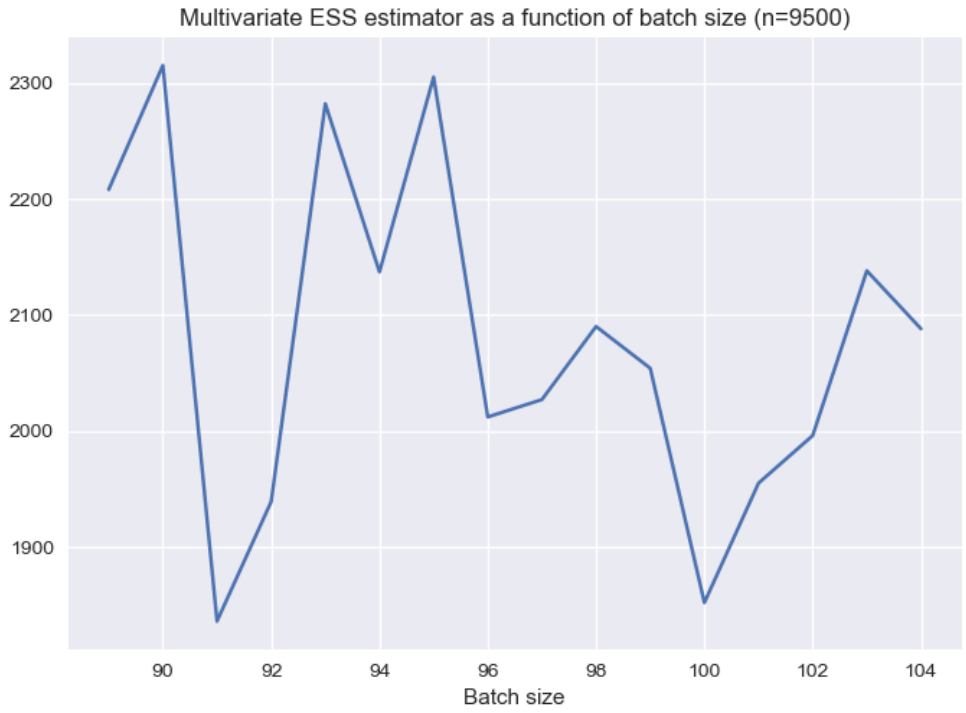


Figure 5: Multivariate ESS estimator as a function of batch size 9500

23. We use the "" sign since it is not exactly a variance-covariance matrix, as explained above.

Figures

Figure 6: Proportions from the Simulated Data used in Section (2).

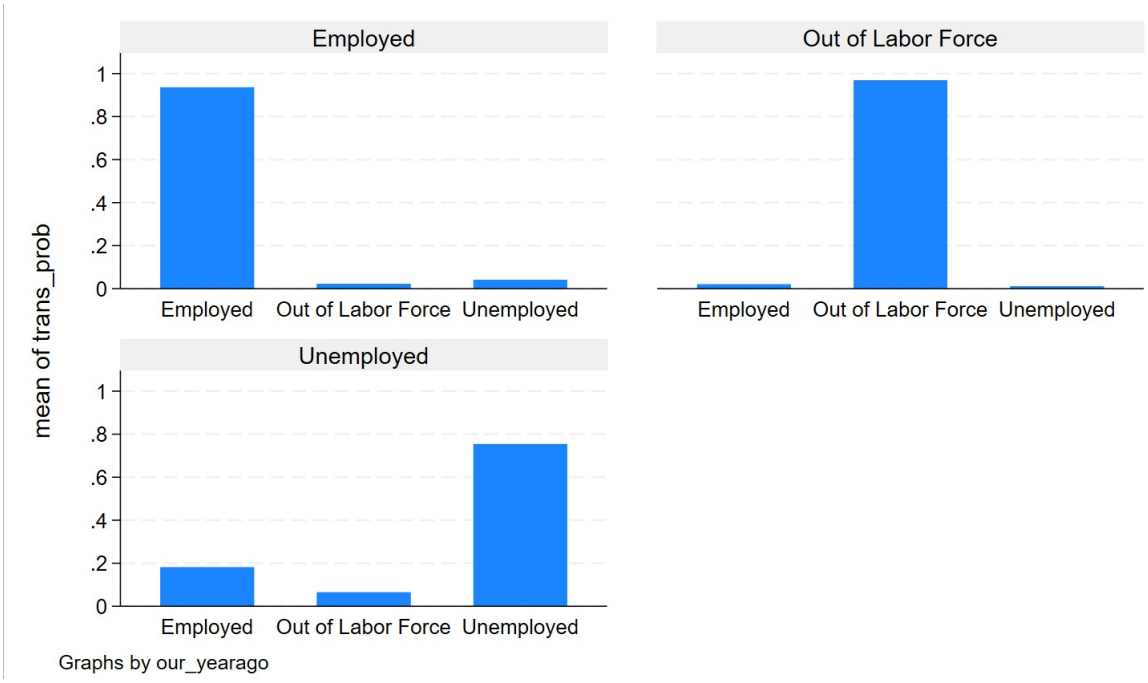


Figure 7: Mean of transition probabilities, estimated from the Labor Force Survey (I sem. 2014 - IV sem. 2022), exploiting the question about the perceived employment situation one year before the survey.

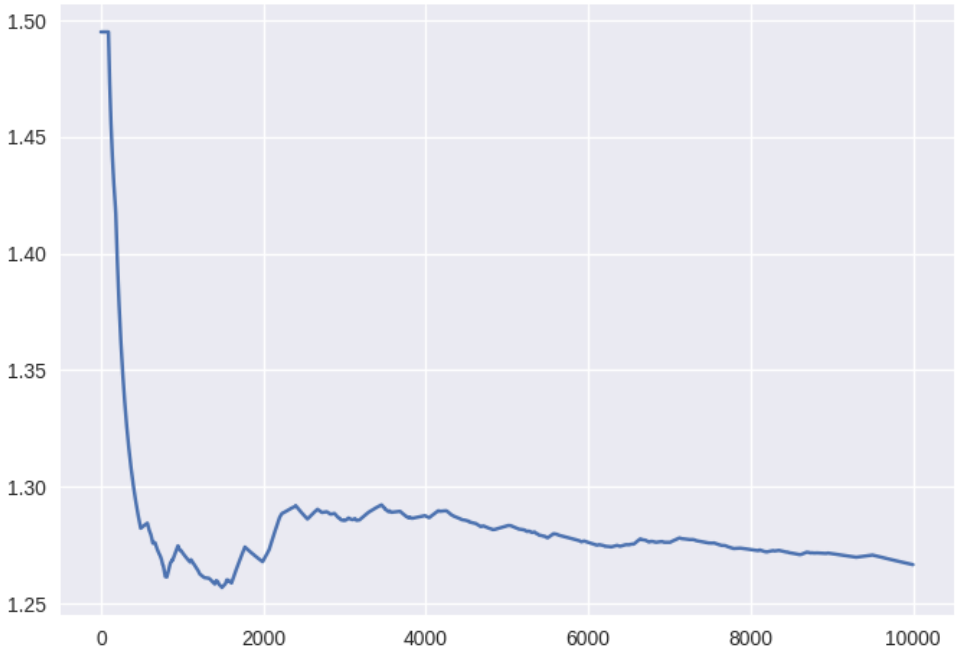


Figure 8: Frobenius Norm for the sampled observations in the Labor Economics Application.

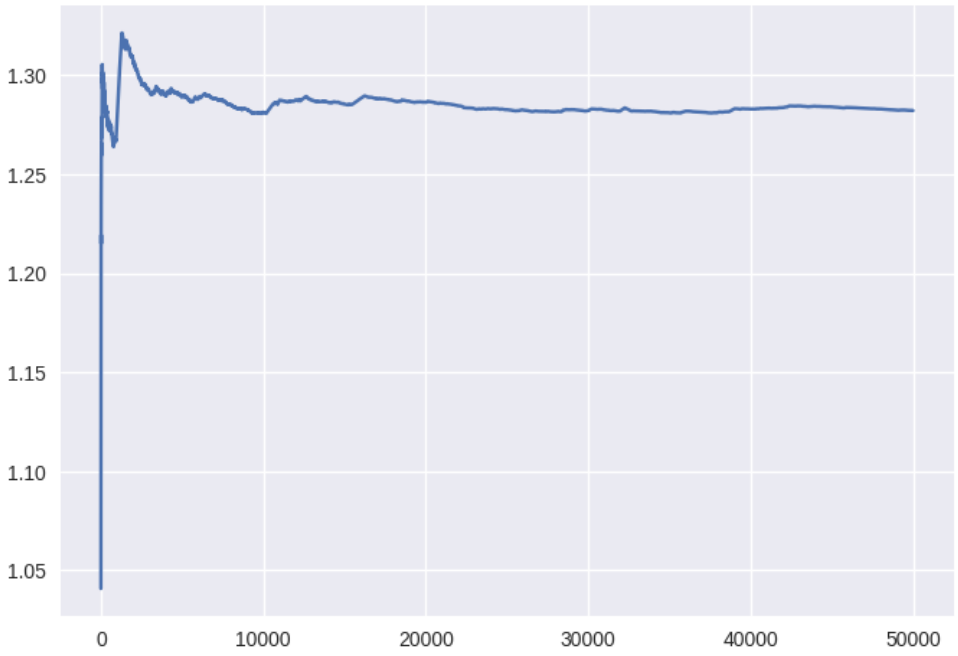


Figure 9: Frobenius Norm for the sampled observations in the Political Economics Application