Documentation for CHANGEPOINT.c

Author: Xiaolin Fan

Updated: 8/1/2008

Questions or bug reports can be sent to xfan@mcw.edu

Description

This program is for implementation of the Cox-type regression on cumulative incidence function with a time change-point, in the context of violation of the proportionality assumption under the competing risks setting. Methods, described in Section 5.3.1 of Fan (2008), use the Mixture of Polya Trees (MPT) process prior and are based on the full likelihood.

Input File Format

The program requires some of the GSL subroutines and GSL thus needs to be installed on your system (download GSL for free from http:// www.gnu.org/software/gsl/). Before running the program, you need to set up two input files in the same directory as CHANGEPOINT.c. One file, named as *parameter.txt*, sets up the parameters and the other file, *data.txt*, contains the observed competing risks data.

1. **Parameter data** *parameter.txt*: The file is constructed as follows:

Line	Description	Example
1	Level of partitions in MPT	5
2	Smoothing parameter in MPT	1.0
3	Sample size for competing risks data	200
4	Number of MCMC iterations	10000
5	Tuning parameters for sampling Polya Trees	0.6 0.6
6	Tuning parameters for centering distributions in MPT	0.1 0.1
7	Tuning parameters for updating coefficients in cause 1	0.5 0.5
8	Tuning parameters for updating coefficients in cause 2	0.2
9	Parameters of beta distribution for updating normalizing constant	6.0 4.0
10	Distance between two predictive points	0.03
11	Initial values for parameters of centering distribution	1.0 1.0
12	Value of time change-point	0.8

The first two lines are for the practical setting in MPT. According to Hanson (2006), level in MPT can be approximately equal to $log_2(n/N)$, where n is the sample size of observed data and N is a typical number of observations falling into each partition at the bottommost level, such as 10. Smoothing parameter is considered to be 1, as a sensible canonical choice in Lavine (1992). However, sensitivity analysis should be considered via several different values. Line 3 denotes the sample size of your data. Line 4 is the total number of MCMC iterations, including the number for burn-in. The updating scheme of all the parameters in this method relies on the Metropolis-Hastings Algorithm (Chib and Greenberg, 1995). The corresponding tuning parameter for each of them needs to be manually adjusted in line 5-9. The acceptance rate should be typically around 20%-40%. The number of acceptances is listed in the output file *accept.txt* (see below). However, Hanson (2006) recommended the acceptance rate for updating Polya trees could be about 40% to 60% and may increase as the level of partitions increases. In this program, MPT priors are assigned on the normalizing baseline cumulative incidence function. The centering distributions of MPT priors are chosen to be exponential. Line 6 represents the tuning parameters for updating the mean of exponential distributions. Line 9 denotes the two parameters of beta distribution for sampling the normalizing constant. Mean of the beta distribution can be initially set as the percentage of failure due to cause of interest among the exact (not censored) observations. In the program, the predictive cumulative incidence functions are produced only for 100 equally spaced time points. The range of these points starts at 0 and ends at 99 times the distance between two points. Users thus need specify the distance between two points in line 10. The predictive cumulative incidence functions are implemented in the presence of discrete covariate. Line 11 denotes a reasonable initial guess for the mean parameters of centering distributions which are the exponential distribution in this implementation. The time change-point specifies in line 12.

2. **Competing risks data** *data.txt*: Each row contains failure time, one covariate and failure cause for each individual. Under the competing risks setting, the cause of interest is coded as 1 and failure due to other causes as 2. In the presence of right censoring, the failure cause is coded as 0. For example,

Time	Covariate	Cause
0.4152	0.0	1
2.2329	1.0	0
0.7134	0.0	2
	.	
	.	

Note that in the case of discrete covariate, 0 stands for the baseline value.

Output File Format

Output files will be sent to a directory called *output*. Users need to create such a subdirectory under the directory containing the CHANGEPOINT.c and the input files. The *output* directory has the acceptance file (*accept.txt*), the files containing the samples from MCMC chains (*coef1.txt*, *coef2.txt*, *mu.txt* and *p.txt*), the files containing the predictive distributions (*pred0.txt* and *pred1.txt*) and the file containing LPML values (*LPML.txt*).

1. *accept.txt*: The file contains the numbers of acceptances for all the updated parameters. The acceptance rates can be calculated via the numbers divided by the number of MCMC iterations. The first part is the numbers for updating Polya trees, from the partitions at the bottommost level to ones at the uppermost level and from right to left at each level. The total number of partitions is $2^{M+1} - 2$, where M is the level specification. Since the updates are only required for the partitions with odd numbers, the numbers of acceptances are applied to these odd numbers. The columns next to the label are the acceptance numbers for cause 1 and 2, respectively:

Label	Cause 1	Cause 2
Polya trees 1	5218	4834
Polya trees 3	5398	3742
Polya trees 5	5896	4018
	.	
	.	

   The next lines are the numbers for updating the normalizing constant (p), parameters (mu) in the centering distributions for cause 1 and cause 2, coefficients for cause 1 (*coef1*) and coefficient for cause 2 (*coef2*).

2. *coef1.txt*: The file contains two columns of samples over the MCMC iterations for cause of interest. The first column is the samples for the coefficient fit before the time change-point and the second column is for the coefficient fit after the time change-point.

3. *coef2.txt*: The file contains the coefficient samples for the secondary cause.

4. *mu.txt*: The file contains the samples of parameters in the centering distributions. The first column is for the mean parameters of exponential distributions for cause 1 and the second column is for the ones for cause 2.

5. *p.txt*: The file contains the samples of the normalizing constant.

6. *pred0.txt*: The file contains the predicted baseline cumulative incidence function for cause 1. Since 100 grid points are used in the calculation for each iteration, the file has 100 columns. At each grid point, the mean of the iterations after a burn-in can be treated as the estimated cumulative probability and 2.5th percentile to 97.5th percentile as the pointwise 95% credible interval. One can also compute a simultaneous confidence band from the posterior samples.

7. *pred1.txt*: The file contains the predicted cumulative incidence function for cause 1 when the value of the discrete covariate is 1. It also has 100 columns. Similar calculations can be made as for *pred0.txt*.

8. *LPML.txt*: The file contains the value of LPML calculated at each iteration. The optimal change-point can be found based on the LPMLs by fitting a set of change-points, described in Section 5.3.2.

References

Chib, S. and Greenberg E. (1995). Understanding the Metropolis-hastings Algorithm. *The American Statistician* **49**, 327-335.

Fan, X. (2008). Bayesian Nonparametric Inference for Competing Risks Data. Ph.D. Thesis, Medical College of Wisconsin, Milwaukee.

Hanson, T. (2006). Inference for Mixtures of Finite Polya Tree Models. *Journal of the American Statistical Association* **101**, 1548-1565.

Lavine, M. (1992). Some Aspects of Polya Tree Distributions for Statistical Modeling. *The Annals of Statistics* **20**, 1222-1235.