Ewa Genge

Uniwersytet Ekonomiczny w Katowicach

ZASTOSOWANIE UKRYTYCH MODELI MARKOWA W ANALIZIE OSZCZĘDNOŚCI WŚRÓD POLAKÓW

Wprowadzenie

Modele mieszanek, których składowe są charakteryzowane przez rozkłady prawdopodobieństw (tzw. rozkłady składowe mieszanki) od dawna znajdują swoje zastosowanie w taksonomii. Modele mieszanek rozkładów dla zmiennych jakościowych (mierzonych na skalach słabych), zwane są również modelami (*latent class models*) lub analizą klas ukrytych (*latent class analysis*).

W ostatnim czasie coraz bardziej na popularności zyskują modele klas ukrytych dla danych panelowych czy szeregów czasowych, gdzie celem jest już nie tylko podział obserwacji na homogeniczne grupy, ale również pewna analiza zmian w czasie *t*. W tym przypadku są stosowane ukryte modele Markowa (*latent Markov model*), które bardzo często wykorzystywane są w naukach społecznych. Modele te są stosowane również w psychologii do modelowania procesów uczenia się [zob. np. (Wickens, 1982; Schmittmann et al. 2006, s. 2079-2091)]. W ekonomii modele te zwane są również modelami o zmiennych reżimach [zob. np. (Kim, 1994, s. 1-22; Ghysels, 1994, s. 289-298)]. Wśród innych zastosowań należy również wymienić rozpoznawanie mowy (Rabiner, 1989, s. 267-295) oraz różnego rodzaju badania genetyczne [zob. np. (Krogh, 1998, s. 45-63)]. W tego rodzaju aplikacjach modele te są nazywane modelami ukrytego łańcucha Markowa (*hidden Markov models*). W modelach LMM są najczęściej analizowane krótkie, wielowymiarowe szeregi czasowe o dużej liczbie obserwacji (dane panelowe), natomiast modele HMM są stosowane głównie do długich, jednowymiarowych szeregów czasowych pojedynczych procesów lub jednostek. Bardziej szczegółowe informacje na temat tego rodzaju modeli można znaleźć w pracach Zucchini i Mac Donald (2009), a także Frühwirth-Schnatter (2006), jak również Cappe, Moulines i Ryden (2005).

W literaturze (Visser i Speekenbrink, 2010, s. 1-2) można również spotkać się z terminem zależne modele mieszanek (*dependent mixture model*) na określenie zarówno modeli LMM, jak i HMM.

1. Ukryty model Markowa − definicja

W ukrytym modelu Markowa badana jest funkcja gęstości wielowymiarowego szeregu czasowego $f(\mathbf{Y}_t)$, w którym ukryta struktura przejścia jest zdefiniowana za pomocą procesu Markowa. W modelu tym dyskretna zmienna losowa X_t nie jest bezpośrednio mierzalna, a stany łańcucha nazywa się ukrytymi.

Ukryty model Markowa można zapisać jako:

$$f(\mathbf{Y}_t) = \sum_{X_1=1}^{u} \cdots \sum_{X_T=1}^{u} f(X_1) \prod_{t=2}^{T} f(X_t \mid X_{t-1}) \prod_{t=1}^{T} f(\mathbf{Y}_t \mid X_t), \tag{1}$$

gdzie:

$f(X_1)$ – funkcja gęstości rozkładu początkowego,

$f(X_t \mid X_{t-1})$ – prawdopodobieństwo przejścia, które określa prawdopodobieństwo bycia w ukrytym stanie w czasie t, pod warunkiem bycia w tym stanie w czasie $t - 1$. Ukryta macierz przejścia $\mathbf{A}$ o elementach a_{sr} oznacza prawdopodobieństwo przejścia z ukrytego stanu s do stanu r, tj. $a_{sr} = P(X_t = r \mid X_{t-1} = s)$:

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1u} \\ \vdots & \ddots & \vdots \\ a_{u1} & \cdots & a_{uu} \end{bmatrix} \tag{2}$$

Suma prawdopodobieństw w każdym wierszu macierzy $\mathbf{A}$ jest równa jeden.

$f(\mathbf{Y}_t \mid X_t)$ – funkcja gęstości rozkładu wielowymiarowego.

Ukryty model Markowa opiera się na dwóch głównych założeniach:

Ciąg ukrytych stanów jest zgodny z łańcuchem Markowa, tj. X_t jest zależny jedynie od stanu poprzedniego (wystąpienie każdego kolejnego stanu ukrytego łańcucha Markowa zależy wyłącznie od jego poprzednika).

Obserwacje w każdym czasie są niezależne pod warunkiem znajomości ukrytego stanu X_t. Oznacza to, że obserwacja w czasie t zależy tylko od ukrytego stanu w czasie t, co bardzo często odnosi się do założenia o lokalnej niezależności, która jest głównym założeniem całej grupy modeli ze zmiennymi ukrytymi.

2. Estymacja parametrów i wybór modelu

Estymacja ukrytych modeli Markowa sprowadza się m.in. do oszacowaniu parametrów rozkładów składowych, jak również parametrów macierzy ukrytych prawdopodobieństw przejść. Parametry wspomnianych modeli szacowane są za pomocą funkcji największej wiarygodności. Popularną metodą szacowania parametrów największej wiarygodności jest algorytm EM (Dempster et al., 1997, s. 1-38) lub zmodyfikowana wersja tego algorytmu, tj. algorytm Bauma−Welcha (Baum, Petrie, 1966, s. 1554-1563), wykorzystywana w pakiecie depmixS4. W algorytmach tych jako brakujące dane traktuje się informację o przynależności obiektów do poszczególnych stanów*. W kroku E są wyznaczane oceny parametrów ukrytego modelu Markowa, wykorzystując prawdopodobieństwa a posteriori otrzymane w poprzedniej iteracji. W kroku M, na podstawie uzyskanych ocen parametrów (w kroku E), są wyznaczane nowe wartości prawdopodobieństw a posteriori. Procedurę iteracyjną kontynuuje się do momentu, gdy prawdopodobieństwa a posteriori otrzymane w kolejnych dwóch iteracjach będą różnić się nieznacznie, np. dla $\varepsilon = 0{,}01$.

W modelach klas ukrytych na początku sprawdza się dopasowanie dla liczby stanów równej jeden. W kolejnych krokach zwiększa się liczbę stanów o jeden, tak długo aż model osiągnie najlepsze dopasowanie. Należy jednak pamiętać, że wraz z dodatkową liczbą stanów, wzrasta liczba szacowanych parametrów modelu. Dlatego są najczęściej wykorzystywane kryteria informacyjne, będące wyrazem kompromisu pomiędzy jakością dopasowania a złożonością modelu. Do najbardziej popularnych kryteriów informacyjnych zaliczane są: Bayesowskie kryterium informacyjne Schwarza BIC (*Bayesian Information Criterion*), kryterium informacyjne Akaike AIC (*Akaike Information Criterion*). Kryteria te mogą dawać niejednoznaczne wskazania co do oceny modeli klas ukrytych.

Porównania różnych kryteriów informacyjnych można znaleźć m.in. w pracach McLalchlan i Peel (2000, s. 81-116), Biernacki et al. (1999, s. 49-71), Bozdogan (2000, s. 62-91), Witek (2011, s. 191-197). W części empirycznej pracy wykorzystano dwa najbardziej popularne kryteria, tj. BIC oraz AIC. Kryteria te są stosowane w celach porównawczych modeli o różnej liczbie klas.

* W ukrytych modelach Markowa liczba stanów i przynależność obiektów początkowo jest nieznana, dlatego też informacje te są traktowane jako brakujące dane. Zastosowanie znajdują tu dlatego algorytmy, tj. EM, czy algorytm Bauma−Welcha.

3. Analiza empiryczna

Analizę klas ukrytych przeprowadzono na podstawie danych panelowych pochodzących z Polskiego Generalnego Sondażu Społecznego (2013). W niniejszym artykule rozważano dane z dwóch ostatnio opublikowanych lat, tj. z roku 2009 oraz 2011. Analiza została przeprowadzona z uwzględnieniem siedmiu zmiennych i z pominięciem odpowiedzi „nie wiem". Badana próba liczyła 1509 obserwacji w każdym z analizowanych lat (3018 łącznie).

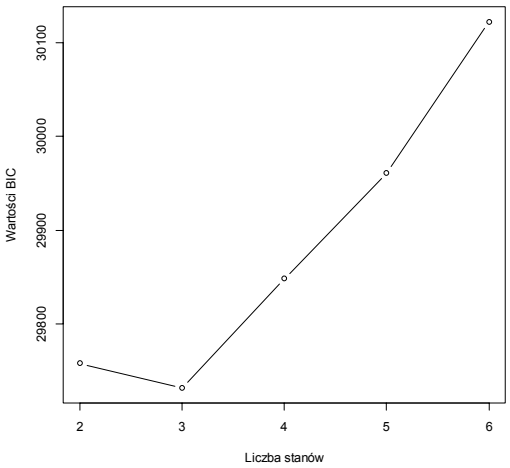
W przykładzie wykorzystano siedem zmiennych $Y_1, \ldots, Y_7$. W nawiasie podano oryginalne nazwy ze zbioru udostępnianego na stronie internetowej, litera *e* poprzedzająca symbol zmiennej (np. a5) dotyczy badania w roku 2009, zaś litera *f* – badania w roku 2011.

1. Y_1 (*ep101_1, fp98_1, gp_104_1*): Źródło utrzymania gospodarstwa (1 − pracowników; 2 − rolników; 3 − pracujących na własny rachunek; 4 − emerytów; 5 − rencistów; 6 − utrzymujących się z niezarobkowych źródeł innych niż emerytura czy renta; 7 − kilka równie ważnych źródeł utrzymania).
2. Y_2 (*e*f2, *f*F2): Łączna wartość posiadanych przez gospodarstwo domowe oszczędności (1 − do wysokości miesięcznych dochodów gospodarstwa; 2 − powyżej miesięcznych do 3 miesięcznych dochodów; 3 − powyżej 3 miesięcznych-do półrocznych dochodów; 4 − powyżej półrocznych do rocznych dochodów; 5 − powyżej rocznych dochodów gospodarstwa domowego, 6 − trudno powiedzieć).
3. Y_3 (*e*f3_1, *f*F3_1): Lokaty w bankach w złotych (1 − tak; 2 − nie).
4. Y_4 (*e*f3_2, *f*F3_2): Lokaty w bankach w walutach obcych (1 − tak; 2 − nie).
5. Y_5 (*e*f3_3, *f*F3_3): Obligacje (1 − tak; 2 − nie).
6. Y_6 (*e*f3_6, *f*F3_6): Inwestycje w papierach wartościowych (1 − tak; 2 − nie).
7. Y_7 (*e*f_10, *f*F10): Sytuacja materialna gospodarstwa domowego (1 − pogorszyła się; 2 − poprawiła się; 3 − nie zmieniła się)*.

W badaniach wykorzystano pakiet depmixS4 programu **R**.

Chcąc wybrać optymalną ukrytą liczbę stanów, obliczono wartości kryteriów informacyjnych AIC oraz BIC. Kryterium BIC wskazało minimalną wartość dla liczby stanów równej trzy (rysunek 1). Niewiele większą wartość otrzymano dla czterech stanów, w przypadku kryterium AIC. W takich sytuacjach często wybierane są modele mniej złożone (Collins i Lanza, 2011, s. 100-103), dlatego też w dalszej części pracy analizowano model o trzech stanach ukrytych.

* Zmienne $Y_3 - Y_6$ dotyczą pytań o posiadanie np. lokat bankowych czy obligacji. Zmienna Y_7 odnosi się do pytania o sytuację materialną gospodarstwa domowego, w porównaniu do sytuacji sprzed dwóch lat.



Rys. 1. Wartości kryterium BIC dla różnej liczby stanów

Następnie szacowano ukryty model Markowa dla zmiennych $Y_1 - Y_7$.

W stanie pierwszym, stanowiącym 35% wszystkich respondentów (1071 respondentów), największą część stanowią emeryci (49%) oraz gospodarstwa pracowników (33%). Można tu znaleźć również 6% odsetek rolników. Dominują osoby posiadające oszczędności o wartości od trzech miesięcy do półrocznego dochodu (33%), niewiele mniej − 22% osób posiada oszczędności o wartości wyższej niż półroczny dochód gospodarstwa, prawie tyle samo (21%) respondentów posiada również oszczędności o wartości niższej niż ich trzymiesięczny dochód. Aż 17% badanych zgromadziło oszczędności powyżej rocznych dochodów gospodarstwa domowego. 91% osób posiada oszczędności na lokatach w złotówkach, a 7% ankietowanych ma również oszczędności na lokatach w walutach obcych. 6% badanych w tym stanie stanowią osoby posiadające obligacje oraz 4% respondentów posiada jakiekolwiek inwestycje w papierach wartościowych. Największy odsetek w tej grupie stanowią gospodarstwa domowe, które oceniają, że w porównaniu do sytuacji materialnej sprzed dwóch lat ich stan nie uległ zmianie (78%), 12% respondentów twierdzi, że sytuacja materialna ich gospodarstwa domowego uległa pogorszeniu, natomiast 10% jest zdania, że ich sytuacja materialna była gorsza niż obecnie.

Stan drugi jest stanem najmniej licznym – należy tu 27% wszystkich ankietowanych (816 respondentów). Największy odsetek w tej grupie osób stanowią gospodarstwa pracujące (79%). W odróżnieniu od stanu pierwszego niemałą część tej grupy stanowią również osoby pracujące na własny rachunek (prawie 16%). Jeśli chodzi o łączną wartość zgromadzonych oszczędności, grupa ta jest

zdecydowanie bardziej zróżnicowania. Największą część tej grupy stanowią gospodarstwa o oszczędnościach powyżej 3-miesięcznych do półrocznych dochodów (25%), jak również o oszczędnościach do 3-miesięcznych dochodów (25%). 17% ankietowanych posiada oszczędności w wysokości do rocznych dochodów gospodarstwa, a 11% powyżej rocznych dochodów gospodarstwa domowego. Osoby posiadające oszczędności o wartości poniżej jednego miesiąca dochodów stanowią 12% tej grupy. Jeżeli chodzi o lokowanie zgromadzonych oszczędności, grupa ta cechuje się również większym zróżnicowaniem. 75% ankietowanych posiada oszczędności zgromadzonego na rachunkach w polskiej walucie. Ankietowani tego stanu wypadają lepiej, w przypadku oszczędności zgromadzonych na rachunkach w walutach obcych − 10 % gospodarstw domowych. Ponad 5% ankietowanych w tej grupie posiada obligacje oraz 5% inwestuje również w papiery wartościowe. W stanie tym najmniejszy odsetek respondentów (8%) twierdzi, że sytuacja materialna ich gospodarstwa domowego jest gorsza od sytuacji sprzed dwóch lat. Prawie połowa (45%) jest zdania, że ich sytuacja materialna jest lepsza niż przed dwoma laty. Nieco więcej ankietowanych (47%) w tej klasie nie odnotowało żadnej zmiany w sytuacji materialnej ich gospodarstw domowych.

Do stanu trzeciego zaliczono najwięcej badanych, tj. 1131 osób (37%). Największy odsetek stanowią tu gospodarstwa pracujące (46%) oraz osoby przebywające na emeryturze (34%). Znaleźć tu również można 6% osób pracujących na roli. Ponad 40% badanych posiada oszczędności o wartości nie wyższej niż trzymiesięczny oraz jednomiesięczny dochód gospodarstwa (stosownie 43% i 41%). Tylko 10% ankietowanych posiada oszczędności nie wyższe niż wartość półrocznego dochodu, 6% gospodarstw nie sprecyzowało wartości posiadanych przez nie oszczędności. W stanie tym znajduje się najniższy odsetek, w porównaniu do stanu pierwszego i drugiego, osób posiadających oszczędności w walucie polskiej (57%). Jeżeli chodzi o inne formy oszczędności, grupa ta również najmniej chętnie odkłada środki na kontach w walucie obcej (0,2%) czy zakupuje obligacje (0,1%). Znikomy jest również odsetek osób inwestujących w papiery wartościowe. W stanie tym najwięcej ankietowanych, bo aż 26% gospodarstw domowych twierdzi, że ich sytuacja materialna jest gorsza od sytuacji sprzed dwóch lat. Tylko 7% twierdzi, że ich sytuacja materialna jest lepsza niż przed dwoma laty. Natomiast również największy odsetek tej grupy stanowią gospodarstwa, które nie odnotowały poprawy ich sytuacji materialnej (67%).

Oszacowane prawdopodobieństwa przejścia ukrytego modelu Markowa, obrazujące stabilność pozostania w danym stanie w kolejnych okresach zostały przedstawione w tabeli 1.

Tabela 1

Prawdopodobieństwa przejścia dla trzech stanów

Stan *s*/Stan *r*	Stan 1	Stan 2	Stan 3
Stan 1	0,46	0,23	0,31
Stan 2	0,34	0,40	0,26
Stan 3	0,27	0,25	0,48

Wynika z tego, że gospodarstwa będące w danym stanie mają największą tendencję do pozostawania w nim w kolejnym okresie (na co wskazują najwyższe prawdopodobieństwa na głównej przekątnej macierzy przejścia). Największe prawdopodobieństwo pozostania w tym samym stanie ma stan trzeci, kolejno pierwszy i drugi. Niestety osoby, które cechują się najmniejszą skłonnością do oszczędzania (stan trzeci) najprawdopodobniej nie zmienią swych przyzwyczajeń. Jeżeli jednak uda im się zmienić swoje nastawienie czy pokonać różnego rodzaju bariery, to bardziej prawdopodobne okazuje się przejście do stanu pierwszego, tj. $a_{31} = 0,27$, aniżeli do stanu drugiego ($a_{32} = 0,25$). Niewiele mniejsze szanse na pozostanie w danym stanie w następnym okresie mają również respondenci należący do stanu pierwszego ($a_{11} = 0,46$). Jeżeli zmienią swoje nastawienie, to również dla nich mniej prawdopodobne okazuje się przejście do stanu drugiego (do stanu o największym odsetku osób najlepiej postrzegających sytuację materialną swego gospodarstwa), chętnie oszczędzających w obcej walucie. Ankietowani należący do stanu osób dywersyfikujących swoje oszczędności i których sytuacja materialna uległa poprawie w ciągu ostatnich lat (do stanu drugiego), cechują się najniższą tendencją pozostania w tym stanie ($a_{22} = 0,40$). Dla tej grupy osób istnieje jednak większe prawdopodobieństwo przejścia do stanu o większej skłonności do oszczędzania (tj. stanu 1), aniżeli do grupy osób najmniej chętnie oszczędzających (stan 3).

Podsumowanie

W artykule przedstawiono przykład zastosowania ukrytego modelu Markowa w ocenie skłonności do oszczędzania w naszym kraju. Przedstawiona analiza empiryczna umożliwiła podział respondentów na podstawie odpowiedzi udzielonych w badaniu panelowym prowadzonym co dwa lata przez prof. Czapińskiego. Wyodrębniono trzy stany o podobnych wzorcach zachowań i postaw dla polskich respondentów. Oszacowano również ukryte prawdopodobieństwa pozostania w wyodrębnionych stanach.

Do stanu pierwszego zaliczono głównie emerytów oraz osoby pracujące, oceniające swoją sytuację materialną jako stabilną (największy odsetek osób, twierdzących, że sytuacja materialna ich gospodarstwa nie uległa zmianie). Są to osoby zabezpieczone finansowo, oszczędzające głównie w tradycyjny sposób. Stan drugi wyróżnia pewna odwaga w inwestowaniu swych oszczędności (inne formy aniżeli oszczędności na rachunkach w walucie polskiej) oraz pozytywna ocena swej sytuacji materialnej. Stan ten tworzą przede wszystkim osoby aktywne zawodowo, pracujące w różnego rodzaju przedsiębiorstwach oraz prowadzące własną działalność gospodarczą. Jeżeli chodzi o stan trzeci, to stanowią go również głównie gospodarstwa osób pracujących, emerytów i rolników, posiadających najmniejszy odsetek lokat bankowych oraz innych form finansowego zabezpieczenia swej przyszłości. Respondenci tego stanu swoją sytuację materialną postrzegają jako najgorszą. Niestety grupa ta wykazuje największe skłonności utrzymania swego nastawienia do oszczędzania w kolejnych okresach (największa wartość prawdopodobieństwa w ukrytej macierzy przejścia).

Literatura

Baum L.E., Petrie T. (1966), Statistical Inference for Probabilistic Functions of Finite State Markov Chains, „Annals of Mathematical Statistics", No. 67, s. 1554-1563.

Biernacki C., Celeux G., Govaert G. (1999), Choosing Models in Model-based Clustering and Discriminant Analysis, „Journal of Statistical Computation and Simulation", No. 64, s. 49-71.

Bozdogan H. (2000), Akaike's Information Criterion and Recent Developments in Information Criterion, „Journal of Mathematical Psychology", No. 44, s. 62-91.

Cappe O., Moulines E., Ryden T. (2005), Inference in Hidden Markov Models, Springer Verlag, New York.

Collins L.M., Lanza S.T. (2011), Latent Class and Latent Transition Analysis with Applications in the Social, Behavioral, and Health Sciences, John Wiley & Sons, New York.

Diagnoza społeczna 2013. Warunki i jakość życia Polaków, Raport, red. J. Czapiński, T. Panek, Rada Monitoringu Społecznego, Warszawa 22.08.2013.

Dempster A.P., Laird N.P., Rubin D.B. (1977), Maximum Likelihood for Incomplete Data via the EM Algorithm (with Discussion), „Journal of the Royal Statistical Society", No. 39, ser. B, s. 1-38.

Frühwirth-Schnatter S. (2006), Finite Mixture and Markov Switching Model. Springer Verlag, New York.

Ghysels E. (1994), On the Periodic Structure of the Business Cycle, „Journal of Business and Economic Statistics", No. 12 (3), s. 289-298.

Kim C.J. (1994), Dynamic Linear Models with Markov-Switching, „Journal of Econometrics", Vol. 60, s. 1-22.

Krogh A. (1998), An Introduction to Hidden Markov Models for Biological Sequences [w:] Computational Methods in Molecular Biology, red. S.L. Salzberg, D.B Searls, S. Kasif, Elsevier, Amsterdam, s. 45-63.

McLachlan G.J., Peel D. (2000), Finite Mixture Models, Wiley, New York, s. 81-116.

Rabiner L.R. (1989), A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, „Proceedings of IEEE", Vol. 77, No. 2, s. 267-295.

Schmittmann V.D., Visser I., Raijmakers M.E.J. (2006), Multiple Learning Modes in the Development of Rule-Based Category-Learning Task Performance, „Neuropsychologia", Vol. 44, No. 11, s. 2079-2091.

Visser I., Speekenbrink M. (2010), depmixS4: An R Package for Hidden Markov Models, „Journal of Statistical Software", Vol. 36, No. 7, s. 1-21.

Witek E. (2010), The Comparison of Model-Based Clustering with Heuristic Clustering Methods [w:] Folia Oeconomica 255, Methodological Aspects of Multivariate Statistical Analysis, Statistical Models and Applications, red. Cz. Domański, J. Białek, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, s. 191-197.

Wickens T.D. (2010), Models for Behavior: Stochastic Processes in Psychology, W.H. Freeman and Company, San Francisco.

Zucchini W., MacDonald I. (2010), Hidden Markov Models for Time Series: An Introduction Using **R**. Monographs on Statistics and Applied Probability, CRC Press, Boca Raton.

AN APPLICATION OF LATENT MARKOV MODELS IN POLISH SAVING ATTITUDE

Summary

In latent class analysis it is assumed that each observation comes from one of a number of classes (groups) and models each with its own probability distribution. When longitudinal data are to be analyzed, the research questions concern some form of change over time. Latent transition analysis (LTA) also known as latent Markov model, is a variation of the latent class model that is designed to model not only the prevalence of latent class membership, but the incidence of transitions over time in latent class membership.

We used latent class analysis for grouping and detecting inhomogeneities of Polish attitude to saving money. We analyzed data collected as part of the Social Diagnosis, based on panel research using depmixS4 package of R.