

spm : an R-package that implements a Stochastic Process Model (SPM)

Ilya Y. Zhbannikov

2015-10-10

Overview

The R-package `spm` (<https://github.com/izhbannikov/spm>) is developed for modeling aging-related changes and it allows (1) data simulation and (2) estimating the process parameters using maximum likelihood estimation by optimizing parameters used in the model. Specifically, developed R-package `spm` allows (i) one-dimensional SPM; (ii) multiple dimensional SPM; (iii) data simulation for one- and multiple dimensions.

Data description

Data represents a typical clinical trait data and must be presented in form of two datasets: longitudinal dataset (follow-up studies), in which one record represents a single observation, and vital statistics, where one record represents all patient information. Longitudinal dataset must contain person ID (Identification number), status (dead(1)/alive(0)), time and measurements across the covariates. The `spm` can handle an infinite number of covariates but in practice, 5-7 covariates is enough.

Below there is an example of clinical data that can be used in `spm` and we will discuss the field later. Longitudinal studies:

##	X	ID	IndicatorDeath	Age	AgeNext	DBP	BMI
##	1	1	1	0	30	32	80.00000 25.00000
##	2	2	1	0	32	34	80.51659 26.61245
##	3	3	1	0	34	36	77.78412 29.16790
##	4	4	1	0	36	38	77.86665 32.40359
##	5	5	1	0	38	40	96.55673 31.92014
##	6	6	1	0	40	42	94.48616 32.89139

Vital statistics:

##	X	ID	IsDead	LSmort
##	1	1	1	85.34578
##	2	2	1	80.55053
##	3	3	1	98.07315
##	4	4	1	81.29779
##	5	5	1	89.89829
##	6	6	1	72.47687

Data fields description

Longitude studies

- ID - subject unique identificatin number.
- IndicatorDeath - 0/1, indicates death of a subject.
- Age - current age of subjects.
- AgeNext - next age of subject he will attend to the survey/exam.
- DBP, BMI - covariates, here “DBP” represents a diastolic blood pressure, “BMI” a body-mass index.

Vital statistics

- ID - subject's unique ID.
- IsDead - death indicator, 0 - alive, 1 - dead.
- LSmort - age at death of stopping observations.

Discrete and Continuous cases

There are two main SPM types in the package: discrete model and continuous model. Discrete model assumes equal intervals between follow-up observations. The example of discrete dataset is given below.

```
library(spm)
data <- sim_discrete(N=10, ystart=c(80), k=1)
head(data)
```

```
##      id xi t1 t2  par1_1  par1_2
## [1,]  1  0 30 31 80.00000 77.77912
## [2,]  1  0 31 32 77.77912 72.74410
## [3,]  1  0 32 33 72.74410 71.42812
## [4,]  1  0 33 34 71.42812 67.89342
## [5,]  1  0 34 35 67.89342 77.49928
## [6,]  1  0 35 36 77.49928 78.17800
```

In this case there are equal intervals between t1 and t2 (Age and Age.next).

The opposite is continuous case, in which intervals between observations are not equal. The example of continuous case dataset is shown below:

```
library(spm)
data <- simdata_cont(N=5, ystart = c(50))
head(data)
```

```
##   id xi      t1      t2      y  y.next
## 1  1  0 73.00218 74.04419 43.84085 46.50996
## 2  1  0 74.04419 74.74969 46.50996 44.97078
## 3  1  0 74.74969 74.95910 44.97078 45.87578
## 4  1  0 74.95910 75.02787 45.87578 45.82403
## 5  1  0 75.02787 75.13046 45.82403 45.64709
## 6  1  0 75.13046 76.17987 45.64709 46.38737
```

Discrete case

In discrete case, we use the following assumptions:

$$\bar{y}(t+1) = \bar{u} + \bar{R} \times \bar{y}(t) + \bar{\epsilon}$$
$$\mu(t) = \mu_0(t) + \bar{b}(t) \times \bar{y}(t) + \bar{Q} \times \bar{y}(t)^2$$

Where:

$$\mu_0(t) = \mu_0 e^{\theta t}$$
$$\bar{b}(t) = \bar{b} e^{\theta t}$$
$$\bar{Q}(t) = \bar{Q} e^{\theta t}$$

Continuous case

$$\mu(u) = \mu_0(u) + (\bar{m}(u) - \bar{f}(u)^* \times \bar{Q}(u) \times (\bar{m}(u) - \bar{f}(u)) + Tr(\bar{Q}(u) \times \bar{\gamma}(u))$$

$$\begin{aligned} dm(t)/dt &= \bar{a}(t) \times (\bar{m}(t) - \bar{f}_1(t)) - 2\bar{\gamma}(t) \times \bar{Q}(t) \times (\bar{m}(t) - \bar{f}(t)) \\ d\bar{\gamma}(t)/dt &= \bar{a}(t) \times \bar{\gamma}(t) + \bar{\gamma}(t) \times \bar{a}(t)^* + \bar{b}(t) \times \bar{b}(t)^* - 2\bar{\gamma}t \times \bar{Q}(t) \times \bar{\gamma}(t) \end{aligned}$$

Coefficient conversion between continuous and discrete cases

$$\begin{aligned} Q &= \bar{Q} \\ \bar{a} &= \bar{R} - diag(k) \\ \bar{b} &= \bar{c} \\ \bar{f}1 &= -1 \times \bar{u} \times a^{-1} \\ \bar{f} &= -0.5 \times \bar{b} \times Q^{-1} \\ mu_0 &= mu_0 - \bar{f} \times \bar{Q} \times t(\bar{f}) \\ \theta &= \bar{\theta} \end{aligned}$$

Case with time-dependent coefficients

In two previous cases, we assumed that coefficients is sort of time-dependant: we multiplied them on to

$$e^{\theta t}$$

. In general, this may not be the case. We extend this to a general case, i.e. (we consider one-dimensional case):

$$a(t) = par_1 t + par_2$$

- linear function.

The corresponding equations will be equivalent to one-dimensional continuous case described above.

Simulation

We added one- and multi- dimensional simulation to be able to generate test data for hyphotesis testing. Data, which can be simulated can be discrete (equal intervals between observations) and continuous (with arbitrary intervals).

Discrete

For discrete case:

```
library(spm)
data <- sim_discrete(N=100, ystart=c(75, 94), k=2)
head(data)
```

```
##      id xi t1 t2  par1_1  par1_2  par2_1  par2_2
## [1,]  1  0 30 31 75.00000 70.01428 94.00000 85.32073
## [2,]  1  0 31 32 70.01428 69.46131 85.32073 82.81002
## [3,]  1  0 32 33 69.46131 67.54828 82.81002 80.03232
## [4,]  1  0 33 34 67.54828 62.52220 80.03232 85.69856
## [5,]  1  0 34 35 62.52220 76.13648 85.69856 84.48328
## [6,]  1  0 35 36 76.13648 64.57883 84.48328 86.73917
```

Continuous

For continuous case:

```
library(spm)
data <- simdata_cont(N=100)
head(data)
```

```
##   id xi      t1      t2      y  y.next
## 1  1  0 37.91704 39.47524 77.06048 75.00499
## 2  1  0 39.47524 39.97166 75.00499 75.95150
## 3  1  0 39.97166 41.62206 75.95150 84.24697
## 4  1  0 41.62206 42.01856 84.24697 84.59256
## 5  1  0 42.01856 43.05420 84.59256 84.46657
## 6  1  0 43.05420 44.27984 84.46657 84.66441
```

More Examples

[TODO]