twins

A two-component model for counts of infectious diseases

Leonhard Held[1], Mathias Hofmann, Michael Höhle, Volker Schmid
Department of Statistics, Ludwig-Maximilians-Universität München,
Ludwigstr. 33, 80539 München, Germany

Version 1.0

twins is avalable at `http://www.stat.uni-muenchen.de/~mhofmann`

[1]Corresponding Author. E-mail: leonhard.held@stat.uni-muenchen.de

Contents

1 Licence agreement

2 Starting *twins*

After you have installed twins, edit the twins.ini file or write your own ini-file. Then start the program by typing

```
twins ini-file
```

or (if you use twins.ini) just type

```
twins
```

in the command line.
On window-based OS you can also start twins with a double click on the twins-icon. Twins then will use the `twins.ini` file.

3 General information

Twins is a software for estimating the stochastic model for time series count data of infectious diseases proposed in Held et al. (2005). The model is based on a Poisson or negative binomial observation model with two components: A parameter-driven component that relates the disease incidence to latent parameters describing endemic seasonal patterns, which are typical for infectious disease surveillance data, and an observation-driven or epidemic component that should be suitable to explain for possible outbreaks.

4 Theoretical background

Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ denote a time series of counts of infectious diseases. Our model is specified through the conditional distribution of $Z_t|Z_{t-1}$, so we also need observed counts Z_0 at time $t = 0$ to condition on. We assume that Z_t follows a generalized Poisson branching process with immigration,

$$\begin{aligned} Z_t &= X_t + Y_t, \quad t = 1, \ldots, n \text{ with} \\ X_t &\sim \text{Po}(\omega_t \nu_t), \text{and} \\ Y_t|Z_{t-1} &\sim \text{Po}(\omega_t \lambda_t Z_{t-1}), \end{aligned}$$

where the observed number of counts Z_t is decomposed into two (unknown) components: X_t and Y_t, which are assumed to be independent. We call those two quantities the *endemic* and

epidemic components respectively. By introducing the effects $\omega_t \sim \text{Ga}(\psi, \psi)$, we get a negative binomial distribution for the Z_t,

$$Z_t|Z_{t-1} \sim \text{Nb}(\nu_t + \lambda_t Z_{t-1}, \psi),$$

with mean $\mu_{Z_t} = \nu_t + \lambda_t Z_{t-1}$ and variance $\sigma_{Z_t} = \mu_{Z_t}(1+\mu_{Z_t}/\psi)$, where ψ is called the dispersion parameter.

4.1 The endemic component

The *parameter-driven* or *endemic* component of the process is driven by the parameter ν_t. Most data on infectious disease surveillance data exhibit strong seasonality. We therefore model $\log \nu_t$ as the sum of L harmonic waves of different frequencies plus an intercept,

$$\log \nu_t = \gamma_0 + \sum_{l=1}^{L} \Big(A_l \sin\big(\rho l t + \phi_l\big) \Big), \tag{1}$$

where A_l is the amplitude of the corresponding sine curve, ϕ_l the phase shift, and ρ is the base frequency. For weekly data, $\rho = 2\pi/52$ is the obvious choice. It is well known that (1) can be rewritten as

$$\log \nu_t = \gamma_0 + \sum_{l=1}^{L} \Big(\gamma_{2l-1} \sin\big(\rho l t\big) + \gamma_{2l} \cos\big(\rho l t\big) \Big), \tag{2}$$

so with $s_{t0} = 1$ and

$$s_{tj} \quad = \quad \begin{cases} \sin(\frac{\rho t(j+1)}{2}) & \text{for } j = 1, 3, \ldots, 2L-1 \\ \cos(\frac{\rho t j}{2}) & \text{for } j = 2, 4, \ldots, 2L \end{cases},$$

equation (2) can be reduced to a simple linear regression form $\log \nu_t = \sum_{j=0}^{J} \gamma_j s_{tj}$, where $J = 2L$.

4.2 The epidemic component

The *observation-driven* or *epidemic* component of the process is driven by the parameter sequence $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, which is assumed to be piecewise constant with unknown number of changepoints K and unknown location of the changepoints $\theta_1 < \ldots < \theta_K$. More specifically,

we assume the following model:

$$\lambda_t = \begin{cases} \lambda^{(1)} & \text{if} \quad t = 1, 2, \ldots, \theta_1 \\ \lambda^{(k)} & \text{if} \quad t = \theta_{k-1} + 1, \ldots, \theta_k \\ \lambda^{(K+1)} & \text{if} \quad t = \theta_K + 1, \ldots, n \end{cases}$$

where $\theta_k, k = 1, \ldots, K$ are the K unknown changepoints, so $\theta_k \in \{1, 2, \ldots, n-1\}$.

4.3 Prior assumptions

The proposed model is particularly well-suited for Bayesian inference. For the regression coefficients of the endemic component $\boldsymbol{\gamma}$ we set $\boldsymbol{\gamma} \sim \mathrm{N}(0, \sigma 2_\gamma \boldsymbol{I})$ with $\sigma 2_\gamma = 10^6$, which corresponds to highly dispersed independent normal priors for each coefficient.
For the partition model of the parameter of the epidemic component we have used the following settings: The number K of changepoints is assumed to be uniformly distributed among the possible values $\{0, 1, \ldots, n-1\}$, i.e. $Pr(K = k) = 1/n$, $k = 0, 1, \ldots, n-1$. For given $K > 0$, the location of the changepoints $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, where $\theta_1 < \theta_2 < \ldots < \theta_K$, is again uniformly distributed among all possible configurations, i.e.

$$Pr(\boldsymbol{\theta} | K = k) = \binom{n-1}{k}^{-1}.$$

The *unconditional* prior probability for a changepoint at any arbitrary location i, A_i, is hence

$$P(A_i) = \sum_{k=0}^{n-1} \frac{k}{n-1} \cdot \frac{1}{n} = \frac{1}{2}.$$

Finally, for $\lambda^{(k)}$, $k = 1, \ldots, K+1$, we specify independent exponential distributions with mean $1/\xi$ and variance $1/\xi^2$, using a Gamma(α_ξ, β_ξ) distribution for the hyperparameter ξ.

5 Input files

5.1 The data

The data file must containing the number of observations followed by the observations that have to be integers. The entries have to be separated by the new line command.

5.2 The parameters

All parameters for the algorithm are specified in an ini-file, by default called twins.ini, which contains the following information

datafile	path of the data file
logfile	path of the first output file; the estimation results are written to the two output files as explained in section 6.
logfile2	path of the second output file
burnin	burnin; total number of iterations = burnin + filter*sampleSize.
filter	filter
sampleSize	sample size
seed	seed
alpha_xi	first parameter of the prior of xi
beta_xi	second parameter of the prior of xi
season	season of the endemic component X
frequencies	number of frequencies of the endemic component X
psiRWSigma	starting value for tuned standard deviation of the proposal of psi
alpha_psi	first parameter of the prior of psi
beta_psi	second parameter of the prior of psi

The ini-file needs to have 14 lines. The two columns have to be separated by :.

6 Output files

The output is written to three output files:

- The first output file, contains the samples of the posterior distributions of psi, gamma, K, xi, lambda, Z_{n+1} and the Deviance.
- The second output file contains the posterior means of X, Y and omega and the posterior probabilities of the breakpoints.
- The acc file, "output file name".acc, contains the acceptance rates.

For a better handling with R, the time index starts at $t = 1$ instead of $t = 0$.

7 Figures

The R-program figures.R reads the output files of the estimation and creates some figures.

xyz.pdf	Posterior means of the components X and Y
tms-lambda.pdf	Posterior mean and pointwise 95% credibility interval of lambda
tms-nu.pdf	Posterior mean and pointwise 95% credibility interval of nu
theta.pdf	Posterior probabilities of the changepoints
lambdage1.pdf	Posterior probability of lambda > 1
histogram-K.pdf	Posterior probability of K
histogram-psi.pdf	Posterior probability of psi
histogram-Znp1.pdf	Posterior predictive probability of Z_{n+1}
traj-gamma-i.pdf	Trajectory of gamma-i
traj-K.pdf	Trajectory of K
traj-psi.pdf	Trajectory of psi
traj-xi.pdf	Trajectory of xi
autocorrelation.pdf	Autocorrelation of K and psi

References

Held, L., M. Hofmann, M. Höhle, and V. Schmid (2005). A two-component model for counts of infectious diseases. SFB 386 Discussion Paper 424, University of Munich.