Processus latent à mémoire longue pour l'analyse de données de Qualité de Vie longitudinales

Latent Processes with Long Range Memory for Longitudinal Quality of Life Data

Mounir Mesbah[1], Rachid Senoussi[2]

[1]Sorbonne University. Statistique (LPSM, UMR 8001) BC 158, 4 place Jussieu, 75252 Paris, France. mounir.mesbah@sorbonne-universite.fr

[2] Biostatistique et Processus Spatiaux (BioSP), INRAE, Site Agroparc, 84914 Avignon, France. rachid.senoussi@inrae.fr

RÉSUMÉ. Dans cet article, nous présentons une méthode basée sur les modèles à variables latentes pour décrire l'évolution longitudinale de qualité de vie liée à la santé de patients se trouvant dans des conditions spécifiques. Tout d'abord, nous traitons le cas fréquent où des questionnaire différents sont utilisés séquentiellement pour mesurer un même trait latent pendant une longue période de suivi. Dans un second temps, nous proposons des modèles où le processus latent peut tenir compte d'un effet de longue mémoire, la qualité de vie d'un individu pouvant fortement dépendre de ses lointains antécédents. Pour cela, nous avons construit un cadre statistique général et donné la formule de vraisemblance correspondante. Nous proposons ensuite un algorithme d'approximation de cette vraisemblance, écrit dans le langage R, et l'avons appliqué à un ensemble de données réelles. Les résultats statistiques obtenus pour ce jeu de données étayent la pertinence de cette approche concernant certaines hypothèses rationnelles de test, la cohérence des valeurs estimées des paramètres ainsi que sa robustesse vis-à-vis des changements de protocole de mesure.

ABSTRACT. In this paper, we present a latent based method to model the longitudinal evolution of Health related quality of life of patients under specific survey conditions. First of all, we will deal with the frequent issue when different questionnaires are sequentially used to measure the same latent trait during a long follow up time. Secondly, we propose models allowing the latent process to potentially behave under a long range memory constraint as the quality of life of an individual can highly depend on his or her far antecedents. For that purpose, we constructed a general statistical framework and gave the corresponding likelihood formula. Then, we developed an approximation algorithm for the likelihood, within the R-software, and applied it to a real data set. The statistical results obtained for this data set substantiate the following points : The pertinence of this approach concerning some rational testing hypotheses, the compliance of the parameter estimate values as well as it robustness with respect to measurement protocol changes.

MOTS-CLÉS. Qualité de vie liée à la santé, processus latent, mémoire à longue portée, courbe de fiabilité backward, modèle de Rasch, processus d'Ornstein-Uhlenbeck.

KEYWORDS. Health Related Quality of Life, Latent process, Long range memory, Backward reliability curve, Rasch model, Ornstein-Uhlenbeck process.

1. Introduction

Computerized monitoring of patients'health status is an important issue in modern medicine. Aging and medical progress allow more and more people, with one or several chronic conditions, to live longer. Increasing costs of health care in institutions prompt the health officials to find alternative solutions to treat patients in their own homes. However, this alternative requires the monitoring of complex processes to follow a patient's health status. Among the monitoring tools, the patient-reported outcome (PRO) measures any aspect of the health status that comes directly from the patient, without any preliminary interpretation of the patient's responses by a physician or anyone else. PRO measurement system is mostly of the time assessed through a self-questionnaire, where item, question, or variable responses are often categorical. Health related quality of life (HRQoL) measurement is one of the most famous PRO measurement tools.

It is recognized that health is a latent and multi-component concept. HRQoL is one of these components, and has itself a multi-dimensional structure in general. In practice, each dimension is usually assessed through one or, more frequently, several questions. In this paper, we focus on measurements related to a single dimension of the HRQoL component, in the special case of questions with categorical ordinal response choices, e.g. *(Not at all, a little, half and half, a lot of, completely)*. For sake of completeness and comparison, we present in Section 2.1, three measurement models, the classical one for quantitative responses and two more recent models adapted to ordinal responses. These models are parametric ones describing the distribution of the observed items conditionally to the underlying latent value.

In a longitudinal context, the patients are interviewed about their health, during different visits, by fulfilling a questionnaire to provide current information on some latent trait of their health. However, an important issue often arises in long longitudinal studies, where for various reasons, different (or updated) questionnaires, meaning to measure the same latent trait, are sequentially carried out. Among these reasons, individuals used to answering the same questionnaire doubt of its usefulness. Therefore, they can report their last interview answers, distorting the study results with a *memory bias*. Another used reason is the investigator wish to switch to a more responsive/pertinent instrument.

In previous literature, such a situation is tackled with a preliminary step to somewhat equate the successive quality of life scores produced by successive questionnaires, for more details see von Davier & al. (2003), Kolen & Brennan (2004), El Fassi & al. (2009) and Boisson & al. (2015).

We extend in Appendix $A.1$ a non equating method proposed in Mesbah (2010) which is based on the so called Cronbach reliability coefficient (CRC). The latter tackles the issue of selecting a set of items related to the same unidimensional latent process. Actually, different methods mostly based on classical factorial analysis models, exist. The most known one selects in an exploratory factorial analysis including the whole set of items, the more correlated items to the first principal component. Our method is different as it uses the Backward Reliability Curve to select the pertinent items relying on the same latent variable. We performed this method on a real data set for a preliminary identification of the pertinent items.

A second crucial issue for longitudinal studies of the quality of life is that the present status of an individual often depends on his serious past health events and also on his proper social and cultural heritage as the effects of the latter one do not fade quickly over time. It is therefore natural to suppose that the quality of life of an individual at a time t, can be highly correlated with past values taken at times $s = t - \delta$, even for large time lags δ.

Due to their flexibility, Gaussian processes have been widely used as latent processes in many research areas, such as machine learning, finance, neural networks, etc. For example, in epidemiology Hall & al. (2008) proposed a latent Gaussian process to model sparse generalized longitudinal observations and took advantage of smoothing techniques to estimate the underlying mean and covariance functions under specific constraints. Xu and Ji (2014) constructed a latent Gaussian process with cyclic features to model binary outcomes (disease relapses) in monitoring clinical trials. Therein, they gave proof elements of the consistency of the posterior estimates, but also proposed a practical hybrid Monte Carlo procedure to achieve inference computations. In a discrete time context, Mesbah (2015) considered stationary latent processes obeying autoregressive dynamics of order 1 ($\mathrm{AR}_1(1)$), with regression coefficient $|\rho| < 1$, that

only allows short range correlations with exponential decay. In the context of public health surveillance, Morrison & al. (2016) presented a latent process model for forecasting multiple discrete health outcomes, e.g. Poisson mortality counts, arising from a common underlying environmental exposure (e.g. particulate pollution). In that instance, the latent process was assumed to be a discrete time AR process of order p. They also described the implementation of such a model within a hierarchical Bayesian framework using integrated nested Laplace approximations (INLA) method.

A taxonomy of most current models based on latent Gaussian processes, at least in the context of machine learning, was proposed in a well referenced review article (Li and Chen (2016). Many of the methods pointed out therein can be easily adapted to the HRQoL context.

To our knowledge, these applied models do not specify the general form of the covariance structure of the latent processes, but only propose more or less plausible formulas of the latter structure only for the date of observations. This can be a major drawback in terms of a time continuous forecasting of patients' health conditions. For that reason, we present in Section 2.2.1 two stochastic parametric models of latent processes able to adopt various reminiscence properties, namely the Ornstein-Uhlenbeck and the Fractional Brownian ones. To rationalize their introduction we briefly describe their properties, their nature and interpretation, as well as their covariance structures. Thus, we allow the quality of life latent process to be non stationary and able to keep memory of past values for very long and even infinite time lags. A more detailed analysis of the various asymptotic memory behaviors is given in Appendix 3 for Ornstein-Uhlenbeck processes.

In Section 3, we describe in detail the likelihood equations and the statistical methodology used to analyse a data set. For that purpose, we developed an algorithm (in R programming language) to perform the maximum likelihood procedure. Several statistical softwares, such as SAS, Stata, SPSS, and some R (packages) have been proposed to deal with the HRQoL data analysis, but none includes specific programs to tackle the little more complex covariance structure of these latent processes. We detail in Section 3.3 our algorithm which has been tested in Section 4 on a real data set related to an HIV research study. Also, after a brief description of the data set and the chosen measurement instruments, we present in Section 4, the statistical results emerging from the performance of our model. The final Section 5 is devoted to discuss some of the issues related to this approach and of its comparison with existing ones. We future therein some practical interpretations related to this data set analysis, and conclude this section by drawing attention not only to the many potential extensions but also to some limitations of a such HRQoL modeling.

2. Measurement Models for HRQoL Data Analysis

Many measurement models have been proposed in the literature, and the item response theory (IRT) is the field of psychometry devoted to that purpose. Sometimes a preliminary step in the analysis may be necessary in order to select among many available measuring variables (items) the most pertinent ones for sake of a parameter parsimony principle but also to exclude the singular parametric models -see Appendix $A.1$ for the specific context of the parallel model described below. Once chosen the measurement variable set, the specification of models including latent variables is usually carried out by defining two components :

i- a measurement model that describes the conditional probabilities linking observed variables to

latent variables.

ii- a prior probability distribution that sums up our present knowledge on these latent variables.

To be more precise, we adopt in the context of a longitudinal survey, the following notations for questionnaire responses (*i.e.* observations). Let $q_{i,r,v}$ be the response value to the random item $Q_{i,r,v}$ of individual $i \in \{1, .., , N\}$ to the question (item) $Q_r,\ r \in \{1, ..., R\}$ at visit date $v \in \{1, ..., V_i\}$ occurring at time $t_{i,v}$ and let $q_i = (q_{i,r,v})_{r=1,...,R;\ v=1,...,V_i}$ be the whole set of response values attached to individual i, in other words

$$q_i \ = \ (q_{i,1,1}, \cdots, q_{i,R,1}, q_{i,1,2}, \cdots, q_{i,R,2}, \cdots, q_{i,1,V_i}, \cdots, q_{i,R,V_i}). \qquad [1]$$

Without loss of generality, we appropriate the well-founded convenience of assuming a qualitative ordinal item Q_r to be numeric valued in $\{0, 1, ..., m_r\}$ to give it a more effective and operative sense.

Besides, sustaining the responses of an individual i, assumes the existence of a proper latent time varying variable $\theta_i(t)$. In all generality and particularly for HRQoL issues, this latent process may describe several interdependent biological, physiological, psychological aspects of the patient and is likely a multivariate process. In this paper, we however assume that a specific questionnaire is intended to only investigate the characteristics of a single (or uni-dimensional) hidden component $\theta(t)$. To be consistent with this assumption, we give in Appendix 1.2, a criterion allowing the selection of items related to a single hidden variable. We did apply it to our data set in Section 4.

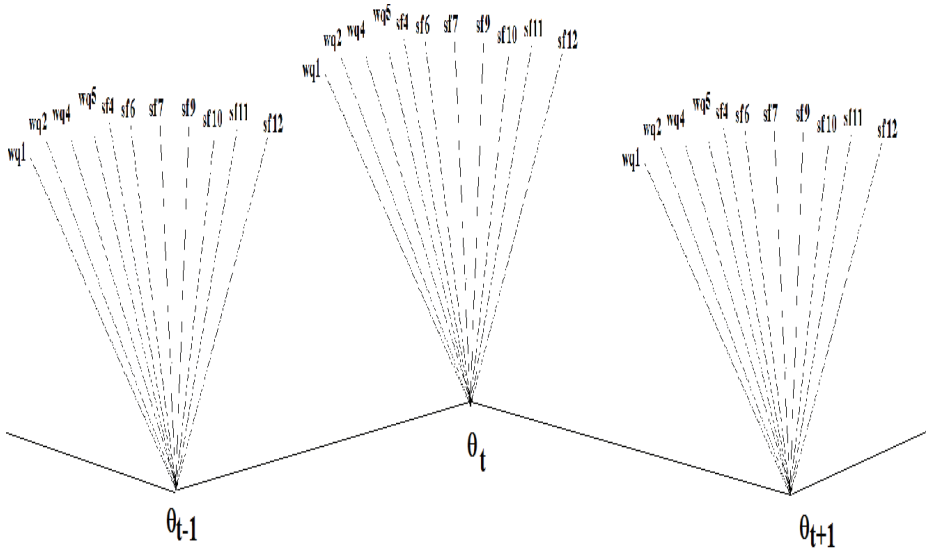

Figure 1. – Graph of the items' independence conditionally to latent process values $\theta(t_j)$ at three visit dates.

Moreover, in the context of HRQoL data analysis, additional assumptions are needed to plainly specify the joint distribution of the latent process $\theta(t)$ and that of the set of items Q_r. For example, the graphical independence model assumes that at each visit time v, conditionally to the longitudinal value $\theta_i(v)$, the items $(Q_{i,r,v}, r = 1, ..., R)$ are independent. This is generally illustrated by a figure such as Figure 1. It is also rational to assume the distribution of the observed items conditionally to the latent value is time

invariant to assert the changeless of the measure instrument. In any case, many appealing models can be shaped according to specific concerns. However in most considered studies, the models

H1- are uni-dimensional, in the sense that θ is a real valued scalar, and that all items $Q_{i,r}$, $(r = 1, ..., R)$ depend on this single latent variable θ_i.

H2- postulate the operative property that, conditionally to the latent variable, the responses are independent (as depicted in Figure 1.

H3- require the conditional expectations of the responses to be monotonic functions with respect to the latent variable.

To sum up, we assume that the individuals behave independently of each other, and that for each individual i, $i = 1, ..., N$ there exists a hidden real valued process $(\theta_i(t), t \in [0, \tau]$ such that conditionally to the event $\theta_i(t) = \theta$, the responses to items $Q_{i,r}(t)$, $r = 1, ..., R$, are also independent.

2.1. *Choosing a Conditional Measurement Model for HRQoL*

In the following, we make explicit three models, the first one is classically used for quantitative responses while the two latter ones are competitors for qualitative ordinal responses. We recall the following widely used notations $\mathbf{E}[X|Y = y]$ for the conditional expectation and the logit function defined by $logit(p) = \log(p/(1-p))$.

2.1.1. The Parallel Model (PM) for Quantitative Responses

The PM -see Lord (1952), stipulates that

$$Q_{i,r}(t) = \theta_i(t) - \beta_r + \epsilon_{i,r}(t) \quad [2]$$

where,

1. β_r stands for a time constant and fixed (non-random) effect,
2. $\theta_i(t)$ corresponds to a time varying random effect with zero mean and standard error $\sigma_\theta(t)$,
3. $\epsilon_{i,r}(t)$ is a centered random effect with a constant standard error σ, and corresponds essentially to uncontrolled measurement errors,
4. the latent $\theta_i(t)$ and the error $\epsilon_{i,r}(t))$ are uncorrelated for all t and r.

Under these assumptions, we have

$$\mathbf{E}[Q_{i,r}(t) \mid \theta_i(t) = \theta] = \theta - \beta_r. \quad [3]$$

2.1.2. The Graded Response Model (GRM) for Ordinal Responses

The GRM - see Samejima (1969), assumes that $\beta_{r,(q+1)} \geq \beta_{r,q}$ for $q = 0, ..., m_r - 1$, and

$$Prob(Q_{i,r}(t) = q \mid \theta(t) = \theta) = \frac{\exp(\theta - \beta_{r,q})}{1 + \exp(\theta - \beta_{r,q})} - \frac{\exp(\theta - \beta_{r,(q+1)})}{1 + \exp(\theta - \beta_{r,(q+1)})} \quad [4]$$

Similarly, if we set $Q^q_{i,r}(t) = \mathbf{1}_{Q_{i,r}(t) \leq q}$, with $q \in \{0, 1..., m_r\}$, we obtain

$$logit\left(\mathbf{E}\left[Q^q_{i,r}(t) \mid \theta_i(t) = \theta\right]\right) = \theta - \beta_{r,q}. \quad [5]$$

2.1.3. The Partial Credit Model (PCM) for Ordinal Responses

The PCM -see Masters (1982), is defined by :

$$Prob(Q_{i,r}(t) = q \mid \theta(t) = \theta) = \frac{\exp\left(\sum_{l=0}^{q}(\theta - \beta_{r,l})\right)}{\sum_{k=0}^{m_r}\exp(\sum_{l=0}^{k}(\theta - \beta_{r,l}))}. \quad [6]$$

Again, setting $\tilde{Q}_{i,r}^{q}(t)$= $\mathbf{1}_{Q_{i,r}(t)=q}$, yields the conditional expectation

$$logit\left(\mathbf{E}\left[\tilde{Q}_{i,r}^{q}(t) \mid \theta_i(t) = \theta \text{ and } (q-1) \leq Q_{i,r}(t) \leq q\right]\right) = \theta - \beta_{r,q}. \quad [7]$$

Note that a parsimonious parametrization of PCM is known as the Polytomous Rasch Model -see Christensen & al. (2013).

2.1.4. General Comments

Note that for both GRM and PCM, the coefficients β's, $\beta_{r,l}$ being attached to the response level l of item Q_r, are assumed to be fixed unknown real parameters. For the statistical model to be identifiable, additional linear constraints on the $\beta' s$ are needed. Usually, these are either $\beta_{r,1} = 0$ or $\sum_{l=1}^{m_r}\beta_{r,l} = 0$ for each $r = 1, \ldots, R$.

GRM and PCM, based on the handy logistic link function, are parametric models which satisfy the above mentioned properties (H1-H3). If $m_r = 1$, both models coincide with the classical binary Rasch model - see Rasch (1960). However, they somewhat differ because, the raw score (sum of item responses) of an individual at any fixed time t, is a sufficient statistic for the latent variable under the Partial Credit Model, but not under the Graded Response Model - see Dossar & Mesbah (2018). For that reason, we simply used the PCM model for our data set, even if both models proved to be both useful in practice.

Considering Formulas 3, 5 and 7 related to conditional expectation, the three models can be termed *"parallel"*, in the sense that the regression functions of responses with respect to the latent variable θ are linear provided that some logit scales to measure θ are considered.

2.2. *Choosing a Latent Process allowing Long Range Memory*

In addition to individual independence implying that latent processes $\theta_i(t)$ of individuals are independent, we assume them to be Gaussian distributed with possibly a long memory range and sharing a common linear trend $\mu(t) = \alpha t$, that is :

$$\theta_i(t) = Z_i^*(0) + \alpha t + Z_i(t). \quad [8]$$

The parameter α, which is the slope of the linear trend is deterministic. $Z_i^*(0)$ which stands for the initial random value of the latent variable is $N(0, \sigma_0^2)$ distributed. Whereas $Z_i(.)$ is a centered Gaussian process, null at time 0. Pertinent processes $Z_i(t)$ should exhibit several types of path behavior (asymptotic

stationarity, autoregressive property, long or short memory range,...). Moreover, the $\theta_i(t)$ being a specific dimension reflecting the Quality of Life of individual i at time t, its modeling is expected to allow the performance of some statistical tests related to its properties.

Since all stochastic properties of centered Gaussian processes are entirely determined by their covariance structures $C(s,t) = E(Z_i(s)Z_i(t))$, these must be flexible and richer enough to encompass a wide range of sensible behaviors with parameters easy to interpret. Among the many competitors, the Ornstein-Uhlenbeck process (OU) and the Fractional Brownian Motion (FBM) both proved their usefulness in many applied domains. In this paper, we only developed programs in the case of OU-processes. The FBM can be dealt with in exactly the same manner -for more in-depth properties of such processes, see Protter (2005), Samorodnitsky & Taqqu (1994) and Ibe (2013).

2.2.1. The Ornstein-Uhlenbeck (OU) Process

The Ornstein-Uhlenbeck process - see Ornstein & Uhlenbeck (1930), is a centered Gaussian process whose covariance function is, for $\gamma \neq 0$,

$$C^{OU}(s,t \mid \sigma,\gamma) = \frac{\sigma^2}{2\gamma}\exp(-\gamma(s+t))(\exp(2\gamma(s \wedge t)) - 1), \quad [9]$$

In the limit case $\gamma = 0$, the covariance function is written

$$C^{OU}(s,t \mid \sigma,0) = (s \wedge t)\sigma^2. \quad [10]$$

The notation $s \wedge t$ refers to the smallest value of s and t. Parameter γ expresses the strength of the autoregressive part with respect to a given time unit. In this example, γ also determines the memory range of the latent process. Parameter σ^2 specifies the importance of the erratic innovation process (white noise). To argue the choice of OU-processes, let us recall some of their properties. An OU-process obeys the stochastic differential equation : $dZ(t) = -\gamma Z(t)dt + \sigma dB(t)$ where $B(t)$ is the standard Brownian motion. This equation is proven to have a unique solution $Z(t) = \sigma e^{-\gamma t}\int_0^t e^{\gamma s}dB(s)$ if $Z(0) = 0$. When $\gamma = 0$, the process has independent increments (no memory), and restricts to the standard Brownian motion if $\sigma = 1$. When $\gamma > 0$, $Z(.)$ has a bounded variance and is asymptotically stationary. When $\gamma < 0$, the process paths explode almost surely to infinity (Figure 2).

Another property is that $Z(.)$ is a Markovian, continuous time autoregressive process, entailing that observations made at discrete times $(t_j)_{j\geq 0}$, form an AR sequence with time varying coefficients :

$$Z(t_{j+1}) = e^{-\gamma(t_{j+1}-t_j)}Z(t_j) + \sigma\eta_t(t_{j+1},t_j), \quad [11]$$

where the innovation component is given by

$$\eta_t(t_{j+1},t_j) = e^{-\gamma(t_{j+1})}\int_{t_j}^{t_{j+1}} e^{\gamma s}dB(s). \quad [12]$$

The last equation shows that innovation after time t is independent from the past values $Z(s)$, $s \leq t$ and has a time independent distribution as well. This autoregressive property is of valuable interest to account for emblematic behaviors of dynamical systems such as those related to quality of life processes.

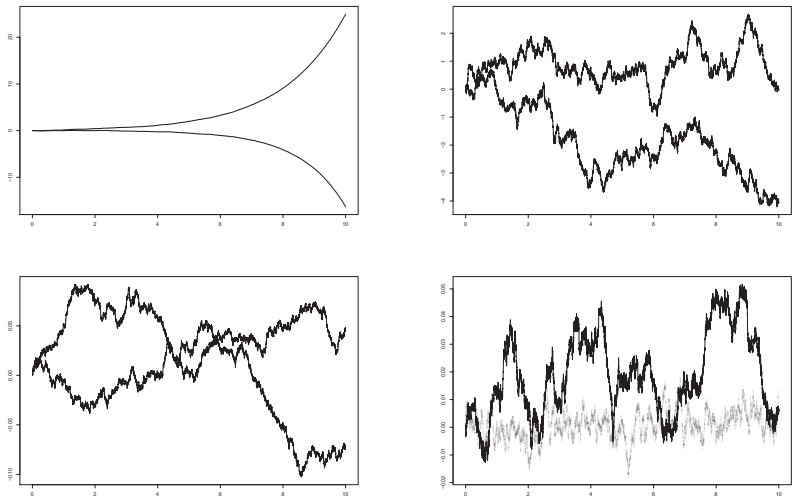

Figure 2. – Realizations of Ornstein Uhlenbeck processes with varying regression coefficients γ but fixed coefficient diffusion $\sigma = 1$. Top right : 2 paths showing an explosive autoregressive behavior $\gamma = -0.5$. Top left : 2 paths of a Brownian motion ($\gamma = 0$ without autoregressive component. Bottom right : 2 paths of an autoregressive stationary component ($\gamma = 0.1$). Bottom left : Comparison between 2 stationary autoregressive components : when γ increases (here from 0.5 to 10) the paths tends to 0.

Classical OU processes are zero valued at initial time $t = 0$, and therefore have a null variance at $t = 0$ and covariance function given by Equations 9 and 10. For the sake of a better parameter interpretation, we re-parametrize the distribution of the latent process by requiring the variance at time $t = 1$ to be σ^2, that is to replace the OU-process $Z(t)$ with $\tilde{Z}(t) = \sqrt{\frac{2\gamma}{(1-e^{-2\gamma})}} Z(t)$ when $\gamma \neq 0$.

2.2.2. The Fractional Brownian Motion (FBM)

Other stochastic models based on pertinent dynamics, eg stochastic differential equations driven by Levy processes and possibly allowing jumps could be of valuable interest. As another example, we point out the Fractional Brownian Motion (FBM) -see Samorodnitsky & Taqqu (1994), a centered and Gaussian process which is likewise totally characterized by its covariance function :

$$C^{FB}(s,t \mid \sigma,\gamma) = \frac{\sigma^2}{2}(t^{2\gamma} + s^{2\gamma} - \mid t-s \mid^{2\gamma}), \qquad [13]$$

with parameters $\gamma \in]0,1]$ and $\sigma > 0$. FBM has still the interesting properties of having stationary increments and of being self-similar i.e. $Z(at)$ and $\mid a \mid^{\gamma} Z(t)$ are identically distributed, meaning that a time unit change (days, months,...) would modify the size but not the topological characteristics of paths. The critical case $\gamma = 1/2$ also coincides with the standard Brownian motion when $\sigma = 1$, while $\gamma > 1/2$ (resp.$\gamma < 1/2$) yields positively (resp. negatively) correlated increments. For these path properties, the FBM process is popular in many applied fields and can also reveal to be very useful for HRQoL data analysis.

2.2.3. Memory Range Indexes, a Contextual Definition

Several notions related to the memory range can be defined for stochastic processes. The most classical ones refer to the various theoretical notions of mixing for stationary Markovian processes. For general Gaussian processes, which are characterized by their covariance structures, it seems natural to directly derive indexes related to the memory range from the correlation function itself $R(s,t) = C(s,t)/\sqrt{C(s,s)C(t,t)}$. For that purpose, we defined in Appendix 3.2, two types of memory indexes when the time lag increases to infinity according to two different asymptotic schemes. The first scheme (S1) identifies the ever persistent part of the correlation when s is fixed and t tends to $+\infty$, namely $\tau_1 = \lim_{t\uparrow+\infty} R(s_0,t)$. The second scheme (S2) identifies the ever persistent part of the correlation when both s and t tend to $+\infty$ with a given ratio $\nu \in]0,1[$, namely $\tau_2 = \lim_{t\uparrow+\infty} R(\nu t, t)$.

If non null, the index τ_1 indicates the correlation between the present latent state and its value at infinite horizon, whereas Index τ_2 quantifies the correlation between latent states at times t and t/ν for large t. We detailed in Appendix 3.2, the exact values of these indexes as functions of the vector parameter $(\sigma_0^2, \gamma, \sigma^2)$.

3. Statistical Inference : Methods

3.1. *Data likelihood over a fixed time interval* $[0,\tau]$

Concerning observations, let $q_{i,r,v}$ be the response of individual i to question r at visit v occurring at time $t_{i,v}$ and let $q_i = (q_{i,r,v})_{r=1,\ldots,R;\ v=1,\ldots,V_i}$ be the whole set of responses of individual i as mentioned in Equation 1. Similarly, for the hidden part, let $\theta_i = (\theta_i(t_{i,v}), v = 1,\cdots,V_i)$ denotes the set of values of the unidimensional latent process for individual i. The component $\beta = (\beta_{r,l})_{r=1,R,l=1,m_r}$ representing the set of parameters related to the questionnaire is assumed to be common to all individuals and not to vary in time.

Under these assumptions, the likelihood of the observations related to an individual i, conditionally to its latent process $\theta_i(.)$ is written :

$$L^C(q_i \mid \theta_i, \beta) = \prod_{v=1}^{V_i}\prod_{r=1}^{R} P(Q_{i,r}(t_{i,v}) = q_{i,r,v} \mid \theta_i(t_{i,v}) = \theta_{i,v}, \beta). \quad [14]$$

In the case of the Partial Credit Model 6, the conditional likelihood can be translated into :

$$L^C(q_i \mid \theta_i, \beta) = \prod_{v=1}^{V_i}\prod_{r=1}^{R} \frac{\exp(q_{i,r,v}\theta_{i,v} - \sum_{l=1}^{q}\beta_{r,l})}{1+\sum_{k=1}^{m_r}\exp(k\theta_{i,v} - \sum_{l=1}^{k}\beta_{r,l})}, \quad [15]$$

while the unconditional likelihood of responses of individual i, called individual marginal likelihood is written :

$$L(q_i \mid (\alpha,\sigma_0,(\sigma,\gamma),\beta) = \int_{\mathbf{R}^{V_i}} L^C(q_{i,r,v} \mid \theta,\beta)\Phi(d\theta \mid \mu_i,\sigma_0,\Sigma_i). \quad [16]$$

$\Phi(d\theta \mid \mu_i, \Sigma_i)$ is the V_i-dimensional Gaussian distribution with mean $\mu_i = (\mu(t_{i,v})_{v=1,V_i} = \alpha(t_{i,1}, ..., t_{i,V_i})$ and covariance matrix $\Sigma_i = (C^*(t_{i,v}, t_{i,v'} \mid \sigma, \gamma), 1 \leq v, v' \leq V_i)$, where * stands either for OU (Eq. 9) or FB (Eq. 13). Since individuals are independent and identically distributed, the overall data likelihood sums up to :

$$L(q_1, \cdots, q_i, \ldots, q_N \mid \alpha, \sigma_0, (\sigma, \gamma), \beta) = \prod_{i=1}^{N} \int_{\mathbf{R}^{V_i}} L^C(q_{i,r,v} \mid \theta, \beta)\Phi(d\theta \mid \mu_i, \sigma_0, \Sigma_i). \quad [17]$$

3.2. *Maximizing the Likelihood Function*

There is no analytical form for multiple integrals in Equation 17, hence the maximization procedure of the log-likelihood should be performed using numerical approximations or Monte Carlo (MC) and Markov Chain Monte Carlo (MCMC) related procedures. For a general presentation of these techniques, one can refer to Robert & Casella (1999) and to Spiegelhalter & al. (1995) and Plummer (2016) for operational MCMC tools such as WinBUGS and JAGS. The computation of the multiple integrals in the likelihood can be very time consuming as they essentially depend on the maximum number of patients' visits. For relatively small dimensions (7 for our data set), simple Monte Carlo procedures can be used. For higher dimensions, one has to make use of more sophisticated algorithms such as those based on importance sampling methods -see Richard & Zhang (2007), on Hamiltonian Monte Carlo approach -see Stan (2015), or on the powerful deterministic integrated nested Laplace approximation (INLA methodology) -see Rue and Helde (2005) and Rue & al. (2009).

3.3. *R Implementation*

Let $\Lambda = (\alpha, \sigma_0, (\sigma, \gamma), (\beta_{r,l}, r = 1...R, l = 1, ..., m_r))$ denote the whole parameter set of the model and set the whole data set as a matrix X of size $(\sum_{i=1}^{N} V_i) \times (2 + R)$. Each row includes for an individual, its index i, its visit dates $t_{i,v}$ and its respective responses ($q_{i,r,v}$ to the R items of the questionnaire at theses dates.

3.3.1. Step 1

Define the covariance function $C(\Lambda)$ depending on the parameter $(\sigma_0, (\sigma, \gamma))$ that computes for any set of times $\mathbf{t} = (t_1, \ldots, t_J)$ the $J \times J$ covariance matrix $\Sigma_{\mathbf{t}} = (C(t_j, t_{j'}|\sigma_0, \sigma, \gamma))_{j,j'}$ according to formula 23 (see Appendix 4.1).

3.3.2. Step 2

Define the log-likelihood function (Appendix 4.2) depending on Λ (and on the fixed data set X) with the help of the Fortran routine ("log.lik.f" in Appendix 4.2) for the multiple iterations (individual,visits, items) within a Monte Carlo approximation of the multiple integral using K samples.

1. For each individual $i = 1, ..., n$, initialize $L_i = L_i^C = 0$, and

(a) consider the observation dates $\mathbf{t}_i$ of size V_i, then iteratively for $k = 1, ..., K$, say $K = 2.10^4$.

i. simulate a V_i-multivariate Gaussian variable $\theta_{i,k}$, with mean $\mu_i = \alpha \mathbf{t}_i$ and covariance matrix Σ_i of size V_i, independent of $\theta_{i,k-1}$, taking $\theta_{i,0} = (0, \ldots, 0)$.

ii. compute and sum up $L_i^C \leftarrow L_i^C + L^C(q_i|\theta_{i,k}, \beta)$, till $k = K$.

(b) Approximate (by MC rule) $ll \leftarrow ll + L_i^C/K$, till $i = n$.

2. Return ll as the full log-likelihood function

3.3.3. Step 3

Initialize the parameter $\Lambda \leftarrow \Lambda_0$, say $(0, 1, (1, 0.5), (0, \ldots, 0))$, and then use the function "$mle2$" of the R-library "bbmle", which is a customized version of the "optim" function for MLE issues (see Appendix 4.4).

4. Statistical Analysis of HRQoL HIV Research Survey

Advances in the field of highly active antiretroviral combination therapies have transformed HIV infection into a chronic disease requiring patients to undergo long-term treatment. However, there are serious limitations to these treatments, such as the occurrence of numerous and important adverse events that can affect the quality of life of patients. As a result, the interest of the medical profession no longer focuses solely on the therapeutic action and the state of health of the patients, but also on how the patients themselves feel about their quality of life. In order to study quality of life, at the interface of medicine and psychology, databases are being set up. The data used in this work is an extraction from the Co-pilot French ANRS CO-8 study, a longitudinal cohort study setup in 1997, by the ANRS (French National Agency for Aids Research)- see Mesbah (2015) for more details. This study aimed at describing clinical, immunological, virological and socio-behavioral characteristics of HIV-1-infected patients who were beginning combination antiretroviral therapy (HAART) that included a protease inhibitor (PI). In this study two different instruments were sequentially used.

i- The Short-Form twelve (SF-12), is a shortened version (12 items) of the well known questionnaire 36-items Short-Form Health Survey (SF-36). It was then superseded by.

ii- The WHOQOL HIV Brief (WHB), a shortened version of the WHOQOL HIV which is an HIV (Human Immunodeficiency Virus) specific questionnaire developed by the WHO (World Health Organization).

The data set consists of 2090 questionnaires issued from 342 different patients (275 males and 67 females). Patients retained in this work answered both questionnaire types with a number of interviews that never exceeded 7 dates. At start (month=0), there were 322 patients. 20 more patients were included later : 10 with first questionnaire at month 28, 3 at month 44, 3 at month 60 and the latest 4 at month 72. Other details are given in Table 2.3 of Appendix B.

4.1. *The Unidimensionnality Issue*

Regarding the item selection issue for the analysis, the items $sf10$, $sf11$ and $wq5$ have been inversely recoded in order to be positively correlate with the HRQoL latent variable likewise the other items, -

confer to assumption H3 in Section 2. Let us also mention that only 226 out of 339 patients interviewed at month 72, which is the date of the questionnaire substitution, responded to all items of the two questionnaires.

Assuming that the two questionnaires essentially measure the same unidimensional latent variable, one can denominate *the psychological dimension of quality of life*, we used the Backward Reliability curve (BRC) approach -see Appendix 1.2 for more details. For that purpose, the BRC was performed on the raw 226 responses collected at month 72 and yielded a unidimensional set of 11 out of 16 items -see Table 2.4 in Appendix 2, for their exact contents. The Cronbach alpha coefficient computed for these 226 questionnaires is equal to 0.88.

4.2. *Parameter estimates*

For this data set, the maximization procedure of the log-likelihood is performed using a Monte Carlo procedure since the maximum number of visits is relatively low, here 7. $K = 20000$ simulations of multidimensional Gaussian variables $\Phi(d\theta \mid \mu_i, \Sigma_i)$, at each maximization iteration of the likelihood, were used. Actually, $K = 10000$ simulations revealed to be sufficient to yield identical efficiency. Moreover, the estimation of the forty parameters, 36 $\beta' s$ for the two combined questionnaires and 4 parameters for the underlying OU-process $(\alpha, \sigma_0^2, \sigma_0^2, \gamma)$, was performed using the mle2 procedure of the R-software (www.R-project.org), a version of the "optim" function suited to statistical purposes. It is worth noticing that even for integrals of low dimension (e.g. 7), a Fortran routine has proved to be a compulsory recourse to achieve maximization in a reasonable time period.

It's also worth noticing that the β estimates in Table 4.1 show increasing values for most questionnaire responses, save that $\beta_{8,1} > \beta_{8,2}$; $\beta_{9,1} > \beta_{9,2}$; $\beta_{10,1} > \beta_{10,2}$ and $\beta_{11,2} > \beta_{11,3}$. We also observe that the latter values only involve the second questionnaire (WQ). These results suggest that the Rasch model is well adapted to translate the very nature of the ordinal responses of items with respect to assumption H3 of Section 2 in this study.

Regarding the estimation of parameters of the OU latent process in Table 4.2, the slope α of the linear trend and the standard deviation σ_0 of the initial value are very well estimated. Contrariwise, the regression and diffusion coefficients (γ, σ) and the Cronbach coefficient are poorly estimated. Therefore, we tested the judiciousness of the introduction of such a type of latent process in our analysis.

Note that the mle2 routine of R is based upon numerical approximations of both the gradient and the second order partial differentials and thereby yields not only estimates of the parameters but also gives their covariance matrix described by the Fisher information matrix.

Regarding the choice of the starting values for the algorithm, several attempts were made and lead to similar results save that completely random values leads to endless calculation. In our opinion, the choice of initial values should correspond to the "minimal regular" model, that corresponds to null covariate effects, null trend and positive values for variances to avoid NAN (not a number) outputs.

4.3. *Hypothesis testings*

We performed simple hypothesis testing (i.e., one hypothesis at a time). Since the β's parameter are numerous (36 parameters) and neither crucial nor always judicious for all steps of the analysis, we considered them as well estimated and therefore restricted to the testing of the OU latent characteristics.

For that purpose, we used the classical asymptotic log likelihood ratio, namely the Wilks formula, to precisely assess the following pertinent hypotheses. In general terms, to test a specific hypothesis H_0 (corresponding to a sub-model) vs. a global alternative (within a global model), Wilks formula asserts that under H_0, the statistic $\Delta(p) = -2(ln(L_{H_0}) - ln(L_{H_1}))$ is asymptotically $\chi^2(p)$ distributed, p being the dimensions' difference of the parameter spaces under H_1 and H_0 -see Cox & Hinkley (1974).

4.3.1. Testing the existence of a linear trend

The assumption : H_0 : "$\alpha = 0$" vs H_1 : "$\alpha \neq 0$" is strongly rejected. The slope is significantly positive and is about 0.4 with respect to time counted in day units. The asymptotic likelihood ratio amounts to $\Delta(1) = 2(12334.91 - 12314) = 41.48$.

One can of course assume more general forms of trend e.g. quadratic, logarithmic or periodic. However having already a great number of parameters to estimate to deal with, the authors did not consider a more refined model that adds more parameters to estimate. However, for other future studies one may consider that beta coefficients are already known (or well estimated), and then can advantageously test more specific forms for the trend.

4.3.2. Testing for a deterministic null initial value for the latent process

Testing that the latent process is null at survey start for all individuals amounts to test $\sigma_0^2 = 0$ again $\sigma_0^2 > 0$. As a matter of fact, this is a tricky task since $\sigma_0^2 = 0$ corresponds to a frontier value of the parameter set that requires a more elaborate theory -see e.g. Greven & al. (2008). In this study, we circumvent this issue by testing the nearby simple hypothesis H_0 : "$\sigma_0^2 \leq 10^{-5}$" v.s. $\sigma_0^2 > 10^{-5}$. The asymptotic likelihood ratio amounts to $\Delta(1) = 2(12706.66 - 12314) = 785.32$. So, the null hypothesis is strongly rejected.

4.3.3. Testing for a simple Brownian latent process (without autoregressive component)

We test here the possibility for data to support the assumption H_0 : "$\gamma = 0$" v.s. H_1 : "$\gamma \neq 0$", meaning that the autoregressive part does not exist and that the latent process turns into a Brownian motion. The asymptotic likelihood ratio amounts to $\Delta(1) = 2(12327.32 - 12314) = 26.64$. Hence, the null hypothesis is rejected suggesting that an autoregressive behavior of the latent process can be plainly argued.

4.3.4. Testing for the almost absence of the diffusion component

This is equivalent to say that the latent process actually reduces to a simple deterministic autoregressive component. Again, the null hypothesis $\sigma^2 = 0$ corresponds to a frontier value of the parameter set, and in that respect we tested the nearby assumption $H_0 : \sigma \leq 10^{-5}$ v.s. $H_1 : \sigma > 10^{-5}$. This revealed to be a judicious questioning since the he asymptotic likelihood ratio amounts to $\Delta(1) = 2(12315.01-12314) = 2.02$, so that we can not sensibly reject H_0.

5. Conclusion and Perspectives

The initial objective of this work is to propose some tools to analyze longitudinal studies, where participants are interviewed at regular fixed dates of visit about their health-related quality of life (HRQOL) via potentially changing questionnaires over time. Fairly, the probabilistic models presented here allow relevant statistical inference on supposed latent processes underlying the evolving status of patients. Within the proposed framework, the answer to some epidemiological issues is made possible and may even help to predict at the cohort level how quality-of-life can evolve regarding the parameter estimates such as trends. A by product of the proposed framework, is that it also tackles the issue of now and then substitution of a measurement instrument by another during the study period. This situation however necessarily increases the number of parameters and adds new computational challenges.

When mentioning discrete time context, it is good acknowledged that time varying latent processes were considered by many authors e.g. Meiser & Langeheine (1998), Lee & Daniels (2008, 2013), Hunger &al. (2012)). Atkinson & Schiffrin (1968) already pointed out the need of mathematical models referring to long-term memory processes, a well known notion in psychology research today. However, most of proposed longitudinal models are stationary, Markovian and presume that the correlation function $R(s,t)$ between HRQOL variables at time s and time t decreases to zero when time lag $t-s$ increases. The model presented in this research assumes that individuals evolve within a continuous time context and allow the latent component to exhibit a wide range of enlightening behaviors, such as long range reminiscence, stationarity, trends and autoregressive dynamics. Some of these behaviors could be explosive, decaying or just behaving as a white noise process.

Furthermore, the latent process is assumed to be Gaussian distributed for it has a reduced number of parameters while offering a wide range of local and asymptotic path behaviors, that are actually uniquely determined by correlation function $R(s,t)$.

From the statistical point of view, a simple model was provided enabling the latent process to incorporate and quantify the effect of some judicious components, such as trends, memory range parameter, etc. Moreover, unidimensional models were focused on, but their extension to multidimensional cases can be straightforward.

As regards to the computation issues related to likelihood function, it is important to notice that neither the GLIMMIX procedure of SAS for the GR models, nor the NLMIXED procedure for PC models were directly available for the framework proposed here - see Mesbah (2015). Therefore, an ad hoc program in R statistical language was developed. To which, an efficient Fortran routine was added to overtake iterations and multiple integral calculations. It is worth emphasizing that programs given in Appendix 4.

Tableau 4.1. – β estimates

Item	Parameter	Estimate	Std. Error	P_{value}
sf4	$\beta_{1,1}$	-2.0459	0.1602	$\leq$ 2.2e-16 ***
	$\beta_{1,2}$	-1.5502	0.0760	$\leq$ 2.2e-16 ***
sf6	$\beta_{2,1}$	-0.5356	0.0524	$\leq$ 2.2e-16 ***
sf7	$\beta_{3,1}$	-1.2389	0.0670	$\leq$ 2.2e-16 ***
sf9	$\beta_{4,1}$	-0.6225	0.0546	$\leq$ 2.2e-16 ***
sf10	$\beta_{5,1}$	-2.3378	0.1794	$\leq$ 2.2e-16 ***
	$\beta_{5,2}$	-1.1512	0.0990	$\leq$ 2.2e-16 ***
	$\beta_{5,3}$	-0.0540	0.0466	0.2460
	$\beta_{5,4}$	0.8592	0.0694	$\leq$ 2.2e-16 ***
	$\beta_{5,5}$	2.4782	0.12746	$\leq$ 2.2e-16 ***
sf11	$\beta_{6,1}$	-1.6289	0.12451	$\leq$ 2.2e-16 ***
	$\beta_{6,2}$	-0.5632	0.08428	$\leq$ 2.3e-11 ***
	$\beta_{6,3}$	0.6484	0.08173	$\leq$ 2.1e-15 ***
	$\beta_{6,4}$	1.2756	0.09377	$\leq$ 2.2e-16 ***
	$\beta_{6,5}$	3.0571	0.16472	$\leq$ 2.2e-16 ***
sf12	$\beta_{7,1}$	-2.5231	0.2599	$\leq$ 2.2e-16 ***
	$\beta_{7,2}$	-1.4717	0.1608	$\leq$ 2.2e-16 ***
	$\beta_{7,3}$	-1.3604	0.1010	$\leq$ 2.2e-16 ***
	$\beta_{7,4}$	-0.0290	0.0510	0.5695
	$\beta_{7,5}$	0.8252	0.0737	$\leq$ 2.2e-16 ***
wq1	$\beta_{8,1}$	-1.6457	0.2195	$\leq$ 6.5e-14 ***
	$\beta_{8,2}$	-2.3642	0.1384	$\leq$ 2.2e-16 ***
	$\beta_{8,3}$	-0.0428	0.0093	$\leq$ 4.6e-06 ***
	$\beta_{8,4}$	1.5635	0.0700	$\leq$ 2.2e-16 ***
wq2	$\beta_{9,1}$	-1.9903	0.2916	$\leq$ 8.8e-12 ***
	$\beta_{9,2}$	-2.5322	0.1626	$\leq$ 2.2e-16 ***
	$\beta_{9,3}$	-0.5956	0.0668	$\leq$ 2.2e-16 ***
	$\beta_{9,4}$	1.0846	0.0646	$\leq$ 2.2e-16 ***
wq4	$\beta_{10,1}$	-1.9077	0.2228	$\leq$ 2.2e-16 ***
	$\beta_{10,2}$	-2.0267	0.1249	$\leq$ 2.2e-16 ***
	$\beta_{10,3}$	-0.3298	0.0596	$\leq$ 3.2e-08 ***
	$\beta_{10,4}$	1.9262	0.0792	$\leq$ 2.2e-16 ***
wq5	$\beta_{11,1}$	-2.0897	0.1938	$\leq$ 2.2e-16 ***
	$\beta_{11,2}$	-1.0872	0.1191	$\leq$ 2.2e-16 ***
	$\beta_{11,3}$	-1.1371	0.0785	$\leq$ 2.2e-16 ***
	$\beta_{11,4}$	1.7435	0.0709	$\leq$ 2.2e-16 ***

Tableau 4.2. – Parameter estimates of the latent process

Parameter	Estimate	Std. Error	P_{value}	
Slope of the linear drift : α	0.3957	0.0470	$\leq$ ***	2.2e-16
Initial Std coefficient : σ_0	1.1458	0.0408	$\leq$ ***	2.2e-16
Diffusion coefficient : σ	0.0000001	0.0013	0.9999	
Autoregressive parameter : γ	0.1178	0.3125	0.7061	
Likelihood value : $-2logL$	24628.75			

can be easily extended to deal with the Fractional Brownian Motion based models.

Besides, the classical theory of maximum likelihood methods relies on some regularity assumptions such as integrability and smoothness properties of the likelihood function including thereby nice characteristics for the parameter domain - see Cox & Hinkley (1974). The difficult question of testing a boundary value for a parameter, e.g. testing a null value for variance, has not received a clear theoretical response. Its rare addressing depends on the context and often necessitates some extra-hypothesis -see Self & Liang (1987), Grevens & al. (2008). Investigating rigorous conditions for a throughout answer in the context of Sections 4.3.2 and 4.3.4, is clearly beyond the scope of this paper. All the same, one can rightfully consider that assessing the plausibility of a null variance for some component amounts in practice to test this variance to be relatively small.

Equitably, the data analysis presented in this work may appear suffering from a lack of comparison with other methods. Neither a ready to use method to deal with surveys having changed the measurement instrument at the mid-time or a close comparable model for the latent process, were found. A simulation based study may be an alternative to assess the robustness of this approach. However, considering the large number of parameters involved, this is postponed to future work. Besides, even if the present work used all available data, the recurrent issue of individual dropouts has not been addressed explicitly in this paper. This omission corresponds de facto to the implicit assumption that dropouts are non informative and occur at completely random time, tantamount to the situation of independent censoring in survival analysis. Nevertheless, one can precisely tackle this point at the cost of an additional statistical component detailing the conditional distributions of individual absence/presence at each questionnaire date.

OU and fractional Brownian motion are described as two possible models for the longitudinal latent variable. Even if the FBM was not implemented here, one can provide some guidelines on how tho choose, in practice, between both models. For that purpose, one simply recall that the two models describe distinct dynamics and process characteristics. The OU process mainly assume a regular autoregressive behavior possibly leading to explosion or vanishing limit, while the FBM with self similarity properties corresponds to irregular arrival of packages of events of varying intensity. Thus, these two models reflect two ways of interpreting (and thereby quantifying) the evolution of patient statuses that a physician might examine.

As a concluding remark, we can reasonably assert that the framework proposed here is both operative

and enough open to tackle simultaneously important issues such as questionnaire concatenation during time, health reminiscence assessing, data censoring and multi-dimensionality for HRQoL surveys.

Bibliographie

Andersen, E.B. (1977) Sufficient Statistics and Latent trait models. *Psychometrika*, 42,69-81.

Atkinson, R.C., and Shiffrin, R.M. (1968). Human memory : a proposed system and its control processes. *In The psychology of learning and motivation : Advances in research and theory. (Vol. 2). Eds : Spence, K.W. and Spence, J.T. Academic Press, New York, 89-195.*

Bartolucci, F. and Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika, 70, 31-43.*

Bartolucci, F. (2007). A class of Multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika, 72, 141-157.*

Boisson, V., Mesbah, M. and Ying, Z. (2015). Log-rank type test for evolution of health related quality of life. *Statistica Applicata - Italian Journal of Applied Statistics, Vol. 27 1, 75-91.*

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3, 296-322.*

Christensen, K.B., Bjorner, J.B., Kreiner, S. and Petersen, J.H. (2002). Testing unidimensionnality in polytomous Rasch models. *Psychometrika, 67, 563-574.*

Christensen, K.B., Kreiner, S. and Mesbah, M. (2013). Rasch Models in Health. *Iste, London and J.Wiley, New York.*

Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics. *Chapman and Hall, London.*

Dossar, P. and Mesbah, M. (2018). Cumulative or adjacent logits : which choice for an ordinal logistic latent variable model ? *Communication in Statistics. Theory and Methods, 47 :11, 2563-2575.*

El Fassi, K., Abdous, B. and Mesbah M. (2009). Local polynomial fitting of the equipercentile equating function : strong uniform consistency. *Comptes Rendus Mathematique, Vol. 347, Issues 3-4, 195-200.*

Greven, S., Crainiceanu, C.M., Helmut Küchenhoff H. and Peters, A. (2008). Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models. *Journal of Computational and Graphical Statistics, Vol. 17, Issue 4.*

Hall, P., Muller, H.-G., and Yao, F. (2008). Modeling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society : Series B (Statistical Methodology), 70(4) p.703- 723.*

Hunger, M., Doring, A. and Holle, R. (2012). Longitudinal beta regression models for analyzing Health-related quality of life scores over time. *BMC Medical Research Methodology, 12, 144.*

Ibe, O.C. (2013). Markov Processes for Stochastic Modeling. Second Edition. *Elsevier Inc.*

Kolen, M.J. and Brennan R.L. (2004). Test Equating, Scaling, and Linking Methods and Practices. *Series : Statistics for Social and Behavioral Sciences, 2nd ed. Springer-Verlag, New York.*

Lee, K. and Daniels, M.J. (2008). Marginalized Models for Longitudinal Ordinal Data with Application to Quality of Life Studies. *Stat. Med., Sept. 20. 27(21). 4359-4380.*

Lee, K. and Daniels, M.J. (2013). Flexible marginalized models for bivariate longitudinal ordinal data. *Biostatistics, 14, 3, 462-476.*

Li, P. and Chen, S. (2016). A review on Gaussian Process Latent Variable Models *ScienceDirect CAAI Transactions on Intelligence Technology Vol.1, Issue 4 (2016) p.366-376.*

Lord, F.M. (1952). Psychometric Monograph No 7. *Psychometric Society, 1952.*

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47, 149-174.*

Meiser, T., Stern, E. and Langeheine R. (1998) Latent Change in Discrete Data : Unidimensional, Multidimensional, and Mixture Distribution Rasch Models for the Analysis of Repeated Observations. *Methods of Psychological Research Online. Vol.3, No2.*

Mesbah, M. (2010). Statistical Quality of Life. *In "Methods and Applications of Statistics in the Life and Health Sciences". Chap. 74. Eds : Balakrishnan, N. Wiley, New York.*

Mesbah, M. (2012). Measurement and Analysis of Quality of Life in Epidemiology. *In "Handbook of Statistics. Vol 28". Chap. 15. Eds : Chakraborty, R., Rao, C.R., and Sen, P.K. Elsevier, Amsterdam.*

Mesbah, M. (2015). Analysis of a complex Longitudinal Health-Related Quality of Life Data by a mixed Logistic Model. *In "Applied Statistics in Biomedicine and Clinical Trials Design". Chap. 19. Eds : Chen, Z., Liu, A., QU, Y., Tang, L., Ting, N. and Tsong, Y. Springer, New York.*

Morrison K.T., Shaddick G., Henderson S.B., and Bickeridge D. (2016). A latent process model for forecasting multiple time series in environmental public health surveillance. *Statistics in Medicine, 35 3085-3100.*

Ornstein, L. S., and G. E. Uhlenbeck. (1930). On the Theory of the Brownian Motion. *Physical Review, 36, no. 5 : 823.*

Plummer, M. (2016). rjags : Bayesian Graphical Models using MCMC. *R package version 4-6.*

Protter, P.E. (2005). Stochastic Integration and Differential Equations. 2nd Edition. *Springer, London.*

Rasch, G.(1960). Probabilistic models for some intelligence and attainment tests. *Danisch National Institute for Educational Research, Copenhagen, 1960.*

Rasmussen, C., and Williams, C. (2006). Gaussian Processes for Machine Learning. *MIT Press. ISBN 0-262-18253-X.*

Richard, J.F. and Zhang W. (2007). Efficient high-dimensional importance sampling. *Journal of Econometrics, Vol. 141, Issue 2.*

Robert, C.P. and Casella, G. (1999). Monte Carlo Statistical Methods. *Springer Texts in Statistics.*

Rue, H., and Held, L. (2005). Gaussian Markov Random Fields : Theory and Applications. *Monographs on Statistics and Applied Probability, vol. 104. Chapman and Hall, London.*

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society : Series B (Statistical Methodology); 71(2), p.319–392.*

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph, 34 (suppl. 17), 386-415.*

Samorodnitsky, G. and Taqqu, M.S. (1994), Stable non-Gaussian random processes : stochastic models with infinite variance, Stochastic Modeling. *Chapman and Hall, New York.*

Self, S.G. and Liang, K.Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *Journal of the American Statistical Association. Theory and Method, Vol. 82, Issue 398.*

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3, 271-295.*

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B, 64 (2). p.583-639.*

Stan Development Team (2015). Stan Modeling Language User's Guide and Reference Manual. *Version 2.6.1.*

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52, 589-617.*

van Zyl, J.M., Neudecker, H., and Nel, D.G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65, 271-280.*

von Davier, A.A., Holland, P.W., and Thayer, D.T.(2003) The Kernel Method of Test Equating. *Statistics for Social Science and Public Policy, Springer, New York.*

Xu, Y., and Ji, Y. (2014). A latent Gaussian Process Model with Application to Monitoring Clinical Trials. *arXiv 1403.7853v1 [stat.ME], 31 Mar 2014.*

1. Reliability of the Unidimensionality Assumption

1.1. *Checking Instrument Reliability : Cronbach Alpha Coefficient*

A measurement instrument is sensed to provide values we call observations concerning some state variables. In the case of quantitative continuous responses, the reliability ρ of an instrument is defined as the ratio of the variances of the true over the observed measure. In the particular case of the parallel model 2.1.1, the reliability of an item Q, considered as measurement instrument of the hidden state θ is given by :

$$\rho = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma^2}. \qquad [18]$$

Clearly, $0 \leq \rho \leq 1$, values close to zero indicate unreliability whereas values close to 1 mean high reliability. Additionally, it can be shown that ρ equals the correlation coefficient between any two distinct items Q_i, Q_j, known in this case as the intra-class correlation coefficient. The extension of reliability concept to a set of k variables is simply written

$$\tilde{\rho}_k = \frac{k\rho}{k\rho + (1-\rho)}, \quad k \geq 1. \qquad [19]$$

Equation 19 is known as the Spearman-Brown prophecy formula -see Brown (1910), Spearman (1910). It points out the increasing hyperbolic relationship of ρ with respect to the number k of variables. The left graph of Figure 3 shows the reliability curves corresponding to distinct values of ρ, $\rho = 0.1, \ldots, 0.9$.

Under Gaussian assumptions, the maximum likelihood estimator of $\tilde{\rho}_k$, named Cronbach's Alpha Coefficient (CAC), expresses for an N-sample, as

$$\alpha_{N,k} = \frac{k}{k-1}\left(1 - \frac{\sum_{r=1}^{k} S_{N,r}^2}{S_{N,tot}^2}\right), \qquad [20]$$

where

$$S_{N,r}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Q_{i,r} - \overline{Q_r})^2 \text{ and } S_{N,tot}^2 = \frac{1}{Nk-1}\sum_{i=1}^{N}\sum_{r=1}^{k}(Q_{i,r} - \overline{Q})^2.$$

with

$$\overline{Q_{N,r}} = \frac{1}{N}\sum_{i=1}^{N} Q_{i,r} \text{ and } \overline{Q} = \frac{1}{kN}\sum_{i=1}^{N}\sum_{r=1}^{k} Q_{i,r}.$$

Beware that $\alpha_{N,k}$ is also used (with the same appellation) as a natural estimator of another population parameter $\tilde{\alpha}_k$ - see Mesbah (2010).

When the number of patients $N \to \infty$, the following results hold true even for non Gaussian cases -refer to van Zyl & al. (2000) :

$$\alpha_{N,k} \overset{a.s}{\to} \tilde{\rho}_k, \quad E(\alpha_{N,k}) \to \tilde{\rho}_k, \text{ and } NVar(\alpha_{N,k}) \to \frac{2(1-\tilde{\rho}_k)^2 k}{(k-1)}\tilde{\rho}_k.$$

Moreover, we have

$$\frac{\sqrt{n}}{2}\ln(1-\alpha_{N,k}) \sim N\left(\frac{1}{2}\ln(1-\tilde{\rho}_k); \frac{k}{2(k-1)}\right). \qquad [21]$$

1.2. *Ascertaining the Unidimensionality of an Instrument : The Backward Reliability Curve*

For the parallel models and IRT models (Bertolucci & Forcina (2005), a large literature addresses the important issue of statistical validation of unidimensionality through the goodness of fit test approaches. However, the proposed approaches revealed to be powerless because the null hypothesis does not focus on unidimensionality, but also sustains additional important assumptions such as Gaussianity and local independence. Therefore, the departure from the null hypothesis does not correspond specifically to

a departure from unidimensionality assumption - refer to Stout (1987), Christensen & al. (2002) and Bartolucci (2007).

Within the parallel Model in Section 2.1.1, an other approach based on the CAC's values $\alpha_{N,k}, k = 1, ..., R$, estimating the respective reliability coefficients $\tilde{\rho}_k, k \geq 1$ can be considered as well. Indeed, since the Spearman-Brown Formula asserts that symptomatically (when $N \to \infty$), the estimate reliability curve $\alpha_{N,k}, k = 1, ..., R$, is an increasing function of the number k of variables, we can heuristically derive the following practical graphical tool, namely the Backward Reliability Curve (BRC), to check the unidimensionality of R items' set.

BRC consists in recursively drawing optimal CAC values according to the following scheme : First, compute the CAC of the whole set of R available items. Then, remove the item for which the CAC value of the remaining variables is maximum. Repeat the procedure till the latest item.

Therefore any obvious decrease of the curve would entail a strong suspicion that this variable belongs to the unidimensional set of variables already introduced in the set. If this is the case, one (or more variables) has to be removed until an increasing curve is obtained. In some sense, the selected final set of items can be said to be *more valid* in term of unidimensionality than the initial one.

From a robustness point of view, it is worth noticing that a simulation study shows that similar behaviors and conclusions occur for large samples of individuals for GR and PC models defined in Sections 2.1.2 and 2.1.3, -see Mesbah (2012).

The procedure was applied for the ANRS CO-8 data set including all qualitative ordinal items of questionnaires SF12 and WHB at the visit date 72 as illustrated in the graph at the right of Figure 3.

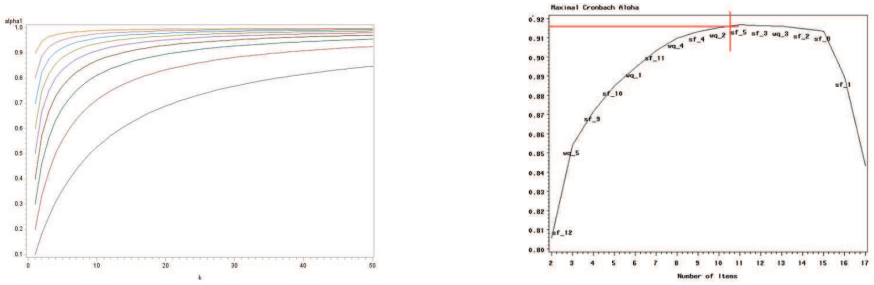

Figure 3. – **Left :** Theoretical Relationship between $\tilde{\rho}_k$ and the number k of items. **Right :** The backward reliability curve (BRC) yielding the final set of items for the ANRS CO-8 data set.

2. ANRS CO8 Study : Design and Questionnaires

Table 2.3 summarizes the main features of the French cohort study CO8 set up in 1997 -see Boisson & al. (2015) and Mesbah (2015), related to HRQoL of HIV-1-infected patients, namely the time intervals between patient interviews, the type of questionnaire and the number of interviewed patients.

Table 2.4 details the exact content of the 11 items selected from both questionnaires SF12 and WHB after performing the backward reliability curve algorithm. The algorithm was initiated with twelve SF12 items and four WHB items.

Tableau 2.3. – The ANRS CO8 Study Design

Time (in months)	Questionnaire	Patient Number
0	SF12	322
28	SF12	266
44	SF12	252
48	SF12	5
60	SF12	74
72	SF12 + WHB	339
84	WHB	342
96	WHB	265
108	WHB	212
120	WHB	13

3. Correlation structure of Ornstein-Uhlenbeck Process

3.1. *Covariance and Correlation Functions*

To be sensible and practical, we required in Equation 8 that the latent $\theta(s)$ to be a Gaussian process being able to exhibit a long memory property and includes as well a linear deterministic trend, that is :

$$\theta(t) = Z^*(0) + \alpha t + Z(t),\ t \geq 0 \qquad [22]$$

The parameter α is deterministic and denotes the slope of the linear trend. $Z^*(0)$ stands for the initial random value of the latent variable and is $N(0, \sigma_0^2)$-distributed. The process $Z(.)$ is assumed to be a centered Ornstein-Uhlenbeck Gaussian process, independent of $Z^*(0)$ and null at time $t = 0$. Depending on the different parameter values, the latent process may more or less enhance different types of path behaviors (stationarity, asymptotic explosion, long/short memory range,...). This is mainly due to the behavior of $Z(.)$ which is totally characterized by its covariance structure $C^{OU}(s,t)$ given in Equations 9 and 10. So, the covariance function of the latent process $C(s,t) = Cov(\theta(s), \theta(t)) = Var(Z^*(0)) + C^{OU}(s,t)$ is written for $s \leq t$:

$$C(s,t) = \begin{cases} \sigma_0^2 + s\sigma^2 & \text{if} \quad \gamma = 0 \\ \sigma_0^2 + \frac{\sigma^2}{2\gamma}\exp(-\gamma(s+t))(\exp(2\gamma s) - 1) & \text{if} \quad \gamma \neq 0 \end{cases} \qquad [23]$$

In particular, we obtain the variance function of the latent process $\theta(.)$

$$C(s,s) = \begin{cases} \sigma_0^2 + s\sigma^2 & \text{if} \quad \gamma = 0 \\ \sigma_0^2 + \frac{\sigma^2}{2\gamma}(1 - e^{-2\gamma s}) & \text{if} \quad \gamma \neq 0 \end{cases} \qquad [24]$$

and its correlation function

Tableau 2.4. – Unidimensional Set of SF12 and WHB Items Selected by the BRC Procedure

Instrument labels	Contents
SF12 (Period 1)	
sf4	During the past four weeks, according to your physical state, have you accomplished less than you would like ? 1-Yes 2-No
sf6	During the past four weeks,according to your emotional state, have you accomplished less than you would like ? 1-Yes 2-No
sf7	During the past four weeks, according to your emotional state, have you had difficulty doing what you had to do with so much care ? 1-Yes 2-No
sf9	Frequency of discomfort due to your health, physical or emotional condition in your life and your relationships with others ? 1-All the time 2-Most of the time 3-Sometimes 4-Seldom 5-Never
sf10	How much time during the past 4 weeks have you felt calm and peaceful ? 1-All the time 2-Very often 3-Often 4-Sometimes 5-Seldom 6-Never
sf11	How much time during the past 4 weeks have you felt energized ? 1-All the time 2-Very often 3-Often 4-Sometimes 5-Seldom 6-Never
sf12	How much time during the past 4 weeks have you felt sad and depressed ? 1-All the time 2-Very often 3-Often 4-Sometimes 5-Seldom 6-Never
WHB (Period 2)	
wq1	How much do you enjoy your life ? 1-Not at all 2-A little 3-A moderate amount 4-Very much 5-An extreme amount
wq2	How satisfied are you with your ability to learn new information ? 1-Not at all 2-A little 3-A moderate amount 4-Very much 5-An extreme amount
wq4	How much do you value yourself ? 1-Very dissatisfied 2-Dissatisfied 3-Neither satisfied nor dissatisfied 4-Satisfied 5-Very satisfied
wq5	How often do you have negative feelings, such a blue mood, despair, anxiety, depression ? 1-Never 2-Seldom 3-Quite often 4-Very often 5-Always

$$R(s,t) = \begin{cases} \sqrt{\frac{\sigma_0^2+s\sigma^2}{\sigma_0^2+t\sigma^2}} & \text{if} \quad \gamma = 0 \\ \frac{\sigma_0^2+\frac{\sigma^2}{2\gamma}\exp(-\gamma(t-s))(1-\exp(-2\gamma s))}{\sqrt{\sigma_0^2+\frac{\sigma^2}{2\gamma}(1-\exp(-2\gamma s))}\sqrt{\sigma_0^2+\frac{\sigma^2}{2\gamma}(1-\exp(-2\gamma t))}} & \text{if} \quad \gamma \neq 0 \end{cases} \qquad [25]$$

3.2. *Long range behavior of R(s,t)*

The asymptotic behavior of the correlation function can be considered as a memory's range indicator of the latent process. Since the asymptotic can be investigated under many manners, two pertinent schemes of asymptotic are proposed below, namely

— S1 : fix a time s_0 and seek for $\tau_1 = \lim_{t\uparrow+\infty} R(s_0, t)$. If the limit exists, it is called the asymptotic memory of type S1 .

— S2 : fix a coefficient $\nu;\ 0 < \nu < 1$ and look at $\tau_2 = \lim_{t\uparrow+\infty} R(\nu t, t)$. If the limit exists, it is called the asymptotic memory of type S2.

For sake of simplicity let us denote $C^{OU}(s,s) = \frac{\sigma^2}{2\gamma}(1 - e^{-2\gamma s})$ and consider the different cases.

1. $\gamma < 0$ (explosive OU process). In this case, Formula 25 can be rewritten

$$R(s,t) = \frac{\frac{\sigma_0^2}{C^{OU}(s,s)} + \exp(-\gamma(t-s))}{\sqrt{\frac{\sigma_0^2}{C^{OU}(s,s)} + 1}\sqrt{\frac{\sigma_0^2}{C^{OU}(s,s)} + \frac{C^{OU}(t,t)}{C^{OU}(s,s)}}},$$

and yields

— under S1, $R(s_0,t) \xrightarrow[t\to\infty]{} \tau_1 = \frac{\exp(\gamma s_0)\sqrt{\exp(-2\gamma s_0)-1}}{\sqrt{1+\frac{\sigma_0^2+2\gamma}{\sigma^2(1-\exp(-2\gamma s_0))}}} > 0$. So, the latent process $\theta(.)$ has a long range asymptotic memory of type S1.

— Under S2; $R(\nu t, t) \approx \frac{\exp(-\gamma(1-\nu t))}{\sqrt{\frac{1-exp(-2\gamma t)}{1-\exp(-2\gamma\nu t)}}} \xrightarrow[t\to\infty]{} \tau_2 = 1$. In this case, the latent process has a "full" long range asymptotic memory of type S2.

2. $\gamma = 0$ (white noise OU process). R is then written $R(s,t) = \sqrt{1 - \frac{(t-s)\sigma^2}{\sigma_0^2 + t\sigma^2}}$, and yields

— Under S1, $R(s_0,t) = \sqrt{1 - \frac{(1-\frac{s_0}{t})\sigma^2}{\frac{\sigma_0^2}{t}+\sigma^2}} \xrightarrow[t\to\infty]{} \tau_1 = 0$. In this case, the latent process can be said to have a short range asymptotic memory of type S1.

— Under S2, $R(\nu t, t) = \sqrt{1 - \frac{(1-\nu)\sigma^2}{\frac{\sigma_0^2}{t}+\sigma^2}} \xrightarrow[t\to\infty]{} \tau_2 = \sqrt{\nu} > 0$, and the latent process has a long range memory of type S2.

3. $\gamma > 0$ (asymptotic stationary OU process). Since $C^{OU}(t,t) \searrow \frac{\sigma^2}{2\gamma}$, we get

— Under S1, $R(s_0,t) \xrightarrow[t\to\infty]{} \tau_1 = \frac{\sigma_0^2}{\sqrt{\sigma_0^2 + C^{OU}(s_0,s_0)}\sqrt{\sigma_0^2+\frac{\sigma^2}{2\gamma}}} > 0$. The latent process exhibits a long range asymptotic memory of type S1.

— Under S2, $R(\nu t, t) \xrightarrow[t\to\infty]{} \tau_2 = \frac{\sigma_0^2}{\sqrt{\sigma_0^2+\frac{\sigma^2}{2\gamma}}\sqrt{\sigma_0^2+\frac{\sigma^2}{2\gamma}}} > 0$. The latent process also exhibits a long range asymptotic memory of type S2.

3.3. *Comparison with stationary AR$_1(1)$ sequences*

In Mesbah (2015), the latent process was assumed to be a discrete time autoregressive sequence (DAR$_1(1)$) obeying the stochastic dynamics $\tilde{\theta}_{n+1} = \rho\tilde{\theta}_n + \epsilon_{n+1},\quad n = 0, 1, ...$ with a white noise sequence $(\epsilon_n)\,, n \geq 1$. It is known that $\tilde{\theta}$ is stationary if and only if it is centered and $|\rho| < 1$. In that case, its correlation function is written $\tilde{R}(k) = E[\tilde{\theta}_n, \tilde{\theta}_{n+k}] = \rho^k$, for any non negative integers k, n. This

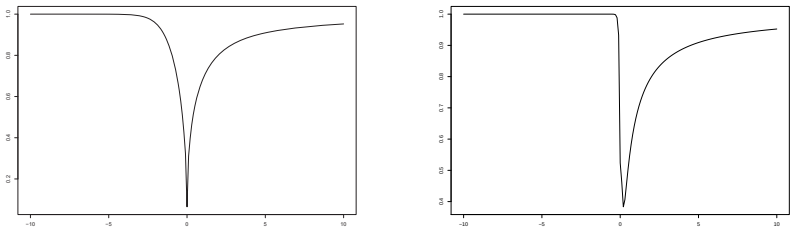

Figure 4. – Right : Graph of the asymptotic memory of type S1, that is $\lim_{t\to+\infty} R(s_0,t)$ of the latent process $\theta(t)$ as a function of the regression coefficient γ in abscissa when $s_0 = 1$ and the other parameters are fixed to $\sigma_0^2 = \sigma^2 = 1$ **Left :** Graph of the asymptotic memory of type S2, that is $\lim_{t\to+\infty} R(\nu t,t)$ of the latent process $\theta(t)$ as a function of the regression coefficient γ in abscissa when $\nu = 0.5$ and the other parameters are fixed to $\sigma_0^2 = \sigma^2 = 1$

means that the correlation decreases exponentially to zero as the time lag k increases and the sequence can be said of (very) short memory range.

The quality of life latent process $\theta(t)$ could have a long memory, that is to get slowly decreasing correlation function with respect to large time lags. For illustration purpose, here are the correlation values for both stationary sequences $\tilde{\theta}$ and Ornstein Uhlenback processes θ for different regression coefficients ρ and γ, for a same time lag $k = 6$, :

$\rho =$	0.1	0.2	0.3	0.4	0.5
$\tilde{R}(6) =$	0.000001	0.000064	0.000729	0.004096	0.015625
$\gamma =$	-1.5	-0.5	0	0.5	0.5
$R(1,7) =$	0.906185	0.650436	0.325000	0.571038	0.754741.

[26]

4. R-Program

4.1. *Covariance function of the latent process*

```
C <- function(a,b0, bt,dates){
    # dates are increasing
    # a = gamma  is the time scale
    # b0 variance of the initial r.v. indpt of the OU process
    # bt = sigma such that bt*[exp(-a(t-s)-exp(-a(t+s)]/(2a)= variance of the
    # latent  at time  t= 1.
    nt <- length(dates)
    C <- matrix(0, nr=nt, nc=nt)
    if(a==0){
      for(i in 1:nt){ for( j in 1:i){
                  C[i,j] <- b0 + bt^2*dates[j]
                  C[j,i] <- C[i,j]  }
                    }
            } else{ # a neq 0
       for(i in 1:nt){ for( j in 1:i){
  C[i,j]<- b0+ bt^2*exp(-a*(dates[i]+dates[j]))*(exp(2*a*dates[j])-1)/(2*a)
  C[j,i] <-   C[i,j]                }
                    }
                  }
    return(C)                    }
```

4.2. *"loglik.f" Fortran file (to compile)*

```
#subroutine iterations sur: individu, integrale MC, temps, variables
*       adresse fichier:   cd ...\...
*       compilation : R CMD SHLIB -o loglik.so loglik.f
*       (compilation under  Windows interface : R CMD SHLIB  loglik.f )
*
     subroutine logv(Nsim,Nicum,X,bbeta,Zsim,lvr)
*
*       # Nsim: nb de tirages pour le calcul de l'integrale
*       # Nicum(Nind+1): début des lignes des données des individus
*       # X: la matrice des données (ici des entiers relatifs)
*       # bbeta: paramètres beta cumulés des covariables(en matrice) et
*       # complété par des zéros quand un paramètre est réduit
*       # Zsim: simulations du processus latent (dépendant de theta)
*       # lvr: loglik à retourner
*       # Nind=342: nb d'individus
*       # Nt=10: nb total de dates de l'étude
*       # p=11: nb de covariables prises en compte
*       # Nimax=8: nb max de dates d'observations par individu
*
      integer Nsim, Nicum(343),  X(2090,15)
      double precision  Zsim(10,Nsim) , bbeta(11,6), lvr
      integer j1, j2, nbeta(11) , ni, xi( 8, 15), xitj, ti(8)
      double precision      wxitj
*       # nbeta indique pour chaque covariable son nombre de paramètres
      DO 1 j1=1,11
      nbeta(j1) = idint(bbeta(j1,1))
1     CONTINUE
*       #boucle sur les individus i pour la logvraisemblance lvr
      lvr = 0.
      DO 10 i=1,342
      vri = 0.
*     # on remplit la matrice des données xi de l'ind. i (datesXvariables)
      ni = Nicum(i+1) - Nicum(i)
      DO 11 j1=1,ni
      DO 12 j2=1,15
      xi(j1,j2)  =   X( Nicum(i)+ j1 , j2 )
12    CONTINUE
11    CONTINUE
*       #ti=x(,15) : numéro des dates concernant i
      DO 13 j1=1,ni
      ti(j1)  =  X(Nicum(i)+j1, 15)
13    CONTINUE
*       #boucle sur les simulations s (du processus latent)
      DO 20 is=1,Nsim
      vris =1.
*       #boucle sur les dates t d'observations de i
      DO 30 it= 1,ni
      vrist =1.
*       #boucle sur les covariables presentes à cette date de l'ind. i
      DO 40 itj= 1,11
*       # initialisation vrais ou si donnée manquante pour variable j
      vristj =1.
      xitj  = xi(it , itj+3 )
      wxitj = real(xitj )
*       # si variable j présente
      IF( wxitj .GE.  0) THEN
*       # calcul et initialisation du dénominateur, numerateur
      vrnum = 1.
      vrden = 1.
*           # sous programme pour calculer le dénominateur
```

```
      DO 50 k=1,nbeta(itj)
      vrden = vrden +  exp( k*Zsim(ti(it),is) - bbeta(itj,k+1))
50    CONTINUE
*                        # fin sous programme pour le dénominateur
*       # sous prog pour le numérateur (si valeur covariable >0)
      IF(wxitj .GT. 0) THEN
      vrnum = exp( wxitj*Zsim(ti(it),is) - bbeta(itj,xitj+1) )
      ENDIF
*                      #fin calcul du numérateur si covariable >0
      vristj= vrnum/vrden
      ENDIF
*       # fin du calcul vrais de la variable j si elle est présente
      vrist=vrist*vristj
*       # fin traitement de la variable j (présente ou non)
40    CONTINUE
*       fin boucle sur   les covariables à cette date de l'individu
      vris = vris*vrist
30    CONTINUE
*       fin boucle sur les dates observees de cet cindividu i
      vri =  vri + vris
20    CONTINUE
*       fin iteration sur les simulations du latent de l'individu i
      lvr = lvr + log( vri/Nsim )
10    CONTINUE
*       fin boucle sur les individus
      RETURN
      END
```

4.3. *Log likelihood Function*

```
 ll <- function(param){ # loglikelihood for all individual paths
   # sample of standard Gaussian r.v. for MC approximation
   # zsim <- matrix(rnorm(Nt*Nsim,0,1), nr=Nt, nc=Nsim)

   lvr <- 0 #  loglik initialization
   # param = c( beta[i], theta_mu,theta_cov) corresponds to 40 parameter

  bbeta <- matrix(0, nr=p, nc=5)  #  col. 1 states the parameter nb  per
  # covariable, void cells are implicitely 0 valued.
  bbeta[ ,1] <- nvx[1:p]
  for(j in 1:p){
  bbeta[j,2:(nvx[j]+1)]<-cumsum(param[(n_param[j]+1):n_param[j+1]])}

  theta    <- param[  (n_param[p+1]+1): (n_param[p+1]+4)  ]
  mu <- theta[1]*tt   # trend of the latent process for all dates
  Sigma  <- C( theta[2],theta[3],theta[4], tt )  # covar of  OU process
  Zsim  <- t(chol(Sigma))%*%zsim + mu

  out<-.Fortran("logv",Nsim=as.integer(Nsim),Nicum=as.integer(Nicum),
     X=as.integer(X), bbeta=as.double(bbeta), Zsim=as.double(Zsim),
 lvr=as.double(lvr)  )
  llk  <- out$lvr
 return(- llk)               }
```

4.4. *Maximum Likelihood Estimators*

```
               # library loading
library(bbmle)
library(stats4)
dyn.load("loglik.dll")
```

```
# Read and define the data set in matrix X (a fixed matrix)
# Draw a large sample of multiv. standard Gaussian iid  r.v.
# to approximate  the multiple integral
Nsim  <- 20 000
zsim <- matrix(rnorm(Nt*Nsim,0,1), nr=Nt, nc=Nsim)

# Optimization
 fit <- mle2( minuslogl=ll,
start=list(Lambda=Lambda0),method="L-BFGS-B",upper=up_val,lower=low_val)
# Results
  summary(fit)
```