Latent Class and Genetic Stochastic Process Models: Implications for Analyses of Longitudinal Data on Aging, Health, and Longevity

Konstantin G. Arbeev[1], Liubov S. Arbeeva[1], Igor Akushevich[1], Alexander M. Kulminski[1], Svetlana V. Ukraintseva[1], Anatoliy I. Yashin[1]

[1] Biodemography of Aging Research Unit, Center for Population Health and Aging, Social Science Research Institute, Duke University, 2024 W. Main Street, Durham, NC 27705, USA

Abstract

Specific applications of statistical methods for joint analyses of longitudinal and time-to-event information in the context of studies on aging can benefit from incorporation of knowledge and theories about mechanisms and regularities of aging-related changes into respective analytic approaches. A conceptual analytic framework for these purposes, the stochastic process model of aging (SPM), has been recently developed in the biodemographic literature. Here we present two modifications of such models: the latent class SPM (LCSPM) and the genetic SPM (GenSPM). The LCSPM allows applications to populations consisting of latent subpopulations with distinct patterns of longitudinal trajectories of biomarkers that can also have different effects on the time-to-event outcome in each subpopulation. The GenSPM aims at applications analyzing genetic effects on the longitudinal trajectories and time-to-event outcomes taking into account observed characteristics affecting the probability of the presence of an allele/genotype in the genome of an individual. This case assumes that genetic information is available for a sub-sample of participants of the longitudinal study or for the entire sample. The GenSPM allows joint analyses of information from genotyped and non-genotyped subsamples which results in an increase in the power compared to analyses of the genotyped subsample alone. We present simulation studies and discuss practical applications of these approaches.

Key Words: stochastic process model, mortality, health, aging, longitudinal data

1. Introduction

An important point to consider in applications to research on aging is how to incorporate knowledge and theories about mechanisms and regularities of aging-related changes that accumulate in the research field into respective analytic approaches. In the absence of specific observations of longitudinal dynamics of relevant biomarkers manifesting such mechanisms and regularities (which is a typical situation in a contemporary longitudinal studies), traditional approaches may have a rather limited utility to estimate respective parameters that can be meaningfully interpreted from the biological point of view. A conceptual analytic framework that incorporates available knowledge about mechanisms of aging-related changes which may be hidden in the individual longitudinal trajectories of physiological variables and that allows for analyzing their indirect impact on the risks of diseases and death has been recently developed in the biodemographic literature. This approach, the stochastic process model (SPM), has its roots in the random walk model by Woodbury and Manton (1977). The SPM by Yashin et al. (2007a) incorporates

substantive knowledge about different aging-related concepts (such as the notions of physiological norm, allostatic adaptation, measures of stress resistance, and adaptive capacity) and it has been extended in various ways and applied in different contexts; see, e.g., the review paper Yashin et al. (2012). The advantage of the SPM is that it provides an approach to work with such "hidden components of aging" indirectly and to estimate parameters relevant to research on aging. In this paper, we focus on two aspects of developments in the SPM framework. The first deals with incorporation of latent classes in the SPM (the latent class SPM, LCSPM). The second deals with extensions to analyze genetic data (or, more broadly, any variable with missing observations for a sub-sample of participants of the longitudinal study) in the context of SPM (the genetic SPM, GenSPM).

The original SPM (Yashin et al. 2007a) assumes that all components of the model have similar patterns in all individuals in a population. However, a population may consist of latent subpopulations with distinct patterns of longitudinal trajectories with different effects on the time-to-event outcome in each such subpopulation. The presence of such heterogeneity is a realistic scenario which cannot be simply ignored in statistical analyses of longitudinal data. One example could be that carriers of some alleles or genotypes can have distinct patterns of aging-related characteristics. If the corresponding genetic marker is not available in the data, then evaluation of the true characteristics from the data can be performed indirectly in the model that incorporates such hidden heterogeneity. The extension of the SPM (Yashin et al. 2007a) to accommodate such hidden heterogeneity was suggested in Yashin et al. (2008). In this paper, we present a modified version of this model, LCSPM, which includes dependence of the probability of the latent class membership and other components of the model on observed covariates. We formulate the discrete-time specification of the model. Continuous-time version is presented elsewhere, see, e.g., Arbeev et al. (2014).

The LCSPM is designed for applications where the variable defining the latent structure is completely unobserved. In the particular case of genetic markers, it may happen that respective information is actually available for at least a sub-sample of participants of a longitudinal study because many longitudinal studies collecting data on biomarkers started including genetic information as well. In this case, such genetic informatio can be used and one can apply traditional SPM treating the genetic covariate as any other covariate included in the model. Follow-up data on mortality and longitudinal measurement of biomarkers for non-genotyped individuals provide an additional source of information which can be used in analyses. The group of non-genotyped individuals is a mixture of carriers/non-carriers of the same alleles/genotypes collected in the genetic data and a similar effect of the alleles/genotypes on the mortality rate and the age trajectory of biomarkers can be assumed in both genotyped and non-genotyped parts of the entire sample. An approach for joint analysis of longitudinal and time-to-event outcomes for genotyped and non-genotyped participants of longitudinal studies has been presented recently within the SPM framework (see Arbeev et al. 2009; Arbeev et al. 2012). As the original SPM, this modification (the GenSPM) is especially relevant in the context of research on aging as it incorporates several essential mechanisms of aging-related changes in organisms and it allows for evaluating genetic effects on such characteristics and their influence on mortality or onset of a disease. In this paper, we formulate the discrete-time specification of the model modified to include the dependence of the probability of having an allele/genotype in the genome of an individual and other components of the model on observed covariates. Continuous-time version is presented elsewhere, see, e.g., Arbeev et al. (2014). We also describe

simulation studies which illustrate the increase in power of joint analyses of genotyped and non-genotyped participants in a longitudinal study compared to analyses of genotyped participants alone.

2. The Latent Class Stochastic Process Model

2.1 Specification of the Model

We present a one-dimensional specification of the model here. The situation when several longitudinal variables need to be analyzed jointly can be accommodated as well (Yashin et al. 2008).

Consider a population of N independent individuals at the baseline that can belong to a finite number K of latent subpopulations or latent classes. One specific example of such subpopulations could be carriers of some alleles or genotypes at some gene or single nucleotide polymorphism (SNP) when the corresponding genetic information is not collected in the data. Denote by Z_i a random variable identifying the latent class membership for i^{th} individual, that is, $Z_i = k$ if the individual belongs to the class $k = 1 \ldots K$. We can specify the probabilities of the latent class membership conditional on observed covariates. Following a common practice in the joint latent class models (Lin et al. 2002; Proust-Lima et al. 2009; Proust-Lima et al. 2014; Proust-Lima and Taylor 2009), we represent this probability using a multinomial logistic regression:

$$P(Z_i = k \mid X_i^0) = \frac{e^{\beta_{0k} + \beta_{1k}^T X_i^0}}{1 + \sum_{c=1}^{K-1} e^{\beta_{0c} + \beta_{1c}^T X_i^0}}, \tag{1}$$

for $k = 1 \ldots K-1$, and

$$P(Z_i = K \mid X_i^0) = 1 - \sum_{k=1}^{K-1} P(Z_i = k \mid X_i^0) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_{0k} + \beta_{1k}^T X_i^0}}. \tag{2}$$

Here β_{0k} and β_{1k} are the intercept and the column vector of class-specific regression parameters, respectively, for the latent class k, and X_i^0 is the corresponding (column) vector of time-independent covariates for i^{th} individual ("T" denotes transposition).

The longitudinal sub-model in the LCSPM specifies the latent-class specific dynamics of a longitudinal variable measured in individual i, $i = 1 \ldots N$, at age t_j^i, $j = 1 \ldots n_i$, $Y_{t_j^i}$ (omitting its dependence on k for brevity of notations), similar to Yashin et al. (2007b):

$$\begin{aligned} Y_{t_{j+1}^i} = Y_{t_j^i} + a(t_j^i, k, X_{t_j^i})(Y_{t_j^i} - f_1(t_j^i, k, X_{t_j^i}))(t_{j+1}^i - t_j^i) + \\ B(t_j^i, k, X_{t_j^i})\sqrt{t_{j+1}^i - t_j^i}\,\varepsilon(t_j^i) \end{aligned}, \tag{3}$$

with $Y_{t_1^i} \sim N(f_1(t_1^i, k, X_{t_1^i}), \sigma_0^2(t_1^i, k, X_{t_1^i}))$ (note that this specification of the initial values is arbitrary so it can be modified in particular applications, for example $Y_{t_1^i}$ can be

assumed fixed). Here $\varepsilon_{t_j^i} \sim N(0,1)$ and $X_{\cdot}$ is a vector of possibly time-dependent covariates (which may include some variables from X^0).

Equation (3) has some properties useful in applications to research on aging for modelling relevant biological mechanisms (see more discussion on this topic in Arbeev et al. (2014)). Interpretation of $f_1(\cdot)$ and $a(\cdot)$ in terms of aging-related mechanisms (allostatic trajectory and adaptive capacity) is discussed in Yashin et al. (2007a; 2012) and Arbeev et al. (2011).

The time-to-event sub-model specifies the latent class-specific probability of death in the time interval conditional on observations of $Y_{t_j^i}$ and $X_{t_j^i}$ and survival until the beginning of the interval:

$$M(t_j^i, k, Y_{t_j^i}, X_{t_j^i}) = P(D_{t_j^i} = 1 \mid Y_{t_j^i}, Z_i = k, X_{t_j^i}, D_{t_{j-1}^i} = 0) = 1 - e^{-\mu(t_j^i, Y_{t_j^i}, k, X_{t_j^i})(t_{j+1}^i - t_j^i)} \quad (4)$$

where

$$\mu(t_j^i, Y_{t_j^i}, k, X_{t_j^i}) = \mu_0(t_j^i, k, X_{t_j^i}) + Q(t_j^i, k, X_{t_j^i})(Y_{t_j^i} - f_0(t_j^i, k, X_{t_j^i}))^2 \quad (5)$$

and $D_{t_j^i} = 1$ ($D_{t_j^i} = 0$) indicate that the event (death) happened (did not happen) in the interval $(t_j^i, t_{j+1}^i]$.

The interpretation of $\mu_0(\cdot)$, $f_0(\cdot)$, and $Q(\cdot)$ in applications to research on aging (baseline hazard, "physiological norm," and the quadratic term associated with stress resistance) are discussed in Yashin et al. (2007a; 2012) and Arbeev et al. (2011). Note that (5) assumes a symmetric U-shape (as a function of *Y*) so that the same deviation of *Y* from f_0 to the smaller or to the larger values causes the same increase in $\mu(t_j^i, Y_{t_j^i}, k, X_{t_j^i})$. It may be a reasonable assumption in some applications but in general one may expect that this relationship could be non-symmetric. In this case (5) can be generalized as

$$\begin{aligned} \mu(t_j^i, Y_{t_j^i}, k, X_{t_j^i}) = \mu_0(t_j^i, k, X_{t_j^i}) + Q_L(t_j^i, k, X_{t_j^i})(Y_{t_j^i} - f_0(t_j^i, k, X_{t_j^i}))^2 \times \\ I(Y_{t_j^i} \le f_0(t_j^i, k, X_{t_j^i})) + \\ Q_R(t_j^i, k, X_{t_j^i})(Y_{t_j^i} - f_0(t_j^i, k, X_{t_j^i}))^2 I(Y_{t_j^i} > f_0(t_j^i, k, X_{t_j^i})) \end{aligned} \quad (6)$$

where *I*(.) is the indicator function.

2.2 Likelihood Function

Let τ_i be age at death or censoring for i^{th} individual and $\delta_i = 1$ if this individual died at age τ_i and $\delta_i = 0$ if he/she is censored at that age. Note that in real data applications it is possible that follow-up information on some individuals is not available after the age at last observation $t_{n_i}^i$ in which case one can assume $\tau_i = t_{n_i}^i + 1/365.25$ (if the unit of

measurement for age is years) to accommodate the last observation in the likelihood function.

Let us first introduce expressions used in the formula of the likelihood function:

$$\Phi_0(k, t_1^i, X_i^0) = \frac{1}{\sqrt{2\pi}\sigma_0(t_1^i, k, X_i^0)} e^{-\frac{(Y_{t_1^i} - f_1(t_1^i, k, X_i^0))^2}{2\sigma_0^2(t_1^i, k, X_i^0)}}, \quad (7)$$

$$\Phi(k, t_{j-1}^i, t_j^i, Y_{t_{j-1}^i}, X_{t_{j-1}^i}) = \frac{1}{\sqrt{2\pi(t_j^i - t_{j-1}^i)}B(t_{j-1}^i, k, X_{t_{j-1}^i})} e^{-\frac{(Y_{t_j^i} - \bar{Y}(t_{j-1}^i, k, X_{t_{j-1}^i}))^2}{2(t_j^i - t_{j-1}^i)B^2(t_{j-1}^i, k, X_{t_{j-1}^i})}}, \quad (8)$$

$$\bar{Y}(t_{j-1}^i, k, X_{t_{j-1}^i}) = Y_{t_{j-1}^i} + a(t_{j-1}^i, k, X_{t_{j-1}^i})(Y_{t_{j-1}^i} - f_1(t_{j-1}^i, k, X_{t_{j-1}^i}))(t_j^i - t_{j-1}^i). \quad (9)$$

Consider the conditional probability

$$\theta(t_j^i, k) = P(Z_i = k \mid Y_{t_j^i}, X_{t_j^i}, D_{t_j^i}) \quad (10)$$

which is for $j < n_i$

$$\theta(t_j^i, k) = \frac{\theta(t_{j-1}^i, k)\Phi(k, t_{j-1}^i, t_j^i, Y_{t_{j-1}^i}, X_{t_{j-1}^i})(1 - M(t_j^i, k, Y_{t_j^i}, X_{t_j^i}))}{\sum_{c=1}^{K} \theta(t_{j-1}^i, c)\Phi(c, t_{j-1}^i, t_j^i, Y_{t_{j-1}^i}, X_{t_{j-1}^i})(1 - M(t_j^i, c, Y_{t_j^i}, X_{t_j^i}))} \quad (11)$$

and for $j = n_i$

$$\theta(t_{n_i}^i, k) = \frac{\begin{array}{r}\theta(t_{n_i-1}^i, k)\Phi(k, t_{n_i-1}^i, t_{n_i}^i, Y_{t_{n_i-1}^i}, X_{t_{n_i-1}^i})(1 - M(t_{n_i}^i, k, Y_{t_{n_i}^i}, X_{t_{n_i}^i}))^{1-\delta_i} \times \\ M(t_{n_i}^i, k, Y_{t_v^i}, X_{t_v^i})^{\delta_i}\end{array}}{\begin{array}{r}\sum_{c=1}^{K} \theta(t_{n_i-1}^i, c)\Phi(c, t_{n_i-1}^i, t_{n_i}^i, Y_{t_{v-1}^i}, X_{t_{n_i-1}^i})(1 - M(t_{n_i}^i, c, Y_{t_{n_i}^i}, X_{t_{n_i}^i}))^{1-\delta_i} \times \\ M(t_{n_i}^i, c, Y_{t_{n_i}^i}, X_{t_{n_i}^i})^{\delta_i}\end{array}} \quad (12)$$

where the evolution of $\theta(t_j^i, k)$ starts with $P(Z_i = k \mid X_i^0)$ at the baseline. Let also

$$\bar{\Phi}(t_{j-1}^i, t_j^i, Y_{t_{j-1}^i}, X_{t_{j-1}^i}) = \sum_{k=1}^{K} \theta(t_{j-1}^i, k)\Phi(k, t_{j-1}^i, t_j^i, Y_{t_{j-1}^i}, X_{t_{j-1}^i}) \quad (13)$$

The likelihood function is

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n_i} LY_{ij} LD_{ij}, \quad (14)$$

where

$$LY_{ij} = \sum_{k=1}^{K} \Phi(k, t^i_{j-1}, t^i_j, Y_{t^i_{j-1}}, X_{t^i_{j-1}}) \theta(t^i_{j-1}, k), \; j = 2 \ldots n_i, \tag{15}$$

$$LY_{i1} = \sum_{k=1}^{K} \Phi_0(k, t^i_1, X^0_i) P(Z_i = k \mid X^0_i), \tag{16}$$

and

$$LD_{ij} = \sum_{k=1}^{K} (1 - M(t^i_j, k, Y_{t^i_j}, X_{t^i_j})) \frac{\theta(t^i_{j-1}, k) \Phi(k, t^i_{j-1}, t^i_j, Y_{t^i_{j-1}}, X_{t^i_{j-1}})}{\overline{\Phi}(t^i_{j-1}, t^i_j, Y_{t^i_{j-1}}, X_{t^i_{j-1}})}, \; j < n_i, \tag{17}$$

$$LD_{in_i} = \sum_{k=1}^{K} (1 - M(t^i_{n_i}, k, Y_{t^i_{n_i}}, X_{t^i_{n_i}}))^{1-\delta_i} M(t^i_{n_i}, k, Y_{t^i_{n_i}}, X_{t^i_{n_i}})^{\delta_i} \times \frac{\theta(t^i_{n_i-1}, k) \Phi(k, t^i_{n_i-1}, t^i_{n_i}, Y_{t^i_{n_i-1}}, X_{t^i_{n_i-1}})}{\overline{\Phi}(t^i_{n_i-1}, t^i_{n_i}, Y_{t^i_{n_i-1}}, X_{t^i_{n_i-1}})} \ldots . \tag{18}$$

2.3 Simulations

We performed a simulation study to illustrate the situation when the latent structure in the data is taken into account and the situation when such latent structure is ignored in the estimation procedure. We simulated 100 datasets (5,000 individuals in each) using data structure resembling the Framingham Heart Study data (Dawber et al. 1951) and chose parameters producing realistic mortality rates. We used linear functions of age for $\ln \mu_0(\cdot)$, $Q(\cdot)$, $f_0(\cdot)$, $a(\cdot)$, and $f_1(\cdot)$ and constant (i.e., age-independent) $B(\cdot)$ and $\sigma_0^2(\cdot)$. We did not model dependence on observed covariates in these parameters for simplicity. The probabilities of the latent class membership are assumed to be a function of two covariates: a binary (with probability of each outcome 0.5) and a continuous one assumed to have a standard normal distribution.

Figure 1 (left column) displays estimated trajectories of different components of the model in two latent classes for 100 simulated datasets. The results show that the model correctly separates all model components for two latent classes. Figure 1 (right column) shows estimated components in the model that ignores the latent structure. Although in some cases the estimates in the latter model have noticeably smaller standard deviations (because in this case the entire sample is used to estimate them), the estimates themselves do not correspond to the true trajectories in the latent classes. This is a simplified "toy" example and in the reality the situation can be much more complicated and the latent classes can have diverse dynamics of the longitudinal process and its relation to the time-to-event outcome. Therefore, making conclusions based on the "entire sample" or "population" estimates can be risky.

Figure 1 about here

3. The Genetic Stochastic Process Model

3.1 Specification of the Model

We present a one-dimensional specification of the model here. The situation when several longitudinal variables need to be analyzed jointly can be accommodated as well (Arbeev et al. 2009).

Consider a study with N independent individuals at the baseline and let $N = N_{gen} + N_{nong}$, where N_{gen} and N_{nong} are the numbers of genotyped and non-genotyped individuals in the sample, respectively. Denote by Z_i a random variable identifying the presence of allele or genotype k in the genome of i^{th} individual, $k = 1 \ldots K$ (for example, it may be a binary variable coding the presence/absence of minor allele at some locus, or it may variable representing minor allele homozygote, heterozygote and major allele homozygote). For the genotyped individuals, information on a genetic marker is available (i.e., the value k is known) but for the non-genotyped individuals this value is unknown (still, the longitudinal and time-to-event information is available for them).

We can specify the probabilities of having allele or genotype k conditional on some observed (time-independent) covariates as in (1), (2). The longitudinal sub-model in the GenSPM specifies the allele- or genotype-specific dynamics of a longitudinal variable measured in individual i at age t_j^i, $j = 1 \ldots n_i$, $Y_{t_j^i}$ as in (3). The time-to-event sub-model specifies the allele or genotype-specific probability of death in the time interval conditional on observations of $Y_{t_j^i}$ and $X_{t_j^i}$ and survival until the beginning of the interval as in (4), (5) (or (4), (6)).

Note that here we use exactly the same expressions (1)-(6) as in the LCSPM. Nevertheless, construction of the likelihood function is different because in the LCSPM the latent classes are not known for any individual whereas in the GenSPM the genetic data (i.e., the values k) are assumed to be collected for at least a sub-sample of participants of the longitudinal study.

3.2 Likelihood Function

The likelihood for the genotyped individuals is

$$L_{gen} = \prod_{i=1}^{N_{gen}} P(Z_i = k_i \mid X_i^0) \prod_{j=1}^{n_i} LY_{ij}(k_i) LD_{ij}(k_i), \tag{19}$$

where k_i is the value of the allele/genotype variable for i^{th} individual, and (using notations introduced for the LCSPM)

$$LY_{ij}(k) = \Phi(k, t_{j-1}^i, t_j^i, Y_{t_{j-1}^i}, X_{t_{j-1}^i}), \; j = 2 \ldots n_i, \tag{20}$$

$$LY_{i1}(k) = \Phi_0(k, t_1^i, X_i^0), \tag{21}$$

$$LD_{ij}(k) = 1 - M(t_j^i, k, Y_{t_j^i}, X_{t_j^i}) \;, \; j < n_i, \tag{22}$$

$$LD_{in_i}(k) = (1 - M(t_{n_i}^i, k, Y_{t_{n_i}^i}, X_{t_{n_i}^i}))^{1-\delta_i} M(t_{n_i}^i, k, Y_{t_{n_i}^i}, X_{t_{n_i}^i})^{\delta_i} \tag{23}$$

The likelihood for the non-genotyped individuals is

$$L_{nong} = \prod_{i=N_{gen}+1}^{N} \sum_{k=1}^{K} P(Z_i = k \mid X_i^0) \prod_{j=1}^{n_i} LY_{ij}(k) LD_{ij}(k) \tag{24}$$

where the respective expressions are given by (1), (2), and (20) – (23).

The total likelihood for the genotyped and non-genotyped individuals is

$$L = L_{gen} L_{nong} \tag{25}$$

with respective components given by (19) and (24).

Importantly, the likelihood function contains the same parameters for the genotyped and non-genotyped sub-samples. Therefore, the joint analysis of available information for the non-genotyped participants (such as the longitudinal measurements and time-to-event data) along with that for the genotyped sub-sample provides an opportunity for improving power compared to analyses based on the genotyped individuals alone. The advantage of the genetic SPM in applications to research on aging is that it has different components representing specific biological concepts and aging-related mechanisms for which the respective parameters have clear biological interpretations. This allows for testing different hypotheses on the presence of genetic effect of the alleles/genotypes on respective aging-related characteristics (such as stress resistance, adaptive capacity, age-dependent physiological norms, etc.) which is not possible in the traditional analyses.

3.3 Simulations

We performed a simulation study to illustrate the increase in the accuracy and power in joint analyses of genotyped and non-genotyped individuals compared to analyses of genotyped individuals alone. We simulated 100 datasets (2,500 individuals in each) using data structure resembling the Framingham Heart Study data (Dawber et al. 1951) and chose parameters producing realistic mortality rates. We assumed that 25% of the sample is genotyped and these individuals have information on some genetic marker (carriers/non-carriers of some allele/genotype; the proportion of carriers at birth, p_1, is supposed to be 0.25). For the rest of the sample, information on the genetic marker is not available for the estimation procedure but information on the longitudinal variable and follow-up is available for the entire sample. We used linear functions of age for $\ln \mu_0(\cdot)$, $Q(\cdot)$, $f_0(\cdot)$, $a(\cdot)$, and $f_1(\cdot)$ and constant (i.e., age-independent) $B(\cdot)$ and $\sigma_0^2(\cdot)$. We performed several studies using different specifications of the model's components aimed at testing various null hypotheses on the equality of the components in carriers and non-carriers of some allele/genotype. For simplicity, we did not model dependence on observed covariates in all components except for $\mu_0(\cdot)$ in one study. Table 1 summarizes parameters used in the studies.

In each study, we estimated parameters in all datasets using two likelihood functions: 1) joint likelihood for the genotyped and non-genotyped individuals; 2) likelihood for the genotyped sample. As the joint likelihood uses information from non-genotyped individuals (follow-up data and longitudinal measurements), the resulting estimates of parameters are more accurate than in the case of analysis of the genotyped sample alone.

In Study 1, the standard deviations of respective parameters (see highlighted in Table 1) computed from the estimates in 100 simulated datasets using the "joint" likelihood are about 2.2 – 3.4 times smaller than those from the "genotyped-only" version (denote this ratio $s.d._G / s.d._J$). Respectively, the power (for $\alpha = 0.05$ and the effect sizes defined by the highlighted parameters from corresponding rows in Table 1) differs for the "genotyped only" (w_G) and the "joint" likelihoods (w_J): $w_G = 0.34$ and $w_J = 0.81$. Similar results are observed in the other studies: ($s.d._G / s.d._J$ ranges 3.1 – 6.9, $w_G = 0.42$, $w_J = 0.95$ in Study 2; $s.d._G / s.d._J$ ranges 1.9 – 2.2, $w_G = 0.35$, $w_J = 0.81$ in Study 3; $s.d._G / s.d._J$ ranges 1.7 – 2.1, $w_G = 0.48$, $w_J = 0.96$ in Study 4; $s.d._G / s.d._J$ ranges 1.4 – 1.8, $w_G = 0.73$, $w_J = 0.88$ in Study 5; and $s.d._G / s.d._J$ ranges 1.8 – 2.2, $w_G = 0.64$, $w_J = 0.97$ in Study 6). In some studies (1, 2, and 4) a few data sets in the "genotyped-only" version produced estimates at the boundaries set in the constrained maximization procedure. This was not observed in any study and any dataset when the "joint" likelihood was used.

Table 1 about here

4. Conclusions

We presented two variants of the stochastic process models, the LCSPM and GenSPM. Both models can be useful in applications to research on aging as they have components that incorporate several aging-related mechanisms and have a clear interpretation relevant in the field. The models allow formulating and testing relevant biological hypotheses on the presence of these "hidden" components of the process of aging and their impact on the risk of death or developing aging-related diseases.

The LCSPM allows one to investigate the effects of unobserved heterogeneity (latent structure) that may distort conclusions in joint analyses of longitudinal and time-to-event outcomes when such hidden structure is present in the data but ignored in the analyses. Thus analyses by the LCSPM can complement the analyses by the original SPM to test the hypotheses on the presence of hidden heterogeneity in the data and to appropriately adjust the conclusions or analytic approach if such structure is revealed.

The GenSPM can be applied to genetic analyses in the field of research on aging as it introduces dependence of major components representing aging-related mechanisms on genetic markers. The model can be used to test the hypotheses on the presence of genetic effects on different aging-related processed to help determine genetic underpinning of longevity and healthy lifespan. The approach also combines data from genotyped and non-genotyped individuals. Such joint analyses can increase the power compared to analyses based on information from the genotyped subsample alone.

The discrete-time specifications of the model presented here are simplifications of the more comprehensive continuous versions (Arbeev et al. 2014). They, however, have important practical advantage as the likelihood optimization in these models takes considerably less time than in case of the continuous models which require solutions of differential equations at each step of the optimization procedure. Therefore, the discrete-time counterparts can be used at the initial stage of the analyses when, for example, a large number of genetic variants (e.g., SNPs) needs to be analyzed. The continuous

versions can be then applied at the second stage when the candidate variants are selected for more detailed analyses. The discrete-time models are also convenient for use in comprehensive sensitivity analyses when a large number of assumptions should be tested. Also they can be used to quickly estimate the initial value for the likelihood optimization procedure for the continuous model which, especially in the multidimensional case, can significantly improve the speed of convergence to the maximum. Thus, the discrete-time versions of the LCSPM and the GenSPM provide convenient and practical alternative for extensive applications to data analyses.

Acknowledgements

This work was partly supported by the National Institute on Aging of the National Institutes of Health under Award Numbers R01AG046860, P01AG043352, and P30AG034424. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Arbeev KG et al. (2009) Genetic model for longitudinal studies of aging, health, and longevity and its potential application to incomplete data J Theor Biol 258:103-111 doi:10.1016/j.jtbi.2009.01.023

Arbeev KG, Akushevich I, Kulminski AM, Ukraintseva S, Yashin AI (2014) Joint analyses of longitudinal and time-to-event data in research on aging: Implications for predicting health and survival Frontiers in Public Health 2:article 228 doi:10.3389/fpubh.2014.00228

Arbeev KG et al. (2011) Age trajectories of physiological indices in relation to healthy life course Mech Ageing Dev 132:93-102 doi:10.1016/j.mad.2011.01.001

Arbeev KG et al. (2012) Effect of the APOE Polymorphism and Age Trajectories of Physiological Variables on Mortality: Application of Genetic Stochastic Process Model of Aging Scientifica 2012:Article ID 568628 doi:10.6064/2012/568628

Dawber TR, Meadors GF, Moore FE (1951) Epidemiological approaches to heart disease: The Framingham Study Am J Public Health 41:279-286

Lin HQ, Turnbull BW, McCulloch CE, Slate EH (2002) Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer J Amer Statistical Assoc 97:53-65 doi:10.1198/016214502753479220

Proust-Lima C, Joly P, Dartigues J-F, Jacqmin-Gadda H (2009) Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach Comput Stat Data Anal 53:1142-1154 doi:10.1016/j.csda.2008.10.017

Proust-Lima C, Sene M, Taylor JM, Jacqmin-Gadda H (2014) Joint latent class models for longitudinal and time-to-event data: a review Stat Methods Med Res 23:74-90 doi:10.1177/0962280212445839

Proust-Lima C, Taylor JMG (2009) Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach Biostatistics 10:535-549 doi:10.1093/biostatistics/kxp009

Woodbury MA, Manton KG (1977) A random-walk model of human mortality and aging Theor Popul Biol 11:37-48

Yashin AI, Arbeev KG, Akushevich I, Kulminski A, Akushevich L, Ukraintseva SV (2007a) Stochastic model for analysis of longitudinal data on aging and mortality Math Biosci 208:538-551 doi:10.1016/j.mbs.2006.11.006|ISSN 0025-5564

Yashin AI, Arbeev KG, Akushevich I, Kulminski A, Akushevich L, Ukraintseva SV (2008) Model of hidden heterogeneity in longitudinal data Theor Popul Biol 73:1-10 doi:10.1016/j.tpb.2007.09.001|ISSN 0040-5809

Yashin AI, Arbeev KG, Akushevich I, Kulminski A, Ukraintseva SV, Stallard E, Land KC (2012) The quadratic hazard model for analyzing longitudinal data on aging, health, and the life span Physics of Life Reviews 9:177-188 doi:10.1016/j.plrev.2012.05.002

Yashin AI, Arbeev KG, Kulminski A, Akushevich I, Akushevich L, Ukraintseva SV (2007b) Health decline, aging and mortality: how are they related? Biogerontology 8:291-302 doi:10.1007/s10522-006-9073-3|ISSN 1389-5729

Figures:

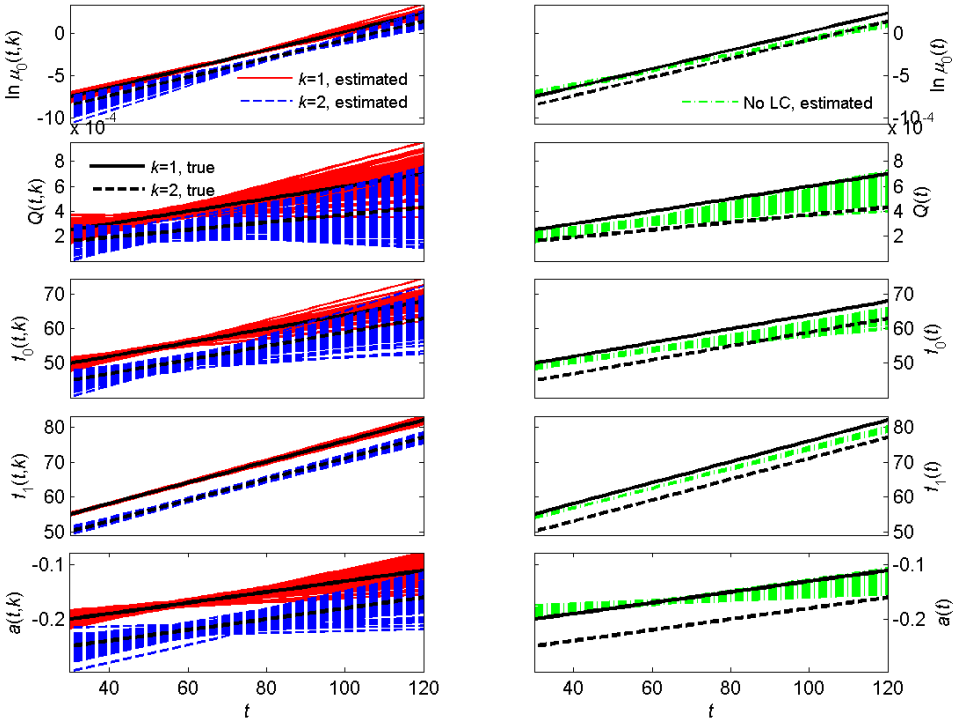

Figure 1. Simulation studies in data with latent structure: latent-class trajectories estimated by the LCSPM (left column) and "population" trajectories estimated by the original SPM (right column) that ignores this latent structure. Thick lines denote true trajectories in two latent classes (k = 1, 2). Thin lines denote estimates in 100 simulated datasets either in the latent classes (left column) or in the general sample ("NoLC", right column).

Tables:

Table 1: Simulation studies of the GenSPM: Parameters used to generate data in different studies (parameters defining the null hypotheses in each specific study are ***highlighted***)

Study	***k***	**Parameters**													
		$\ln a^k_{\mu_0}$	$b^k_{\mu_0}$	β^k_X	a^k_Q	b^k_Q	a^k_Y	b^k_Y	$a^k_{f_1}$	$b^k_{f_1}$	$a^k_{f_0}$	$b^k_{f_0}$	σ^k_0	B^k	p_1
1	1	***-9.0***	***0.080***		0.5	0.10	-0.25	1.0	45.00	0.20	45.0	0.1	4.0	4.0	0.25
	2	***-8.5***	***0.082***		0.3	0.10	-0.20	1.0	50.00	0.25	40.0	0.1	4.0	4.0	
2	1	-9.0	0.080	***-0.04***	0.5	0.10	-0.25	1.0	45.00	0.20	45.0	0.1	4.0	4.0	0.25
	2	-8.5	0.082	***-0.04***	0.3	0.10	-0.20	1.0	50.00	0.25	40.0	0.1	4.0	4.0	
3	1	-9.0	0.080		***0.5***	***0.10***	-0.25	1.0	45.00	0.20	45.0	0.1	4.0	4.0	0.25
	2	-8.5	0.082		***0.5***	***0.23***	-0.20	1.0	50.00	0.25	40.0	0.1	4.0	4.0	
4	1	-9.0	0.080		0.5	0.10	***-0.25***	***1.0***	45.00	0.20	45.0	0.1	4.0	4.0	0.25
	2	-8.5	0.082		0.3	0.10	***-0.23***	***1.0***	50.00	0.25	40.0	0.1	4.0	4.0	
5	1	-9.0	0.080		0.5	0.10	-0.25	1.0	***45.00***	***0.20***	45.0	0.1	4.0	4.0	0.25
	2	-8.5	0.082		0.3	0.10	-0.20	1.0	***45.75***	***0.20***	40.0	0.1	4.0	4.0	
6	1	-9.0	0.080		0.5	0.10	-0.25	1.0	45.00	0.20	***45.0***	***0.1***	4.0	4.0	0.25
	2	-8.5	0.082		0.3	0.10	-0.20	1.0	50.00	0.25	***40.0***	***0.1***	4.0	4.0	

Notes:

Some parameters are rescaled for better visibility in the table: a^k_Q is multiplied by 10^4; b^k_Q is multiplied by 10^5; b^k_Y is multiplied by 10^3; k = 1, 2 denotes carriers and non-carriers of some allele/genotype, respectively. Specifications of components: 1) $\ln \mu_0(t,k,X) = \ln a^k_{\mu_0} + b^k_{\mu_0} t + \beta^k_X X$, (where $X = c - c_0$, c is year of birth (cohort), c_0 = 1890, and X is uniformly distributed over [1890, 1920]) in Study 2, and $\ln \mu_0(t,k,X) = \ln a^k_{\mu_0} + b^k_{\mu_0} t$ in the other studies; 2) $Q(t,k,X) = a^k_Q + b^k_Q t$; 3) $f_1(t,k,X) = a^k_{f_1} + b^k_{f_1} t$; 4) $f_0(t,k,X) = a^k_{f_0} + b^k_{f_0} t$; 5) $a(t,k,X) = a^k_Y + b^k_Y t$, with $a^k_Y \le 0$ and $b^k_Y \ge 0$; 6) $B(t,k,X) = B^k$; 7) $\sigma_0(t,k,X) = \sigma^k_0$; 8) p_1 is the proportion of carriers (we do not assume its dependence on any covariates).