Exploring the dependencies between epigenetic pathways and air pollution with the use of the latent Markov model

Fulvia Pennoni
Department of Statistics and Quantitative Methods[♭]
University of Milano-Bicocca
fulvia.pennoni@unimib.it

&

Andrea Baccarelli[♯], Francesco Bartolucci[♮], Elena Colicino[♯], Giorgio Vittadini[♭]
[♯]*Department of Statistics and Quantitative Methods, Harvard School of Public Health*
[♮]*Department of Economics, University of Perugia, Italy*

Outline

- We illustrate a *state space* model which is tailored for longitudinal data;
- It is conceptualized as something that is *"latent"* beyond the observed data;
- We consider the *missing data pattern* by allowing for the missing at random assumption;
- We propose a suitable parameterization to account for *multiple* binary test response variables and individual specific time-constant and time-varying covariates as well as space factors;
- We consider the *maximum likelihood estimation* of the model parameters by means of the EM algorithm;
- We consider the data of the Veterans Affairs (VA) cohort *Normative Aging Study* with a numver of ≤ 6 visits per participant between 1995 and 2011.

Theoretical model

- We propose a rather new *statistical method* that relies on time dependencies to investigate the epigenetic process leading to etiology of cognitive impairment;
- *The field of epigenetics* originates about 70 years ago to describe various unspecified non-genetic mechanisms influencing phenotype (Cortesis et. al, 2012);
- It refers to *heritable* changes in phenotype unrelated to differences in underlying DNA sequence (Burris and Baccarelli, 2013);
- *Epigenetics marks* change over time and they can be affected in many ways by nutritional and environmental exposures;
- By considering *an observational longitudinal human cohort study* we aim at assessing in which way diseases, environmental exposure and epigenetic phenomena contributes to the cognitive resilience or decline of the elders.

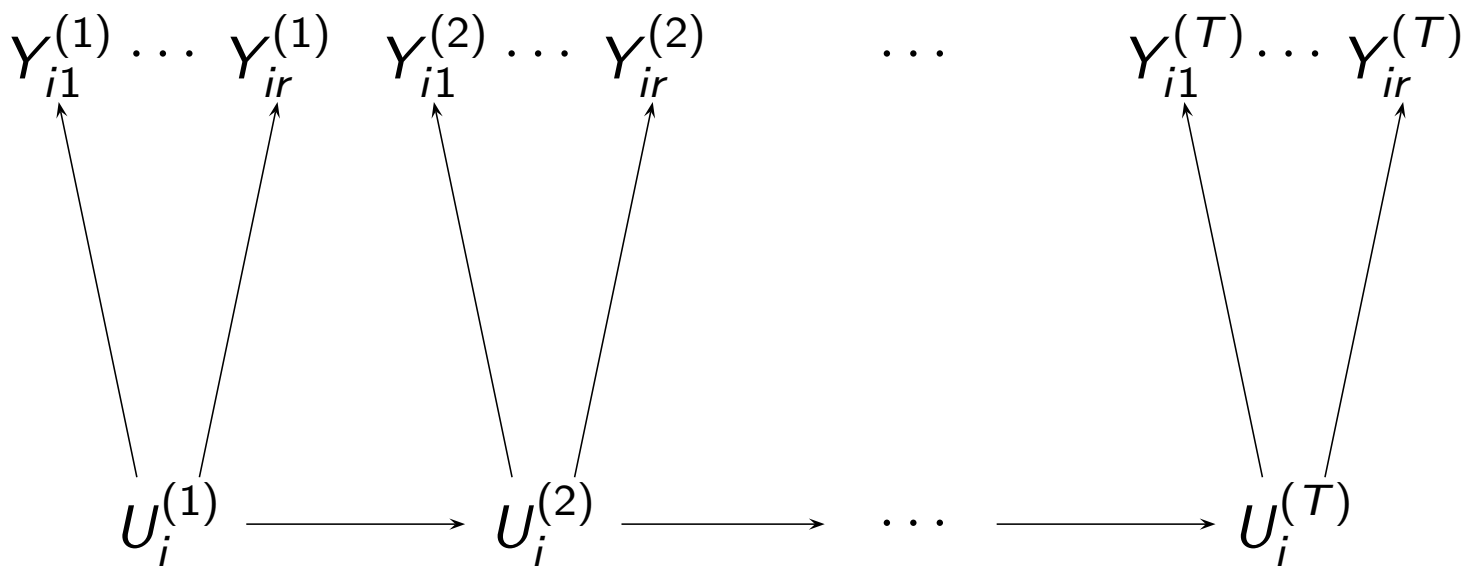
Theoretical model

- *Mitochondrial* (mtDNA) haplogroups are determined by considering the blood DNA using Taqman assays (Applied Biosystems, Foster City, CA);
- The *population* under study is provided by the US Department of Veterans Affairs (VA) Normative Aging Study (NAS) including only men;
- The *compliers* to the study completed several batteries of cognitive tests, starting from 1993;
- In the cohort there are the most common *mtDNA haplogroups* in population of European ancestry;
- They are recorded in 1999; the 9 types are reduced to *4 clusters* according to the phylogenetic evolutionary tree.

Main notation

- Basic *notation*:
  - $\mathbf{Y}^{(t)}$: is the vector of $(r = 3)$ observed categorical *responses* where $t = 1, \ldots, T$: sequences of test results at each time occasion;
  - $Y_{i1}^{(t)}$ is the binary response variable of individual i to *test 1*;
  - $\mathbf{X}^{(t)}$: is the vector of the *intrinsic features* of the individual which can be time-varying;
  - $\mathbf{Z}^{(t)}$: is the vector of time-varying *environmental factors*;
  - $\mathbf{W}^{(t)}$: is the vector of *risk factors*;
  - $\mathbf{U} = (U^{(1)}, \ldots, U^{(T)})$ is the latent process that is an individual and time specific *latent variable* that affects the distribution of the observed variables.

Path diagram of the proposed model Latent Markov model



Directed acyclic graph (DAG):

t, $t = 1, \ldots, n$ denotes the visit occasion for individual i.

Latent Markov (LM) chain approach

- We assume a *discrete distribution* for the *latent* process such that it follows a first-order homogeneous *Markov* chain with state space $1, \ldots, k$ states;
- the model formulation is such that it involves the parameters of the *measurement model* which are: the *conditional response probabilities*

$$\phi_{y|u} = p(Y^{(t)} = y | U^{(t)} = u), \quad y = 0, 1,$$

where we relay on the time homogeneity assumption;

- and the parameters of the *latent process* which are the initial probabilities

$$\pi_{u|\mathbf{xzw}} = p(U^{(1)} = u | \mathbf{X}^{(1)} = \mathbf{x}, \mathbf{Z}^{(1)} = \mathbf{z}, \mathbf{W}^{(1)} = \mathbf{w}), \quad u = 1, \ldots, k,$$

and the transition probabilities

$$\pi^{(t)}_{u|\bar{u}\mathbf{xzw}} = p(U^{(t)} = u | U^{(t-1)} | \mathbf{X}^{(t)} = \mathbf{x}, \mathbf{Z}^{(t)} = \mathbf{z}, \mathbf{W}^{(t)} = \mathbf{w}),$$
$$t = 2, \ldots, T, \ \bar{u}, u = 1, \ldots, k$$

Latent Markov (LM) chain approach

- A *multinomial logit parameterization* is used to account for the covariates on the latent model:
- for the *initial probabilities* we consider the following:

$$\log \frac{\pi_{u|\mathbf{xzw}}}{\pi_{1|\mathbf{xzw}}} = \beta_{0u} + \mathbf{x}'\boldsymbol{\beta}_{1u} + \mathbf{z}'\boldsymbol{\eta}_{1u} + \mathbf{w}'\boldsymbol{\kappa}_{1u}, \quad u \geq 2,$$

- for the *transition probabilities* we consider the following:

$$\log \frac{\pi^{(t)}_{u|\bar{u}\mathbf{xzw}}}{\pi^{(t)}_{\bar{u}|\bar{u}\mathbf{xzw}}} = \gamma_{0\bar{u}u} + \mathbf{x}'\boldsymbol{\gamma}_{1\bar{u}u} + \mathbf{z}'\boldsymbol{\nu}_{1\bar{u}u} + \mathbf{w}'\boldsymbol{\tau}_{1\bar{u}u}, \quad t \geq 2, \;\; \bar{u} \neq u.$$

Missing data

- We use all the available information under the *ignorable* missing data mechanism (Rubin, 1976);
- We take into account the joint distribution of the complete data and the *missing data pattern* for each subject;
- We consider the *binary* indicator $m_{ij}^{(t)}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, r$ equal to 1 if the response $y_{ij}^{(t)}$ of individual i to test j is missing and to 0 otherwise;
- We assume *conditional independence* between the response variables and the missing indicator given the latent process.

Maximum likelihood estimation of the model parameters

- Given a sample of *n independent men*, their response vectors are $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n$;
- The corresponding vectors of observed covariates are $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$, $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n$ and $\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_n$;
- Th LM model *log-likelihood* assumes the following expression

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\tilde{\mathbf{y}}_i \mathbf{m}_i | \tilde{\mathbf{x}}_i \tilde{\mathbf{z}}_i \tilde{\mathbf{w}}_i),$$

where $p(\tilde{\mathbf{y}}_i \mathbf{m}_i | \tilde{\mathbf{x}}_i)$ corresponds to the *joint probability of the responses* provided or not by individual i, given the available covariates.

Maximum likelihood estimation of the model parameters

- The model estimation is performed by the *Expectation-Maximization* algorithm;
- It is based on the *complete data log-likelihood* that with *multivariate* categorical data factorizes in the sum of three components related to the model parameters;
  - **E-step**: compute the *posterior* expected value of each indicator variable (given the observed data) involved in the complete data log-likelihood by suitable forward-backward recursions (Baum et al., 1970):
  - **M-step**: *maximize* the complete data log-likelihood and thus update θ at each step. The maximization is performed according to chosen parameterization for the covariates;
- The *routines* to estimate the proposed model are implemented in the open source language `R` in the `LMest` package (forthcoming Version 2.2).

The applicative example

- We focus on the *probability* of achieving the screening cut-off scores used to asses dementia on three different tests measuring different domains;
- *The mini mental state examination* of *global cognitive* functions. A score lower than 26 is considered indicative of cognitive impairment;
- *The verbal fluency test* that asses *language functions* (vocabulary size, naming). A score lower than 15 is considered indicative of cognitive impairment;
- *The constructional praxis* concerning the *visual and motor abilities*. A score lower then 4 is considered indicative of cognitive impairment.

The applicative example

Covariates (**X**, **W**)	mean or proportions for each covariate	
age in years:	at each visit	77.26
physical activity:	MET-hr/week	16.50
matrilinear ethnicity:	English speaking European ancestry	0.45
	Mediterranean Europe	0.26
	others European countries	0.20
	other countries	0.09
haplogroup:	J or T	0.15
	H or V	0.52
	K or U	0.21
	I, W or X	0.12
years of education:	≤ 12	0.26
	$(12 - 16]$	0.51
	≥ 16	0.23
BMI:	norm	0.22
	over	0.53
	obese	0.25
smoke status:	never	0.30
	former	0.03
	current	0.67

Phylogenitic evolutionary tree of the *mtDNA haplopgroups*

Clusters of *mtDNA haplopgroups*	
1*: J and T*:	Levant, Bedouin, North Eur. Eastern Eur., Indus Mediterranean
2*: H and V*:	Europe , Western Asia, North Africa
3*: K and U*:	West Eurasia, India sub-continent, Algeria, First Extent Middle East
4*: I, X and W*:	Northen Eastern Europe, Amerindians, Southern Siberians, Southern Asians, Eastern Europeans, Southern East Asia

Flow chart of inclusion criteria

- The VA Normative Aging Study is a *longitudinal cohort study* established in 1993 included men from 21 to 81 years of age;
- They are *invited to* medical examination every three to five years;
- The *inclusion criteria* of our sample are the following:
  - 1600 individuals who are in the database with the *first cognitive visit* in 1995 and last cognitive visit in 2011;
  - 782 individuals with *complete record on cognitive date, black carbon concentration and without experience of stroke*;
  - 606 individuals who *did not experience the terminal event*; (*therefore, we condition upon be alive on a given time* (t));
  - 547 with the *haplogroup record*.

Environmental factors and risk factors

- The *BC concentration* (ug/m^3) is predicted according to a spatio-temporal land use regression model (Gryparis at. al. 2007);
- Its measure is available for each individual at each visit; the BC average concetration in the sample is equal to 0.63 ± 0.28;
- Among of the risk factors there is diabetes; the proportion of individuals in the sample with current diabetes medication is equal to 0.08;
- Other *risk factors* may also be available and could be evaluated in the model;

Results of the application

- The *LM model with $k = 2$ latent states* has a log-likelihood equal to -1871.358 and a BIC index equal to 4234.464 with 78 parameters;
- The *estimates of the conditional response probabilities* $\hat{\phi}_{y|u}$:

category : *test 1*	$h = 1$	$h = 2$
0	0.588	0.056
1	0.412	0.944
test 2		
0	0.406	0.120
1	0.594	0.880
test 3		
0	0.361	0.090
1	0.639	0.910

- They show that individuals in the *first latent state* have higher probability to have cognitive impairement in each of the three tests with respect to those in the second latent state.

Results of the application

- The estimated intercept (0.886) referred to the *multinomial logit model for the initial probabilities* shows that there is a general tendency towards good cognitive conditions at the first visit for each participant;
  - The negative parameter estimates for *age* (-0.014) indicates that elder individuals start in the worse cognitive status;
  - The log-odds referred to the four *haplogroups* are all positive (0.67, 0.504, 0.071, 0.312) indicating that individuals in each group show good cognitive conditions at the first visit;
  - The log-odds referred to those individuals *native language is not English* is negative (-0.731) indicating that they show worse cognitive status;
  - The log-odds referred to the *current smokers* is negative (-0.598) indicating that they start in the worse cognitive status.
  - The log-odds referred to the *BC exposure* is negative (-0.851) indicating those with an high exposure start in the worse cognitive status.

Results of the application

- The estimated intercepts referred to the *multinomial logit model for the transition probabilities* are negative (-2.304, -2.284) showing that there is a general tendency towards a cognitive decline after the first visit;
  - The log-odds referred to the four *haplogroups* on the second conditional logit (of begin in latent class 2 given that the individual is in latent class 1) are negative for clusters 1, 2, and 4 and positive for cluster 3 (-0.008, -0.137, 0.219, -0.097);
  - The log-odds on the first logit referred *matrilinear ethnicity* referred to Other European countries is negative (-0.234);
  - The log-odds referred to *BC exposure* is negative (-0.016) on the first logit and on the second logit (-0.01).

Results of the application

- The estimated initial and transition probabilities for groups of individuals of interest may be obtained: for example elders with *haplogroups in cluster 4 (I,X, W), who are current smokers and with an high educational level*:
  - Their *initial probabilities* are $\pi_1 = 0.145$ and $\pi_2 = 0.858$ showing that 14% of them start in the status indicating worse cognitive impairment;
  - Their *estimated transition matrix* is

	$\hat{\pi}_{h\mid\bar{h}}$	
$\bar{h}$	$h=1$	$h=2$
1	0.785	0.215
2	0.084	0.916

  - It shows that there is high persistence in the same latent state and we expect that about 8% of them have the tendency to increase their cognitive decline over time;

Results of the application

- For the elders with haplogroup 4, current smokers and *less educated*:
  - The *initial probabilities* are $\pi_1 = 0.174$ and $\pi_2 = 0.826$ showing that 17% of them start in the status indicating worse cognitive impairment;
  - The *estimated transition matrix* is

	$\hat{\pi}_{h\|\bar{h}}$	
$\bar{h}$	$h = 1$	$h = 2$
1	0.867	0.133
2	0.092	0.908

  - It shows that the percentage in the first latent class is higher than the above as well as the percentage of those switching from the second to the first latent state.

Conclusions

- The *cognitive aging decline* is modelled by conceptualize it as a latent process evolving over time;
- The non-ignorable *missing data* mechanism is taken into account;
- The proposal allows to evaluate the influence of the *covariates* on the latent structure of the model;
- Some *haplogroups* are related with higher rate of impaired cognition;
- The *BC exposure* contributes to worsen the cognitive status over time;
- Currently we are working to enlarge the model to the case of *semi-competitive risks* such as to include the terminal event;
- We are also working to enlarge the model in a *potential outcome context* as it is suitable from a causal perspective involving the marginal distribution of the observed time related to the event of interest.

Main References

- Bartolucci, F., Bacci, S., Pennoni, F. (2014). Longitudinal analysis of self-reported health status by mixture latent auto-regressive models, *Journal of the Royal Statistical Society: Series C*, **63**, 267-288.

- Bartolucci, F. and Farcomeni, A. and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*, Chapman and Hall/CRC press.

- Baum, L.E. and Petrie, T. and Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics*, **41**, 164-171.

- Burris, H. and Baccarelli, A. (2013). Environmental epigenetics: from novelty to scientific discipline. *Journal of Applied Toxicology*, **34**, 113-116.

Main References

- Cortessis, V. K., Thomas, D. C., Levine, A. J., Breton, C. V., Mack, T. M., Siegmund, K. D., ... Laird, P. W. (2012). Environmental epigenetics: prospects for studying epigenetic mediation of exposure-response relationships. *Human genetics*, **131**, 1565-1589.

- Gryparis, A., Coul, B. A., Schwartz, J., Suh H. H. (2007). Semiparametric latent regression models for spatitemporal modelling of mobile source particles in the greater Boston area. *Journal of the Royal Statistical Society, Series C*, **56**, 183-209.

- R CORE TEAM (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.r-project.org/`.

- Rubin, D. B. (1976). *Inference and missing data*. *Biometrika*, **63**, 581-592.

- Wiggins, L.M. (1955). *Mathematical models for the Analysis of Multi-wave Panels*. Ph.D. Dissertation, Columbia University, University microfilms, Ann Arbor.