Politechnika Warszawska
Wydział Matematyki i Nauk Informacyjnych

PRACA DYPLOMOWA MAGISTERSKA
NA KIERUNKU MATEMATYKA
SPECJALNOŚĆ STATYSTYKA MATEMATYCZNA I ANALIZA DANYCH

ESTYMACJA W MODELU COXA METODĄ STOCHASTYCZNEGO SPADKU GRADIENTU NA PRZYKŁADZIE DANYCH Z THE CANCER GENOME ATLAS

Autor:
Marcin Piotr Kosiński

Promotor:
prof. ndzw. dr hab. inż Przemysław Biecek

Warszawa, lipiec 2015

.....

podpis promotora

.....

podpis autora

Spis treści

Wprowadzenie

Rozdział 1

Estymacja metodą największej wiarogodności

The making of maximum likelihood was one of
the most important developments in 20th century statistics.
It was the work of one man but it was no simple process (...).
John Aldrich o R. A. Fisher'ze, 1997 [1]

1.1. Estymacja

1.2. Metoda największej wiarogodności

Metodę największej wiarogodności wprowadził R. A. Fisher w 1922 r. [6], dla której po raz pierwszy procedurę numeryczną zaproponował już w 1912 r. [5]. O burzliwym procesie powstawania metody, o zmianach w jej uzasadnieniu, o koncepcjach, które powstały w obrębie tej metody takich jak parametr, statystyka, wiarogodność, dostateczność czy efektywność oraz o podejściach, które Fisher odrzucił tworząc podstawy pod nową teorię można przeczytać w obszernej pracy dokumentalnej [1].

Metoda ta, jako alternatywa dla metody najmniejszych kwadratów [11], [7], była rozwijana i szeroko stosowana później przez wielu statystyków i wciąż znajduje obszerne zastosowania w wielu obszarach estymacji statystycznej, np. [9], [10], [12].

Aby zdefiniować estymator oparty o metodę największej wiarogodności, należy najpierw wprowadzić pojęcie funkcji wiarogodności.

Definicja 1.1. ***Funkcją wiarogodności*** *nazywamy funkcję* $L : \Theta \to \mathbb{R}$ *daną wzorem*

$$L(\Theta|x_1, \ldots, x_n) = f(\Theta; x_1, \ldots, x_n),$$

którą rozważamy jako funkcję parametru θ *przy ustalonych wartościach obserwacji* $x_1, \ldots, x_n$*, gdzie*

$$f(\Theta; x_1, \ldots, x_n) = \begin{cases} \mathbb{P}_\Theta(X_1 = x_1, \ldots, X_n = x_n), & \text{dla rozkładów dyskretnych,} \\ f_\Theta(x_1, \ldots, x_n), & \text{dla rozkładów absolutnie ciągłych.} \end{cases}$$

Oznacza to, że wiarogodność jest właściwie tym samym, co gęstość prawdopodobieństwa, ale rozważana jako funkcja parametru θ, przy ustalonych wartościach obserwacji $x = X(\omega)$.

Definicja 1.2. ***Estymatorem największej wiarogodności*** *parametru θ, oznaczanym ENW(θ), nazywamy wartość parametru, w której funkcja wiarogodności przyjmuje supremum*

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

Niektóre pozycje w literaturze, w definicji estymatora największej wiarogodności, supremum zastępują wartością największą [13], str 14.

1.3. Asymptotyczne własności estymatorów największej wiarogodności

W tym podrozdziale zostanie wykazane, że estymator największej wiarygodności jest

1. asymptotycznie nieobciążony,
2. zgodny,
3. asymptotycznie normalny.

1.3.1. Zgodność estymatorów największej wiarogodności

1.3.2. Asymptotyczna normalność estymatorów największej wiarygodności

Rozdział 2

Model Coxa

2.1. Estymacja analityczna w oparciu o metodę największej wiarogodności dla funkcji pseudo/sub-wiarogodności

2.2. Estymacja numeryczna w oparciu o metodę stochastycznego spadku gradientu rzędu I dla funkcji pseudo/sub-wiarogodności

Poszukujemy rozwiazan równosci

$$\delta lnLn/\delta\theta = 0.$$

Tym razem w ogólnym przypadku zwykle nie znajdziemy analitycznego rozwiazania. W zwiazku z tym jestesmy zdani na metody iteracyjne. Poza tym, byc moze rozwiazanie problemu nie istnieje albo istnieje ich wiele. Zwykle uzywa sie do tego celu, tj. znalezienia rozwiazania, metody Newtona, zwykle w literaturze statystycznej w zastosowaniu do tego problemu, nazywanej metoda Newtona-Raphsona. W efekcie w zasadzie dla kazdego modelu z osobna nalezy badac własnosci asymptotyczne estymatora najwiekszej wiarygodnosci.

Rozdział 3

Numeryczne metody estymacji

3.0.1. Ogólne pojęcia związane ze zbieżnością algorytmu

Warunki stopu itp

3.1. Algorytmy spadku wzdłuż gradientu

3.1.1. Algorytm Cauchy'ego

3.1.2. Algorytm Raphsona-Newtona

3.2. Algorytmy stochastycznego spadku wzdłuż gradientu

3.2.1. Metoda estymacji stochastycznego spadku gradientu I

Algorytm SGD

3.2.2. Metoda estymacji stochastycznego spadku gradientu II

Rozdział 4

Zaimplementowany algorytm

Rozdział 5

Analiza danych genomicznych - model Coxa z estymacją metodą stochastycznego spadku gradientu

5.1. Opis i pobranie danych

5.2. Analiza

Dodatek A

Wykorzystane narzędzia

Dodatek B

Kody w R

Dodatek C

Dokumentacja pakietu RTCGA

R documentation

of all in 'RTCGA/man'

May 5, 2015

R topics documented:

RTCGA-package *The Caner Genome Atlas data integration*

Description

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes. The key is to understand genomics to improve cancer care. RTCGA package offers download and integration of the variety and volume of TCGA data using patient barcode key, what enables easier data possession. This may have an benefcial infuence on impact on development of science and improvement of patients' treatment. Furthermore, RTCGA package transforms TCGA data to form which is convenient to use in R statistical package. Those data transformations can be a part of statistical analysis pipeline which can be more reproducible with RTCGA

Details

For more detailed information visit **RTCGA** wiki on Github.

Author(s)

Marcin Kosinski [aut, cre] < m.p.kosinski@gmail.com >
Przemyslaw Biecek [aut] < przemyslaw.biecek@gmail.com >

See Also

Other RTCGA: availableDataSets; availableDates; availableGenesNames, checkGenesNamesAvailability; checkDataSetsAvailability; downloadTCGA; infoTCGA; mergeTCGA; read.clinical

---

`availableDataSets` *TCGA datasets' names*

Description

Enables to check TCGA datasets' names for current release date and cohort.

Usage

```
availableDataSets(cancerType, date = NULL)
```

Arguments

`cancerType`	A character of length 1 containing abbreviation (Cohort code) of types of cancers to check for available datasets' names on http://gdac.broadinstitute.org/.
`date`	A `NULL` or character specifying from which date datasets' names should be checked. By default (`date = NULL`) the newest available date is used. All available dates can be checked on http://gdac.broadinstitute.org/runs/ or by using availableDates function. Required format `"YYYY-MM-DD"`.

Value

A vector of available datasets' names to pass to the downloadTCGA function.

See Also

Other RTCGA: RTCGA-package; availableDates; availableGenesNames, checkGenesNamesAvailability; checkDataSetsAvailability; downloadTCGA; infoTCGA; mergeTCGA; read.clinical

Examples

```
## Not run:
availableDataSets( "BRCA" )
availableDataSets( "OV", availableDates()[5] ) # error

## End(Not run)
```

availableDates *TCGA datasets' releases dates*

Description

Enables to check dates of TCGA datasets' releases.

Usage

```
availableDates()
```

Value

A vector of available dates to pass to the downloadTCGA function.

See Also

Other RTCGA: RTCGA-package; availableDataSets; availableGenesNames, checkGenesNamesAvailability; checkDataSetsAvailability; downloadTCGA; infoTCGA; mergeTCGA; read.clinical

Examples

```
## Not run:
availableDates()

## End(Not run)
```

checkDataSetsAvailability *TCGA datasets' names availability*

Description

Enables to check TCGA datasets' names availability for current release date and cancer type.

Usage

```
checkDataSetsAvailability(cancerTypes, pattern = "Merge_Clinical.Level_1",
  date = NULL)
```

Arguments

cancerTypes	A character vector containing abbreviation (Cohort code) of types of cancers to check for availability of datasets' name on http://gdac.broadinstitute.org/.
pattern	A character vector of length 1 containing a part of a dataset's name to be checked for availability for current date parameter. By default phrase "Merge_Clinical.Level_1" is checked.
date	A NULL or character specifying from which date datasets' names should be checked for availability. By default (date = NULL) the newest available date is used. All available dates can be checked on http://gdac.broadinstitute.org/runs/ or by using availableDates function. Required format "YYYY-MM-DD".

Value

A vector of available datasets names to pass to the downloadTCGA function.

See Also

Other RTCGA: RTCGA-package; availableDataSets; availableDates; availableGenesNames, checkGenesNamesAvailability; downloadTCGA; infoTCGA; mergeTCGA; read.clinical

Examples

```
## Not run:
checkDataSetsAvailability( "BRCA" )
checkDataSetsAvailability( c("BRCA", "OV") )
checkDataSetsAvailability( "BRCA", "Mutation_Packager_Calls.Level" )

## End(Not run)
```

---

checkGenesNamesAvailability

TCGA genes' names and availability in Merge_rnaseqv2__... *dataset.*

Description

availableGenesNames returns all available genes' names from genes' expressions dataset, where checkGenesNamesAvailability checks whether genes specified in genes are available in Merge_rnaseqv2__illumina (genes' expressions) dataset.

Usage

```
checkGenesNamesAvailability(rnaseqDir, genes)

availableGenesNames(rnaseqDir)
```

Arguments

rnaseqDir	A directory to a cancerType.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_ file
genes	A character - which genes to check for availability in a dataset.

Value

A vector containing genes' names that matched existing names.

See Also

Other RTCGA: RTCGA-package; availableDataSets; availableDates; checkDataSetsAvailability; downloadTCGA; infoTCGA; mergeTCGA; read.clinical

Other RTCGA: RTCGA-package; availableDataSets; availableDates; checkDataSetsAvailability; downloadTCGA; infoTCGA; mergeTCGA; read.clinical

Examples

```
## Not run:
   checkGenesNamesAvailability( rnaseqDir, "TP53" )

## End(Not run)
```

`downloadTCGA` *Download TCGA data*

Description

Enables to download TCGA data from specified dates of releases of concrete Cohorts of cancer types. Pass a name of required dataset to the `dataSet` parameter. By default the Merged Clinical dataSet is downloaded (value `dataSet = "Merge_Clinical.Level_1"`) from the newest available date of release.

Usage

```
downloadTCGA(cancerTypes, dataSet = "Merge_Clinical.Level_1", destDir,
  date = NULL)
```

Arguments

`cancerTypes`	A character vector containing abbreviations (Cohort code) of types of cancers to download from http://gdac.broadinstitute.org/.
`dataSet`	A part of the name of dataSet to be downloaded from http://gdac.broadinstitute.org/runs/. By default the Merged Clinical dataSet is downloaded (value `dataSet = "Merge_Clinical.Level_` Available datasets' names can be checked using availableDataSets function.
`destDir`	A character specifying a directory into which `dataSets` will be downloaded.
`date`	A `NULL` or character specifying from which date `dataSets` should be downloaded. By default (`date = NULL`) the newest available date is used. All available dates can be checked on http://gdac.broadinstitute.org/runs/ or by using availableDates function. Required format `"YYYY-MM-DD"`.

See Also

Other RTCGA: `RTCGA-package`; `availableDataSets`; `availableDates`; `availableGenesNames`, `checkGenesNamesAvailability`; `checkDataSetsAvailability`; `infoTCGA`; `mergeTCGA`; `read.clinical`

Examples

```
## Not run:

dir.create( "hre")

downloadTCGA( cancerTypes = "BRCA", dataSet = "miR_gene_expression",
destDir = "hre/" )

downloadTCGA( cancerTypes = c("BRCA", "OV"), destDir = "hre/" )

## End(Not run)
```

infoTCGA *Information about cohorts from TCGA project*

Description

Function restores codes and counts for each cohort from TCGA project.

Usage

```
infoTCGA()
```

Value

A list with a tabular information from http://gdac.broadinstitute.org/.

See Also

Other RTCGA: RTCGA-package; availableDataSets; availableDates; availableGenesNames, checkGenesNamesAvailability; checkDataSetsAvailability; downloadTCGA; mergeTCGA; read.clinical

mergeTCGA *Merge Clinical data with genes' Mutations and Expressions data*

Description

mergeTCGA enables to

Usage

```
mergeTCGA(clinicalDir, rnaseqDir = NULL, mutationDir = NULL, genes)
```

Arguments

clinicalDir	A directory to a cancerType.clin.merged.txt file. cancerType might be BRCA, OV etc. Can be checked using infoTCGA function.
rnaseqDir	A directory to a cancerType.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_ file, which is a set with gene's Expressions.
mutationDir	A directory to a Mutation_Packager_Calls.Level folder where are genes' Mutations files.
genes	For rnaseqDir - which genes' expressions to merge with clinical data in clinicalDir. For mutationDir which gene's mutations to merge with clinical data in clinicalDir.

Value

A cancerType.clin.merged.txt file is updated with newline containing informations about genes passed to genes argument.

Note

Original cancerType.clin.merged.txt file will be changed after performing merge operation.

Only one of rnaseqDir and mutationDir can be used at a time.

See Also

Other RTCGA: RTCGA-package; availableDataSets; availableDates; availableGenesNames, checkGenesNamesAvailability; checkDataSetsAvailability; downloadTCGA; infoTCGA; read.clinical

---

read.clinical *Read from txt fo;e*

Description

TODO

Usage

```
read.clinical(clinicalDir, ...)
```

Arguments

clinicalDir	A directory to a cancerType.clin.merged.txt file. cancerType might be BRCA, OV etc.

Value

A data.frame with clinical data.

See Also

Other RTCGA: RTCGA-package; availableDataSets; availableDates; availableGenesNames, checkGenesNamesAvailability; checkDataSetsAvailability; downloadTCGA; infoTCGA; mergeTCGA

Index

Literatura

[1] Aldrich J., (1997) *R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922*, Statistical Science 1997, Vol. 12, No. 3, 162-176.

[2] Biecek P., (2011) *Przewodnik po pakiecie R*, Rozprawa doktorska, Oficyna Wydawnicza GiS, wydanie II.

[3] Bottou L., (2010) *Large-Scale Machine Learning with Stochastic Gradient Descent.*

[4] Bottou L., (2012) *Stochastic Gradient Descent Tricks.*

[5] Fisher R. A., (1912) *An absolute criterion for fitting frequency curves.*

[6] Fisher R. A., (1922) *On the mathematical foundations of theoretical statistics*, Philos. Trans. Roy. Soc. London Ser. A 222 309-368.

[7] Gauss C. F., (1809) *Theoria Motus Corporum Coelestium.*

[8] Gągolewski M., (2014) *Programowanie w języku R*, Wydawnictwo Naukowe PWN.

[9] Hutchinson J. B., (1928) *The Application of the "Method of Maximum Likelihood" to the Estimation of Linkage*, Genetics. 1929 Nov; 14(6): 519–537.

[10] Kenward M. G., Lesaffre E. and Molenberghs G., (1994) *An Application of Maximum Likelihood and Generalized Estimating Equations to the Analysis of Ordinal Data from a Longitudinal Study with Cases Missing at Random*, Biometrics Vol. 50, No. 4 (Dec., 1994), pp. 945-953.

[11] Legendre A. M., (1804) *Nouvelles methods pour la determination des orbites des com'etes.*

[12] Millar R. B., (2011) *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB, chapter 6. Some Widely Used Applications of Maximum Likelihood*, John Wiley & Sons, Ltd.

[13] Rydlewski J., (2009) *Estymatory Największej Wiarogodności w Uogólnionych Modelach Regresji Nieliniowej*, Rozprawa doktorska.

Marcin Piotr Kosiński
Nr albumu 265361

Warszawa, 16 maja 2015

Oświadczenie

Oświadczam, że pracę magisterską pod tytułem „ Estymacja w modelu Coxa
metodą stochastycznego spadku gradientu
na przykładzie danych
z The Cancer Genome Atlas", której promotorem jest prof. ndzw. dr hab. inż Przemysław Biecek
wykonałem samodzielnie, co poświadczam własnoręcznym podpisem.

.....
Marcin Piotr Kosiński