Journal de la Société Française de Statistique
Vol. 155 No. 1 (2014)

Mixed Hidden Markov Model for Heterogeneous Longitudinal Data with Missingness and Errors in the Outcome Variable

Titre: Modèle de Markov caché mixte pour des données longitudinales hétérogènes avec erreurs et données manquantes dans la variable de sortie

Dominique Dedieu[1], Cyrille Delpierre[1], Sébastien Gadat[3], Thierry Lang[1,2], Benoît Lepage[1,2] and Nicolas Savy[3]

Abstract: Analysing longitudinal declarative data raises many difficulties, such as the processing of errors and missingness in the outcome variable. Moreover, long-term monitored cohorts (commonly encountered in life-course epidemiology) may reveal a problem of time heterogeneity, especially regarding the way subjects respond to the investigator. We propose a Mixed Hidden Markov Model which considers several causes of randomness in response and also enables the effect of a past health outcome to act on present responses through a memory state. Hence, we take into account both errors and missing responses, time heterogeneity, and retrospective questions. We thus propose a Stochastic Expectation Maximization algorithm (SEM), which is less time-consuming than usual EM algorithms to perform the estimation of the parameters of our MHMM.
We carry out a simulation study to assess the performances of this algorithm in the context of cancer epidemiology with the British NCDS 1958 cohort. Simulations show that the effect of covariates on the transitions probabilities is estimated with moderate bias. At last, we investigate a brief real data application on the effect of early social class on cancer through a smoking behaviour. It appears that in the female sample we used, the early social class does not mainly act on smoking behaviours. Moreover, more information is needed to compensate for data missingness and declarative errors in the view to improve our statistical analysis.

Résumé : L'analyse de données déclaratives longitudinales fait apparaître de nombreuses difficultés, comme le traitement des erreurs et des données manquantes de la variable de sortie. En outre, les cohortes suivies sur le long terme, telles que celles utilisées en épidémiologie "life-course" peuvent soulever un problème d'hétérogénéité du temps, surtout en ce qui concerne la façon de répondre aux questions de l'enquêteur. Nous proposons dans cet article l'introduction d'un modèle de Markov caché mixte qui comprend les possibilités d'erreur et de non-réponse, et permet également de considérer que l'effet d'un résultat de santé passé peut agir sur les réponses actuelles à travers une mémoire d' état. En ce qui concerne les estimations, nous avons proposé d'utiliser un algorithme EM Stochastique (SEM), qui est moins gourmand en temps de calcul que l'algorithme EM usuel utilisant une intégration sur les effets aléatoires.
Nous avons effectué une étude par simulation afin d'évaluer les performances de cet algorithme dans le contexte de l'épidémiologie du cancer avec les données de la cohorte britanniques "NCDS 1958". Les simulations ont montré que l'effet des covariables sur les probabilités de transitions a été estimée avec un biais modéré. Enfin, nous avons réalisé une application à des données réelles en étudiant l'effet de la classe sociale précoce sur le cancer à travers un comportement tabagique. Il est apparu que, dans l'échantillon de femmes utilisé pour cette enquête, la classe sociale précoce n'agit pas principalement sur l'usage du tabac. Cependant, plus d'information est nécessaire pour compenser les données manquantes et les erreurs de déclaration et obtenir de meilleurs résultats statistiques.

[1] INSERM, U1027, F-31073 Toulouse, France.
[2] Toulouse University Hospital, Department of Epidemiology, F-31062, Toulouse, France.
[3] Institut de Mathématiques de Toulouse, UMR 5219, F-31062, Université Paul Sabatier, Toulouse, France.
E-mail: nicolas.savy@math.univ-toulouse.fr

 ISSN: 2102-6238

Keywords: Longitudinal data, Mixed Hidden Markov Model, Random effects, Stochastic EM
Mots-clés : Données longitudinales, Modèle de Markov caché mixtes, Effets aléatoires, Algorithme EM stochastique
AMS 2000 subject classifications: 62-02, 62F10, 62M05, 62P10

1. Introduction.

Analysing longitudinal data which comes from the declarations of a patient - usually referred as longitudinal declarative data - raises many difficulties, such as the processing of errors and missingness in the outcome variable. Extensive literature is available on the general issue of measurement errors and missingness. Langeheine in Hagenaars and McCutcheon (2002) stresses that latent classes models are a general solution to successfully cope with measurement errors. In such a work, a true latent (or hidden) quantity is distinguished from the measured (or declared) quantity. In a longitudinal framework, observations form a time series (also denoted process in the sequel) and these observations depend on a second hidden process. Hidden Markov Models (HMM) belong to such a type of longitudinal models.

Even if these models have been initially developed in the quite different situation of artificial intelligence, there exist yet a lot of examples of application of HMM for the statistical analysis of problems with measurement errors. For example, Satten and Longini (1996) use some HMM in an analysis of HIV data to consider the CD4 count measurement error. Authors stress the fact that raw observations are useless for the description of the HIV progression and prove that the introduction of a true hidden CD4 count variable improves the estimations of the transition probabilities. Jackson et al. (2003) use a continuous HMM in the context of misclassification in a chronic disease stage diagnostic, with an application to screening for abdominal aortic aneurysms. This is also the case in Bureau et al. (2003) who carry out applications to oral lesion hairy leukoplakia in a cohort of HIV-infected men and to human *papillomavirus* infection in a cohort of young women. In such an application, the real health condition is described through a continuous hidden Markov process, and HMM allows to consider misclassification errors. Furthermore, HMM may help to deal with Missing Non At Random (MNAR) data, which are another source of difficulty in longitudinal studies. This issue has been addressed in Albert (2000). The corresponding model, fitted to clinical trial data, includes two extended Markovian processes for the outcome (which is partially hidden) and for the missingness indicator (which is completely observed), respectively, the latter being related to the former. However, no measurement error is mentioned.

Up to our knowledge, HMM describing both data missingness and error measurement has not been investigated yet. Lastly, monitored cohort over the long term (frequently encountered in life-course epidemiology) may raise the problem of time heterogeneity in the response process as well as in the health condition transition, or even in the outcome definition. The outcome variable may concern present-day health conditions as well as certain past health-related events, which is inconsistent with the usual Markov hypothesis. This is the case for example in the GAZEL cohort Goldberg et al. (2007) or in the NCDS 1958 cohort Power and Elliott (2006). It seems interesting to extend the usual Markov framework to take past events into account, thus allowing for "event history analysis" Aalen et al. (2008) with HMM usual algorithms.

In the multi-state model, which includes HMM, Commenges (2002) notices that the assumption of homogeneous state transitions was very stringent, while in most cases the study population is

 ISSN: 2102-6238

heterogeneous with regard to some relevant characteristics. Then, he defines a model for state transitions involving observed covariates. Vermunt et al. (1999) propose a General Linear Model (GLM) to describe the probabilities related either to the measurement or to the response model, which could be covariate-dependant. However Commenges (1999) also observes that there may remain an unexplained heterogeneity following the adjustment for available covariates. This requires the introduction of random effects into the transition models. In this context, Altman (2007) introduced MHMM (Mixed Hidden Markov Models) which are applied to multiple sclerosis data. As this disease is very sensitive to individual differences, it is necessary to introduce individual random effects into HMM, leading to MHMM. Another example is given in Detilleux (2008), who used random effects within the transition model in a HMM describing the evolution of a biomarker.

Nevertheless, performing estimations for such mixed effects models is challenging, particularly due to the impossibility of computing the expected likelihood in a closed form expression, which generally implies expensive computational methods. Different approaches to the problem of discrete-time MHMM parametric estimation have been developed recently. Altman (2007) performs such estimations by the use of MCEM algorithm. She underlines the slowness of the EM algorithms despite of the well-known good performance of the recursive forward/backward method developed in Baum et al. (1970). She proposes to directly maximise the likelihood function. To this aim Celeux and Diebolt (1992) and Diebolt and Ip (1996); Gilks et al. (1996) develop, on a general framework, a stochastic EM approach (SEM). Using a SEM algorithm instead of performing, for example, a numerical integration as proposed in Zhang et al. (2010) is not only a way of avoiding expensive computation. Indeed EM algorithms may be dependent on the choice of initial values and, in Gilks et al. (1996), authors point out that SEM can be expected to detect the most stable fixed point of EM by random exploration of the parameters' space, which is a great advantage. Lastly, if the SEM framework appears to be an interesting tool for MHMM estimations, convergence and consistency are proved only for specific simple examples Celeux and Diebolt (1992) or on assumptions which are difficult to validate Nielsen (2000). It appears that the case of partial drawing for unobserved variables has not been theoretically explored. Delattre (2010) makes use of SAEM algorithm a variant of SEM algorithm developed in Delyon et al. (1999); Kuhn and Lavielle (2004); Panhard and Samson (2008). The convergence of SAEM has been studied by many different authors Kuhn and Lavielle (2004) and references therein and is established under some assumptions especially in the context of the exponential family. These assumptions are valid for the MHMM emission model used by Delattre (which involves a Poisson distribution) but no more in our multinomial logit setting. Then, in our context, we prefer to perform a punctual SEM estimation by simply averaging on the stochastic estimations. Moreover, Delattre proposes to simulate all "individual" transition parameters (missing covariates and real health states). Here the Metropolis-Hastings sampler may have a heavy cost as regards time computation (as it is an iterative procedure which must be performed for each subject). We prefer to compute and exact integration over the real health state (their number is limited in our model) and to use the simulation step of the SEM for the missing covariates.

The paper organizes as follow. In section 2, we develop a MHMM for longitudinal declarative data. It allows both declaration errors and non-response. The hidden process corresponds to the true health state, and cannot be observed for various potential reasons. We propose to use an

 ISSN: 2102-6238

extension of a state memory which allows each response to be based either on the true health state on the current date or on the existence of some past health-related event. In Section 3, we describe the EM framework for the MHMM parameters estimation and propose to use a Stochastic EM (SEM) algorithm. In Section 4, we consider the example of the use of such a model to deal with a cancer study from the NCDS 1958 cohort Power and Elliott (2006). We briefly analyse on a novel database the effect of early social class on cancer at adulthood and the possible contribution of smoking as a mediator based NCDS data. Section 5 is devoted to a simulation study with different scenarios. First we aim to investigate the quality of the estimation procedure. Second we investigate a sensitivity analysis of the model.

2. The model

We are interested in the time evolution of a particular disease and its associated health state. This health state is thus related to the presence or not of such a disease. Each subject is described by a stochastic process $(S_t)_{t\geq 0}$ indexed by the time parameter t and this process quantifies the health state of the subject. Of course, $(S_t)_{t\geq 0}$ is not directly observed and is only known through the subject's declarations, which is represented in our paper using another stochastic process $(Y_t)_{t\geq 0}$. The process $(Y_t)_{t\geq 0}$ is obviously related to the real hidden health state $(S_t)_{t\geq 0}$. Some notations are provided in the following sections.

2.1. Longitudinal structure

We consider the evolution of N independent subjects, each of them is then referred to with an integer $1 \leq n \leq N$. We assume that the time $0 \leq t \leq T$ is discretely sampled into a finite set of intervals $]t_d;t_{d+1}]$, with $1 \leq d \leq D$ such that $t_0 = 0$ and $t_{D+1} = T$. The intervals are assumed to be known at the beginning of the study. Hence, we denote by $(S_{n,d})_{1\leq n\leq N,1\leq d\leq D}$ (resp. $(Y_{n,d})_{1\leq n\leq N,1\leq d\leq D}$) the real health state (resp. the observed declarations) of subject n at "time" d.

The real health state $S_{n,d}$ is then described according to three possible states $\{0,1,2\}$ which code the situation of subject n at time d:

- if the disease is absent and the subject is alive in $]t_d;t_{d+1}]$, then $S_{n,d} = 0$,
- if the disease is present at any time of $]t_d;t_{d+1}]$, then $S_{n,d} = 1$,
- if the subject dies in $]t_d;t_{d+1}]$, $S_{n,d} = 2$.

Remark 1. *We should remark that up to these rules, when both "disease" and "death" events occur in the same time interval* $]t_d;t_{d+1}]$*, we decide that* $S_{n,d+1} = 2$*. As a result, there is a slight imprecision concerning the date of the subject's death. However, we only aim to study the incidence of the disease and thus accept this loss of information.*

The process $(Y_{n,d})_{1\leq n\leq N,1\leq d\leq D}$ is slightly more complex than $(S_{n,d})_{1\leq n\leq N,1\leq d\leq D}$ since it is described according to four possible states $\{0,1,2,3\}$. Each of these states depends of course of the declaration of the subject n:

- $Y_{n,d} = 0$ if no disease is signalled along $]t_d;t_{d+1}]$,
- $Y_{n,d} = 1$ if the disease has been stated during $]t_d;t_{d+1}]$,
- $Y_{n,d} = 2$ if no response is obtained in $]t_d;t_{d+1}]$,
- $Y_{n,d} = 3$ if the subject dies during $]t_d;t_{d+1}]$.

 ISSN: 2102-6238

For any subject n and any time d, the response $Y_{n,d}$ randomly depends on covariates. The observed ones are denoted $(\mathbf{X}_{n,d})_{1\leq d\leq D,\ 1\leq n\leq N}$ and the unobserved ones $(\mathbf{W}_n)_{1\leq n\leq N}$ since there are supposed homogeneous (independent on d) for the sake of simplicity. It is then natural to assume the following filtration properties

– $S_{n,d}$ is independent of $\{S_{n,d-k},\ 1<k\leq d-1\}$ conditionally to $(S_{n,d-1},\mathbf{X}_{n,d-1},\mathbf{W}_n)$,
– $Y_{n,k}$ is independent of $\{Y_{n,d},\ d\neq k\}$ conditionally to $(S_{n,k},\mathbf{X}_{n,k},\mathbf{W}_n)$.

According to these several assumptions, we then obtain N Markov processes $(S_{n,d})_{d=1,\dots,D}$ which form, along with the $(Y_{n,d})_{d=1,\dots,D}$ processes, a MHMM. In the next paragraph, we define a model for the state transitions and response emissions. We will omit in the sequel the conditioning for covariates $\mathbf{X}_{n,1},\dots,\mathbf{X}_{n,D}$ for clarity purposes.

2.2. *Real state transitions model*

These transitions concern the evolution of the true health state $(S_{n,d})_{1\leq n\leq N,1\leq d\leq D}$ and are embedded in a Markov dynamic. Our model implies a time heterogeneity on stochastic behaviours for each of the subject. Such an heterogeneity is introduced through the use of some covariates $\mathbf{X}_{n,d}$ (observed) and $\mathbf{W}_n$ (unobserved) and these covariates directly influence the formal transition:

$$f_d(s,q,\mathbf{X}_{n,d},\mathbf{W}_n,\theta^{trans})=\mathbb{P}_{\theta^{trans}}(S_{n,d+1}=q\,|\,S_{n,d}=s,\mathbf{X}_{n,d},\mathbf{W}_n). \tag{1}$$

Covariates $\mathbf{X}_{n,d}$ and $\mathbf{W}_n$ are taken into account by the use of a General Linear Model (GLM). More precisely, let us denote θ^{trans} a the set of parameters $\theta^{trans}=(\theta^{trans,\mathbf{X}},\theta^{trans,0})$ which stands for the influence of covariates $\mathbf{X}$ as well as the random effects that do not depend on the covariates. Remark that all transitions may occur between $\{0,1\}$ and $\{0,1,2\}$ but 2 is a fixed point of the dynamic thus we need to describe seven transitions in our model.

Hence, for each admissible transition $s\mapsto q$, $\theta^{trans,\mathbf{X}}_{s,q}$ is an unknown matrix which acts on the observed covariates $\mathbf{X}_{n,d}$ at time d on subject n. Second, the set of parameters $(\theta^{trans,0}_{s,q,d})_{s,q,d}$ stands for the natural transition from state s to state q at time d. At last, the covariates $(\mathbf{W}_n)_{1\leq n\leq N}$ model the individual randomness from one subject to another and each $\mathbf{W}_n$ is also a vector of $\mathbb{R}^7$.

A linear predictor $\eta_{s,q,d}$ is defined as

$$\forall(s,q)\in\{0,1\}\times\{0,1,2\},\forall d\geq 0,\quad \eta_{s,q,d}(\mathbf{X}_{n,d},\mathbf{W}_n)=\theta^{trans,0}_{s,q,d}+\mathbf{X}'_{n,d}\theta^{trans,\mathbf{X}}_{s,q}+(\mathbf{W}_n)_{s,q}, \tag{2}$$

and the transitions probabilities are defined by a multinomial logit model with η:

$$\forall(s,q)\in\{0,1\}\times\{0,1,2\},\forall d\geq 0,\quad f_d(s,q,\mathbf{X}_{n,d},\mathbf{W}_n,\theta^{trans})=\frac{\exp(\eta_{s,q,d}(\mathbf{X}_{n,d},\mathbf{W}_n))}{\sum_i\exp(\eta_{s,i,d}(\mathbf{X}_{i,d},\mathbf{W}_i))}. \tag{3}$$

Since $S_{n,d}=2$ is a cemetery state, we have of course

$$f_d(2,q,\mathbf{X}_{n,d},\mathbf{W}_n,\theta^{trans})=\mathbb{I}_2(q).$$

where $\mathbb{I}_2(q)=1$ if $q=2$ and otherwise $\mathbb{I}_2(q)=0$.

 ISSN: 2102-6238

2.3. Response model

We describe here the probability to obtain response $Y_{n,d} = q$ from a real state $S_{n,d} = s$. The transition probabilities mainly rely on an emission parameter θ^{em}. From any state of $\{0,1\}$, four responses are possible, each of them being related with an emission (or response) probability. Each emission has a specific interpretation :

- Response $Y_{n,d} = 1$ (disease) from a state $S_{n,d} = 0$ (no disease) is considered as an error.
- Response $Y_{n,d} = 0$ from a state $S_{n,d} = 1$ has several interpretations :
  - i) Subject n may not be ill when the question was asked and became sick just after while the collecting of the data. If the question being asked concerns only the current health state, as it could apply to long-term observational cohorts, the information is lost.
  - ii) The diagnostic has not been told to the subject.
  - iii) The subject may present a denial behaviour.
- Non-response $Y_{n,d} = 2$ is only possible from $S_{n,d} \in \{0,1\}$
- Of course, $Y_{n,d} = 3$ if and only if $S_{n,d} = 2$.

The parameter θ^{em} quantify exactly the randomness in the response emission :

$$g_d(s,y,\theta^{em}) := \mathbb{P}(Y_{n,d} = y \,|\, S_{n,d} = s) = \theta^{em}_{s,y,d}. \tag{4}$$

Remark 2. *In the former expression, note that the emission may depend on the time d and a more simple model would assume the transitions independent on the time evolution. It would also be possible to describe a more general emission process which may involve the unobserved covariates $\mathbf{W}_n$ through a GLM following the same strategy already used for the definition of the functions f_d and $\eta_{s,q,d}$ introduced above. This allows more flexibility, including the use of random effects to describe individual non-response behaviours. However, even if the proposed approach is less flexible, estimations are easier and a direct description of the emission response through θ^{em} is therefore quite easy to interpret.*

2.4. Initial state and unobserved covariates

We end the model statement by the description of the initial values of $(S_{n,1})_{1\le n\le N}$. In this view, we assume without loss of generality that for any subject n, $S_{n,1}$ belongs to $\{0,1\}$ (initially dead subject won't be considered!). We then define θ^{ini} as

$$\theta^{ini,0} = \mathbb{P}(S_{n,1} = 0) \qquad \theta^{ini,1} = \mathbb{P}(S_{n,1} = 1) = 1 - \theta^{ini,0}.$$

In the sequel, we then use the definition

$$h(s,\theta^{ini}) = \theta^{ini,0}\mathbb{I}_0(s) + \theta^{ini,1}\mathbb{I}_1(s). \tag{5}$$

Furthermore, covariates $(\mathbf{W}_n)_{1\le n\le N}$ are assumed to be an i.i.d. sample of centered Gaussian laws $\mathcal{N}(0,\Sigma^{rand})$ where the covariance matrix is supposed diagonal with a diagonal equals to θ^{rand}. As already the size of the vector $\mathbf{W}_n$ corresponds to the total number A of random effects ($A = 7$ in our model). The density function $\gamma_{\theta^{rand}}$ of each $\mathbf{W}_n$ is given by

$$\forall u \in \mathbb{R}^A \qquad \gamma_{\theta^{rand}}(u) = (2\pi)^{-A/2}\det(\Sigma^{rand})^{-1}e^{-\frac{1}{2}u'(\Sigma^{rand})^{-1}u}. \tag{6}$$

 ISSN: 2102-6238

Let us briefly comment the structure of the unobserved covariates $(\mathbf{W}_n)_{1\leq n\leq N}$. It is of course quite natural to assume $\mathbf{W}_n$ to be independent to $\mathbf{W}_{n'}$ when $n \neq n'$ since we consider that in our study, the subjects cannot interact each others. Moreover, covariates $(\mathbf{W}_n)_{1\leq n\leq N}$ are assumed to be stationary (independent of time d), there is almost no loss of generality in this assumption since time heterogeneity is already considered in our model in the parameter $\theta^{trans,0}$ which depends on d. The last hypothesis concerns the diagonal structure of Σ^{rand}. It implicitly imposes that each transitions $s \mapsto q$ are independent from any couple of transition to another one. This assumption is natural for the transitions $(0 \mapsto 0), (0 \mapsto 1)$ and $(0 \mapsto 2)$ (the disease is initially absent). Such an independence assumption is also natural for the couple of transitions $(0 \mapsto q)$ and $(1 \mapsto q)$. The most questionable fact is the independence between $(1 \mapsto 1)$ and $(1 \mapsto 2)$ since the absence of healing certainly influences (ans is positively correlated to) the death occurrence. A more general model could take this last point into account.

2.5. Extension to retrospective data.

In some longitudinal studies, the subjects may be asked questions concerning both their present health state and their past health state. In this section, one presents an adaptation to the MHMM described in the previous section in order to analyse such data. We assume that at a certain (but not necessarily any) date t the subjects are asked two questions : "are you ill *now* ?" and "have you *ever* been ill ?".

Let fix $n \in \{1,\ldots,N\}$ a subject. We denote $Y^*_{n,d}$ the random response variable in the date interval d, with $v_R(d)$ possible values. We assume that $Y^*_{n,d}$ may only stand for the first question, or only for the second question, or may gather the response to *both* questions. In this latter case, we assume that the "non-response" level stands for both, and then we obtain $v_R(d) = 6$ levels ((0,0), (0,1), (1,0), (1,1), the non-response levels (.,2) and (2,.) corresponds, by assumption, to one level and level 3 (death)). In the two former cases we had $v_R(d) = 4$. The Markov hypothesis does no longer stand as $Y^*_{n,d}$ depends not only on the current state $S_{n,d}$ but on the complete state history $(S_{n,1},...,S_{n,d})$. Let us assume that $Y^*_{n,d}$ is independent from $Y^*_{n,k}$ $(k \neq d)$ conditionally to $S^*_{n,d} = (S_{n,d}, S'_{n,d-1})$, with $S'_{n,d-1}$ adopting value 1 if there exists some $k < d$ with $S_{n,k} = 1$ (with in addition $S_{n,l} \neq 2$ for all $l < d$). To provide a more accurate definition of S' it is convenient to define a composition law $\bullet$ by:

- $s \bullet q = 0$ if s and q equal 0,
- $s \bullet q = 1$ if s or q equals 1 with s and q different from 2 and 3,
- $s \bullet q = 2$ if s or q equals 2 or 3.

Then we obtain $S'_{n,d-1} = S_{n,1} \bullet S_{n,2} \bullet \cdots \bullet S_{n,d-1}$. With $S_{n,d}$ being the current health state at the date t_{d+1}, we consider $S'_{n,d-1}$ as a state memory. The independence assumption can be interpreted as an absence of causal dependence between Y^*_{n,d_1} and Y^*_{n,d_2} though, due to a retrospective data collection, Y^*_{n,d_1} may give information on Y^*_{n,d_2} $(d_1 < d_2)$. The only causal link lies between the current health state $S_{n,d}$, the state memory $S'_{n,d}$, and the response $Y^*_{n,d}$ in the same date interval. We prove in Appendix A that the processes $(S^*_{n,d})_{d=1,\ldots,D}$ and $(Y^*_{n,d})_{d=1,\ldots,D}$ form a HMM. It is then possible to consider $(S^*_{n,d})_{d=1,\ldots,D}$ as a five-states Markov process taking values in $\{(0,0),(0,1),(1,0),(1,1),(2,0)\}$. Notice that, due to the definition of the state memory, some state transitions are deterministic. Indeed, the transitions $(0,0) \to (\cdot,1)$ and $(0,1) \to (\cdot,0)$ are not

 ISSN: 2102-6238

possible.

We could use the state memory in the transition model and define different incidence parameters for subjects which never had the disease and subjects who have had it. However we decide to favour a more parsimonious approach by assuming that we have, for any $q' \in \{0;1\}$ and $s' \in \{0;1\}$,

$$\mathbb{P}(S^*_{n,d+1} = (q,q') \,|\, S^*_{n,d} = (s,s'), \mathbf{X}_{n,d}, \mathbf{W}_n; \theta^{trans}) = \mathbb{P}(S_{n,d+1} = q, \,|\, S_{n,d} = s, \mathbf{X}_{n,d}, \mathbf{W}_n; \theta^{trans}).$$

Based on this assumption, the parameters vector has the same dimension as in the three-levels state model.

3. Parameters estimation.

3.1. The EM framework

We aim to estimate the several parameters introduced above ($\theta := (\theta^{trans}, \theta^{em}, \theta^{ini}, \theta^{rand})$) using the maximisation of the observed likelihood. The complete log-likelihood ℓ is a random function due to its dependence on the unobserved state variables $(S_{n,d})_{1\leq n\leq N, 1\leq d\leq D}$ and the unobserved random effects variables $(\mathbf{W}_n)_{1\leq n\leq N}$.

3.1.1. Reminders on EM algorithms

We provide here a short summary of the EM principle for sake of completeness. We consider a statistical model parametrised by a family of laws $(\mathbb{P}_\theta)_{\theta\in\Theta}$. Each law $\mathbb{P}_\theta$ produces a couple of variables (U,V) where U is observed and V is a missing unobserved variable. We assume that a close analytic formula is available to compute $\ell(U,V,\theta) := \log \mathbb{P}_\theta[(U,V)]$. Such a formula is known in our case of MHMM described in Section 2. Given i.i.d. observations $(U_i)_{1\leq i\leq N}$, we are looking for an optimal parameter θ^* which maximises the likelihood of the observed variables when $(V_i)_{1\leq i\leq N}$ are unknown. In the view to maximise $\theta \longmapsto \log P_\theta(U)$, the EM algorithm produces a sequence of parameters $(\theta_k)_{k\geq 0}$ which converges under some mild conditions to a local maxima of such a function (see Dempster et al. (1977)). The sequence is defined as follows.

- **E Step** Let be given $\theta_k \in \Theta$, we define the application

$$Q_k(\theta) := \mathbb{E}_{V\sim\mathbb{P}^U_{\theta_k}}[\ell(U,V,\theta)]$$

  where the unobserved variables V follows the conditional law $\mathbb{P}_{\theta_k}[.|U]$ also denoted $\mathbb{P}^U_{\theta_k}$ in the sequel. Such an expectation must be computed for each value of θ and generally requires heavy MCMC computations.
- **M Step** The next value θ_{k+1} is then obtained through the maximisation step:

$$\theta_{k+1} := \arg\max_{\theta\in\Theta} Q_k(\theta).$$

  This maximisation procedure is sometimes possible up to the knowledge of some analytic formulas.

Such an algorithm was applied to HMM by Baum and Welch (see for example Baum et al. (1970); Bartolucci et al. (2007); Zhang et al. (2010)).

 ISSN: 2102-6238

3.1.2. *Application of the EM method to our setting*

Likelihood decomposition Owing to the Markov dynamic of the process $(S_{n,d})_{1\leq n\leq N, 1\leq d\leq D}$, the transitions $(S_{n,d}, \mathbf{W}_n, \mathbf{X}_{n,d}) \longmapsto Y_{n,d}$ and the independence of the N subjects each others, it clearly appears that ℓ can be split into 4 terms, linked to the initial states parameters, transition parameters, emission parameters, and random effects parameters, respectively:

$$\begin{aligned}\ell(\theta) = &\sum_{n=1}^{N}\sum_{d=1}^{D-1} \ln f_d(S_{n,d}, S_{n,d+1}, \mathbf{X}_{n,d}, \mathbf{W}_n, \theta^{trans}) + \sum_{n=1}^{N}\sum_{d=1}^{D} \ln g_d(S_{n,d}, Y_{n,d}, \theta^{em}) \\ &+ \sum_{n=1}^{N} \ln h(S_{n,1}, \theta^{ini}) + \sum_{n=1}^{N} \ln \gamma_{\theta^{rand}}(\mathbf{W}_n). \end{aligned} \tag{7}$$

Transition parameters Each set of parameter is estimated independently by maximising the corresponding term of the log-likelihood ℓ. We first consider the transition parameters θ^{trans}. Let us assume that at iteration k we have a certain estimation $\hat{\theta}_k$ of the parameters vector $\theta = (\theta^{ini}, \theta^{trans}, \theta^{em}, \theta^{rand})'$. In the expectation (E) step of the EM algorithm, we define the objective function

$$Q_k^{trans}(\theta^{trans}) = \mathbb{E}_{\hat{\theta}_k}\left[\sum_{n,d} \ln f_d(S_{n,d}, S_{n,d+1}, \mathbf{X}_{n,d}, \mathbf{W}_n, \theta^{trans}) \mid Y_{n,1}, \ldots, Y_{n,D}\right].$$

If we now decompose our randomness structure, we have

$$\begin{aligned}&Q_k^{trans}(\theta^{trans}) \\ &= \sum_{n,d}\sum_{s,q}\int_{\mathbb{R}^A} \mathbb{P}_d^{trans}(s, q | Y_{n,1}, \ldots, Y_{n,D}; \hat{\theta}_k, \mathbf{w}, \mathbf{X}_{n,d}) \gamma_{\hat{\theta}_k^{rand}}(\mathbf{w}) \ln f_d(s, q, \mathbf{X}_{n,d}, \mathbf{w}, \theta^{trans}) \, \mathrm{d}\tilde{\mathbf{w}}. \end{aligned} \tag{8}$$

The computation of $\mathbb{P}_d^{trans}(s, q | Y_{n,1}, \ldots, Y_{n,D}; \hat{\theta}_k, \mathbf{w}, \mathbf{X}_{n,d})$ may be performed using the well-known forward / backward algorithm (see Appendix B). However, unfortunately no closed-form expression exists for the integration over $\gamma_{\hat{\theta}_k^{rand}}(\mathbf{w})\,\mathrm{d}\tilde{\mathbf{w}}$. This raises some practical issues. Zhang et al. (2010) performed a numeric integration through a Gaussian quadrature method, but this approach may be expensive. The integration through Monte Carlo approximation does not prove to be less expensive. We therefore propose a stochastic approach, according to the framework described in Gilks et al. (1996), which will be detailed further in this paper.

Emission parameters We now focus on the estimation of emission parameters. The objective function has a similar form given by

$$\begin{aligned}&Q_k^{em}(\theta^{em}) \\ &= \sum_{n,d}\sum_{s,y}\int_{\mathbb{R}^A} \mathbb{P}_d^{em}(s, y | Y_{n,1}, \ldots, Y_{n,D}, \mathbf{w}; \hat{\theta}_k, \mathbf{X}_{n,d}) \gamma_{\hat{\theta}_k^{rand}}(\mathbf{w}) \ln g_d(s, y, \theta^{em}) \, \mathrm{d}\tilde{\mathbf{w}}. \end{aligned} \tag{9}$$

 ISSN: 2102-6238

Let us remark that if the response model does not include any random effect, the integration disappears, the computation of such an objective function is easier and the maximisation step can be performed numerically. It is also the case when the response model does not include any random effects or covariates effects, according to the standard HMM method (see Appendix C). Now, if the response model includes some random effects, we propose a stochastic approach, such as the one we will develop for the transition model.

Other parameters Finally, we have to estimate the initial state parameter θ^{ini} and the random effect parameters θ^{rand}. Regarding the former θ^{ini}, we use the standard HMM method (see Appendix C). As for the latter, the objective function is given by

$$Q_k^{rand}(\theta^{rand}) = \sum_n \int_{\mathbb{R}^A} \gamma_{\hat{\theta}_k^{rand}}(\mathbf{w}) \ln \gamma_{\theta^{rand}}(\mathbf{w}) \, \mathrm{d}\tilde{\mathbf{w}}. \tag{10}$$

Let us recall that $\gamma_{\theta^{rand}}$ is a Gaussian law whose covariance matrix is diagonal and described by θ^{rand}. Denoting $\hat{\sigma}_{i,k}$ the i-th component of estimated parameter vector $\hat{\theta}^{rand}$ at iteration k, cancellation of the gradient of Q_k^{rand} leads to

$$\sigma_{i,k+1}^2 = \sigma_{i,k}^2 . \frac{1}{N} . \sum_n \frac{1}{\mathbb{P}(Y_{n,1}, \ldots, Y_{n,D}; \hat{\theta}_k)} \int_{\mathbb{R}^A} \mathbf{w}_i^2 \, \mathbb{P}(Y_{n,1}, \ldots, Y_{n,D} | \mathbf{w}; \theta_k) \, \mathrm{d}\tilde{\mathbf{w}}. \tag{11}$$

with $\hat{\sigma}_{i,0} = 1$ for instance. Once more, we use a stochastic approach (detailed in the following section) rather than performing a numeric integration.

3.2. Stochastic EM algorithm

Reminders on SEM In order to avoid a difficult integration step in the E step of the EM algorithm described in Subsection 3.1.1, we have applied the Stochastic EM (SEM) method. Such an improvement of the initial EM is described for instance in Gilks et al. (1996) or in Nielsen (2000) in the context of MHMM.

We keep the simple notations of Subsection 3.1.1, the algorithm produce a sequence of parameters $(\theta_k)_k$. Such an enhancement concerns the situation when there is no close formula to compute at step k the function $\theta \longmapsto Q_k(\theta) = \mathbb{E}_{V \sim \mathbb{P}_{\hat{\theta}_k}^U}[\ell(U, V, \theta)]$. Instead of using a costly integration over the whole space of unobserved data V, the idea is to use a stochastic draw with respect to a suitable probability distribution. The SEM algorithm then exploits this idea and produces at step k a single realisation of a missing variable $V_k \sim \mathbb{P}_{\hat{\theta}_k^U}$. We then compute the estimate function $SQ_k(\theta)$ as

$$SQ_k(\theta) := \ell(U, V_k, \theta). \tag{12}$$

Hence, the E step is replaced by a Stochastic Expectation SE step.

Application to MHMM We here manage the SEM algorithm as follows: we make a single drawing of the random effects at each step of the SEM algorithm. Note that unlike the general SEM approach, we do not randomly draw all the missing data, which include the hidden states

 ISSN: 2102-6238

$(S_{n,d})_{1\le n\le N,1\le d\le D}$. Given any current value of the parameter $\hat{\theta}_k$, we then sample some possible values of the unobserved covariates $(\mathbf{W}_n)_{1\le n\le N}$. We denote $\tilde{\mathbf{W}}_k := (\tilde{\mathbf{W}}_{k,n})_{1\le n\le N}$ i.i.d. samples of the conditional law $\gamma_{\tilde{\theta}_k^{rand}}$ given realisations $(Y_{n,d})_{1\le n\le N,1\le d\le D}$. Then, we compute

$$SQ_k^{trans}(\theta^{trans},\tilde{\mathbf{W}}_k)=\sum_{n,d}\sum_{s,q}\mathbb{P}_d^{trans}(s,q|Y_{n,1},\ldots,Y_{n,D};\tilde{\mathbf{W}}_{k,n};\hat{\theta}_k,\mathbf{X}_{n,d})\ln f_d(s,q,\mathbf{X}_{n,d},\tilde{\mathbf{W}}_{k,n},\theta^{trans}), \tag{13}$$

$$SQ_k^{em}(\theta^{em},\tilde{\mathbf{W}}_k)=\sum_{n,d}\sum_{s,y}\mathbb{P}_d^{em}(s,y|Y_{n,1},\ldots,Y_{n,D};\tilde{\mathbf{W}}_{k,n};\hat{\theta}_k,\mathbf{X}_{n,d})\ln g_d(s,y,\theta^{em}), \tag{14}$$

$$SQ_k^{rand}(\theta^{rand},\tilde{\mathbf{W}}_k)=\sum_{n}\ln\gamma_{\theta^{rand}}(\tilde{\mathbf{W}}_{k,n}). \tag{15}$$

For the simulation of $\tilde{\mathbf{W}}_k$, we use the Metropolis-Hastings algorithm (see Chib and Greenberg (1995) and Appendix D). It consists of constructing a Markov chain $(\mathbf{Z}_p)_{p\ge0}$ with a succession of acceptances or rejections from random proposals and then choose $\tilde{\mathbf{W}}_k$ as a realization of $\mathbf{Z}_p$ for large values of p once the Markov chain has reached its steady regime. Given any value of $\mathbf{Z}_p$, we simulate the Markov dynamic as follows: sample first a new proposition $\mathbf{z}$ from the distribution $\gamma_{\hat{\theta}_k^{rand}}$ and compute the ratio

$$q_p=\frac{\mathbb{P}(Y_{n,1},\ldots,Y_{n,D}|\mathbf{z},\hat{\theta}_k)\gamma_{\hat{\theta}_k}(\mathbf{z})}{\mathbb{P}(Y_{n,1},\ldots,Y_{n,D}|\mathbf{Z}_p,\hat{\theta}_k)\gamma_{\hat{\theta}_k}(\mathbf{Z}_p)}.$$

The next state $\mathbf{Z}_{p+1}$ is then chosen as $\mathbf{Z}_{p+1}=\mathbf{z}$ with a probability $1\wedge q_p$. Remark that in the former acceptance ratio, the computation of $\mathbb{P}(Y_{n,1},\ldots,Y_{n,D};\hat{\theta}_k)$ is not necessary, which is a great advantage. The Metropolis-Hastings Markov chain converges exponentially fast provided that the acceptance rate is suitably chosen. In practice, we have stopped the iterations of the Metropolis-Hastings Markov chain algorithm when the acceptance rate becomes close to, say, 0.3 Chib and Greenberg (1995).

Once we draw $\tilde{\mathbf{W}}_k$ with a value of $\mathbf{Z}_p$ for a large value of p, the computations and maximizations of the objective functions is possible. For the transition and the emission score functions, we perform a numeric maximization, and denote

$$\hat{\theta}_{k+1}^{trans}=\underset{\theta^{trans}}{\operatorname{argmax}}\left[SQ_k^{trans}(\theta^{trans},\tilde{\mathbf{W}}_k)\right]\qquad\text{and}\qquad\hat{\theta}_{k+1}^{em}=\underset{\theta^{em}}{\operatorname{argmax}}\left[SQ_k^{em}(\theta^{em},\tilde{\mathbf{W}}_k)\right].$$

For the random effects objective function, it can be easily seen that vanishing the gradient of the score functions leads to

$$\hat{\sigma}_{i,k}^2=\frac{1}{N}\sum_n\tilde{\mathbf{W}}_{i,k}^2,$$

and the updated random effect parameters appear to be a simple average of observed variances.

It is to be noted that unlike the usual EM estimator $\hat{\theta}_k$, the stochastic EM estimator $\hat{\theta}_k=(\hat{\theta}_k^{ini},\hat{\theta}_k^{trans},\hat{\theta}_k^{em},\hat{\theta}_k^{rand})$ is no longer consistent since SEM involves a variability of the simulation step (we draw at each iteration k a new realisation of $\tilde{\mathbf{W}}_k$). Figure 3 displays some examples of fluctuations of $\hat{\theta}_k$ components with $1\le k\le 700$. We thus consider an average of such estimators

 ISSN: 2102-6238

and use $\frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k$. Ergodic results (Cesaro convergence) on Markov chains imply the converge when $K \to \infty$ and we denote

$$\hat{\theta} = \lim_{K\to+\infty} K^{-1}\sum_{k=1}^{K}\hat{\theta}_k.$$

Then we use $\hat{\theta}$ as an estimator of θ (see Gilks et al. (1996)). In practice, we compute an approximation of $\hat{\theta}$ with a finite number of successive iterations of the SEM algorithm after a burn-in period.

Remark 3. *Note that in the implementation of the SEM method, we do not have use a complete stochastic strategy since we only draw an unobserved state for the missing covariates but still use a global summation over the hidden states of $(S_{n,d})_{n,d}$. This is reasonable in our case because the state space for $(S_{n,d})_{n,d}$ is not so large: three times of observations and three health states. In a more complex setting of large N and D and large number of possible states, a complete SEM on unobserved variables should be better but remark that convergence of SEM towards its stationary distribution becomes slower when the number of unobserved simulated data increase.*

3.3. Standard errors

Classical approach The observed information matrix $\mathbf{J}_{obs}(\theta)$ is usually used to derive standard errors of the SEM estimators Diebolt and Ip (1996). This matrix is given by the opposite of the second derivative of the observed log-likelihood. A classical result Louis (1982) states that, ℓ being the log-likelihood of the complete data and ℓ_{obs} being the observed log-likelihood, we have

$$\frac{\partial \ell_{obs}}{\partial \theta} = \frac{\partial \mathbb{E}_\theta[\ell(\theta)|Y]}{\partial \theta}.$$

For example, if we assume that derivation under the integration sign is possible, the estimated covariance matrix for the transition parameters is given by:

$$\Sigma^{trans} = -\left[\sum_{n,d}\sum_{s,q}\int_{\mathbb{R}^A} \mathbb{P}_{n,d}(s,q|Y_{n,1},\ldots,Y_{n,D};\tilde{\mathbf{w}};\hat{\theta},\mathbf{X}_{n,d}) \right.$$
$$\left. \nabla^2 \ln f(s,q,\mathbf{X}_{n,d},\mathbf{Y}_n,\tilde{\mathbf{w}},\theta^{trans})\frac{\gamma_{\hat{\theta}^{rand}}(\tilde{\mathbf{w}}))}{\mathbb{P}(Y_{n,1},\ldots,Y_{n,D};\hat{\theta})}\,\mathrm{d}\tilde{\mathbf{w}}\right]^{-1}. \quad (16)$$

Since the Hessian $\nabla^2 \ln f$ can be computed in a closed form expression, we can perform a numerical integration over the random effects. Although computationally intense, this operation is only performed once the parameter $\tilde{\theta}$ is estimated by our SEM. Such a method is known to underestimate the standard errors as pointed in our simulation studies and a bootstrap approach may also be possible Efron and Tibshirani (1994) but this option generally yields also important computational costs.

Stochastic strategy It is possible to improve our initial SEM algorithm to obtain on-line estimation of the Fisher matrix. Such an improvement is very similar to the stochastic modification of the initial EM algorithm to the SEM one following the initial strategy of Louis (1982). We

 ISSN: 2102-6238

still use the notations of Subsection 3.1.1 and remind that the model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ yields a couple of random variables (U,V) where U is observed and V is hidden. We are interested in the Fisher matrix associated to the observed variable U. As pointed by Louis (1982), if we denote the log-likelihood of complete data by $\ell(U,V,\theta)$ and of observed data by $\ell(U,\theta)$, such a matrix may be written as

$$J_{obs}(\theta) = \mathbb{E}_{V\sim\mathbb{P}_\theta^U}\left[-\nabla^2\ell(U,V,\theta) - \nabla\ell(U,V,\theta)\nabla\ell(U,V,\theta)'\right] + \nabla\left(\ell(U,\theta)\right)\nabla\left(\ell(U,\theta)\right)' \quad (17)$$

Now, remark that when θ is close to the maximum likelihood estimator of observed values, the third term of (17) is close to zero. Hence, if θ is near the ML of observed values, it is enough to approximate the two first terms of (17).

We can now improve our SEM to obtain in a clear and simple way an estimator of $J_{obs}(\hat{\theta})$.

- SE Step: consider the initial sequence of parameters $\hat{\theta}_k$ built by the SEM procedure and consider at step k the current estimator $\hat{\theta}_k$. At step k, draw a simulation of the unobserved random variable V_k according to the conditional law $\mathbb{P}^U_{\hat{\theta}_k}$. This can be obtained using again an acceptance/rejection algorithm described in the paragraph above.
- M Step: Compute the new parameter $\hat{\theta}_{k+1}$ as a maximum of the estimated complete log likelihood $\hat{Q}_k(.)$, as well as the average of estimates at step k:

$$\tilde{\theta}_k := \frac{1}{k}\sum_{j=1}^{k}\hat{\theta}_j.$$

The estimated Fisher information matrix is then approached using

$$\tilde{J}_k := -\nabla^2\ell(U,V_k,\tilde{\theta}_k) - \nabla\ell(U,V_k,\tilde{\theta}_k)\nabla\ell(U,V_k,\tilde{\theta}_k)'.$$

Using again ergodic convergence results on Markov chain to steady regime, a suitable estimator of the observed information matrix using

$$\tilde{J} = \lim_{k\longmapsto+\infty} k^{-1}\sum_{j=1}^{k}\tilde{J}_j.$$

Such an improvement of the algorithm may be included in a very simple way in our initial algorithm by using at iteration k both a stochastic drawing of the unobserved covariates $(\mathbf{W}_{k,n})_{1\le n\le N}$ and of unobserved real state $(\hat{S}_{n,d,k})_{1\le n\le N,1\le d\le D}$, each of them being simulated with respect to the suitable conditional law.

4. Application to real data.

4.1. Study of cancer in the NCDS cohort.

Kelly-Irving et al. (2012) use the British NCDS 1958 cohort Power and Elliott (2006) to study the relationship between cancer and adversity at an early life stage. However, analysing these data raises many statistical difficulties, especially due to time heterogeneity.

- It seems impossible to assume that the behaviour of the subjects with regards to self-reports does not vary in a 30-years time gap.

 ISSN: 2102-6238

– We have to face with the presence of different types of outcomes: the cohort members may be asked questions concerning their current health state as well as about certain past health-related event.
– There is also data missingness and declarative errors (due for instance to a possible recovery at the date of interview, denial, or wrong health representations ; Manjer *et al.* , Navarro *et al.*, Cho *et al.* stress that in the field of cancer epidemiology, under- and over-reporting are quite frequent Manjer et al. (2004); Navarro et al. (2006); Cho et al. (2009).

In this section we describe a model according to the NCDS 1958 design, which integrates all those aspects.

In the NCDS 1958 study, the subjects are aged of 23 (t_1), 33 (t_2), 42 (t_3), 47 (t_4) and 51 years old (t_5) and are questioned bout cancer. They must answer to the question "Do you have cancer ?" on dates t_1, t_4, t_5. On date t_2 they also answer to the question "have you *ever* had cancer ?". Lastly on date t_3 they only answer to the question "have you ever had cancer", including the possibility to be ill at the very date on which the question is asked. Then we assume that on dates t_1, t_3, t_4, t_5 the outcome has four levels ("yes", "no", non-response, or dead). On date t_2 we assume that the outcome has six levels (the non-response being common to both questions) depending on the state on the current date and also on the state memory. We illustrate in Figure 1 the corresponding longitudinal scheme.

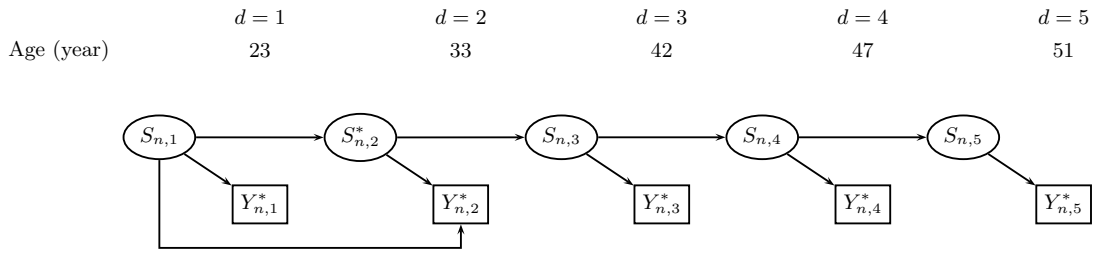

FIGURE 1. *Study of cancer with the NCDS 1958 cohort.* $Y^*_{n,d}$*: response of the subject n at date d and* $S_{n,d}$ *or* $S^*_{n,d}$*: true health state of the subject n on each time interval.*

We aim to estimate the effect of the early social class of cohort members on cancer during adulthood.

Social epidemiologists pinpoint an association between social class and cancer. However little is known about the mechanisms which lead from early life context to a future health event. We propose to use a MHMM to test the hypothesis of a pathway leading from a given early life context to cancer and involving smoking behaviour.

For computational reasons, we only use a sub-sample of 1,000 subjects from the 8,959 female NCDS cohort members: we only consider females since one of the most early frequent cancer events are breast cancer. It is also a simple way to reduce heterogeneity especially in the response

 ISSN: 2102-6238

behaviours.

A binary variable X_1 is derived from the NCDS six level "social class" variable which describe the parental social class at birth. We get 5,158 subjects with $X_1 = 0$ (social class from I to III-manual levels) and 1,760 subjects with $X_1 = 1$ (social class from III- non manual to V levels). We exclude the 2,041 subjects with missing values for X_1. We derive a quantitative variable X_2 by considering the cumulative number of cigarettes smoked from the age of 23 years old. For smoking subjects with no more precision, we impute the median level. Moreover, we impute values for people with missing X_2 on a certain date, by using the previous or next declaration when it exists. Then we obtain 7,289 subjects with data at ages 23, 33, 42 47 and 51 years old. As regards cancer, we have 39 events declared at age 23 years old (with 2,689 missing data), 135 at age of 33 years old (with 3,243 missing data), 202 at age 42 years old (with 3,186 missing data), 61 at age 47 years old (with 4,071 missing data) and 67 at age 51 years old (with 4,003 missing data). The missingness patterns is non-monotonous, and we consider missingness as a response level. At last, we make estimations with sub-samples from 5,704 subjects.

4.2. The MHMM of interest.

In order to investigate the possibility of a mediated effect of an early social class on cancer through a smoking behaviour, we estimate parameters from two models : a univariate model explaining cancer from X_1 and a bivariate model explaining cancer from X_1 and X_2. The evolution of the coefficient associated with X_1 will give an indication on the mediation hypothesis.
We propose a model (M) taking into account time heterogeneity in transitions by the use of as many intercepts as different dates (Table 1). We use covariates $\mathbf{X}_{n,d}$ only for the transition $(0,.) \to (1,.)$ which is the transition of interest, and we assume that the covariate effect on the disease is time homogeneous.

TABLE 1. *Structure of the GLM transition model.*

Transition	Linear predictor η
$(0,\cdot) \to (0,\cdot)$	0 (reference)
$(0,\cdot) \to (1,\cdot), \quad d=1$	$\theta_7^{trans} + \theta_4^{trans} X_1 + \theta_5^{trans} X_2 + \theta_6^{trans} X_3 + W_1$
$(0,\cdot) \to (1,\cdot), \quad d=2$	$\theta_8^{trans} + \theta_4^{trans} X_1 + \theta_5^{trans} X_2 + \theta_6^{trans} X_3 + W_2$
$(0,\cdot) \to (1,\cdot), \quad d=3$	$\theta_9^{trans} + \theta_4^{trans} X_1 + \theta_5^{trans} X_2 + \theta_6^{trans} X_3 + W_3$
$(0,\cdot) \to (1,\cdot), \quad d=4$	$\theta_{10}^{trans} + \theta_4^{trans} X_1 + \theta_5^{trans} X_2 + \theta_6^{trans} X_3 + W_4$
$(0,\cdot) \to (2,\cdot)$	θ_3^{trans}
$(1,\cdot) \to (0,\cdot)$	0 (reference)
$(1,\cdot) \to (1,\cdot)$	θ_2^{trans}
$(1,\cdot) \to (2,\cdot)$	θ_1^{trans}

We assume that emission probabilities are time-dependent, but homogeneous regarding individuals, and we use an identity link multinomial emission model with 28 parameters θ_i^{em} (Table 2).

 ISSN: 2102-6238

TABLE 2. *Structure of the multinomial emission model. On the left : the emission, on the right, the parameter.*

date 1				date 2				date 3				date 4 or 5			
$(0,\cdot)$	$\to$	0	–	$(0,0)$	$\to$	$(0,0)$	–	$(0,0)$	$\to$	0	–	$(0,\cdot)$	$\to$	0	–
$(0,\cdot)$	$\to$	1	θ_1^{em}	$(0,0)$	$\to$	$(0,1)$	θ_5^{em}	$(0,0)$	$\to$	1	θ_{25}^{em}	$(0,\cdot)$	$\to$	1	θ_{21}^{em}
$(0,\cdot)$	$\to$	2	θ_2^{em}	$(0,0)$	$\to$	$(1,0)$	θ_6^{em}	$(0,0)$	$\to$	2	θ_{26}^{em}	$(0,\cdot)$	$\to$	2	θ_{22}^{em}
$(1,\cdot)$	$\to$	0	θ_3^{em}	$(0,0)$	$\to$	$(1,1)$	θ_7^{em}	$(1,\cdot)$	$\to$	0	θ_{27}^{em}	$(1,\cdot)$	$\to$	0	θ_{23}^{em}
$(1,\cdot)$	$\to$	1	–	$(0,0)$	$\to$	$(2,0)$	θ_8^{em}	$(\cdot,1)$	$\to$	0	θ_{27}^{em}	$(1,\cdot)$	$\to$	2	θ_{24}^{em}
$(1,\cdot)$	$\to$	2	θ_4^{em}	$(0,1)$	$\to$	$(0,0)$	θ_9^{em}	$(1,\cdot)$	$\to$	1	–	$(1,\cdot)$	$\to$	1	–
				$(0,1)$	$\to$	$(0,1)$	–	$(\cdot,1)$	$\to$	1	–				
				$(0,1)$	$\to$	$(1,0)$	θ_{10}^{em}	$(1,\cdot)$	$\to$	2	θ_{28}^{em}				
				$(0,1)$	$\to$	$(1,1)$	θ_{11}^{em}	$(\cdot,1)$	$\to$	2	θ_{28}^{em}				
				$(0,1)$	$\to$	$(2,0)$	θ_{12}^{em}								
				$(1,0)$	$\to$	$(0,0)$	θ_{13}^{em}								
				$(1,0)$	$\to$	$(0,1)$	θ_{14}^{em}								
				$(1,0)$	$\to$	$(1,0)$	–								
				$(1,0)$	$\to$	$(1,1)$	θ_{15}^{em}								
				$(1,0)$	$\to$	$(2,0)$	θ_{16}^{em}								
				$(1,1)$	$\to$	$(0,0)$	θ_{17}^{em}								
				$(1,1)$	$\to$	$(0,1)$	θ_{18}^{em}								
				$(1,1)$	$\to$	$(1,0)$	θ_{19}^{em}								
				$(1,1)$	$\to$	$(1,1)$	–								
				$(1,1)$	$\to$	$(2,0)$	θ_{20}^{em}								

4.3. Implementation issues.

Implementation carries out in C language (with GNU GCC compilator, using the MT19937 random number generator). The maximization of the objective function in a high dimension appears to be a critical part of the algorithm. We adopt a three-step approach. First, a local exploration of the parameters' space is made by randomly drawing a few maximization directions. Then, we use the analytical expression of the Hessian and perform a Newton-Raphson algorithm limited to a few number of iterations. On this occasion, problems of local identification may arise. So we compute the condition number of the Hessian matrix at each step. If it appears to be too large, we remove a certain critical parameter before the inversion of the matrix and this parameter cannot be identified at this step. At last, if the Newton-Raphson algorithm fails to converge towards a zero-gradient point, we carry out a simple maximization in the direction of the gradient.

4.4. Results and discussion.

The univariate analysis with 100 sub-samples of 1,000 female subjects suggests the initial social level has an effect on cancer, with a protection effect for upper social classes (see Figure 2). Unfortunately, the empirical distribution does not allow us to conclude to a 95 % meaningful effect. An interesting result is that, when a second variable (cumulated number of cigarettes) is introduced, no noticeable modification is observed concerning the effect of social class, while the number of cigarettes smoked appears to have an effect as expected (though the empirical distribution does not allow to conclude to a 95% meaningful effect). This result leads to the idea that social class has an influence on cancer which is not necessarily directly related to health

 ISSN: 2102-6238

behaviours. This should be confirmed with further epidemiological investigations, especially with other types of behaviour, such as alcohol consumption.

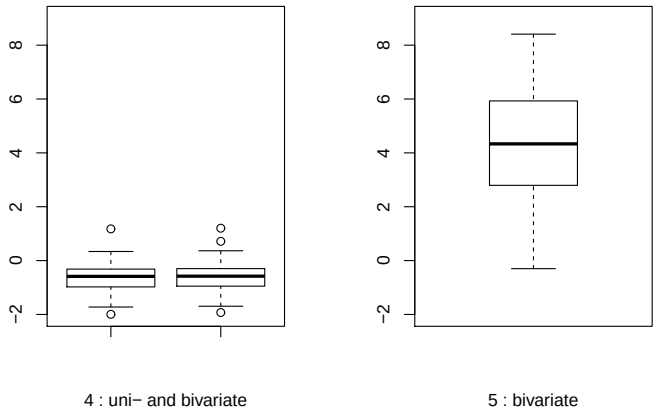

FIGURE 2. *Uni- and bivariate models for mediation investigation (NCDS data, 100 sub-samples with 1000 subjects ; 4: parameters for early social level ; 5: parameter for cumulative number of smoked cigarettes).*

We are fully aware of the methodological limits of our statistical interpretation : estimations are not enough meaningful regarding the empirical distribution; the number of sub-sample is low, and re-sampling is not made with the same number of subjects as in the initial data (as it should be in a bootstrap approach), which could lead to an underestimation of the variability of the estimates. A solution for improving such shortcomings would be to improve the compensation for errors and missing data in the cancer outcome by introducing information about the disease incidence. Such an information is available in the cancer registers held in Great Britain. However, although the idea appears to be simple, such an improvement is difficult to implement, essentially for practical reasons, and requires further investigation.

Although not the aim of this work, goodness-of-fit indications would be interesting in this section. However, assessing goodness-of- fit of MHMM is difficult (as noticed in Altman (2004)). This topic is addressed in Lystig (2001), or in Titman and Sharples (2008). For lack of being able to rigorously test the model goodness of fit, visual predictive checks (VPCs) could be used (see Holford (2005); Post et al. (2006) for a concise presentation of VPCs and Delattre and Lavielle (2012) for its application in MHMM setting).

5. Simulation study.

In this section, we perform a simulation study in order to assess, first the estimation procedure and second the robustness of the model (M). We keep the design of the NCDS 1958 cohort and we still work with the MHMM of Figure 1.

 ISSN: 2102-6238

5.1. Data generation and model assessment.

In order to empirically assess the properties of the estimation algorithm, we generate 100 simulated samples from (M) with 1000 subjects. We use three covariates in the transition model, among which two have an effect on the disease incidence. We set the disease incidence parameters so that we have enough subjects in each sample (arbitrarily around 200 to be coherent with real data). The true values of the response parameters are set so that the individual response behaviour could be considered as realistic. For instance, we fix $\theta_{22}^{em} = 0.20$ which corresponds to 10 % of non-response among people who are not affected by the disease on dates $d = 4$ or $d = 5$. We generate a first data set based on a "reliable response" assumption (with low errors probabilities and low non-response rates) and another dataset based on a "non-reliable response" assumption (with raised error and non-response probabilities). The true values of the parameters for each scenario are pointed in Table 4.

In order to empirically assess the robustness of (M) as regards misspecification on the response model, we generate samples from a model denoted (M_σ) with individual fluctuations in response probabilities. The transition part of (M_σ) is the same as the one of M and thus we know the true value of the parameters and could estimate biases. The emission part is a GLM model with individual random error terms $V_{i,n,d}$ (i being a parameter index, n a subject index and d a time index). We denote $\sigma^2 = \text{var}(V_{i,n,d})$ and we make σ vary within the set $\{0.5, 1, 2, 4, 8\}$. This emission model includes an additive error assumption, which is described in detail in Table 3.

TABLE 3. *Structure of the simulation GLM emission model (with error additivity assumption). On the left : the emission, on the right, the parameters.*

Date 1		Date 2		Date 3		Date 4 or 5	
$(0,\cdot) \to 0$	0	$(0,0) \to (0,0)$	0	$(0,0) \to 0$	0	$(0,\cdot) \to 0$	0
$(0,\cdot) \to 1$	$\theta_1^{em} + V_1$	$(0,0) \to (0,1)$	$\theta_9^{em} + V_5$	$(0,0) \to 1$	$\theta_{15}^{em} + V_{21}$	$(0,\cdot) \to 1$	$\theta_{11}^{em} + V_{25}$
$(0,\cdot) \to 2$	$\theta_2^{em} + V_2$	$(0,0) \to (1,0)$	$\theta_5^{em} + V_6$	$(0,0) \to 2$	$\theta_{16}^{em} + V_{22}$	$(0,\cdot) \to 2$	$\theta_{12}^{em} + V_{26}$
$(1,\cdot) \to 0$	$\theta_3^{em} + V_3$	$(0,0) \to (1,1)$	$\theta_5^{em} + \theta_9^{em} + V_7$	$(1,\cdot) \to 0$	$\theta_{17}^{em} + V_{23}$	$(1,\cdot) \to 0$	$\theta_{13}^{em} + V_{27}$
$(1,\cdot) \to 1$	0	$(0,0) \to (2,0)$	$\theta_6^{em} + V_8$	$(\cdot,1) \to 0$	$\theta_{17}^{em} + V_{23}$	$(1,\cdot) \to 1$	0
$(1,\cdot) \to 2$	$\theta_4^{em} + V_4$	$(0,1) \to (0,0)$	$\theta_{10}^{em} + V_9$	$(1,\cdot) \to 1$	0	$(1,\cdot) \to 2$	$\theta_{14}^{em} + V_{28}$
		$(0,1) \to (0,1)$	0	$(\cdot,1) \to 1$	0		
		$(0,1) \to (1,0)$	$\theta_5^{em} + \theta_{10}^{em} + V_{10}$	$(1,\cdot) \to 2$	$\theta_{18}^{em} + V_{24}$		
		$(0,1) \to (1,1)$	$\theta_5^{em} + V_{11}$	$(\cdot,1) \to 2$	$\theta_{18}^{em} + V_{24}$		
		$(0,1) \to (2,0)$	$\theta_6^{em} + V_{12}$				
		$(1,0) \to (0,0)$	$\theta_7^{em} + V_{13}$				
		$(1,0) \to (0,1)$	$\theta_7^{em} + \theta_9^{em} + V_{14}$				
		$(1,0) \to (1,0)$	0				
		$(1,0) \to (1,1)$	$\theta_1^{em} + \theta_9^{em} + V_{15}$				
		$(1,0) \to (2,0)$	$\theta_8^{em} + V_{16}$				
		$(1,1) \to (0,0)$	$\theta_7^{em} + \theta_{10}^{em} + V_{17}$				
		$(1,1) \to (0,1)$	$\theta_7^{em} + V_{18}$				
		$(1,1) \to (1,0)$	$\theta_{10}^{em} + V_{19}$				
		$(1,1) \to (1,1)$	0				
		$(1,1) \to (2,0)$	$\theta_8^{em} + V_{20}$				

Since we aim to compare the estimation efficiency with different level of σ, we adapt the true

 ISSN: 2102-6238

values of the parameters so that frequencies of observed responses are approximatively identical for all possible values of σ. We empirically manage to obtain around 30 false cancers, 10 false non-cancer, and 600 non-response for the reliable response scenario and around 130 false cancers, 60 false non-cancer, and 1000 non-response for the non-reliable response scenario. We perform 700 SEM iterations for each of the 100 samples of 1000 subjects.

5.2. Results.

The pointwise estimation for each sample derives from average on the last 150 stochastic estimations. Our results are represented as box-plots for transition parameters. We represent in Figure 3 the SEM iterations of stochastic estimates for standardized parameters θ_4, θ_5 and θ_6 which are the parameters of interest in the direct causal inference approach. The stochastic estimator for θ_6 has a high variability, due to the fact that this parameter has no effect on the outcome (θ_6 has a null true value).

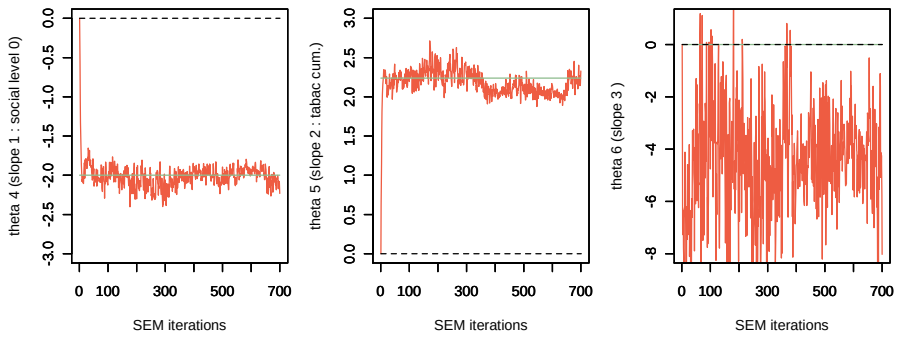

FIGURE 3. *Examples of SEM iterations for one sample (standardized parameters, reliable response simulated data).*

Estimations of the parameters and their 95% confidence intervals from 100 simulated population samples are collected in Table 4. The computational time is long (a few days). The θ_6 estimation appears to have a very large confidence interval including zero, and to be meaningless. Convergence for θ_4 and θ_5 seems to be achieved in a few iterations. We can notice that the biases are low. Estimations for the response model for parameters θ_3^{em}, θ_4^{em}, θ_9^{em}, θ_{10}^{em}, θ_{11}^{em}, θ_{12}^{em}, θ_{17}^{em}, θ_{19}^{em}, θ_{20}^{em} has a confidence interval from 0 to 1, which make the estimations unreliable. However, note that θ_3^{em} and θ_4^{em} correspond to responses from state 1 for $d = 1$, and at this time most subject are not affected by the disease as we fix low initial probabilities θ^{ini}. This explains the bad estimations of the initial emission parameters. Moreover, θ_9^{em}, θ_{10}^{em}, θ_{11}^{em}, θ_{12}^{em}, θ_{17}^{em}, θ_{19}^{em}, θ_{20}^{em} correspond to responses for $d = 2$ from states with a memory of some past disease. For the same reason, this only concerns a few subjects. This also explains the bad estimations obtained for these parameters.

 ISSN: 2102-6238

TABLE 4. *Real values, estimations and 95 % percentile-based confidence interval (I_{95}) from simulated data (100 samples).*

	M_1 (reliable responses data)			M_1 (non reliable responses data)		
param.	true value	estim.	I_{95}	true value	estim.	I_{95}
θ_1^{ini}	0.0001	0.0040	(0.0000, 0.0620)		0.0098	(0.0009, 0.0598)
θ_2^{ini}	0.0002	0.0000	(0.0000, 0.0015)		0.0000	(0.0000, 0.0015)
θ_1^{trans}	−1.0	−0.8	(−1.3, 2.9)		−0.7	(−1.6, 2.9)
θ_2^{trans}	−1.0	−0.8	(−1.5, 4.7)		−0.1	(2.0, 4.9)
θ_3^{trans}	−3.8	−3.9	(−4.2, −3.6)		−3.9	(−4.4, −3.6)
θ_4^{trans}	−1.0	−0.8	(−1.6, −0.3)		−0.7	(−1.4, −0.2)
θ_5^{trans}	1.15	1.00	(0.44, 1.62)		0.85	(0.29, 1.51)
θ_6^{trans}	0	−0.3	(−2.4, 1.7)		−0.1	(−2.3, 2.3)
θ_7^{trans}	−4.1	−3.1	(−4.4, −1.8)		−2.6	(−3.8, −1.4)
θ_8^{trans}	−4.2	−3.5	(−8.4, −2.7)		−3.3	(−7.3, −2.0)
θ_9^{trans}	−4.3	−3.1	(−3.9, −1.9)		−3.0	(−6.7, −2.0)
θ_{10}^{trans}	−4.4	−3.3	(−4.8, −2.2)		−3.0	(−7.2, −2.0)
θ_1^{rand}	2.5	1.9	(0.6, 2.8)		0.9	(0.6, 2.2)
θ_1^{em}	0.005	0.004	(0.000, 0.010)	0.005	0.004	(0.000, 0.008)
θ_2^{em}	0.15	0.149	(0.126, 0.169)	0.200	0.198	(0.169, 0.223)
θ_3^{em}	0.01	0.000	(0.000, 1.000)	0.001	0.000	(0.000, 1.000)
θ_4^{em}	0.1	0.3	(0.0, 1.0)	0.10	0.33	(0.00, 1.00)
θ_5^{em}	0.005	0.0047	(0.0000, 0.0102)	0.0500	0.0503	(0.0318, 0.0653)
θ_6^{em}	0.005	0.0039	(0.0000, 0.0115)	0.060	0.057	(0.028, 0.075)
θ_7^{em}	0.005	0.004	(0.001, 0.010)	0.003	0.001	(0.000, 0.007)
θ_8^{em}	0.2	0.203	(0.177, 0.230)	0.2500	0.2507	(0.2199, 0.2798)
θ_9^{em}	0.01	0.00	(0.00, 1.00)	0.10	0.00	(0.00, 1.00)
θ_{10}^{em}	0.01	0.00	(0.00, 1.00)	0.04	0.00	(0.00, 1.00)
θ_{11}^{em}	0.005	0.000	(0.000, 1.000)	0.03	0.00	(0.00, 0.99)
θ_{12}^{em}	0.1	0.00	(0.00, 1.00)	0.30	0.00	(0.00, 1.00)
θ_{13}^{em}	0.05	0.088	(0.000, 0.281)	0.15	0.23	(0.00, 0.42)
θ_{14}^{em}	0.01	0.004	(0.00, 0.037)	0.03	0.02	(0.00, 0.09)
θ_{15}^{em}	0.005	0.000	(0.000, 0.026)	0.08	0.06	(0.00, 0.20
θ_{16}^{em}	0.2	0.186	(0.022, 0.331)	0.35	0.33	(0.16, 0.49)
θ_{17}^{em}	0.01	0.00	(0.00, 1.00)	0.03	0.00	(0.00, 1.00)
θ_{18}^{em}	0.015	0.000	(0.000, 0.040)	0.20	0.00	(0.00, 1.00)
θ_{19}^{em}	0.005	0.000	(0.000, 1.000)	0.01	0.00	(0.00, 1.00)
θ_{20}^{em}	0.1	0.00	(0.00, 1.00)	0.30	0.00	(0.00, 1.00)
θ_{21}^{em}	0.005	0.00	(0.000, 0.003)	0.010	0.007	(0.000, 0.29)
θ_{22}^{em}	0.1	0.101	(0.085, 0.116)	0.200	0.202	(0.177, 0.228)
θ_{23}^{em}	0.02	0.06	(0.00, 0.57)	0.20	0.30	(0.00, 0.54)
θ_{24}^{em}	0.05	0.066	(0.000, 0.140)	0.25	0.24	(0.02, 0.45)
θ_{25}^{em}	0.005	0.000	(0.000, 0.260)	0.01	0.00	(0.00, 0.03)
θ_{26}^{em}	0.1	0.09	(0.08, 0.11)	0.200	0.198	(0.165, 0.237)
θ_{27}^{em}	0.03	0.033	(0.000, 0.153)	0.20	0.25	(0.00, 0.45)
θ_{28}^{em}	0.05	0.048	(0.000, 0.107)	0.25	0.23	(0.05, 0.40)

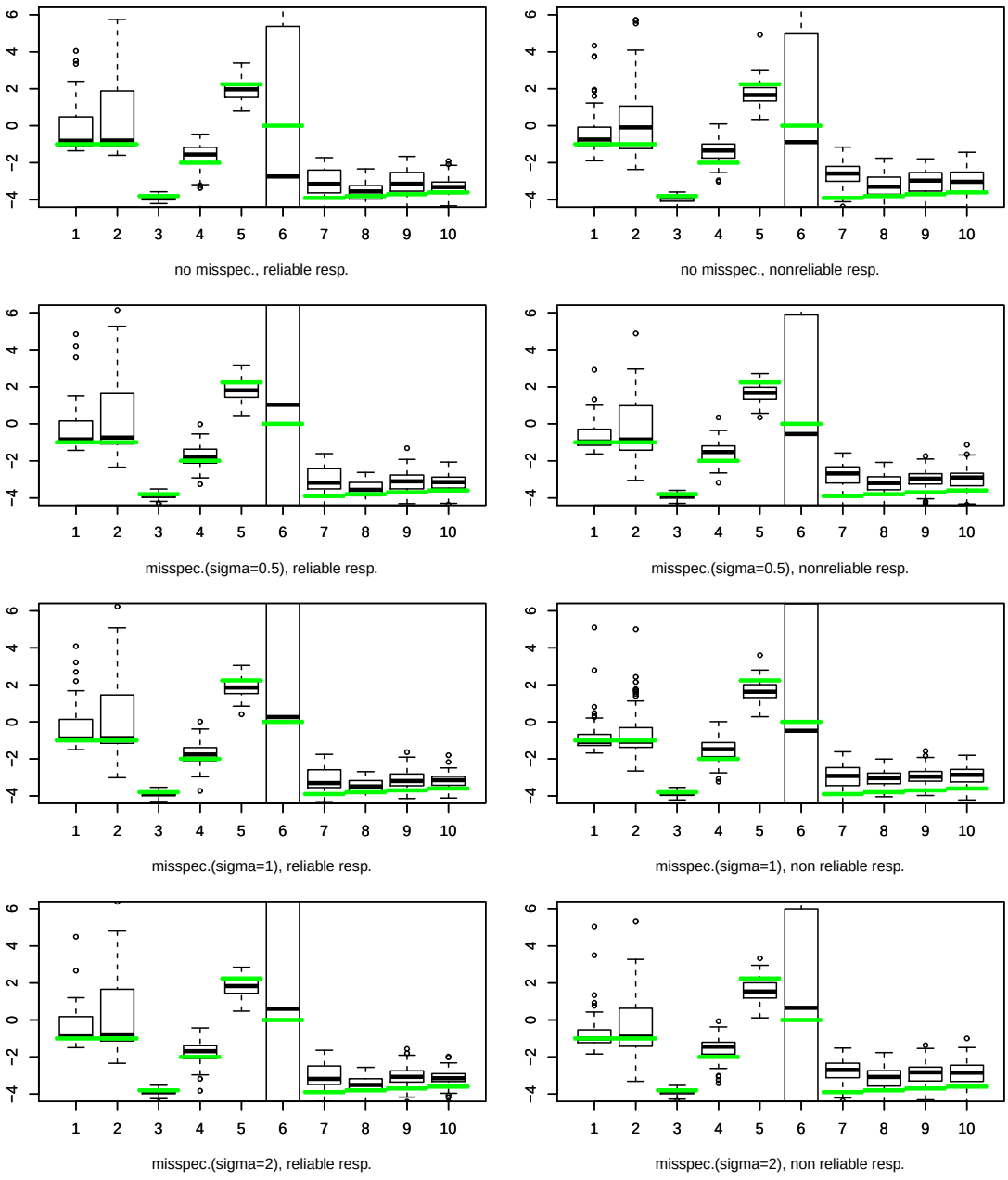

FIGURE 4. *Box-plots of the transition parameters estimations from 100 simulated samples and different scenarios: (heterogeneity in response probabilities (no misspec., misspec. and its intensity (sigma=0.5; 1; 2) and reliability or not of the response). The true value of the parameter is indicated by the coloured horizontal bar.*

 ISSN: 2102-6238

Box-plots (Figure 4) illustrate to which extend misspecification (shaped as individual random terms with variance σ^2 in a GLM response model) and reliability of the responses has an effect on estimations. It appears that the variations of σ have a very little effect on the transition probabilities estimations. It is important to recall that for each value of σ we adapt the true value of the response parameters so as to achieve the same proportions of errors and non-responses as in the M simulated data (which is necessary in order to make meaningful comparisons). Then σ only introduces heterogeneity in the response probabilities. We conclude that (M) is robust as regards this type of heterogeneity. Concerning the effect of response reliability, as expected, we observe that raised rates of error and non-response have an impact on the quality of transition parameters estimations. However, even with rates such as 13 % of "false negatives" (subjects declaring they have no cancer at t_4 or t_5 though they actually have had cancer) estimations for transition parameters remain acceptable.

5.3. Discussion.

We present some simulations with a model (M) designed to fit the longitudinal structure of the NCDS 1958 study. Completing the objective function maximization appears to be a difficult task due to the high dimension of the parameters' space. As the Quasi-Newton algorithm fails, we propose a three-step algorithm which appears to be robust. However, its execution is somewhat expensive from a computational viewpoint. That is why we avoid using a GLM model for emissions.

A solution for improving estimations would be to inject some information about the disease incidence on the general population (assumed to be known by external ways). We observe (through simulations which are not described in this paper) that if intercepts of the GLM transition model are known, then the biases and dispersions of the estimation can be drastically reduced as expected. If we assume (without loss of generality) that the observed and unobserved covariates $\mathbf{X}_{n,d}$ and $\mathbf{W}_n$ have a null population mean value, and if we assume that for each individual n, $f_d(s,q,\mathbf{X}_{n,d},\mathbf{W}_n\theta^{trans})$ is close to its population mean value $\bar{f}_d(s,q,\theta^{trans})$, then the intercept term is close (with respect to a first order approximation) to $\ln \frac{\bar{f}_d(s,q,\theta^{trans})}{\bar{f}_d(s,q_0,\theta^{trans})}$ (where q_0 is some reference value). Then if the population incidence is known, and if the individual probabilities of illness are close to the population incidence, we directly inject this additional information into the model. However, outside the context of such a (questionable) approximation it seems difficult to link these intercepts with incidence indicators, due to the non-linearity of the multinomial logit. A constrained maximization of the objective function could be imagined, but it would involve major practical difficulties. The use of a linear transition additive model inspired from the additive Aalen model (see Aalen et al. (2008)) could be used, but adaptations are needed to integrate individual random effects. Further research is needed on this topic.

Our approach is based on the assumption of the Mixed Markov Hidden Model. The dynamic described by this model may be a rough first approximation of the real dynamic of the system. In fact, the assumptions of Markov transitions as well as the independence of the covariates may be violated. Nevertheless the results obtained on a real case study are relevant at least for the

 ISSN: 2102-6238

estimation of the transition parameters.

At this stage we want to draw attention to the poor results (not presented here) provided by the asymptotic estimations of the standard errors given in our estimated information matrix, although it is the standard way of deriving confidence intervals for MHMM estimations. Indeed, most of the estimators in our model present an asymptotic computed variance around 10^{-2} or even 10^{-3}, which is critically underestimated as shown by the empirical confidence intervals obtained with 100 samples. We cannot recommend strongly enough to use an alternative approach such as bootstrap. It may also be possible to use an online computation of the estimation of the Fisher matrices in the SEM algorithm as mention in Paragraph 3.3.

Acknowledgements.

Authors thank anonymous referees for their comments which widely improve the quality of the paper.
This research is supported by the joint support of "La Ligue Contre le Cancer", of the "Direction Générale de la Santé" (DGS), of the "Caisse Nationale d'Assurance Maladie des Travailleurs Salariés" (CNAMTS), of the "Régime Social des Indépendants" (RSI), of the "Caisse Nationale de Solidarité pour l'Autonomie" (CNSA), of the "Mission Recherche" of the "Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques" (MiRe-DREES) and of the "Institut national de prévention et de promotion de la santé" (Inpes), under the call for research launched by the "Institut de Recherche en Santé Publique" in 2010.

Appendix A. Validation of HMM hypotheses for the state memory model.

Fix $n \in \{1,\dots,N\}$ a subject. We are to prove that the two processes $(S^*_{n,d})_{d=1,\dots,D}$ and $(Y^*_{n,d})_{d=1,\dots,D}$ form a HMM. First, $(S_{n,d})_{d=1,\dots,D}$ is a Markov process, and then $S_{n,d}$ is independent from $S_{n,d-k}(1<k\leq d-1)$ given $S_{n,d-1}$. Secondly, as we have $S'_{n,d-1} = S'_{n,d-2} \bullet S_{n,d-1}$, the state memory $S'_{n,d-1}$ depends on $S^*_{n,d-1}$ in a deterministic way. Then it is made clear that $S^*_{n,d} = (S_{n,d}, S'_{n,d-1})$ is independent from $S^*_{n,d-k} = (S_{n,d-k}, S'_{n,d-(k+1)})$ for any $k = 1,\dots,d$ given $S^*_{n,d-1} = (S_{n,d-1}, S'_{n,d-2})$. So we have proved that $(S^*_{n,d})_{d=1,\dots,D}$ is a Markov process. Moreover given the previous assumption on $Y^*_{n,d}$, the two processes $(S^*_{n,d})_{d=1,\dots,D}$ and $(Y^*_{n,d})_{d=1,\dots,D}$ form a HMM. Lastly, it is to be observed that whenever $S'_{n,d-1} = 2$ we have $S_{n,d} = 2$. It is then possible to consider $(S^*_{n,d})_{d=1,\dots,D}$ as a five states Markov process taking values in $\{(0,0),(0,1),(1,0),(1,1),2\}$.

Appendix B. Baum - Welch Forward and Backward algorithms.

Fix $n \in \{1,\dots,N\}$ a subject. Let us recall that the following probabilities are implicitly conditioned on covariates $\mathbf{X}_{n,1},\dots,\mathbf{X}_{n,D}$. We denote $\alpha_{n,d}(s) = \mathbb{P}(S_{n,d} = s, Y_{n,1},\dots,Y_{n,d}|\mathbf{W}_n;\theta)$ and $\beta_{n,d}(s) = \mathbb{P}(Y_{n,d+1},\dots,Y_{n,D}|S_d = s,\mathbf{W}_n;\theta)$, with $\beta_{n,D}(s) = 1$. Using the local independence property, we

 ISSN: 2102-6238

have

$$\alpha_{n,d}(s) = \mathbb{P}(Y_{n,d}|S_{n,d} = s, \mathbf{W}_n; \theta) \sum_q \alpha_{n,d-1}(q)\mathbb{P}(S_{n,d} = s|S_{n,d-1} = q, \mathbf{W}_n; \theta),$$
$$\beta_{n,d}(s) = \sum_q \mathbb{P}(S_{n,d+1} = q|S_{n,d} = s, \mathbf{W}_n; \theta)\mathbb{P}(Y_{n,d+1}|S_{n,d+1} = q, \mathbf{W}_n; \theta)\beta_{n,d+1}(q),$$

which makes possible to perform a recursive computation. We can now provide the joint probabilities which appear in the definition of the objective functions :

$$\mathbb{P}^{trans}_{n,d}(s, q|Y_{n,1}, \ldots, Y_{n,D}, \mathbf{W}_n; \hat{\theta})$$
$$= \frac{\alpha_{n,d}(s) f_d(s, q, \mathbf{X}_{n,d}, \mathbf{W}_n, \hat{\theta}^{trans}) g_d(q, Y_{n,d+1}, \hat{\theta}^{em})\beta_{n,d+1}(q)}{\mathbb{P}(Y_{n,1}, \ldots, Y_{n,D}|\mathbf{W}_n; \hat{\theta})},$$
$$\mathbb{P}^{em}_{n,d}(s, y|Y_{n,1}, \ldots, Y_{n,D}, \mathbf{W}_n; \hat{\theta}) = \delta_y(Y_{n,d}) \frac{\alpha_{n,d}(s)\beta_{n,d}(s)}{\mathbb{P}(Y_{n,1}, \ldots, Y_{n,D}|\mathbf{W}_n; \hat{\theta})},$$

with $\mathbb{P}(Y_{n,1}, \ldots, Y_{n,D}|\mathbf{W}_n; \hat{\theta}) = \sum_k \alpha_{n,d}(k)\beta_{n,d}(k)$ for any d (and then $\mathbb{P}(Y_{n,1}, \ldots, Y_{n,D}|\mathbf{W}_n; \hat{\theta}) = \sum_k \alpha_{n,D}(k)$).

Appendix C. Parameters re-estimation Rabiner (1989).

Recall that θ^{em}_{s,y,d_0} is the probability of emission $s \to y$ at date d_0. With $\mathscr{D}$ being a set of time indexes including d_0 for which the $s \to y$ emissions are homogeneous, and $\mathscr{N}$ being the set of subjects indexes, the reestimation of θ^{em}_{s,y,d_0} from $\hat{\theta}$ is given by :

$$\theta^{em}_{s,y,d_0} = \frac{\sum_{d\in\mathscr{D}} \sum_{n\in\mathscr{N}} \mathbb{P}^{em}_{n,d}(s, y|Y_{n,1}, \ldots, Y_{n,D}, \mathbf{W}_n; \hat{\theta})}{\sum_{d\in\mathscr{D}} \sum_{n\in\mathscr{N}} \mathbb{P}(S_{n,d} = s|Y_{n,1}, \ldots, Y_{n,D}; \mathbf{W}_n; \hat{\theta})}.$$

This quantity is computed using the forward and backward quantities:

$$\mathbb{P}(S_{n,d} = s|Y_{n,1}, \ldots, Y_{n,D}; \mathbf{W}_n; \hat{\theta}) = \frac{\alpha_{n,d}(s)\beta_{n,d}(s)}{\sum_k \alpha_{n,d}(k)\beta_{n,d}(k)},$$

$$\text{and} \quad \mathbb{P}^{em}_{n,d}(s, y|Y_{n,1}, \ldots, Y_{n,D}, \mathbf{W}_n; \hat{\theta}) = \mathbb{1}_y(Y_{n,d}) \frac{\alpha_{n,d}(s)\beta_{n,d}(s)}{\sum_k \alpha_{n,d}(k)\beta_{n,d}(k)}.$$

As for the initial probabilities, we denote θ^{ini}_s the probability of state s on date $d = 1$. The re-estimation of θ^{ini}_s from $\hat{\theta}$ is given by :

$$\theta^{ini}_s = \frac{1}{\text{card}(\mathscr{N})} \sum_{n\in\mathscr{N}} \mathbb{P}(S_{n,1} = s|Y_{n,1}, \ldots, Y_{n,D}, \mathbf{W}_n; \hat{\theta}) = \frac{1}{\text{card}(\mathscr{N})} \sum_{n\in\mathscr{N}} \frac{\alpha_{n,1}(s)\beta_{n,1}(s)}{\sum_k \alpha_{n,1}(k)\beta_{n,1}(k)}.$$

 ISSN: 2102-6238

Appendix D. The Metropolis-Hastings (MH) algorithm.

The MH algorithm is applied to each subject n so as to draw a random effect vector $\tilde{\mathbf{W}}_{k,n}$ (k being the index of SEM iterations) from the conditional law $\mathbb{P}(\tilde{\mathbf{W}}_{k,n}|Y_{n,1},\ldots,Y_{n,D};\hat{\theta}_k)$. We denote $\mathbf{R}_{j,n}$ the vector drawn at the jth iteration of the MH algorithm. At step $j+1$, we first draw a random effects vector $\mathbf{R}^*_{j+1,n}$ from a Gaussian "proposal" density with 0 mean and some fixed variance. Then we compare the proposition $\mathbf{R}^*_{j+1,n}$ with $\mathbf{R}_{j,n}$ using the following quotient (r being the multinormal density) :

$$a = \frac{\mathbb{P}(Y_{n,1},\ldots,Y_{n,D}|\mathbf{R}^*_{j+1,n};\hat{\theta}_k)r(\mathbf{R}^*_{j+1,n},\hat{\theta}^{rand}_k)}{\mathbb{P}(Y_{n,1},\ldots,Y_{n,D}|\mathbf{R}_{j+1,n};\hat{\theta}_k)r(\mathbf{R}_{j+1,n},\hat{\theta}^{rand}_k)}$$

We compare a with a drawing α from the uniform density over $[0;1]$. If $a > \alpha$ then we accept the proposition $\mathbf{R}^*_{j+1,n}$ and we have $\mathbf{R}_{j+1,n} = \mathbf{R}^*_{j+1,n}$. In the contrary ($a < \alpha$) we reject the proposition and we have $\mathbf{R}_{j+1,n} = \mathbf{R}_{j,n}$. The recursion may be initialized with a null vector $R_{0,n}$. After a certain burn-in period of B iterations, we consider $\mathbf{R}_{B+1,n}$ as a drawing of $\tilde{\mathbf{W}}_{k,n}$. We define B so that the acceptance rate of the algorithm be around 0.3 for example Chib and Greenberg (1995).

References

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis*. Statistics for Biology and Health. Springer, New York.

Albert, P. (2000). A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics*, 56(2):602–608.

Altman, R. (2004). Assessing the goodness-of-fit of hidden Markov models. *Biometrics*, 60(2):444–450.

Altman, R. (2007). Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477):201–210.

Bartolucci, F., Pennoni, F., and Francis, B. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 170(1):115–132.

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Bureau, A., Shiboski, S., and Hughes, J. (2003). Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine*, 22(3):441–462.

Celeux, G. and Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports*, 41(1-2):119–134.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):27–335.

Cho, L., Lian, L., JaeJeong, Y., SoungHoon, C., KeunYoung, Y., and Park, S. (2009). Validation of self-reported cancer incidence at follow-up in a prospective cohort study. *Annals of Epidemiology*, 19(9):644–Ű646.

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis*, 5(4):315–327.

Commenges, D. (2002). Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research*, 11(2):167–182.

Delattre, M. (2010). Inference in mixed hidden Markov models and applications to medical studies. *Journal de la Société Française de Statistique*, 151(1):90–105.

Delattre, M. and Lavielle, M. (2012). Maximum likelihood estimation in discrete mixed hidden Markov models using the SAEM algorithm. *Comput. Statist. Data Anal.*, 56(6):2073–2085.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128.

 ISSN: 2102-6238

Dempster, A., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1).

Detilleux, J. (2008). The analysis of disease biomarker data using a mixed hidden Markov model. *Genetics, Selection, Evolution*, 40(5):491–509.

Diebolt, J. and Ip, E. (1996). *A stochastic EM algorithm for approximating the maximum likelihood estimate, in Markov chain Monte Carlo in practice*. Chapman and Hall, Dordrect, The Netherlands.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman and Hall, Dordrect, The Netherlands.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Chapman and Hall, Dordrecht, The Netherlands.

Goldberg, M., Leclerc, A., Bonenfant, S., Chastang, J., Schmaus, A., Kaniewski, N., and Zins, M. (2007). Cohort profile: the GAZEL cohort study. *International journal of epidemiology*, 36(1):32–9.

Hagenaars, J. A. and McCutcheon, A. L., editors (2002). *Applied latent class analysis*. Cambridge University Press, Cambridge.

Holford, N. (2005). The visual predictive check superiority to standard diagnostic plots. In *Proccedings of the "Population Approach Group in Europe" meeting*.

Jackson, C., Sharples, L., Thompson, S., Duffy, S., and Couto, E. (2003). Multistate Markov models for disease progression with classification error. *Journal of The Royal Statistical Society Series D (the Statistician)*, 52:193–209.

Kelly-Irving, M., Lepage, B., Dedieu, D., Lacey, R., Cable, N., Bartley, M., Blane, D., Grosclaude, P., Lang, T., and Delpierre, C. (2012). Childhood adversity as a risk for cancer. Findings from the 1958 british birth cohort study. Under review for BMC Public Health.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

Lystig, T. (2001). *Evaluation of hidden Markov models*. PhD thesis, University of Washington.

Manjer, J., Merlo, J., and Berglund, G. (2004). Validity of self-reported information on cancer: Determinants of under- and over-reporting. *European Journal of Epidemiology*, 19(3):239–247.

Navarro, C., Chirlaque, M., Tormo, M., Pérez-Flores, D., Rodríguez-Barranco, M., Sánchez-Villegas, A., Agudo, A., Pera, G., Amiano, P., Dorronsoro, M., Larrañaga, N., Quirós, J., Ardanaz, E., Barricarte, A., Martínez, C., Sánchez, M., Berenguer, A., and González, C. (2006). Validity of self reported diagnoses of cancer in a major spanish prospective cohort study. *Journal of Epidemiology and Community Health*, 60(7):593–599.

Nielsen, S. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli*, 6(3):457–489.

Panhard, X. and Samson, A. (2008). Extension of the SAEM algorithm for nonlinear mixed models with 2 levels of random effects. *Biostatistics*, 10(1):121–135.

Post, T., Freijer, J., Winter, W., and Ploeger, B. (2006). Accurate interpretation of the visual predictive check in order to evaluate model performance. In *Proccedings of the "Population Approach Group in Europe" meeting*.

Power, C. and Elliott, J. (2006). Cohort profile: 1958 british birth cohort (national child development study). *International journal of epidemiology*, 35(1):34–41.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Satten, G. and Longini, I. (1996). Markov chains with measurement error: Estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(3):275–309.

Titman, A. and Sharples, L. (2008). A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine*, 27(12):2177–2195.

Vermunt, J., Langeheine, R., and Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24(2):179–207.

Zhang, Q., Snow J., A., Rijmen, F., and Ip, E. (2010). Multivariate discrete hidden Markov models for domain-based measurements and assessment of risk factors in child development. *Journal of Computational and Graphical Statistics*, 19(3):746–765. With supplementary material available online.

 ISSN: 2102-6238