Stockholms universitet

Bayesian back-projection and its application to foodborne disease outbreaks

Bei Yang

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2016:3
Matematisk statistik
September 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2016:3**
http://www.math.su.se

Bayesian back-projection and its application to foodborne disease outbreaks

Bei Yang*

September 2016

Abstract

Back-projection is a statistical method for determining the unknown exposure time in an outbreak data set containing onset time. For an individual, the exposure time is the time period from being infected to symptom occurrence, also called incubation time. Bayesian back-projection is an approach for unknown exposure time estimation. In this thesis, we applied this method to foodborne disease outbreaks data. In our mathematical modelling, informative and non-informative priors have been tested. Uniform and flat Gamma distributions were implemented as non-informative priors. We also tested parametric and non-parametric Empirical Bayesian approaches on data originating from large gastrointestinal disease outbreak, which occurred in Germany 2011. The data come from Robert Koch Institute, the federal public health institute in Germany. The disease incubation period probability distribution was also given by estimates from the Robert Koch Institute. Our data analysis in 686 adult HUS patients indicates that Bayesian approaches lead slightly different results from the EM back-projection method for the point estimate. Under Bayesian approaches, MCMC simulation enable us directly obtain a credible interval.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: beiyang2016@gmail.com. Supervisor: Michael Höhle.

I would like to express my gratitude to my supervisor Michael Höhle for the support of my master of biostatistics degree thesis study. His guidance has been invaluable to me during the whole process of researching and writing this thesis. I am grateful for his inspiration, patience and encouragement.

Contents

Chapter 1

Introduction

Back-projection is a method to estimate the exposure time of an infectious source leading to a disease outbreak. It can be inferred either in real-time or shortly after the end of the outbreak. The method has been used to estimate the unobserved past incidence of infection with the human immunodeficiency virus (HIV)(Becker, Niels. G. and Watson, Lyndsey. F., 1991).

HIV is the virus that causes acquired immunodeficiency syndrome (AIDS). HIV destroys CD4 positive T cells, which are white blood cells crucial to maintaining the function of the human immune system. Most people infected with HIV can carry the virus for years before developing any symptoms. Usually, there are few or no symptoms at first, but the patient later experiences fever, weight loss, gastrointestinal problems and muscle pains. ("https://aidsinfo.nih.gov/education-materials/fact-sheets/19/45/hiv-aids –the-basics")

The data used in applying the back-projection method are the numbers of symptomatic AIDS patients by time period such as month, week or day. The only additional information required is the distribution of time from infection to clinical diagnosis of AIDS. In the end of chapter 1, we give an example of AIDS data (Becker, Niels. G. and Watson, Lyndsey. F., 1991).

Similarly, Hemolytic-uremic syndrome (HUS) occurs after ingestion of a type of E. coli, for example E. coli O157:H7. E. coli produces the stx1 and/or stx2 shiga toxins, which is easier received by children than adults. Diarrhea is the initial symptom of bacteria colonization. HUS develops about 5-10 days after onset of diarrhea. It is followed by decreased urine output, blood in the urine, kidney failure and destruction of red blood cells. The infection is acquired by contaminated water or food ("https://en.wikipedia.org/wiki/Hemolyticuremic/syndrome"). In May, 2011 an epidemic of bloody diarrhea hit Germany. It was caused by Escherichia coli O104:H4 contaminated fenugreek seeds. There were more than 3,800 infected cases, with more than 800

cases developing to HUS patients; including 36 fatal cases. Nearly 90% of the HUS cases were in adults (Buchholz U et al. 2011)

The aim of this thesis is to apply the Bayesian back-projection method to estimate the incidence of Escherichia coli O 104:H4 infection. We estimated the exposed individuals in order to predict the future infected trend. By a simulation study, the uncertainty of the estimate was quantified.

Compared to frequentist evaluates procedures, Bayesian approach requires a prior distribution which is constituted by describing distributions for the involved random variables. The Empirical Bayesian approaches allow the observed data to play some roles in determining the prior distribution. The penalized spline regression was also used for prior estimate in our study. We performed statistical analysis in R environment ("https://www.r-project.org"). Further more, Surveillance (Michael Höhle, 2007) and JAGS (Plummer et al. 2006) packages were used for back-projection and Bayesian calculations.

Our results indicate that the Bayesian back-projection is a reliable method for infected incidence estimation. The MCMC simulation was implemented to quantify the uncertainty of estimation.

1.1 Layout overview of the thesis

Chapter 2 introduces the method of back-projection, including its mathematical theory and application. It also describes the likelihood function of data, EM algorithm, smoothing step, and the EM algorithm's convergence. In Chapter 3, we focus on the difference of concepts between frequentist and Bayesian approaches. We also illustrate point and confidence interval estimates, and discuss the bias of point estimate with two approaches. Chapter 4 introduce the simulation technique. Bootstrap and MCMC simulation, and Gibbs sampling applied to Bayesian approaches. Chapter 5 gives an example for back-projection application. Using HUS outbreak data, we estimated the incidence curve by frequantist and Bayesian back-projection approaches. Various priors testing method is included in this section. Chapter 6 draws conclusions, discusses the difference of results between Bayesian approaches and frequentist methods as well. Finally, it also discusses the distinction between prior estimate using parametric, and semi-parametric methods, within Bayesian approach.

1.2 A back-projection application

The earliest Australian HIV-positive case was found on 9 October 1980 from tests of stored blood samples. Two years later, the first Australian case of AIDS was diagnosed (Becker, Niels. G. and Watson, Lyndsey. F., 1991). Table 1 gives AIDS cases from January, 1982 to December 1989. It was reported to the National Centre in HIV Epidemiology and Clinical Research by the end of March 1990. EM back-projection method has been used to estimate the HIV infected incidence which are unobserved data.

Table 1.1: Reported Australian AIDS incidence data from 1989 to1992 (Becker, Niels. G. and Watson, Lyndsey. F., 1991)

	Months												
Year	J	F	M	A	M	J	J	A	S	O	N	D	Total
1982	0	0	0	0	0	0	0	0	0	0	0	1	1
1983	0	0	0	1	0	0	0	1	1	0	1	2	6
1984	0	0	1	2	0	2	3	6	7	6	5	11	43
1985	11	11	6	9	19	8	9	4	11	10	9	11	118
1986	15	14	13	14	18	18	16	23	22	30	25	13	221
1987	30	25	31	17	46	34	24	27	37	29	44	26	370
1988	39	42	27	28	34	41	49	45	41	53	58	40	497
1989	53	47	32	25	39	42	38	48	48	50			422
													sum 1678

Chapter 2

The back-projection method

The Infectious disease progressing in an individual may have infection, onset of symptoms and recovered stages. For example HIV infection may progress three stages: 1) acute HIV infection 2) clinical latency and 3)AIDS. Acute HIV infection stage is the body's natural response to the HIV infection. It's within 2-4 weeks after HIV infection. Many people develop flu-like symptoms, including swollen glands, sore throat, rash, muscle and joint aches and pains, and headache. After the acute stage of HIV infection, the disease moves into a clinical latency stage meaning that a virus is living or develop in a person without producing symptoms. When immune system is badly damaged, opportunistic infections become vulnerable and AIDS is considered to have been progressed. ("https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/hiv-in-your-body/stages-of-hiv/"). The methodology essentially follows from the concept that for each infected person the time of symptom onset is equal to the sum of exposure time and the incubation (clinical latency) period, all infected persons eventually developing to patients (Becker, Niels. G. and Watson, Lyndsey. F., 1991). A crucial aspect in the back-projection method is the incubation period distribution. Usually, parametric and non-parametric methods are utilized to estimate this distribution (Joseph R. Egan and Lan M. Hall 2015). Here we will model the distribution with non-parametric method and the simple temporal model.

2.1 Incubation period and likelihood function

The incubation period is defined as the time from pathogen exposure to onset of symptoms in an individual. Many factors such as host susceptibility, strain virulence and chance can contribute to diverse incubation periods. Therefore, it is more appropriately characterized by a distribution rather than by a single

point. As the disease's aetiology is unknown, we assume that the time period between exposure and onset of the symptom is independently and identically distributed. Some distribution is used to model the incubation period for a number of acute infectious diseases. The likelihood function is established by the event's incubation period distribution. Estimating the parameters of this distribution is to determine the parameter value that maximize the likelihood function L, shown below.

$$L(\alpha; y^n) = \prod_{i=1}^{n} f(y_i; \alpha). \tag{2.1}$$

Where $y^n = (y_1, y_2, \ldots, y_n)$ represents a vector of incubation periods for each person i, n is the total number of exposed persons who eventually succumbed to the infection in the absence of intervention. These data are used to estimate α, a vector of parameters of the probability density function (PDF) f. The f is the incubation period probability distribution function, for example gamma or log-normal distribution (Joseph R. Egan and Lan M. Hall 2015).

We formulate a model based on the fact that AIDS is the result of HIV infection followed by an incubation period, which is the time from its infection to clinical onset of AIDS.

A time point earlier than the introduction of the virus into the community is chosen as the time origin. In general, we choose a month as the time unit and formulate the likelihood in terms of discrete time, because incidence of AIDS is monthly reported. Therefore the observation is discrete. Infections are assumed to arise according to random processes.

Following the notation and approach from the article written by Becker and Watson 1991, we can determinate the mean number of clinical AIDS cases in month t. let n_t denote the number of individuals infected during month t. The number of AIDS cases diagnosed in month t is denoted by $Y_t, t = 1, 2, ..., T$, where t is the month beyond which no reliable AIDS incidence data are available. Let f_d be the probability that the duration of the incubation period is d months, d=0,1,2...D, where D is the maximal incubation time; Under the assumption that the distribution of the incubation period is the same irrespective of when the individual is infected, we have

$$E[Y_t \mid n_1, n_2, \ldots, n_t] = \sum_{i=1}^{t} n_i f_{t-i}. \tag{2.2}$$

Then the mean number of clinical AIDS cases in month t will be

$$\mu_t = \sum_{i=1}^{t} \lambda_i f_{t-i}. \tag{2.3}$$

where $\mu_t = E[Y_t]$ and $\lambda_i = E[n_i]$. The probability mass function of the incubation period distribution f_d is assumed known in the method of back-projection.

2.2 Back-projection method

Assume $n_{t,d}$ is the number of individuals exposed in interval t = 1,...,T having an incubation time d (i.e. onset symptom was observed at timepoint t + d). n_t is the number of individuals infected in interval t, i.e.

$$n_t = \sum_{d=0}^{\infty} n_{t,d}. \tag{2.4}$$

Furthermore,

$$n_t \sim \text{Poisson}(\lambda_t), \tag{2.5}$$

$$n_{t,d} \sim \text{Poisson}(f(d)\lambda_t) \tag{2.6}$$

where f(d), d=0, 1, 2, ..., D is the PMF of the incubation time.
Assume

$$y_t \sim \text{Poisson}(\mu_t), \tag{2.7}$$

where y_t is the number of patients diagnosed in month t and

$$\mu_t = \sum_{i=1}^{t} E(n_{i,t-i}) = \sum_{i=1}^{t} f_{(t-i)}\lambda_i. \tag{2.8}$$

we can estimate the λ_t under assumption that $n_1, n_2, \ldots, n_t$ are independent Poisson variates using the likelihood function

$$\prod_{t=1}^{T}(\sum_{i=1}^{t} \lambda_i f_{t-i})^{y_t} \exp(-\sum_{i=1}^{t} \lambda_i f_{t-i}). \tag{2.9}$$

2.3 The EM algorithm

The EM algorithm is a technique for obtaining a maximum likelihood estimate in the situation where only incomplete data are observed. When an ML estimate would have been easy to compute, a larger complete data set had been obtained by EM algorithm. Moreover, a completely unobserved data set also can be obtained.

In the AIDS case, the monthly number of new HIV-exposures are unobserved data since the $n_{t,d}$ are independent of Poisson variates with mean $\lambda_t f_d$, we estimate λ_t by EM algorithm. The following formula combines both the E step and the M step of the EM algorithm.

$$\lambda_t^{\text{new}} = \frac{\lambda_t^{\text{old}}}{F_{T-t}} \sum_{d=0}^{T-t} \frac{Y_{t+d} f_d}{\sum_{i=1}^{t+d} \lambda_i^{old} f_{t+d-i}}, \tag{2.10}$$

where

$$F_{T-t} = \sum_{d=0}^{T-t} f_d. \tag{2.11}$$

The incomplete data likelihood function is increased at each step when we use the iterative equation. In the above formula, the maximum likelihood estimate of λ_t is $\sum_{d=0}^{T-t} \frac{N_{td}}{F_{T-t}}$if all the N_{td} are observed. As only the Y_t are observed, we replace the $N_{t,d}$ by $\sum_{i=1}^{t} n_i f_{t-i}$.

$$E[N_{td} \mid Y_1, Y_2, \ldots, Y_t] = Y_{t+d} \frac{\lambda_t^{\text{old}} f_d}{\sum_{i=1}^{t+d} \lambda_i^{old} f_{t+d-i}} \tag{2.12}$$

2.4 Incorporating a smoothing step

As prior knowledge, the infection intensity should be a smooth curve because haphazard jumps in the infection intensity are improbable. Therefore the smoothing step is incorporated after each application of the equation above. Let

$$\Phi_t^{new} = \frac{\lambda_t^{old}}{F_{T-t}} \sum_{d=0}^{T-t} \frac{Y_{t+d} f_d}{\sum_{i=1}^{t+d} \lambda_i^{old} f_{t+d-i}} \tag{2.13}$$

and then let

$$\lambda_t^{new} = \sum_{i=0}^{k} W_i \Phi_{t+i-k/2}^{new} \tag{2.14}$$

where λ_t new is a weighted average of the new parameter value which the E and M steps produce near t, when applied to old parameter values. The value of k determines the "window width" for the weighted average and should be an even integer. Therefore, we choose a value k as a parameter as a smoothing step.

$$W_i = \frac{\binom{k}{i}}{2^k} \tag{2.15}$$

$i = 0, 1, \ldots, k$.

2.5 Convergence

Each iteration of the EM algorithm is known to increase the likelihood. As a result, the algorithm will converge. In practice, convergence of the EM algorithm to a maximum likelihood estimate requires a great number of iterations in the present type of application. We choose a time T'<T and a small positive ϵ and stop iteration when

$$\sum_{t=1}^{T'} \frac{\mid \lambda_t^{new} - \lambda_t^{old} \mid}{\lambda_t^{old}} < \epsilon \tag{2.16}$$

(Becker, Niels. G. and Watson, Lyndsey. F., 1991).

Chapter 3

Frequentist and Bayesian approaches

3.1 Definitions

3.1.1 Frequentist method

In frequentist approach, the sampling model is given in the form of a probability distribution f(y| Θ). This distribution is called likelihood, represented by likelihood function L($\Theta; y$). Giving particular data values y, it is very possible to find the value for parameter Θ that maximizes the likelihood function, i.e maximum likelihood estimate (Fisher, 1922 and Stigler, 2005).

3.1.2 Bayesian approach

As frequentist analysis, the Bayesian approach samples observed data y= ($y_1, \ldots, y_n$), given a vector of unknown parameters θ. In the Bayesian approach, we think of θ as a random quantity instead of supposing it to be a fixed (though unknown) parameter. This approach adopts the prior distribution, a probability distribution for θ that summarizes any information we have. Normally, the prior distribution is not related to the data y. The prior may depend on additional parameters η, referred to as hyperparameters. We assume here that the hyperparameters η are known. Inference concerning θ is based on its posterior distribution, given by

$$p(\theta \mid y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y,\theta)}{\int p(y,\theta)\mathrm{d}\theta} = \frac{f(y \mid \theta)\pi(\theta)}{\int f(y \mid \theta)\pi(\theta)\mathrm{d}\theta} \tag{3.1}$$

where y represents data and $\pi(\theta)$ represents prior. (Carlin and Louis, 2008, Bayesian methods for data analysis, Chapter 2)

3.1.3 The prior for Bayes and Empirical Bayes approaches

As formula 3.1 indicated, in Bayesian analysis, the parameter estimation partially depends on the prior distribution. In Bayes approach, prior distribution estimation is unrelated to the observed data. This is in contrast to the Empirical Bayes (EB) approach, that uses the observed data to estimate parameters of prior. In Bayes approach, the priors can be informative or noninformative, either parametric or nonparametric method can be applied to prior estimation in Emperica Bayes approach.

Noninformative prior in Bayes approach

When no reliable prior information exists, one can use noninformative prior. The noninformative prior is often a flat curve, meaning that the parameter does not vary that much from time to time. Using the noninformative prior, the inference is relatively objective because it is completely based on data information.

Parametric and semiparametric estimated priors in Empirical Bayes approach

Empirical Bayes is another approach to treat the no reliable prior information problem. The method is to estimate the prior distribution using the observed data. The parametric, non-parametric and their combination semiprametric estimated methods can be implemented in prior estimation.

The parametric estimation

As an example, consider the Gamma model,

$$\lambda_i \mid \alpha, \beta \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta), i = 1, \ldots, k. \tag{3.2}$$

Here Gamma (α, β) is the prior distribution of λ_i. One can estimate the parmeter α, β using the observed data.

The Y_i are independent and identically distributed Poisson (λ_i) random variables. We can estimate λ_i according to the data $y = (y_1, y_2, y_3 \ldots\ldots y_k)$ using back projection method. When λ_i is known, the α, β can be estimated using the MLE based on gamma distribution

$$\frac{\alpha}{\beta} = \frac{1}{k} \sum_{i=1}^{k} \lambda_i \tag{3.3}$$

$$\frac{\alpha}{\beta^2} = \frac{1}{k}\sum_{i=1}^{k}(\lambda_i - \bar{\lambda})^2 \tag{3.4}$$

The semiparametric estimation

The prior estimation can also be performed by Penalized Spline Regression, a semiparametric estimation. Semi parametric regression is a combination of parametric and nonparametric regression techniques. In nonparametric regression, there is a nonlinear relationship between outcome and covariates. The semi parametric regression model is constituted of both parametric and nonparametric components. The nonparametric part can adjust to capture the features of data; for example smoothing. When data distribution can not be modelled directly with parametric regression, a semiparametric regression is to better fit data. It also can be viewed as a mixed model because the model consists on fixed and random parts.

Consider the regression model

$$\log(\lambda_i) = m(x_i) + \epsilon_i \tag{3.5}$$

where ϵ_i are i.i.d. N(0, $\sigma^2_{\epsilon_i}$)

$$m(x, \theta) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k \mid x - \kappa_k \mid^3, \tag{3.6}$$

where θ=$(\beta_0, \beta_1, u_1, \ldots, u_k)^T$ is the vector of regression coefficients and $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ are fix knots (Ciprian M. Crainiceanu, et al 2005).

The function (x-κ_k) is called a linear spline basis function. A set of such functions is called a linear spline basis. Any linear combination of linear spline basis function 1, x, (x-κ_1)$_+$, $\ldots$, (x$-\kappa_k$)$_+$ is a piecewise linear function with knots at $\kappa_1, \ldots, \kappa_k$. The linear combination is called a spline.

Considering the vector B=$(\beta_0, \beta_1)^T$ as fixed parameter and vector U as a set of random parameter, the spline regression can also be viewed as a particular case of mixed model.

We minimize

$$\sum_{i=1}^{n} \log(\lambda_i) - m(x_i, \theta)^2 + \frac{1}{\lambda}\theta^T D\theta \tag{3.7}$$

where λ is the smoothing parameter that controls the amount of smoothing and D is a known positive semi-definite penalty matrix that penalizes the coefficient of $\mid x - \kappa_k \mid^3$ (David Ruppert, et al, 2003). Therefore, Penalized Spline Regression is a type of ridge regression. The ridge regression is used to reduce the variability of estimated coefficients, i.e. penalizes the coefficient, in

order to improve the overall prediction accuracy. We also name it "shrinkage" (Trevor Hastie, et al, 2008).

3.2 Point estimate

3.2.1 Frequentist Point estimate

In frequentist method, we use the sample information to estimate one or more population parameters. It is natural to expect that a sample is at all representative of the population. Since an estimate is calculated from sample data alone, it is a statistics (T). Moreover, if the sample is obtained by random sampling, an estimate is a random variable. Its value varies from one sample to another according to its sampling distribution, which is derived from the population distribution. Because of sampling variability, sample information is not totally reliable, and parameter estimate based on sample information are typically in error. The error in any particular estimate is unknown, depending on the true parameter value (θ).

$$MSE(T) = E(T-\theta)^2 = var(T) - (E(T)-\theta)^2 \tag{3.8}$$

The bias in T as an estimate of θ is

$$b_T(\theta) = E(T) - \theta \tag{3.9}$$

The estimate is said to be unbiased when $b_T(\theta)=0$, or $E(T)=\theta$ (Lindgren B. W. 1993).

3.2.2 Bayesian Point estimate

In the Bayesian approach to inference, we treat unknown parameters as random variables. The current distribution of a parameter θ -whether a prior or a posterior- can be used to indicate a value as an estimate of θ. In principle, a value of θ somewhere near the middle of its distribution should be of a good estimate. We assume that the value μ is the estimating of θ, so the "error" in estimating θ to have the value μ to be μ - θ. The absolute error will be a minimum, on average, if we choose μ to be the median of our current distribution for θ. In parallel, the squared error will be a minimum, if we choose μ to be the mean of that distribution. Therefore the value of θ can be estimated by the posterior median or posterior mean (Lindgren B. W. 1993).

3.3 Confidence interval estimate

3.3.1 Frequentist confidence interval

For the usual frequentist CI, if we are able to recompute C for a large number of datasets collected in the same way, about (1-α) x 100% of them contain the true value of θ. In other words, C is the random interval for fixed value of θ with the probability (1-α), called the coverage probability of the interval. It is the probability of a correct guess of where the unknown parameter is (Pawitan, 2001). This is not a very satisfactory statement, since we may not be able to repeat our experiment over a large number of times. Therefore sometimes we construct the confidence interval using bootstrap simulation. Usually, we are in physical possession of only one dataset; our computed C will either contain θ or it won't, so the actual coverage probability will be either 1 or 0. "Thus for the frequentist, the confidence lever (1-α) is only a tag that indicates the quality of the procedure" (Carlin Bradley P.and Louis Thomas A., 2008). For example, we have two same confidence intervals, for 90% and 95% confidence. The 95% confidence should be better than the 90% one, because 95% of repetition will contain true value.

3.3.2 Bayesian credible interval

The Bayesian confidence interval is a credible set, i.e., "The probability that θ lies in C given the observed data y is at least (1-α)" (Carlin Bradley P.and Louis Thomas A., 2008). More formally, a 100 x (1-α)% credible set for θ is a subset C of Θ such that

$$1 - \alpha \leq P(C \mid y) = \int_c p(\theta \mid y) d\theta \tag{3.10}$$

where integration is replaced by summation for discrete components of θ. Hence, the credible set provides an actual probability statement, based only on the observed data and whatever prior opinion we have added.

Highest posterior density

We wish credible sets to have an exactly correct coverage. In order to obtain a more precise estimate, a technique for minimize the credible sets named Highest Posterior Density (HPD) is defined as the set

$$C = \{\theta \in \Theta : p(\theta \mid y) \geq \kappa(\alpha)\} \tag{3.11}$$

where $\kappa(\alpha)$ is the largest constant satisfying

$$P(C \mid y) \geq 1 - \alpha. \tag{3.12}$$

For example, giving the $\kappa(\alpha)$=0.1, we have a 87% HPD interval of (0.12, 3.59) for posterior distribution, if Gamma (2,1) is prior distribution (Carlin and Louis, 2008).

Chapter 4

Simulation technique

4.1 Bootstrap simulation

The explicit recognition of uncertainty is central to the statistical sciences. The uncertainty of statistical inference from sample data representing population may be gauged by analytical calculation based on an assumed probability model for the available data. However, in more complicated problems the mathematical modelling can be difficult, and its results are potentially misleading if inappropriate assumptions or simplifications have been made. Bootstrapping is a computer-intensive method to enabling us obtain reliable standard errors, confidence intervals, and other measures of uncertainty for a wide range of problems. The method is performed by resampling from original data to create replicate dataset. Direct resampling is named nonparametric bootstrapping, while resampling via a fitted model is parametric bootstrapping (A.C.Davison and D.V. Hinkley, 1997).

4.1.1 Nonparametric bootstrap

Let $x_1, x_2, \ldots, x_n$ be a independent and identically random sample from a unknown probability distribution F. One way to estimate SE($\bar{x}$) would be to draw a large number of random samples of size n from F, calculate $\bar{x}$ for each sample, and then use the standard deviation of these $\bar{x}$ values as the desired estimate. The steps in the bootstrap algorithm are as follows:

1. Draw N random samples with replacement from $\mathrm{x}_1, x_2, \ldots, x_n$. Denote these bootstrap samples by $\mathrm{x}^*_{i1}, x^*_{i2}, \ldots, x^*_{in}$, i=1,2,„. . . ,N.

2. Calculate the sample mean of each bootstrap sample and the overall sample mean:

$$\bar{x^*_i} = \frac{\sum_{j=1}^{n} x^*_{ij}}{n} \tag{4.1}$$

$$\bar{\bar{x^*}} = \frac{\sum_{i=1}^{N} \bar{x_i^*}}{N} \tag{4.2}$$

3. Calculate

$$SE(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{N} (\bar{x_i}^* - \bar{\bar{x}}^*)^2}{N-1}} \tag{4.3}$$

(Ajit C.Tamhane and Dorothy D. Dunlop, 2000).

4.1.2 Parametric bootstrap

Let $\mathrm{x}_1, x_2, \ldots, x_n$ be an independent and identically random sample from a known probability distribution F with parameter Ψ. We estimate the Ψ using the $\mathrm{x}_1, x_2, \ldots, x_n$ fitted distribution. Then generate the random variable $\mathrm{X}^* = x_1, x_2, \ldots, x_n$ according to the fitted distribution.The confidence intervals or quantiles of X^* can be obtained based on the calculation of the expectation and variance. The steps in the bootstrap algorithm are as follows:

1. Assume the F is Poisson distribution, and we estimate the parameter λ of Poisson distribution according to samples $\mathrm{x}_1, x_2, \ldots, x_n$.

2. Generate the N random samples $\mathrm{x}_{i1}^*, x_{i2}^*, \ldots, x_{in}^*$, i=1,2,...N using Poisson distribution with parameter λ.

3. Calculate

$$SE(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{N} (\bar{x_i}^* - \bar{\bar{x}}^*)^2}{N-1}} \tag{4.4}$$

(A.C.Davision and D.V. Hinkley, 1997, Bootsrtap methods and their application)

4.2 MCMC simulation

For a continuous parameter space, Bayes's formula for the posterior distribution is presented by formula 3.1. To use this equation we need to determine the normalizing constant (the denominator). If there are k unknown parameters $(\theta_1, \ldots \theta_k)$, then the denominator involves an integration over the k-dimensional parameter space which becomes intractable for large values of k. (Dobson, Annette J. and Barnett Adrian G. 2008)

Markov chain Monte Carlo (MCMC) is a numerical method for calculating complex integrals. This method is operated by sequentially sampling

parameter values from a Markov chain whose stationary distribution is exactly the desired joint posterior distribution of interest. (Carlin Bradley P. and Louis Thomas A., 2008)

Gibbs sampling is one of the algorithms of Bayesian MCMC computation. Given an arbitrary set of starting value $\theta_2^0, \ldots, \theta_k^0$, the algorithm proceeds as follows:

For (t=1,..., T), repeat:
Step 1: Draw θ_1^t form p($\theta_1 \mid \theta_2^{t-1}, \theta_3^{t-1}, \ldots, \theta_k^{(t-1)}, Y$)
Step 2: Draw θ_2^t form p($\theta_2 \mid \theta_1^{t-1}, \theta_3^{t-1}, \ldots \theta_k^{(t-1)}, Y$)
...
Step k: Draw θ_k^t form p($\theta_k \mid \theta_1^t, \theta_2^t, \ldots \theta_{k-1}^t, Y$)

From above algorithms one can show that the k-parameter obtained at iteration t, ($\theta_1^t, \theta_2^t, \ldots, \theta_k^t$), converges in distribution to a draw from the true joint posterior distribution p($\theta_1, \theta_2, \ldots, \theta_k \mid Y$). The time from t=0 to t=t_0 is commonly known as the burn-in period, i.e. chain convergence period. We usually apply multiple chains to assess convergence. By starting start multiple chains at widely varying starting values, each china converges to the same solution. This would increase our confidence in this solution.

Chapter 5

An application: HUS outbreak

5.1 Data sources and assumption

5.1.1 HUS outbreak data

The largest ever documented outbreak of hemolytic uremic syndrome (HUS) occurred in 2011. The disease spread from the outbreak area, Northern Germany, to many other countries including Sweden via persons travelling to or through the outbreak area. The causative agentA was a Shiga toxin -producing Escherichis coli (STES) of the rare serotype O 104:H4. The infection most likely spread by Fenugreek sprouts (Buchholz U. et al 2011).

At the Robert Koch Institute - the federal public health institute in Germany - there were 3,793 registered outbreak cases until March 1, 2012. Among them 827 cases (22%) of STEC gastroenteritis were diagnosed as HUS patients. Comparing to the historical occurrences, the large proportion(>88%) of adult patients was a unique feature to this outbreak. The data used in our analysis are 686 adult HUS patients from the 827 cases.

The incubation period was assumed to be bounded above by a maximum period of 17 days and its distribution was also estimated from symptomatic individuals within the cohorts at the Robert Koch Institute, using Turnbull's method (Werber,D. et al. 2013, and epidemiology course lecture 4 slide 26, Michael Höhle, 2014). Figure 1 displays the time series of daily onsets of the 686 adult HUS patients with information about the onset time.The probability density function of the incubation period distribution can be seen in figure 2.

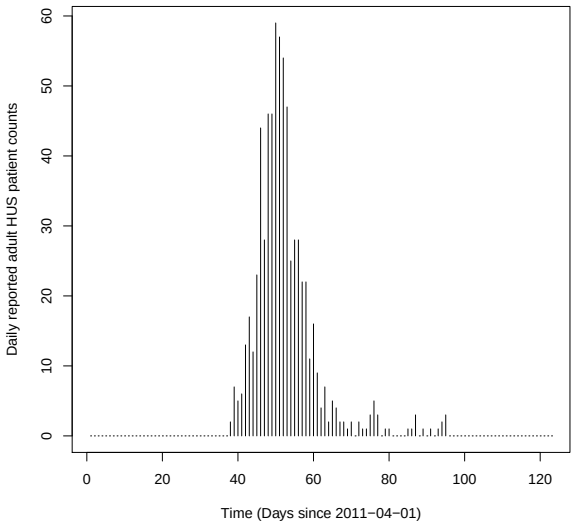

Figure 1. The distribution of 686 adult onset hemolytic uremic syndrome (HUS) patients over time.

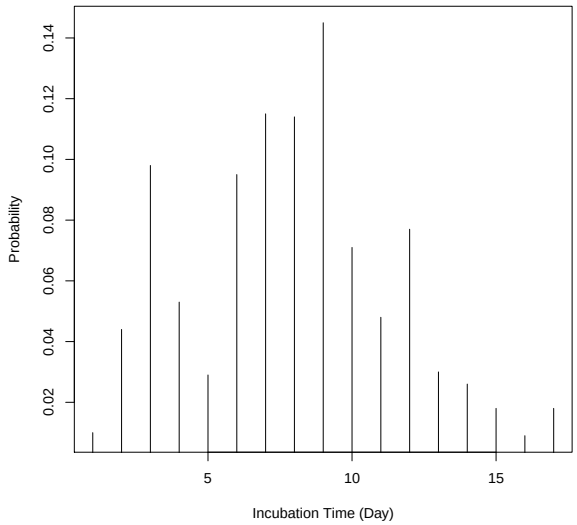

Figure 2. Probability distribution over incubation time estimated from symptomatic individuals within cohorts at the Robert Koch Institute.

5.1.2 Assumption

The purpose of the data analysis is to estimate the infected intensities λ_t, $t = 1, 2...T$, given observed onset patient counts. Using the nonparametric back-projecting method, STES coli O104:H4 outbreak data and the incubation period distribution, we estimated the λ_t, $i = 1, 2...T$, as Scenario 2. Alternatively, we also estimated the λ_t, $i = 1, 2...T$, using the average onset intensity during the occurrence period to represent the infected intensities, pretending that the observation is not available . It's named Scenario 1.

Scenario 1

Assume apriori for infected intensity λ_t is constant (686/100) during the occurrence period, time points 30 to 61, at the rest of time points, $\lambda_t = 0$.

Scenario 2

Assume apriori for infected intensity λ_t is the exposure curve of Escherichia coli O104:H4 (HUS) outbreak data, i.e the infected intensity curve from EM back-project method.

5.2 EM back-projection method application

We applied the EM back-projection method to estimate the infected intensity λ using generated symptom onset patient counts. Therefore we need λ_t to calculate μ in formula 2.8, and then generate symptom onset patient counts according to Poisson distribution.

With λ_t from scenario 1 and 2, and a probability distribution function, we obtained generated symptom onset patient counts. The probability distribution function (probability mass function) was generated by discretizing a continuous random variable with positive support and cumulated distribution function at the Robert Koch Institute, Germany. We assumed that the probability mass function have finite support 1...123 days. By plugging the generated y_t in the function `backprojNP`, the estimated λ_t curves were obtained. The smoothing parameters $k = 2, k = 4$ were used.

The reason that we will generate new onset data y_t instead of observation is that we wish to have a more general estimation of λ_t. This λ_t will be used as a prior in the later Empirical Bayesian approach. Furthermore, we also wish to have a accurate estimation of λ_t despite of its smaller sample size . We used the unique sample λ_t per time point as a mean of Poisson

regression, then generated the onset data randomly. By parametric bootstrap simulation, 25th and 95th quantiles were obtained as well.

Software

R version 3.1.1 combined with RStudio was used for statistical analysis. The packages "survival" and "surveillance" were used for the EM back-projection method (https://cran.r-project.org/web/packages/surveillance/surveillance.pdfand https://cran.r-project.org/web/packages/survival/index.html).

The result

By using the back-projection method, we found the estimated infected curve, shown in the Figure 3. The curve showed that the main infection occurred from the time point 37 to 57 and the highest peak reached up to 73 infected individuals at time point 44. In parallel, the majority of the onset patient incidence appeared from the time point 38 to 70 and, a maximal number of 59 patients was diagnosed at time point 50. These results show that main distribution of onset is slightly delayed compared to the exposure. The Figure 4 displays the λ_t over the delayed time used in a simulation strategy scenario 1. The Figure 5 shows a similar curve generated used in scenario 2.

We first had λ_t from scenario 1, and then generated a series of onset counts based on λ_t. Finally, we estimated a new λ_t using the EM back-projection function. In addition, using the same strategy, we also tested for different smoothing parameters ($(k = 2, 4)$ for newly estimated λ_t. However the residuals between original and newly estimated λ_t did not increase proportionally with the increase of smoothing parameters. Therefore we chose the curve with the smoothing parameter $k = 4$ (Figure 6) as a reference. Then we generated the prior distribution for Empirical Bayesian approach, assuming that the prior is a gamma distribution. The parameters of gamma distribution were estimated according to the epidemic curve based on the infection's curve.

In parallel, we found similar results when we tested scenario 2. In Figure 7, the curve represents infected intensity used as the reference for prior generation of the Empirical Bayesian approach. We selected the smoothed curve with smoothing parameter $k = 4$. With 1000 times regeneration of Poisson regression, we obtained the 25 and 95 quantiles. See displayed results in Figure 8. Nonparametric bootstrap was also applied for bias and confidence interval estimation of the infected intensities' mean. The tabel 5.1 displays the bias adjusted mean and confidence interval for time series infected intensity.

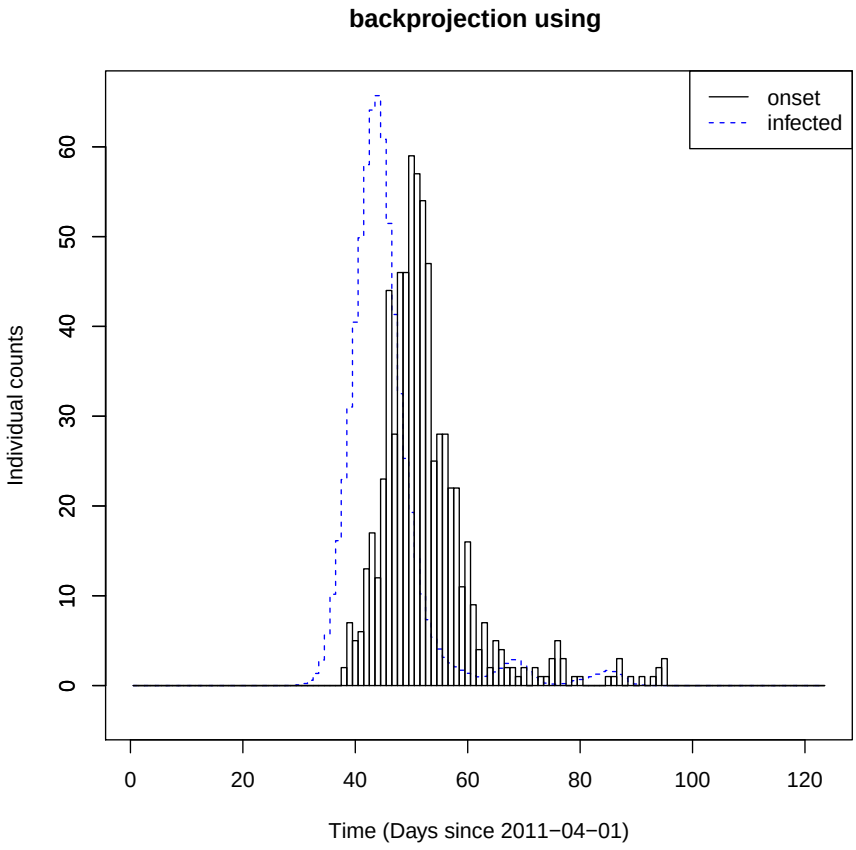

Figure 3. Backprojection using for adult HUS onset data, smoothing parameter k=4.

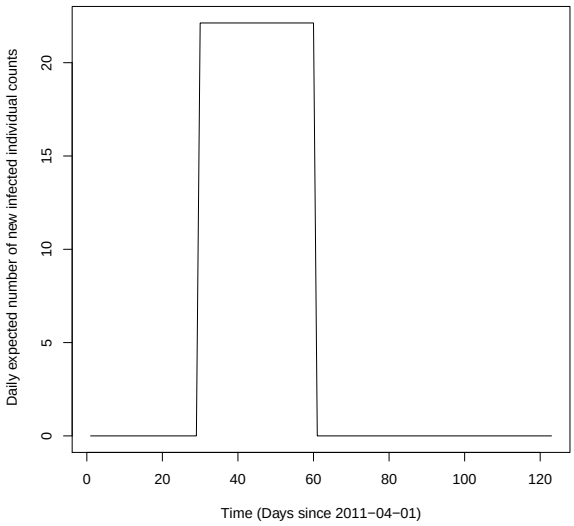

Figure 4. $\lambda(t)$ used in scenario 1.

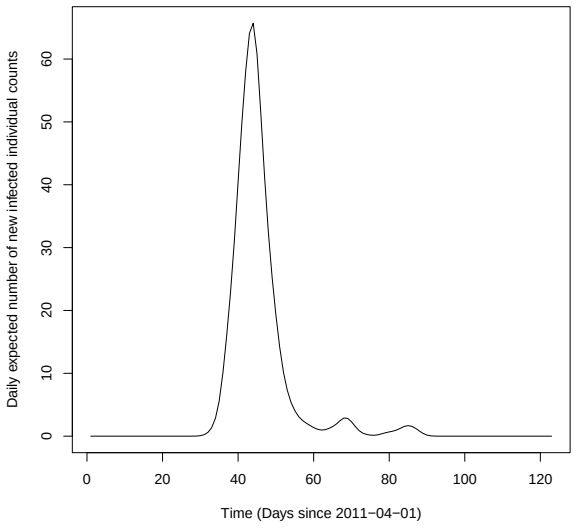

Figure 5. $\lambda(t)$ used in scenario 2.

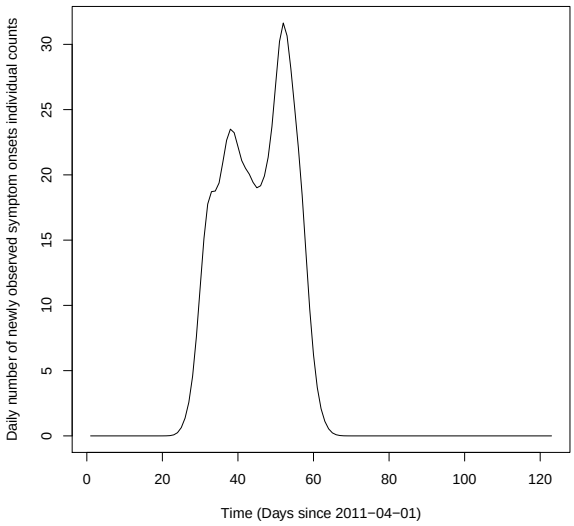

Figure 6. Estimated $\lambda(t)$ according to Scenario 1.

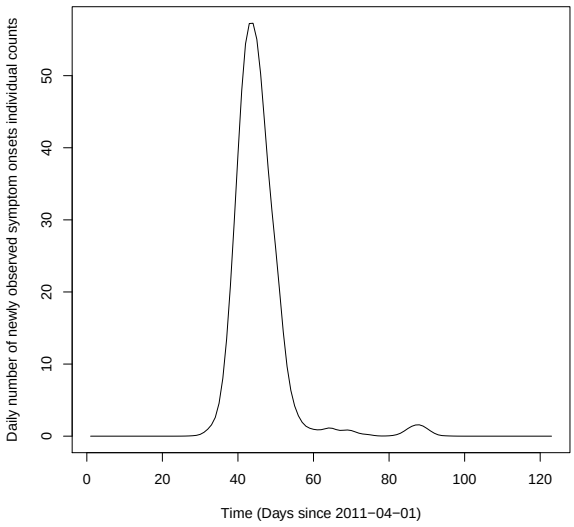

Figure 7. Estimated $\lambda(t)$ according to Scenario 2.

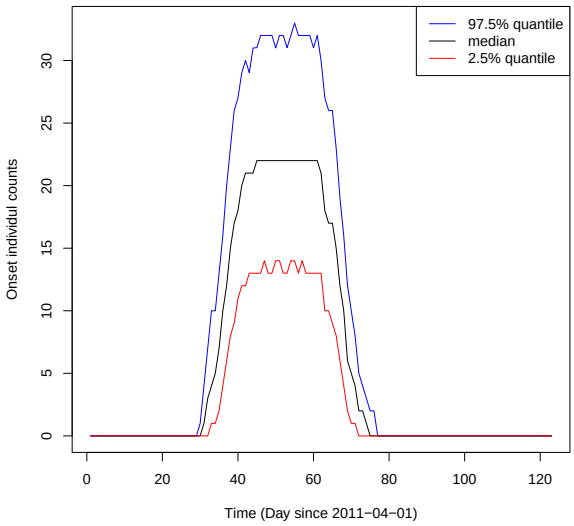

Figure 8. Bootstrap quantiles of onset counts from scenario 1.

Table 5.1: Bias corrected mean and 95% confidence interval of infected individual intensities from scenario 1 and 2.

	Scenario 1	Scenario 2
mean	5.43	5.44
low 95% CI	5.28	5.22
high 95% CI	5.58	5.65

5.3 Bayesian back-projection method application

In order to estimate the uncertainty of parameter λ_t, we applied 6 Bayes approaches for statistical inference. The essential elements for the Bayes and Empirical Bayesian approach are prior distributions. Using observed datasets and models, the prior parameters could be estimated in Emperical Bayesian approach.

The prior distribution for Bayesian approach

Assume the prior of infected intensity λ_t is gamma or uniform distribution

$$\lambda_t \overset{\text{iid}}{\sim} \text{Gamma}(0.01,\ 0.01) \tag{5.1}$$

$$\lambda_t \overset{\text{iid}}{\sim} \text{Uniform}(0,\ 5000) \tag{5.2}$$

$$\lambda_t \overset{\text{iid}}{\sim} \text{Uniform}(0,\ 250) \tag{5.3}$$

The prior distribution for Empirical Bayesian approach

From the back-projection method, we obtained estimated infected intensity curve λ_t. Therefore, the expectation and variance of λ_t can be obtained. Assume the prior is gamma distribution with parameters α, β, therefore we could calculate the α, β according to the formulas (3.4) and (3.5) respectively and then we generated the priors under assumption of gamma ditribution with parameter α, β, i.e.

$$\lambda_t \overset{\text{iid}}{\sim} \text{Gamma}(\alpha,\ \beta) \tag{5.4}$$

Here we corrected the expectation of λ_t using bootstrap simulation.

Assume the prior of infected intensity λ_t is a penalized spline regression model

$$\ln \lambda_t = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k z_{ik} \tag{5.5}$$

$$b_k \sim \text{N}(0,\ \sigma_b^2) \tag{5.6}$$

$$\epsilon_i \sim \text{N}(0,\ \sigma_\epsilon^2) \tag{5.7}$$

$$\beta_0, \beta_1 \sim \text{N}(0,\ 10^6) \tag{5.8}$$

$$\sigma_b^{-2}, \sigma_\epsilon^{-2} \sim \text{Gamma}(10^{-6},\ 10^{-6}) \tag{5.9}$$

where z_{ik} is the (i,k)th entry of the design matrix

$$Z = Z_K \Omega_K^{-1/2} \tag{5.10}$$

The data and method

As description in 5.1.1, the HUS outbreak data, i.e. the time series HUS onset count data were applied. Using Jags model we performed Bayesian approach with 3 chains. With Gibbs sampling, MCMC was completed by `coda.sample`. Based on the test for convergence, we chose 3000 iterations as burn in. We then displayed the median, 2.5% and 97.5 quantiles according to the outcome of MCMC simulation.

Software

The package " rjags" in R environment was implemented for Bayesian approach. The coda package also provides convergence diagnostics to check if the output is valid for analysis. (Plummer et al. 2006).

The result

Based on the observed HUS data and the Poisson model, we performed Bayesian and Empirical Bayesian analyses using different priors. We used Gamma (0.01, 0.01), Uniform (0, 5000) and Uniform (0, 250) as prior distributions and displayed the estimated results for lambda and N. Figure 9 is a posterior distribution for lambda with Gamma (0.01, 001) as a prior. Figure 10 and 11 are posterior distributions for lambda with Uniform (0, 5000) and Uniform (0, 250) as priors, respectively. Meanwhile, a penalized spline regression was used as another approach for prior selection, where 20 knots were used for smoothing. The results are shown in Figure 12. For the Empirical Bayesian method, the priors were estimated by the back-projection method. The posterior distributions of λ_t from scenario 1 and scenario 2 are displayed in Figure 13 and Figure 14, respectively. In all Figures, the estimated λ_ts are presented with medium of HPD. The 2.5 and 97.5 quantiles of HPD are also indicated. We have also obtained similar results for N_t, the infected count. The result figures are not showed in this thesis but available upon request.

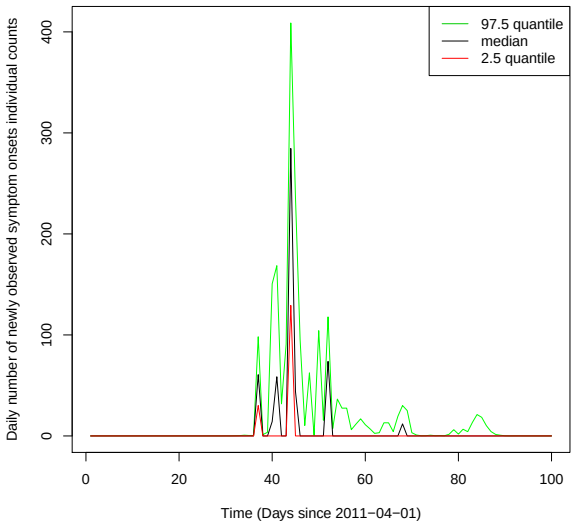

Figure 9. The posterior distribution of λ, with Ga (0,01 0,01) distributed prior.

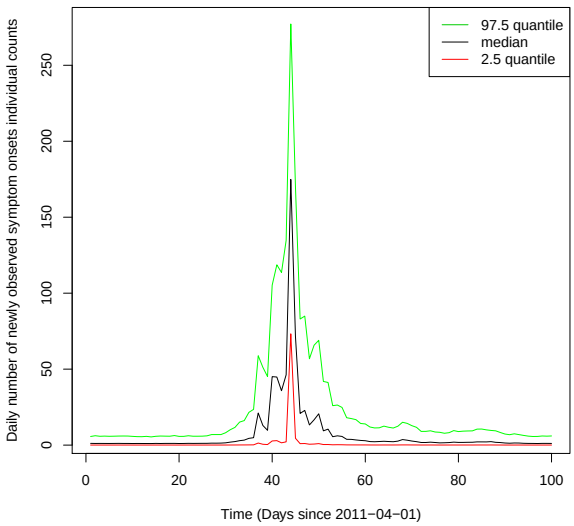

Figure 10. The posterior distribution of λ, with unif (0, 5000) distributed prior.

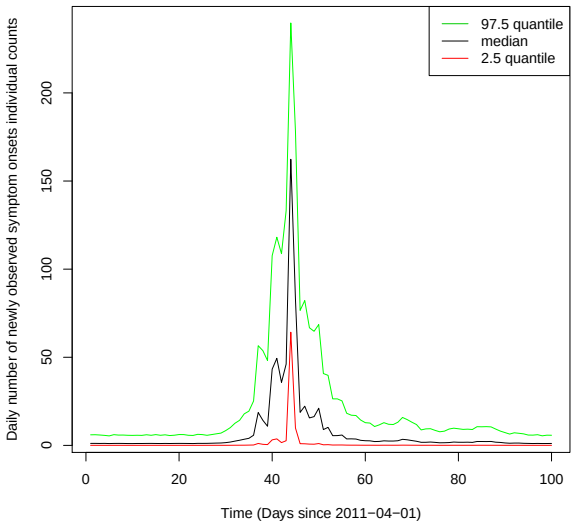

Figure 11. The posterior distribution of λ, with unif (0, 250) distributed prior.

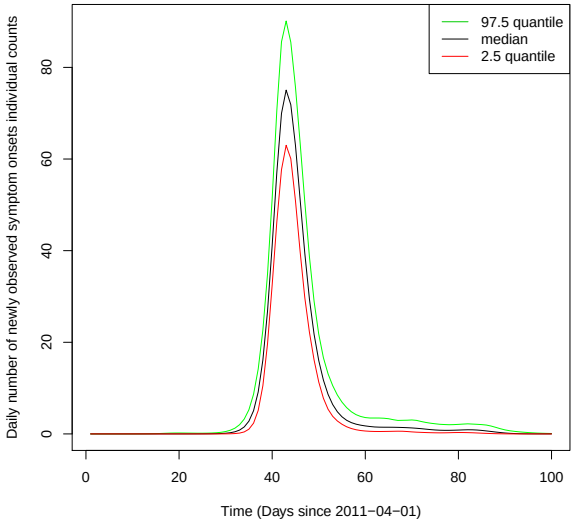

Figure 12. The posterior distribution of λ, with penalized spline regression estimated prior.

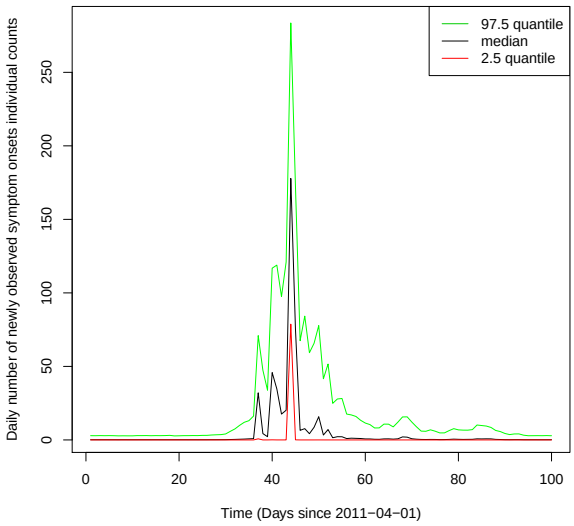

Figure 13. The posterior distribution of λ, with Ga ($\alpha 1$ $\beta 1$) distributed prior.

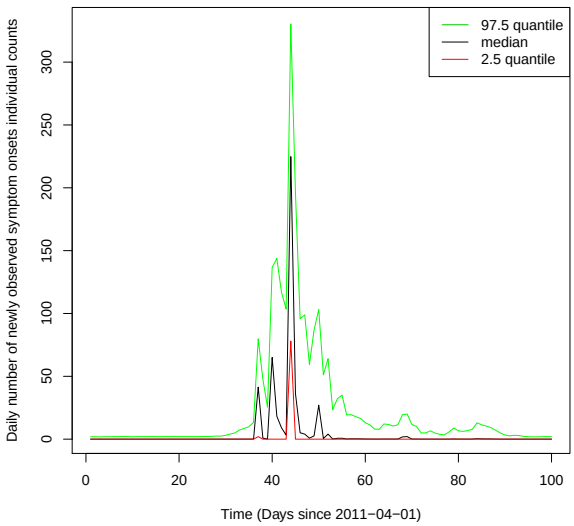

Figure 14. The posterior distribution of λ, with Ga ($\alpha 2$ $\beta 2$) distributed prior.

5.4 The result comparison between EM and Bayesian back projection methods

Mean and standard error of the prior and estimated result of posterior by the Bayesian and EB methods for λ_t and N_t are shown in Table 5.2, 5.3 and 5.4. In general, posterior mean of λ_t is comparable in Table 5.3, though different priors were applied. Variation of the prior distribution may be responsible for a minor shift of the posterior mean. The parameter's difference in the identical prior distribution led to the lowest impacts on the posterior mean.

The prior by spline regression gives the posterior mean (about 6.85, depended on simulation) that is the nearest to the result (6.86) estimated by the frequentist approach which is calculated by maximal likelihood estimate (MLE). The two Empirical Bayesian methods also led to similar results. The scenario 2 produced the result more approximate to the MLE than the scenario 1, since its prior was highly relevant to the real data. The Uniform distribution led to slightly higher estimations with 8.6. Using the Bayesian approaches, the the unknown variable N_t, i.e. the numbers of individual infected in the time interval T were estimated as well. The results are showed on table 5.4.

The mean square error (MSE) of Bayesian approaches and EM in comparison with the original HUS patient data from scenario 1 and scenario 2 are shown in Table 5.5. The EM method gives the lowest MSE. Among Empirical Bayesian aproaches, penalized spline regression showed the best modelling result. The uniform distribution shows a lower MSE than gamma distribution. In general, the scenario 1 resulting less MSE than scenario 2 suggest that the flat prior has some advantage.

Table 5.2: Mean and SE of priors.

	lambda prior mean	lambda prior SE
B Gamma(0.01 0.01)	1.00	100.00
B Uniform(0 5000)	2500.00	2080000.00
B Uniform(0 250)	125.00	5200.00
B Spline	6.86	0.00
EBS1 Gamma(alpha1 beta1)	5.41	9.38
EBS2 Gamma(alpha2 beta2)	5.38	13.61

Table 5.3: Posterior mean and SE of infected intensities.

	lambda posterior mean	lambda posterior SE
B Gamma(0.01 0.01)	6.75	28.29
B Uniform(0 5000)	8.62	20.76
B Uniform(0 250)	8.62	19.94
B Spline	6.85	16.60
EBS1 Gamma(alpha1 beta1)	6.65	19.80
EBS2 Gamma(alpha2 beta2)	6.76	23.11
Freqentist	6.86	13.91

Table 5.4: Posteria mean and SE of infected individual counts.

	N posterior mean	N posterior SE
B Gamma(0.01 0.01)	6.80	29.96
B Uniform(0 5000)	7.62	20.03
B Uniform(0 250)	7.62	19.94
B Spline	6.86	16.49
EBS1 Gamma(alpha1 beta1)	6.73	21.97
EBS2 Gamma(alpha2 beta2)	6.79	23.87

Table 5.5: Mean squear error of EM and Bayesian with 6 different priors.

	MSE
EM Back-projection Scenario 1	8895.84
EM Back-projection Scenario 2	19356.49
Empirical Bayesian Back-projection Scenario 1	46231.63
Empirical Bayesian Back-projection Scenario 2	61020.59
Bayesian Spline prior	24683.41
Bayesian Gamma(0.01 0.01) prior	84666.40
Bayesian Uniform(0 5000) prior	44590.26
Bayesian Uniform(0 250) prior	42843.34

Chapter 6

Discussion and conclusions

The present thesis gave an example for solving unobserved data problem with both Bayesian and traditional frequentist methods. We are interested in inferring the time point of disease exposure/transmission using individual's symptom onset data. The results clearly show that the Bayesian back-projecting method is a reliable method for point estimation of infected incidence. Our conclusion is based on the identity of the results from Bayesian and standard frequentist methods.

We also found an almost equal posterior mean by testing different parameters of the Uniform distribution. It indicates that the noninformative prior is especially robust when the distribution has been determined. In addition, the gamma distribution also leads to a similar posterior mean compared to the frequentist method.

The penalized spline regression for prior estimation leading to the lowest MSE implies that the semi-parametric regression applied to prior estimation has some advantages. The minimal MSE may be due to its close relevance to the data. This method can be considered as a nonparametric Empirical Bayesian approach. In our analyses, the optimal number of knots for smoothing used in penalized spline regression remains to be determined, which could be studied by using cross-validation in the future.

In addition, we performed the EM back-projection method for lambda estimation, and then generated data. The bootstrap simulation has been introduced for re-sampling in order to determine a confidence interval. Both parametric and nonparametric bootstrapping were tested. By parametric bootstrapping of λ, we generated the onset individual data 1000 times with replacement, then estimated λ and, a confidence interval of estimation expected. However, the step-by-step bootstrap simulation process is complicated for the time series data. As a consequence, we did not reach the λ estimation step, and terminated the process after generating the onset indi-

vidual data . In contrast, the Bayesian back-projection approach has some advantage because its MCMC simulation simplifies the complex integration. We obtained the posterior confidence interval in one step.

To summarize, this thesis establishes the Bayesian back-projection approach for unobserved data estimation by applying HUS outbreak time series data. By comparing the EM and Bayesian back projection methods, as well as the different priors of the Bayesian approach, we obtained some experience on the prior selection when no prior information could be obtained. It can be summarized on three aspects. Firstly, the flat distribution of prior gives the better modeling effect, indicated by comparing uniform to gamma distribution. It is also determinated by comparison of scenario 1 and 2 in Empirical Bayesian approach. As a prior, scenario 1, giving a relative flatter curve than scenario 2, leads to a better modeling result. Secondly, the simulation technique applying to prior generation leads to similar posterior means in scenario 1 and 2. It indicates that the simulation technique generalized the data relevant prior, giving a stable result. Thirdly, comparing the result using the Bayesian Spline priora and noninformative prior, the Bayesian Spline priora shows the less MSE than noninformative prior wihich indicates a better result. However, the noninformative prior gives a more objective modeling because the prior information is unrelated to the analyzed data. The advantage and limitation of informative and noninformative prior are topics to discuss in the future.

Chapter 7

References

https://aidsinfo.nih.gov/education-materials/fact-sheets/19/45
/hiv-aids–the-basics

Becker, Niels. G., et al. A method of non-parametric back-projection and its application to AIDS data. Stat Med. 1991;10(10):1527-1542.

https://en.wikipedia.org/wiki/Hemolyticuremic/syndrome

https://www.r-project.org

https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids
/hiv-in-your-body/stages-of-hiv/

Joseph R. Egan and Lan M. Hall, A review of back-calculation techniques and their potential to inform mitigation strategies with application to non-transmissible acute infectious diseases. J. R. Soc. Interface 2015 May 6;12(106).

Isham,V. Estimation of the incidence of HIV infection,', Philosophical Transactions of the Royal Society of London, B,325,113-121 (1989).

Carlin Bradley P. and Louis Thomas A., Bayesian Methods for Data Analysis. Chapman and Hall/CRC Press, 3rd edition. 2008.

Pawitan Yudi, In all likelihood: statistical modelling and inference using likelihood, by Oxford University press Inc., New York, 2001.

Lindgren Bernard W. Statistical theory, by Chapman Hall/CRC. 1993.

Ciprian M. Crainiceanu, et al. Bayesian analysis for penalized splline regression using WinBUGS. 2005

David Ruppert,et al. Semiparametric regression, by Cambridge University Press, 2003. Trevor Hastie, et al. The elements of statistical learning, data mining, inference, and prediction, by Springer Science+Business Media, LLC 2009.

https://en.wikipedia.org/wiki/Bootstrapping$_{(}$*statistics*)

Ajit C.Tamhane and Dorothy D. Dunlop, Statistics and data analysis from elementary to intermediate, by Prentice-Hall, Inc. 2000.

Package 'boot' , Author Angelo Canty [aut], Brian Ripley [aut, trl, cre] (author of parallel support), Version 1.3-17 Date 2015-06-29. https://cran.r-project.org/web/packages/boot/boot.pdf

A.C.Davision and D.V. Hinkley, Bootsrtap methods and their application, by Cambridge University Press, 1997.

Dobson, Annette J. and Barnett Adrian G., An introduction to generalized linear models, by Taylor Francis Group, LLC 2008.

Werber D. et al. Associations of age and sex with the clinical outcome and incubation period of Shiga toxin-producing Escherichia coli O104:H4 infections, 2011. Am J Epidemiol. 2013 Sep 15;178(6):984-92.

Buchholz U et al. German outbreak of Escherichia coli O104:H4 associated with sprouts. N Engl J Med. 2011 Nov 10;365(19):1763-70

https://cran.r-project.org/web/packages/surveillance/surveillance.pdf

Plummer M, Best N, Cowles K, Vines K (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC, R News, 6:7-11.

Package 'rjags',August 30, 2013, Version 3-11, Title Bayesian graphical models using MCMC, Author Martyn Plummer [aut, cre],Alexey Stukalov [ctb]. URL http://mcmc-jags.sourceforge.net

Chapter 8

Appendix

8.1 R code for EM back projection

```
require("surveillance")
require("xts")
require("xtable")

rm(list=ls(all=TRUE))
load("~/project of master degree/O104Outbreak.RData")
ls()
incu.pmf
ts.hus
plot(ts.hus,xlab="Time Day since 2011-04-01 ",type="h")
##timepoint 40-80 fig1
plot(incu.pmf,xlab="incubation Time (Day)", type="h")
##incubation time ditribution fig2

##back-projection with ts.hus real data
sts <- new("sts",epoch=1:length(ts.hus),observed=matrix(ts.hus,ncol=1))
bp.control <- list(k=2,eps=1e-4,iter.max=1000,verbose=TRUE,eq3a.method="C")
sts.bp.k0 <- backprojNP(sts, incu.pmf=incu.pmf, control=bp.control)
plot(sts.bp.k0,xlab="Time (Day since 2011-04-01)",
xaxis.labelFormat=NULL,legend.opts=NULL)##fig3
upperbound(sts.bp.k0)
plot(upperbound(sts.bp.k0),xaxis.labelFormat=NULL,legend.opts=NULL,
type="h")timepoint 30-60

##Create a curve (as modelling result) of lambda(t)
##scenario 1
##all of the patients infected during the time interval t0 to t0+L+1
t0<-30
L<-30
T <-length(ts.hus)
```

```
average<-sum(observed(sts))/(L+1)
lambda1<-c((rep(0,(t0-1))),(rep(average,(L+1))),(rep(0,(T-t0-L))))
plot(lambda1,type="l",xlab="Time (Day since 2011-04-01)",
xaxis.labelFormat=NULL, ylab="No. infected", main="lambda1 for scenario 1")#Fig 5

#scenario 2
## Directly use the result of exposure curve from real data
lambda2<-upperbound(sts.bp.k0)
plot(lambda2,type="l",xlab="Time (Day since 2011-04-01)",xaxis.labelFormat
3=NULL, ylab="No. infected", main="lambda2 for scenario 2")#Fig 6

#create a new incu.pmf for newy function
T <-length(ts.hus)
d.max <- T-1
d.grid <- seq(0,d.max,by=1)
Len<-length(incu.pmf)
incu.pmfnew<-c(incu.pmf,(rep(0,(T-Len))))
barplot(height=incu.pmfnew,names.arg=d.grid,width=1,space=0,
xlab="Delay D (in days)",ylab="Probability")#figure4
I<-length(incu.pmfnew)
#newy function
newy<-function(lambda){
flambda<-matrix(0,nrow=length(ts.hus),ncol=length(ts.hus))
for(i in 1:length(ts.hus)){
 for(j in 1:length(ts.hus)){
   if (i >= j) {
     flambda[i,j]<-incu.pmfnew[i-j+1]*lambda[j]
   }
 }
}
mut<-rowSums(flambda)
set.seed(1234)
y<-vector("list",length(ts.hus))
for (i in 1:length(ts.hus)){
  y[[i]]<-rpois(1,mut[i])
}
y<-as.numeric(y)
return(y)
}
#calculate newy with lambda 1
newy(lambda1)
stssimulationS1 <- new("sts",epoch=1:length(newy(lambda1)),
observed=matrix(newy(lambda1),ncol=1))
bp.control <- list(k=4,eps=1e-4,iter.max=1000,verbose=TRUE,eq3a.
                   method="C")
sts.bp.k0.simS1 <- backprojNP(stssimulationS1, incu.pmf=incu.pmf,
                              control=bp.control)
plot(sts.bp.k0.simS1,xlab="Time (Day since 2011-04-01)",
```

```
     xaxis.labelFormat=NULL,legend.opts=NULL)
lines(lambda1,xaxis.labelFormat=NULL,legend.opts=NULL,col="red")
legend(x="topright",c("onset","infected","lambda1"),lty=c(1,2,1),
       lwd=c(1,1,1),col = c(1,4,2))
infectedS1<-upperbound(sts.bp.k0.simS1)#[27:67]>1
plot(infectedS1,xlab="Time (Day since 2011-04-01)",
     xaxis.labelFormat=NULL,legend.opts=NULL,type="l")
lines(lambda1,xaxis.labelFormat=NULL,legend.opts=NULL,col="red")
legend(x="topright",c("infected","lambda1"),lty=c(1,1),
       lwd=c(1.5,1.5),col = #c(1,2))
residuals1<-sum(lambda1-infectedS1)##residuals1=20 for k=0,k=2,k=4
MSE.EM.lambda1<-sum((lambda1-infectedS1)^2)
## estimate lambda acording to Scenario 1
estimatedlambdaS1<-infectedS1  ##[27:67]
plot(estimatedlambdaS1,type="l",xlab="Time (Day since 2011-04-01)",
ylab="No.infected")#Fig 10
##prior parameters for Scenario 1
aS1<-mean(estimatedlambdaS1)#alpha/beta
bS1<-var(estimatedlambdaS1)
alphaS1<-aS1^2/bS1
betaS1<-aS1/bS1
# calculate newy with lambda 2
newy(lambda2)
stssimulationS2 <- new("sts",epoch=1:length(newy(lambda2)),
                       observed=matrix(newy(lambda2),ncol=1))
bp.control <- list(k=4,eps=1e-4,iter.max=1000,verbose=TRUE,
                   eq3a.method="C")
sts.bp.k0.simS2 <- backprojNP(stssimulationS2, incu.pmf=incu.pmf,
                              control=bp.control)
plot(sts.bp.k0.simS2,xlab="Time (Day since 2011-04-01)",
main="Backprojection using for generated onset with lambda 2, k=4",
xaxis.labelFormat=NULL,legend
.opts=NULL)
lines(lambda2,xaxis.labelFormat=NULL,legend.opts=NULL,col="red")
legend(x="topright",c("onset","infected","lambda2"),lty=c(1,2,1),
       lwd=c(1,1,1),col = c(1,4,2))
infectedS2<-upperbound(sts.bp.k0.simS2)##[35:71]>1
plot(infectedS2,xlab="Time (Day since 2011-04-01)",
     xaxis.labelFormat=NULL,legend.opts=NULL,type="l")
lines(lambda2,xaxis.labelFormat=NULL,legend.opts=NULL,col="red")
legend(x="topright",c("infected","lambda2"),lty=c(1,1),
       lwd=c(1.5,1.5),col = c(1,2))
residuals2<-sum(lambda2-infectedS2)##residuals2=19 for k=0,k=2,k=4,
MSE.EM.lambda2<-sum((lambda2-infectedS2)^2)

## estimate lambda according to Scenario 2
estimatedlambdaS2<-infectedS2 ##[35:71]
plot(estimatedlambdaS2,type="l",xlab="Time (Day since 2011-04-01)",
```

```
     ylab="No.infected")#Fig 14
##prior parameters for Scenario 2
aS2<-mean(estimatedlambdaS2)#alpha/beta
bS2<-var(estimatedlambdaS2)
alphaS2<-aS2^2/bS2
betaS2<-aS2/bS2 sqrt(bS2)
```

8.2 R code for bootstrap

8.2.1 Parametric bootstrap

```
##newy function without seed
newy.f<-function(lambda){
flambda<-matrix(0,nrow=length(ts.hus),ncol=length(ts.hus))
for(i in 1:length(ts.hus)){for(j in 1:length(ts.hus))
  {if (i >= j) {
    flambda[i,j]<-incu.pmfnew[i-j+1]*lambda[j]
  }
 }
}
mut<-rowSums(flambda)
y<-vector("list",length(ts.hus))
for (i in 1:length(ts.hus)){
  y[[i]]<-rpois(1,mut[i])
}
y<-as.numeric(y)
return(y)
}
#begin bootstrap resampling
d<-replicate(1000,newy.f(lambda1))
i<-sample(1000,replace=T)
qt<-vector("list",length(ts.hus))
for(i in 1:length(ts.hus)){
  qt[[i]]<-quantile(d[i,],c(0.975,0.5,0.025))
}
ul<-unlist(qt)
ul97.5<-vector("list",length(ts.hus))
for(i in 1:length(ts.hus)){
  ul97.5[[i]]<-ul[3*i-2]
```

```
}
ul50<-vector("list",length(ts.hus))
for(i in 1:length(ts.hus)){
  ul50[[i]]<-ul[3*i-1]
}
ul25<-vector("list",length(ts.hus))
for(i in 1:length(ts.hus)){
  ul25[[i]]<-ul[3*i]}
##plot for quantiles
plot(unlist(ul97.5),type="l",col="blue",
xlab="Time (Day since 2011-04-01)",ylab="No. onset",
main="Bootstrap quantiles of onset counts")
lines(unlist(ul50),typ="l")
lines(unlist(ul25),typ="l",col="red")
legend(x="topright",c("97.5 quantile","median","2.5 quantile"),
lty=c(1,1,1),lwd=c(1,1,1),col = c(4,1,2))
```

8.2.2 Nonparametric bootstrap

```
library(boot)
i<-sample(123,replace=T)
#function
infected.fun<-function(data,i){
  d<-data[i,]c(mean(d))}
##begin bootstrap
infectedS1.boot<-boot(data=infectedS1, statistic=infected.fun, R=1000)
AS1<-apply(infectedS1.boot$t, 2, sd)
MeanS1<-mean(apply(infectedS1.boot$t, 1, mean))
CIHS1<-MeanS1+1.96*AS1/sqrt(T)
CILS1<-MeanS1-1.96*AS1/sqrt(T)
```

8.3 R code for Bayesian approach

```
install.packages("rjags")
library("rjags")
#Bayesian method
```

```
n.chains<-3
init <- lapply(1:n.chains,function(i) {
  list(.RNG.name="base::Mersenne-Twister",.RNG.seed=i*10)
  })
Yt<-as.vector(ts.hus)
m <- jags.model('C:/Users/Bei/Documents/project of master degree/
                R code/model13.bug',
data = list(y=Yt,n=length(Yt),d.pmf=incu.pmfnew), n.chains = n.chains,
n.adapt=1000)
list.samplers(m)
m$state()
samplesGa <- coda.samples(m, c('lambda'),n.iter=4000)
plot(samplesGa,density=FALSE)
gelman.plot(samplesGa,ylim=c(1,2))
Based on the above plot we decide to remove the first 3000 as burn-
in.

summary(window(samplesGa,start=3001))
SGa<-as.matrix(samplesGa)
median1Ga<-apply(SGa,2,median)[1:100]#cut after 100
quantile1Ga<-apply(SGa,2,quantile,0.025)[1:100]#cut after 100
quantileGa<-apply(SGa,2,quantile,0.975)[1:100]#cut after 100
plot(quantileGa,typ="s",col="blue",xlab="Time (Day since 2011-04-
01)",
main="Posterior lambda for gamma (0,01 0,01) as prior")#fig.13
lines(median1Ga,typ="s")
lines(quantile1Ga,typ="s",col="red")
legend(x="topright",c("97.5 quantile","median","2.5 quantile"),
lty=c(1,1,1),lwd=c(1,1,1),col = c(4,1,2))
mean1Ga<-apply(SGa,2,mean)[1:100]
averageGa<-mean(mean1Ga)
stGa<-sqrt(var(mean1Ga))
MSE.Ga.lambda1<-sum((mean1Ga-lambda1[1:100])^2)
MSE.Ga.lambda2<-sum((mean1Ga-lambda2[1:100])^2)

#model
##gamma(0.01,0.01) as prior
model{
for(i in 1:n){
  lambda[i] ~ dgamma(0.01,0.01)#Prior for lambda
```

```
  N[i] ~ dpois(lambda[i])}
for(i in 1:n){
  for(j in 1:i){d.pmf.f.N[i,j]<-d.pmf[i-j+1]*N[j]
  }
}
for(i in 1:n){
  cum.d.pmf.N[i]<-sum(d.pmf.f.N[i,1:i])
  mu[i]<-cum.d.pmf.N[i]
  y[i] ~ dpois(mu[i])}
}
#Model uniform (0,5000) as prior
model{
for(i in 1:n){
  lambda[i] ~ dunif(0,5000)#Prior for lambda
  N[i] ~ dpois(lambda[i])}
for(i in 1:n){
  for(j in 1:i){
    d.pmf.f.N[i,j]<-d.pmf[i-j+1]*N[j]
  }
}
for(i in 1:n){
  cum.d.pmf.N[i]<-sum(d.pmf.f.N[i,1:i])
  mu[i]<-cum.d.pmf.N[i]
  y[i] ~ dpois(mu[i])
}
}
#Model for Empirical Bayesian
model{for(i in 1:n){
  lambda[i] ~ dgamma(alpha,beta) #Prior for lambda
  N[i] ~ dpois(lambda[i])}
  for(i in 1:n){
    for(j in 1:i){
      d.pmf.f.N[i,j]<-d.pmf[i-j+1]*N[j]
    }
  }
  for(i in 1:n){
    cum.d.pmf.N[i]<-sum(d.pmf.f.N[i,1:i])
    mu[i]<-cum.d.pmf.N[i]
    y[i] ~ dpois(mu[i])
  }
```

```
}
#Model for penalized spline regression as prior
model{
  for(i in 1:n){
    y[i] ~ dpois(mu[i])
    mu[i]<-cum.d.pmf.N[i]
    cum.d.pmf.N[i]<-sum(d.pmf.f.N[i,1:i])
}
for(i in 1:n){
  for(j in 1:i){
    d.pmf.f.N[i,j]<-d.pmf[i-j+1]*N[j]
  }
}
for(i in 1:n){
  N[i] ~ dpois(lambda[i])
##penalized spline regression
log(lambda[i])<-mfe[i]+mre10[i]+mre20[i]
mfe[i]<-X[i,1]+betaS*X[i,2]
mre10[i]<-b[1]*Z[i,1]+b[2]*Z[i,2]+b[3]*Z[i,3]+b[4]*Z[i,4]+b[5]*Z[i,5]
         +b[6]*Z[i,6]+b[7]*Z[i,7]+b[8]*Z[i,8]+b[9]*Z[i,9]+b[10]*Z[i,10]
mre20[i]<-b[11]*Z[i,11]+b[12]*Z[i,12]+b[13]*Z[i,13]+b[14]*Z[i,14]+b[15]*
Z[i#,15]
         +b[16]*Z[i,16]+b[17]*Z[i,17]+b[18]*Z[i,18]+b[19]*Z[i,19]+b[20]*
Z[i,20]
}

betaS ~ dnorm(0,1.0E-6)
for (k in 1:20) {
  b[k]~dnorm(0,taub)
 }
taub ~ dgamma(1.0E-6,1.0E-6)
}


##data
Yt<-as.vector(ts.hus)
n<-length(Yt)
## Smoothing
num.knots<-20
covariate<-seq(1,n,by=1)
```

```
X<-cbind(rep(1,n),covariate)
knots<-quantile(unique(covariate),seq(0,1,length=(num.knots+2))[-
c(1,(num
.knots+2))])
Z_K<-(abs(outer(covariate,knots,"-")))^3
OMEGA_all<-(abs(outer(knots,knots,"-")))^3
svd.OMEGA_all<-svd(OMEGA_all)
## Bayesian  jags model
mP <- jags.model('C:/Users/Bei/Documents/project of master degree/
R code/model17.bug',
data = list(y=Yt,n=n,X=X,Z=Z,d.pmf=incu.pmfnew), n.chains = n.chains,
n.adapt=1000)
list.samplers(mP)
mP$state()
samplesmPlambda <- coda.samples(mP, c('lambda'),n.iter=5000)##n.iter=50000
plot(samplesmPlambda,density=FALSE)
gelman.plot(samplesmPlambda,ylim=c(1,2))
Based on the above plot we decided to remove the first 3000 as burn-
in.
summary(window(samplesmPlambda,start=3000))
SmPlambda<-as.matrix(samplesmPlambda)
median1mPlambda<-apply(SmPlambda,2,median)[1:100]#cut after 100
quantile1mPlambda<-apply(SmPlambda,2,quantile,0.025)[1:100]#cut af-
ter 100
quantilemPlambda<-apply(SmPlambda,2,quantile,0.975)[1:100]#cut af-
ter 100
plot(quantilemPlambda,typ="s",col="blue",
     xlab="Time (Day since 2011-04-01)",
main="Posterior lambda for prior with penalized spline regression" )
lines(median1mPlambda,typ="s")
lines(quantile1mPlambda,typ="s",col="red")
legend(x="topright",c("97.5 quantile","median","2.5 quantile"),
lty=c(1,1,1),lwd=c(1,1,1),col = c(4,1,2))
mean1mPlambda<-apply(SmPlambda,2,mean)[1:100]
averagemPlambda<-mean(mean1mPlambda)
stmPlambda<-sqrt(var(mean1mPlambda))
MSE.mP.lambda1<-sum((mean1mPlambda-lambda1[1:100])^2)
MSE.mP.lambda2<-sum((mean1mPlambda-lambda2[1:100])^2)
```