Stochastic Process Model and Its Applications to Analysis of Longitudinal Data

ACM-BCB 2017, Boston, MA

08/20/2017

Ilya Zhbannikov

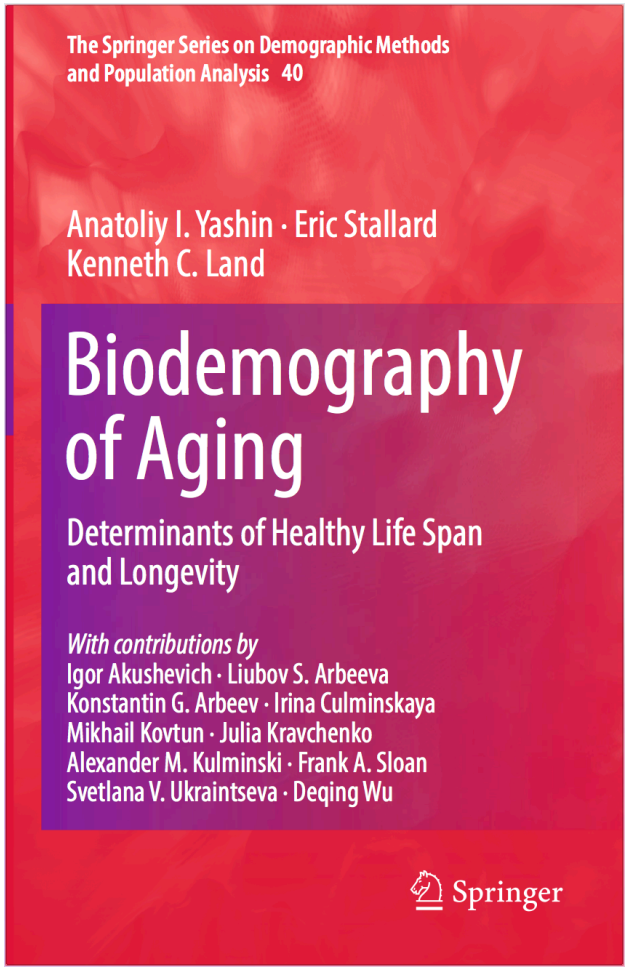
Konstantin Arbeev

Anatoliy Yashin

Duke University, Social Science Research Institute

Synopsis

- Goals and objectives
- Part 1
  - Background
  - Methods
- Part 2
  - R package *stpm*
  - Examples
- Conclusions



This tutorial is also available from: https://github.com/izhbannikov/ACMBCB2017_SPM_Tutorial

Goals and objectives

- Present the Stochastic Process Model (SPM) and discuss its applicability to different research areas.
- Describe software tools and provide practical examples of joint analysis of time-to-event outcomes and longitudinal measurements.
- Present several model extensions: partially observed covariates (e.g. genetics), multiple imputation.
- Establish connections with those who are interested in working with Stochastic Process Model.

Goals and objectives

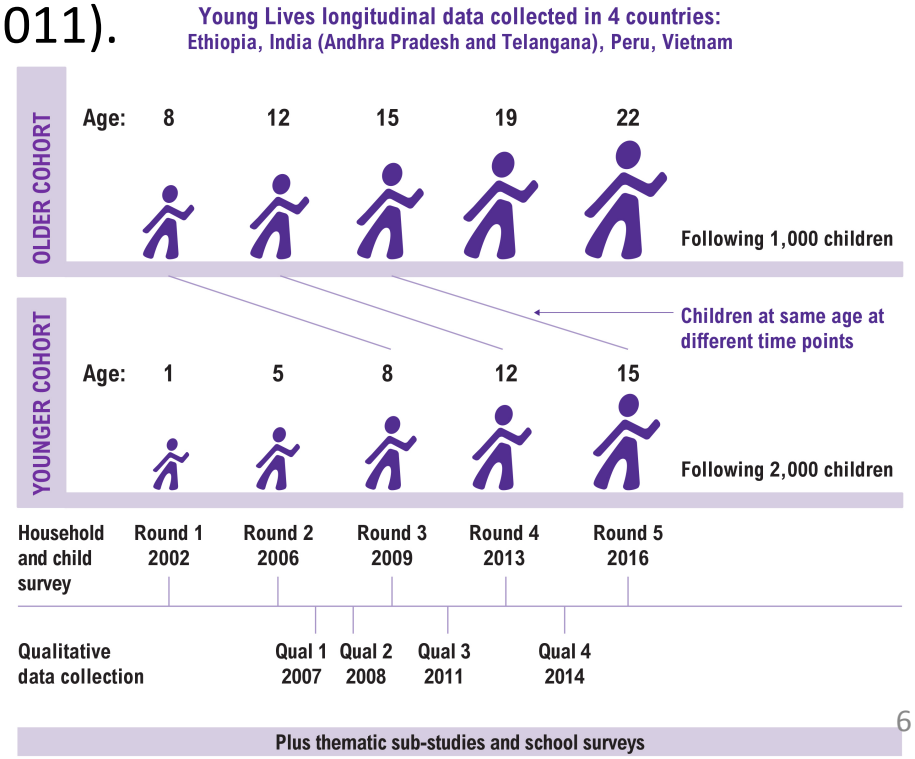
- After attending this tutorial session, participants:
  - Will learn how to jointly analyze time-to-event and longitudinal outcomes using the Stochastic Process Model approach.
  - Will be able to work with the R package *stpm* − a tool to estimate model parameters.
  - Will be provided with examples of analyses of data using SNPs from genes showing pleiotropic effects on different aging-related traits.
  - Will be provided with examples of techniques of longitudinal data imputation with *stpm* R package.

Part 1

- Background
- Methods

Stochastic Process Model: background

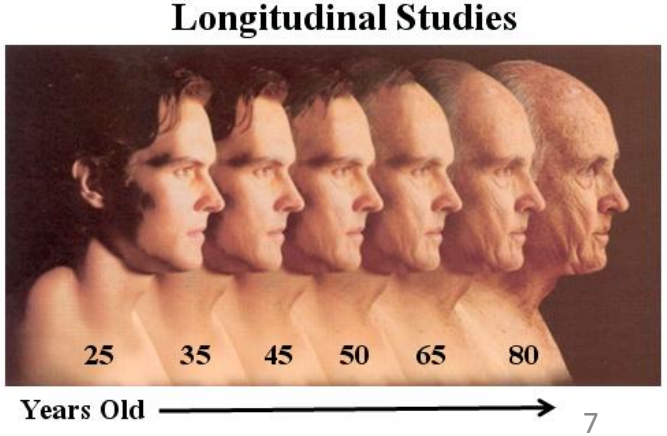
- Longitudinal studies collect various types of outcomes (time-to-event, longitudinal outcomes, such as biomarkers, and other covariates).
- To analyze this information, special statistical methods are needed.
- One of the standard approaches would be the Cox model with time-dependent covariates but it may provide biased results (Prentice, 1982; Sweeting & Thompson, 2011).



© https://www.younglives.org.uk/content/our-research-methods

Stochastic Process Model: background

- The Stochastic Process Model (SPM) is a general framework for joint analysis of time-to-event outcomes and repeatedly measured variables (covariates).
- SPM incorporates knowledge about processes developing over time in system under study which involves deterioration, aging and other kinds of decay and related time-to-event.
- SPM allows for evaluating mechanisms that indirectly affect longitudinal trajectories of physiological covariates and health and survival outcomes.



Stochastic Process Model: background

- History
  - Woodbury & Manton (1977): "Random Walk Model of Human Mortality and Aging".

THEORETICAL POPULATION BIOLOGY 11, 37–48 (1977)

A Random-Walk Model of Human Mortality and Aging

MAX A. WOODBURY

Department of Community Health Sciences,
Duke University Medical School, and
Department of Computer Science, Duke University, Durham, North Carolina 27710

AND

KENNETH G. MANTON

Center for Demographic Studies, Duke University,
Durham, North Carolina 27710

Received September 23, 1975

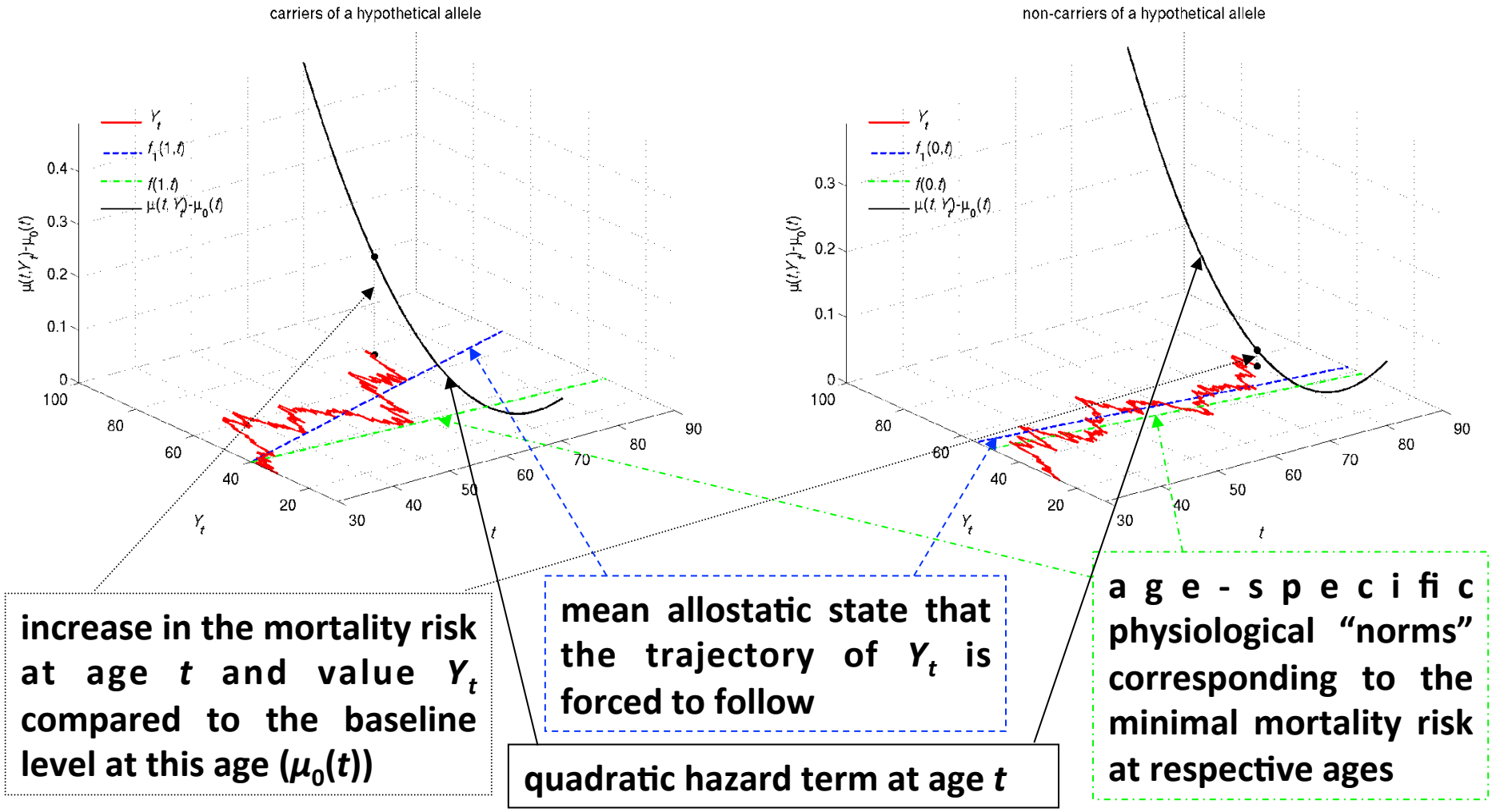
Stochastic Process Model: background

$$d\boldsymbol{x}_{\omega}(t) = \boldsymbol{u}(\boldsymbol{x}_{\omega}, t)dt + d\boldsymbol{\xi}(\boldsymbol{x}_{\omega}, t) \quad (1)$$

$$dP(x_{\omega}) = \mu(\boldsymbol{x}_{\omega}, t)P(\boldsymbol{x}_{\omega})dt \quad (2)$$

$d\boldsymbol{x}_{\omega}(t)$ – changes in physiological variable during time dt
$\boldsymbol{u}(\boldsymbol{x}_{\omega}, t)$ – deterministic effects
$d\boldsymbol{\xi}(\boldsymbol{x}_{\omega}, t)$ – random walk term
$P(x_{\omega})$ – probability of surviving

- The organism is positioned in a state space of physiological variables.
- This model can also be used for description and analysis of changes in technical systems.
- The process of aging is represented by changes in an organism's position in physiological state space.
- Such changes tend to carry the organism into regions of state space which are characterized by probabilities of death $\mu(x_{\omega}, t)dt$ greater than in its previous position.



Schematic illustration of SPM: Increase in mortality risks compared to the baseline mortality (corresponds to zero level) for carriers and non-carriers of a hypothetical allele (t is age and Y_t is the value of a hypothetical physiological variable at this age)

Stochastic Process Model: background

- Woodbury & Manton introduced the quadratic hazard.
- This was convenient because the system dynamics can be described in terms of multi-dimensional Gaussian distribution under some conditions (initial: Gaussian distribution, linear stochastic process and Wiener random process).
- Later this model was extended to take into account some fundamental processes related to system deterioration, aging or decay.
- In application to analyzing human aging, health and longevity, these fundamental processes include age-dependent physiological norm, stress-resistance, adaptive capacity, allostatic adaptation and allostatic load.

Stochastic Process Model: methods

- Age dynamics of a physiological variable ***Y*** (a dynamic component of the extended model*):

$$d\boldsymbol{Y}(t) = \boldsymbol{a}(t)(\boldsymbol{Y}(t) - \boldsymbol{f}_1(t))\mathrm{d}t + \boldsymbol{b}(t)\mathrm{d}\boldsymbol{W}(t),\ \boldsymbol{Y}(t = t_0)$$

$\boldsymbol{a}(t)$ − adaptive capacity
$f_1(t)$ − average allostatic trajectory
$\boldsymbol{b}(t)$ − strength of random disturbances
$\boldsymbol{W}(t)$ − Wiener process

* © Yashin A.I., et al., Stochastic model for analysis of longitudinal data on aging and mortality, Mathematical Biosciences, Volume 208, Issue 2, August 2007, Pages 538-551, ISSN 0025-5564.

Stochastic Process Model: methods

- A hazard component (e.g. a risk to die, mortality risk):

$$\mu(t, \boldsymbol{Y}(t)) = \mu_0(t) + (\boldsymbol{Y}(t) - \boldsymbol{f}(t))^{\mathsf{T}} \boldsymbol{Q}(t)(\boldsymbol{Y}(t) - \boldsymbol{f}(t))$$

$\boldsymbol{f}(t)$ − optimal trajectory that minimizes the risk.
$\boldsymbol{Q}(t)$ − stress resistance (represents deviation from the norm).
$\mu_0(t)$ − baseline hazard.

Stochastic Process Model: background

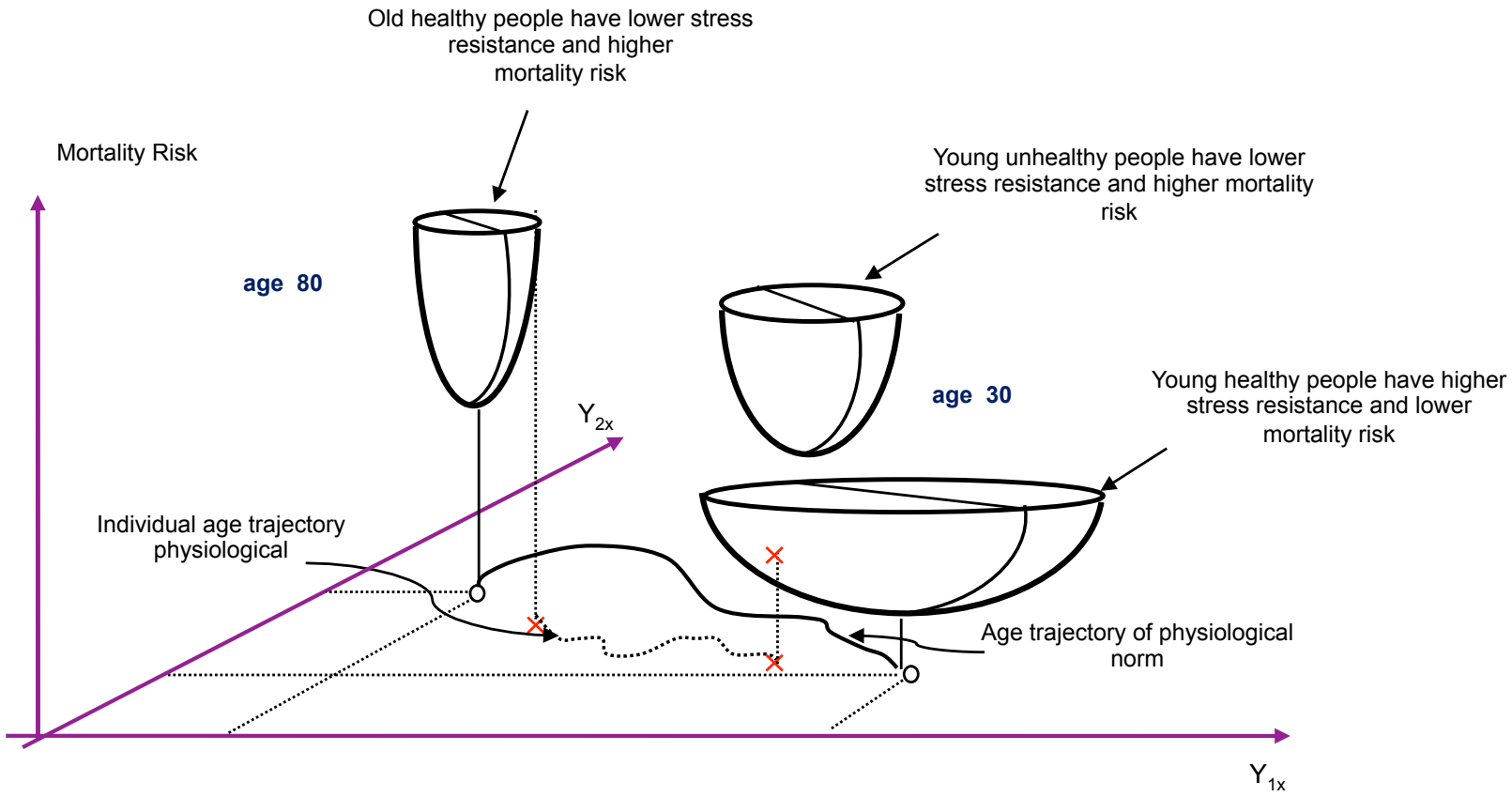


Illustration of hypothetical two-dimensional U-shaped mortality risks considered as a function of two risk factors Y1x and Y2x for the 30 and 80 years old individuals. (by A.I. Yashin, 2015)

Part 2

- Software
  - R package *stpm*
- Practical examples
  - Quick start
  - Data simulation
  - Genetic SPM
  - Multiple imputation of longitudinal data

R-package *stpm*

- Developed for:
  - Data simulation and estimating the model parameters under different scenarios (no genetic effects considered)
  - Estimating genetic effects using joint analysis of longitudinal, mortality and genetic data.
  - Multiple imputation of longitudinal data.
  - First publicly available software that implements Stochastic Process Model methodology.

R-package *stpm*

- R-package *stpm* is available from the following links:
- https://cran.r-project.org/web/packages/stpm/index.html (stable)
- https://github.com/izhbannikov/spm/ (most-recent)

R-package `stpm`

https://cran.r-project.org/package=stpm

stpm: Stochastic Process Model for Analysis of Longitudinal and Time-to-Event Outcomes

Utilities to estimate parameters of the models with survival functions induced by stochastic covariates. Miscellaneous functions for data preparation and simulation are also provided. For more information, see: (i)"Stochastic model for analysis of longitudinal data on aging and mortality" by Yashin A. et al. (2007), Mathematical Biosciences, 208(2), 538-551, <doi:10.1016/j.mbs.2006.11.006>; (ii) "Health decline, aging and mortality: how are they related?" by Yashin A. et al. (2007), Biogerontology 8(3), 291(302), <doi:10.1007/s10522-006-9073-3>.

Version:	1.6.6
Depends:	R (≥ 2.10), Rcpp (≥ 0.11.1), mice
Imports:	sas7bdat, stats, nloptr, survival, tools
LinkingTo:	Rcpp, RcppArmadillo
Suggests:	knitr (≥ 1.11)
Published:	2017-04-07
Author:	I. Y. Zhbannikov, Liang He, K. G. Arbeev, A. I. Yashin.
Maintainer:	Ilya Y. Zhbannikov <ilya.zhbannikov at duke.edu>
License:	GPL-2 | GPL-3 [expanded from: GPL]
NeedsCompilation:	yes
Materials:	README NEWS
CRAN checks:	stpm results

Downloads:

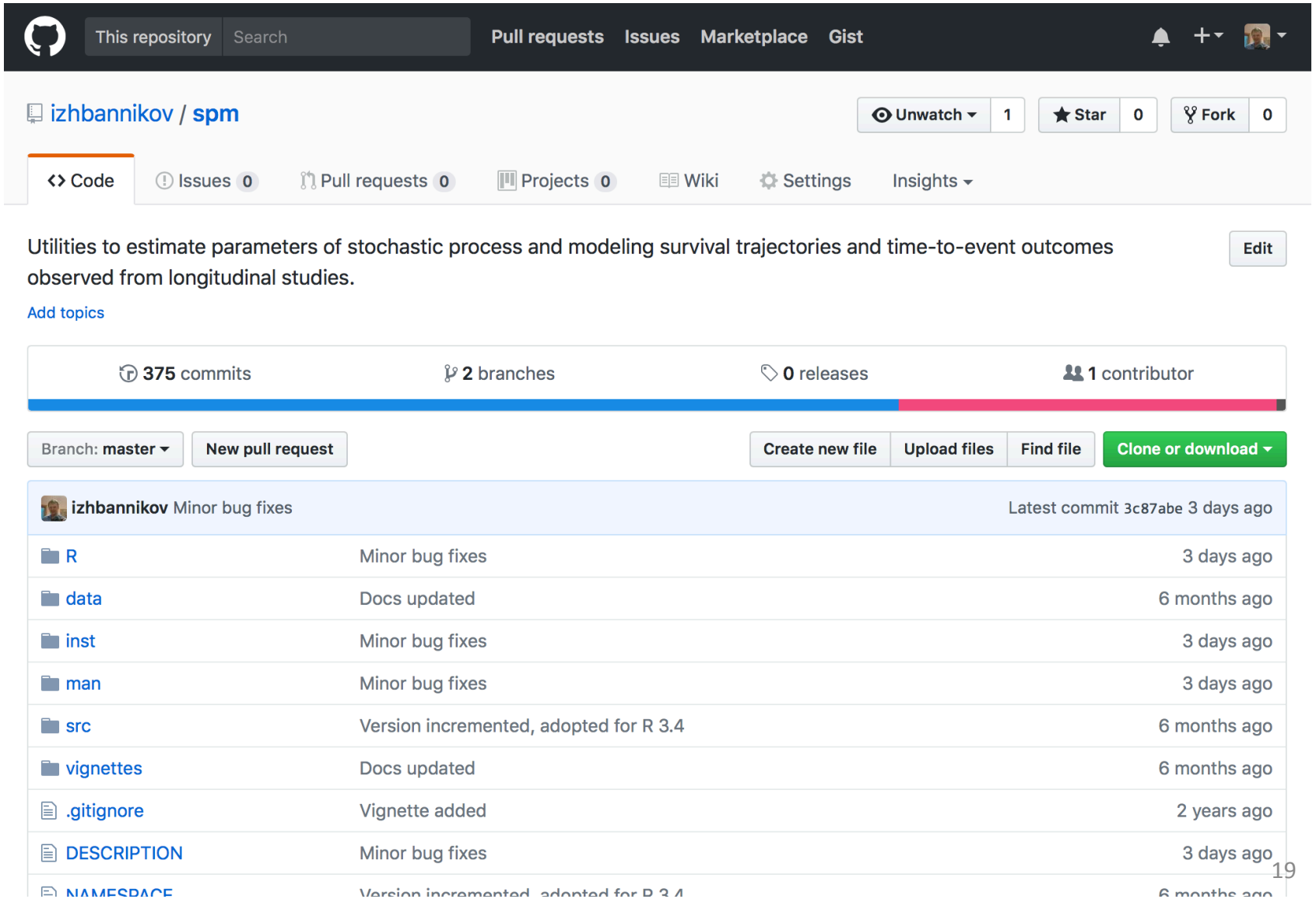
Reference manual:	stpm.pdf
Vignettes:	stpm
Package source:	stpm_1.6.6.tar.gz
Windows binaries:	r-devel: stpm_1.6.6.zip, r-release: stpm_1.6.6.zip, r-oldrel: stpm_1.6.6.zip
OS X El Capitan binaries:	r-release: stpm_1.6.6.tgz
OS X Mavericks binaries:	r-oldrel: stpm_1.6.6.tgz
Old sources:	stpm archive

Linking:

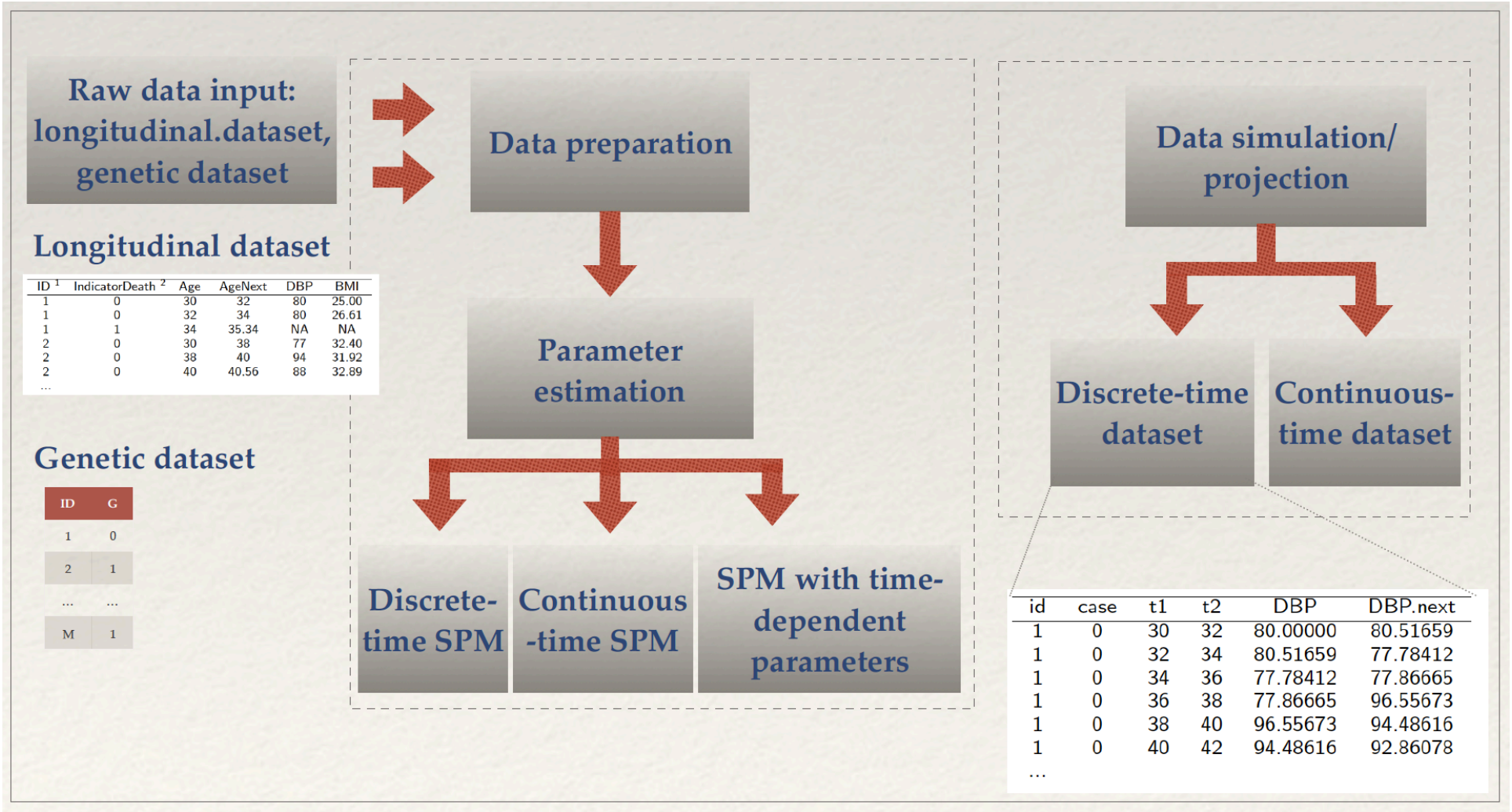
Please use the canonical form `https://CRAN.R-project.org/package=stpm` to link to this page.

R-package `stpm`

https://github.com/izhbannikov/spm



R-package *stpm*



ID [1]	IndicatorDeath [2]	Age	AgeNext	DBP	BMI
1	0	30	32	80	25.00
1	0	32	34	80	26.61
1	1	34	35.34	NA	NA
2	0	30	38	77	32.40
2	0	38	40	94	31.92
2	0	40	40.56	88	32.89
...					

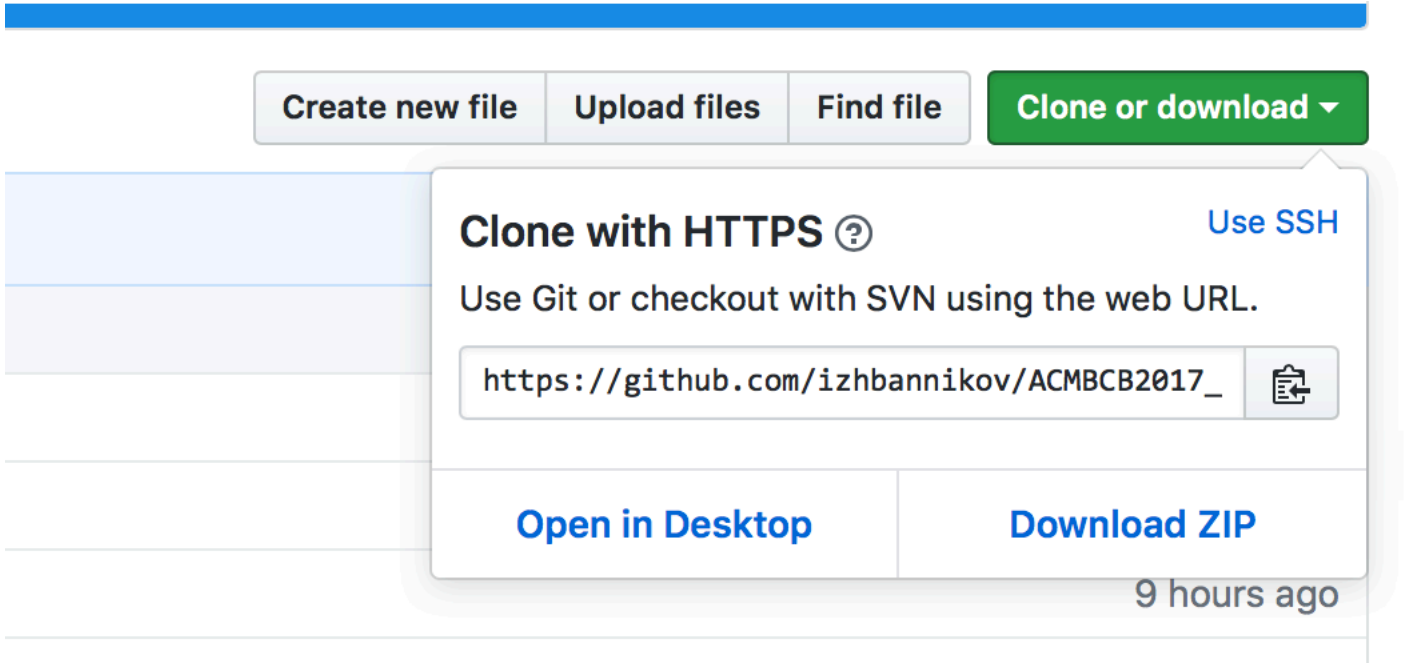
ID	G
1	0
2	1
...	...
M	1

id	case	t1	t2	DBP	DBP.next
1	0	30	32	80.00000	80.51659
1	0	32	34	80.51659	77.78412
1	0	34	36	77.78412	77.86665
1	0	36	38	77.86665	96.55673
1	0	38	40	96.55673	94.48616
1	0	40	42	94.48616	92.86078
...					

Examples

- Please clone or download them from the tutorial page:

https://github.com/izhbannikov/ACMBCB2017_SPM_Tutorial



Package installation

```
#Example1.R

install.packages("stpm")

#OR

library(devtools)
install_github("izhbannikov/spm")
```

Examples

- Typical analysis workflow:
  - Prepare data
  - Estimate parameters
  - Make conclusions

Input data

ID	IndicatorDeath	Age	AgeNext	DBP	BMI
1	0	30	32	80	25
1	0	32	34	80	26.61
1	1	34	35.34	NA	NA
2	0	30	38	77	32.40
2	0	38	40	94	31.92
2	0	40	40.56	88	32.89
...	...	...	...	...	...
2	0	80	80.55	83	26.71
...					

Discrete-time SPM*

$$\mathbf{Y}(t+1) = \mathbf{u} + \mathbf{R}\mathbf{Y}(t) + \epsilon$$

$$\mu(t, \mathbf{Y}(t)) = [\, \mu_0 + \mathbf{b}\mathbf{Y}(t) + \mathbf{Y}(t)^* \mathbf{Q}\mathbf{Y}(t)] \, e^{\theta t}$$

- **u**, **R**, μ_0, **b**, **Q** and **ε** (t) are assumed to be constant and reflect coefficients in the first model form.
- The Gompertz hazard is used and the parameter θ is to be estimated along with other parameters.

* Akushevich, I., et al., Life Tables with Covariates: Dynamic Model for Nonlinear Analysis of Longitudinal Data, Mathematical Population Studies, 12(2):51-80, 2005.

Quick start

```
#Example2.R
#Quick start

library(stpm)

# Let's take a look at raw data
raw.data <- read.csv(system.file("extdata", "longdat.csv",
package="stpm"))

#Prepare data for optimization
data <- prepare_data(x=system.file("extdata", "longdat.csv",
package="stpm"))

#Estimate parameters
#(default model: discrete-time):
p.discr.model <- spm(data)

# Continuous-time model:
p.cont.model <- spm(data, model="continuous”)
```

Code example

```
#Example3.R
#Model with time-dependent parameters

library(stpm)

# Reading raw data
raw.data <- read.csv(system.file("extdata", "longdat.csv", package="stpm"))
head(raw.data)

#Prepare data for optimization
data <- prepare_data(x=system.file("extdata", "longdat.csv",
package="stpm"),
                     col.id = "ID", col.status = "IndicatorDeath",
                     col.age = "Age",
                     covariates = "DBP",
                     impute=FALSE)

pars <- spm_time_dep(x=data[[1]],
                     start = list(a = -0.05, f1 = 80, Q = 2e-08, f = 80,
                                  b = 5, mu0 = 0.001),
                     frm = list(at = "a", f1t = "f1", Qt = "Q",
                                ft = "f", bt = "b", mu0t = "mu0"))
```

Code example

```
#Example4.R

library(stpm)
# Simulation
# Data simulation:
data <- spm_projection(model.par, N=5000,ystart=80,
model="discrete”)
# Print some data:
head(data$data)
# Mean of covariates by age:
data$stat$mean.by.age
# Plot survival probabilities:
plot(data$stat$srv.prob, xlab=“Years",ylab="Percent survival")
```

Code example (cont.)

```
#Example4.R (cont)

ff <- list(at="-0.001*t+.05",
           f1t="60",
           Qt="2e-8",
           ft="80",
           bt="5",
           mu0t="1e-3")

dat <- simdata_time_dep(N=1000, f=ff)
```

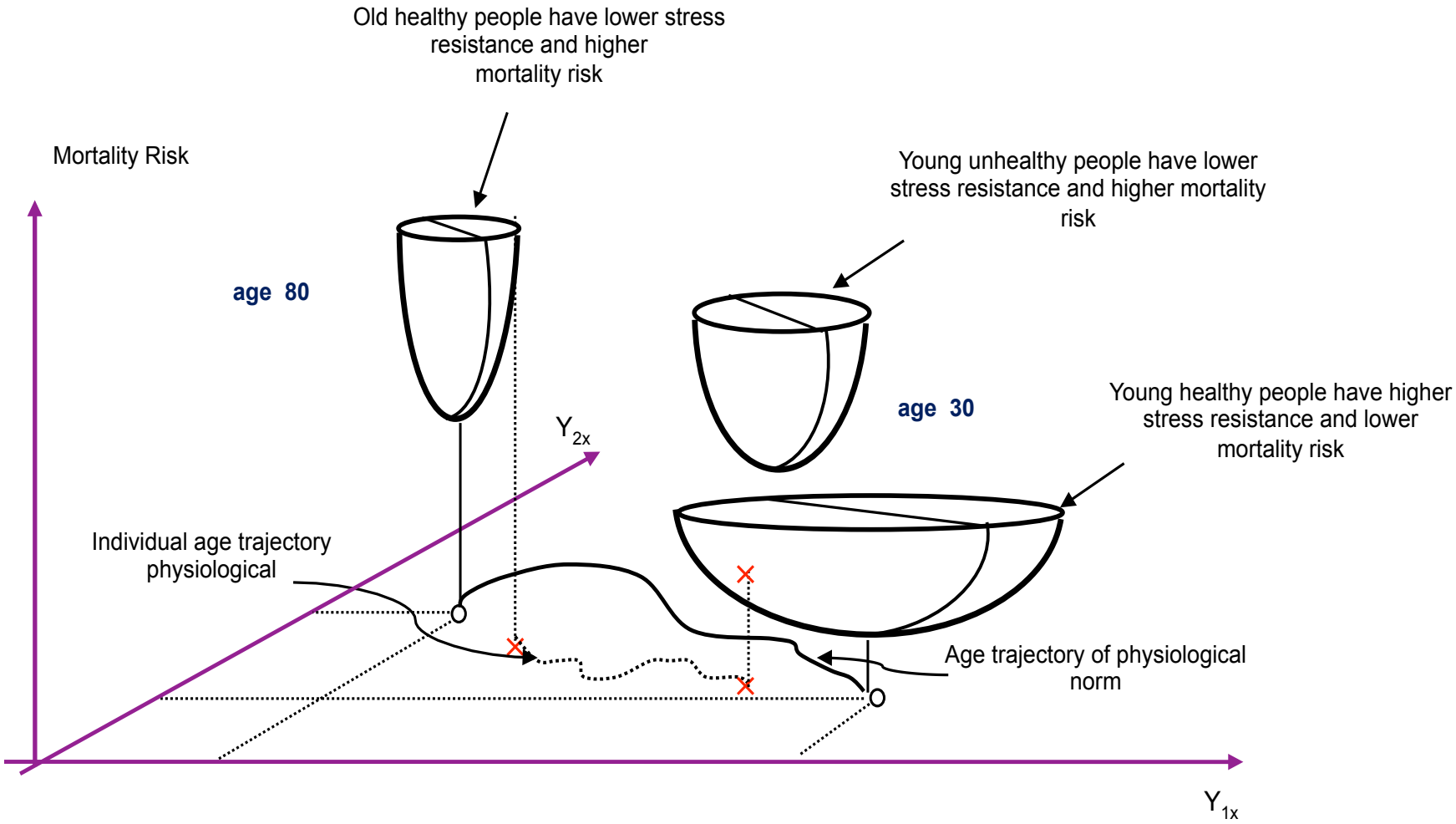


Illustration of hypothetical two-dimensional U-shaped mortality risks considered as a function of two risk factors Y1x and Y2x for the 30 and 80 years old individuals. (by A.I. Yashin, 2015)

Extensions: Genetic SPM (GSPM)

- A categorical variable $Z = \{0, 1\}$, which represents carriers/non-carriers of some allele.
- Assume that $P(Z=1) = p$, $p \in [0, 1]$, where p is is the proportion of carriers and non-carriers of an allele in a population.

$$\begin{cases} d\boldsymbol{Y}(t) = \boldsymbol{\alpha}(\mathrm{Z}, t)(\boldsymbol{Y}(t) - \boldsymbol{f}_1(\mathrm{Z}, t))\mathrm{d}t + \boldsymbol{b}(\mathrm{Z}, t)\mathrm{d}\boldsymbol{W}(t), \boldsymbol{Y}(t = t_0) \\ \mu(t, \boldsymbol{Y}(t)) = \mu_0(t) + (\boldsymbol{Y}(t) - \boldsymbol{f}(\mathrm{Z}, t))^{*}\boldsymbol{Q}(\mathrm{Z}, t)(\boldsymbol{Y}(t) - \boldsymbol{f}(\mathrm{Z}, t)) \end{cases}$$

Extensions: Genetic SPM (GSPM)

ID	Status	Age	Age.next	Z	DBP	DBP.next
1	0	96.61	97.59	0	94.62	100.68
1	0	97.59	98.67	0	100.68	100.59
1	0	98.67	99.67	0	100.59	102.31
1	1	99.67	100.70	0	102.31	NA
2	0	64.78	65.78	1	81.77	80.62
2	0	65.78	66.78	1	80.62	70.49
2	0	66.78	67.68	1	70.49	69.20
2	0	67.68	68.66	1	69.20	67.74
…	…	…	…	…	…	…

Extensions: Genetic SPM (GSPM)

$$a^i = a_{G_0} + \beta_a \cdot G_i$$

a_{G_0} - the rate of the adaptive regulation for those individuals with $G_i = 0$

β_a - effect size (slope)

$G_i = 1$ if $\text{sum}(\text{SNP}_i)/M > T$ else 0

$T = 0.6 \ldots 0.9$

$G_i = 1$ - *carrier*; $G_i = 0$ - *non-carrier*

Dichotomised 'genetic dose' ($\text{sum}(\text{SNP}_i)/M$, $i=1\ldots M$)

ID	SNP1	SNP2	…	SNPM
1	0	0	…	1
2	0	1	…	1
…	…	…	…	…
N	1	1	…	0

ID	G
1	0
2	1
…	…
N	1

($Z = G$)

Extensions: Genetic SPM (GSPM)

Received: 6 March 2017 | Revised: 6 May 2017 | Accepted: 17 May 2017

DOI: 10.1002/gepi.22058

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

A genetic stochastic process model for genome-wide joint analysis of biomarker dynamics and disease susceptibility with longitudinal data

Liang He | **Ilya Zhbannikov** | **Konstantin G. Arbeev** | **Anatoliy I. Yashin** | **Alexander M. Kulminski**

Biodemography of Aging Research Unit, Social Science Research Institute, Duke University, Durham, NC, USA

Correspondence
Liang He, Alexander M. Kulminski, Biodemography of Aging Research Unit, Social Science Research Institute, Duke, University, Erwin Mill Building, 2024 W. Main St., Durham, NC 27705, USA.
Email: lianghe@mit.edu, kulminsk@duke.edu

ABSTRACT

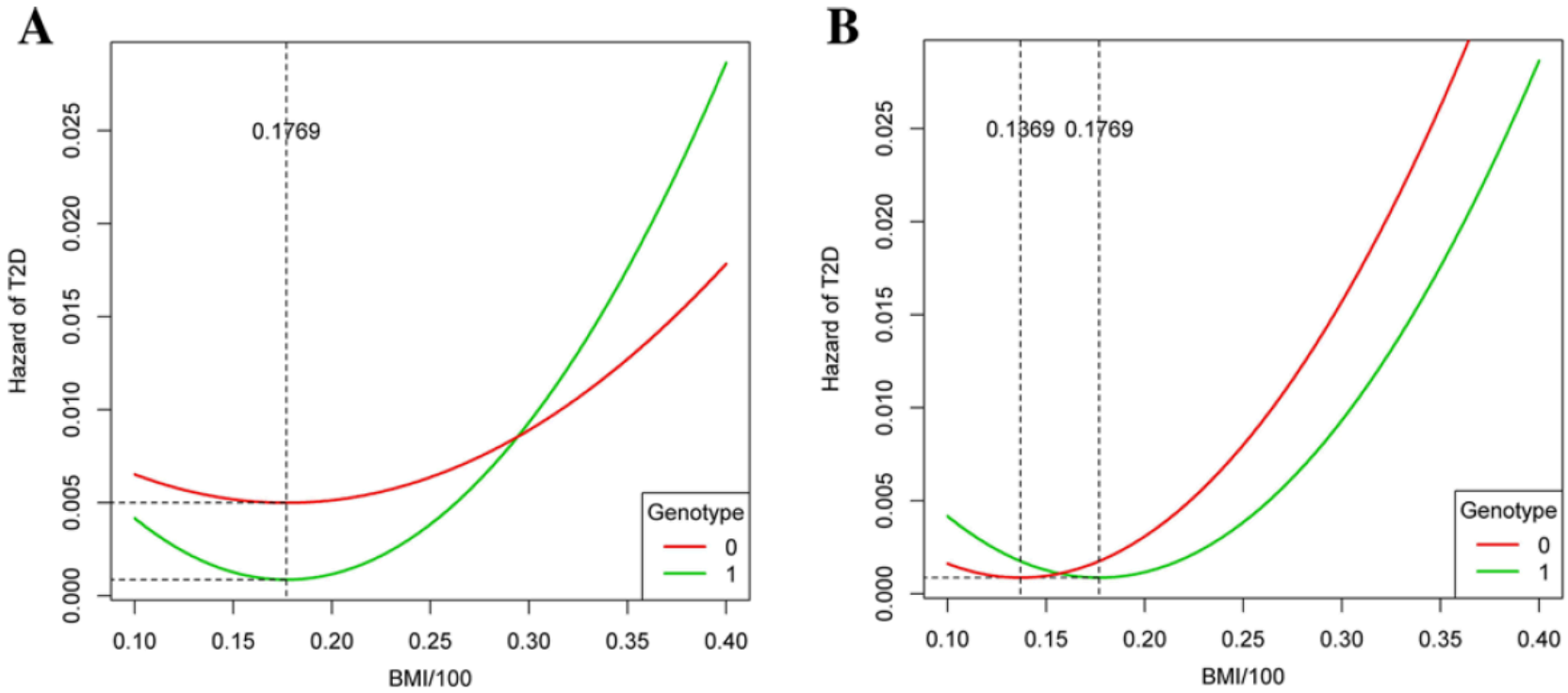
Unraveling the underlying biological mechanisms or pathways behind the effects of genetic variations on complex diseases remains one of the major challenges in the post-GWAS (where GWAS is genome-wide association study) era. To further explore the relationship between genetic variations, biomarkers, and diseases for elucidating underlying pathological mechanism, a huge effort has been placed on examining pleiotropic and gene-environmental interaction effects. We propose a novel genetic

Code example

```
#Example5.R
#Evaluating genetic effects
library(stpm)

data(ex_spmcon1dg)
res <- spm_con_1d_g(ex_data$spm_data, ex_data$gene_data,
                    a = -0.02, b=0.2, q=0.01, f=3, f1=3,
                    mu0=0.01, theta=1e-05,
                    upper=c(-0.01,3,0.1,10,10,0.1,1e-05),
                    lower=c(-1,0.01,0.00001,1,1,0.001,1e-05),
                    effect=c('q'))

res
```



Comparison of the hazard functions in the two genotype groups with different f, but with the same μ_0 and q. This is the model used in the GWAS for detecting the BMI frailty threshold on T2D. This is performed by testing whether f is different across genotype groups with μ_0 and q shared by the genotype groups (SNPs: rs1551133 and rs2613310; Data: Atherosclerosis Risk in Communities study (ARIC), 8,234 individuals).

@ He et al, "A genetic stochastic process model for genome-wide joint analysis of biomarker dynamics and disease susceptibility with longitudinal data", 2017

Multiple imputation

- The SPM offers longitudinal data imputation
- It “preserves” data structure, i.e. relationships between Y(t) and mu(Y(t),t).
- Below there are two examples of multiple data imputation with function `spm.impute(...)`.

Stochastic Process Model and its Applications to Imputation of Censored Longitudinal Data

Ilya Y. Zhbannikov, Konstantin G. Arbeev, Anatoliy I. Yashin

Abstract

Longitudinal data are widely used in medicine, demography, sociology and other areas related to population studies. Incomplete observations in such data often confound the results of analysis. A plethora of data imputation methods have already been proposed to alleviate this problem. The Stochastic Process Model (SPM) represents a general framework for modeling joint evolution of repeatedly measured variables and time-to-event outcome typically observed in longitudinal studies of aging, health and longevity. It is perfectly suitable for imputing missing observations in longitudinal data. We applied SPM to the problem of imputation of missing longitudinal data. This model was applied both to the Framingham Heart Study and Cardiovascular Health Study data as well as to simulated datasets. Comparing to the other best available tools, we show that our proposed methodology in many cases outperforms the current best available solutions both on simulated and real-world censored longitudinal data. R package *stpm* is available under the following link: https://cran.r-project.org/package=stpm

Zhbannikov et al (2017), Stochastic Process Model and its Applications to Imputation of Censored Longitudinal Data, PAA 2017, Chicago, USA.

Code example

```
#Example6.R
library(stpm)

############### One dimensional case ###################
# Data preparation #
data <- simdata_discr(N=1000, dt = 2)
miss.id <- sample(x=2:dim(data)[1], size=round(dim(data)[1]/4)) # ~25% missing data
incomplete.data <- data
incomplete.data[miss.id,5] <- NA; incomplete.data[miss.id-1,6] <- NA
# End of data preparation #
# Estimate parameters from the complete dataset #
p <- spm_discrete(data, theta_range = seq(0.075, 0.09, by=0.001))
##### Multiple imputation with SPM #####
imp.data <- spm.impute(dataset=incomplete.data, minp=5, theta_range=seq(0.075, 0.09,
by=0.001))$imputed
# Estimate SPM parameters from imputed data and compare them to the p:
pp.test <- spm_discrete(imp.data, theta_range = seq(0.075, 0.09, by=0.001))
pp.test
```

Code example

```
#Example7.R
library(stpm)
################## Two-dimensional case ####################
# Parameters for data simulation #
a <- matrix(c(-0.05, 0.01, 0.01, -0.05), nrow=2); f1 <- matrix(c(90, 30), nrow=1, byrow=FALSE)
Q <- matrix(c(1e-7, 1e-8, 1e-8, 1e-7), nrow=2); f0 <- matrix(c(80, 25), nrow=1, byrow=FALSE)
b <- matrix(c(5, 3), nrow=2, byrow=TRUE)
mu0 <- 1e-04; theta <- 0.07
ystart <- matrix(c(80, 25), nrow=2, byrow=TRUE)
# Data preparation #
data <- simdata_discr(N=1000, a=a, f1=f1, Q=Q, f=f0, b=b, ystart=ystart, mu0 = mu0, theta=theta, dt=2)
```

Code example (cont.)

```
# Delete some observations in order to have approx. 25% missing data
incomplete.data <- data
miss.id <- sample(x=dim(data)[1], size=round(dim(data)[1]/4))
incomplete.data <- data
incomplete.data[miss.id, c(5,7)] <- NA
incomplete.data[miss.id-1,c(6,8)] <- NA
# End of data preparation #
# Estimate parameters from the complete data:
p <- spm_discrete(data, theta_range = seq(0.06, 0.08, by=0.001))
p

##### Multiple imputation with SPM #####
imp.data <- spm.impute(dataset=incomplete.data, minp=5, theta_range=seq(0.060, 0.08, by=0.001))$imputed

# Estimate SPM parameters from imputed data and compare them to the p:
pp.test <- spm_discrete(imp.data, theta_range = seq(0.060, 0.08, by=0.001))
pp.test
```

Conclusions

- Stochastic Process Model is a powerful method for joint analysis of time-to-event and longitudinal outcomes.
- R package *stpm* is first publicly available tool for which implements this methodology.
- The approach found useful applications in research on aging and other fields in which time-to-event longitudinal outcomes are involved.
- Applications of the model to genetic analysis are also available and is an active area of our research.

Acknowledgements

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

References

[1] Akushevich I., Kulminski A. and Manton K. 2005. Life tables with covariates: Dynamic model for Nonlinear Analysis of Longitudinal Data. Mathematical Population Studies, 12(2), pp.: 51-80.

[2] Arbeev, K.G., et al. 2009. Genetic model for longitudinal studies of aging, health and longevity and its potential application to incomplete data. Journal of Theoretical Biology. 258(1): p. 103-111.

[3] Manton K.G., Stallard E., Singer B.H. 1992. Projecting the future size and health status of the U.S. elderly population. Int J Forecast 8:433–458.

[4] Witteman J.C.M., Grobbee D.E., Valkenburg H.A. et al.,1994. J-shaped relation between change in diastolic blood pressure and aortic atherosclerosis. Lancet 343:504–507.

[5] Woodbury M.A., Manton K.G. 1977. Random-Walk of Human Mortality and Aging.. Theoretical Population Biology, 11:37-48.

[6] Yashin, A.I., Manton K.G., Vaupel J.W. 1985. Mortality and aging in a heterogeneous population: a stochastic process model with observed and unobserved variables. Theor Pop Biology, 27.

[7] Yashin, A.I., Manton, K.G., Stallard, E. 1986. Dependent competing risks: a stochastic process model, J. Math. Biol. 24, p: 119.

[8] Yashin, A.I. et al. 2007(a). Health decline, aging and mortality: how are they related? Biogerontology, 8(3), 291–302.

[9] Yashin, A.I. et al. 2007(b). Stochastic model for analysis of longitudinal data on aging and mortality. Mathematical Biosciences, 208(2), 538 – 551.

[10] Yashin A.I. et al. 2008. Model of hidden heterogeneity in longitudinal data. Theoretical Population Biology 73, pp. 1–10.

Contact

- Ilya Zhbannikov, ilya.zhbannikov@duke.edu
- Konstantin Arbeev, konstantin.arbeev@duke.edu
- Anatoliy Yashin, aiy@duke.edu