Hindawi Publishing Corporation
BioMed Research International
Article ID 751738

Research Article

A Bayesian Outbreak Detection Method for Influenza-Like Illness

Yury E. García, J. Andrés Christen, and Marcos A. Capistrán

Centro de Investigación en Matemáticas, A.C., Jalisco S/N, Colonia Valenciana, 36240 Guanajuato, GTO, Mexico

Correspondence should be addressed to Yury E. García; yury@cimat.mx

Received 27 November 2014; Revised 24 March 2015; Accepted 26 March 2015

Academic Editor: Farai Nyabadza

Epidemic outbreak detection is an important problem in public health and the development of reliable methods for outbreak detection remains an active research area. In this paper we introduce a Bayesian method to detect outbreaks of influenza-like illness from surveillance data. The rationale is that, during the early phase of the outbreak, surveillance data changes from autoregressive dynamics to a regime of exponential growth. Our method uses Bayesian model selection and Bayesian regression to identify the breakpoint. No free parameters need to be tuned. However, historical information regarding influenza-like illnesses needs to be incorporated into the model. In order to show and discuss the performance of our method we analyze synthetic, seasonal, and pandemic outbreak data.

1. Introduction

An important issue in public health is timely epidemic outbreak detection. Outbreak surveillance and monitoring are usually done by gathering official data reported by hospitals and clinics through medical consultation. One of the most frequent causes of medical consultation in all countries is influenza-like illness (ILI) or acute respiratory infection (ARI) [1–3]. ILI are responsible for substantial morbidity and mortality each year [3]. Seasonal influenza occurs throughout the world, and it is ranked as a leading cause of death for people below 4 and above 65 years of age and it is among the 10 top causes of death in almost all age groups [4, 5].

Early outbreak detection is necessary in order to take suitable control measures. Outbreaks correspond to breakpoints in surveillance data sets. Substantial research efforts have been devoted to this topic, inspired by a variety of statistical techniques including regression methods, time-series models, and statistical process control approaches and extensions to those fields that involve space-temporal studies and multivariate methods or techniques that include Bayesian inference [6, 7]. Comprehensive reviews of the field are presented by Unkel et al. [8], Sonesson and Bock [9], Brookmeyer and Stroup [10], and Watkins et al. [11]. Each of these papers presents a classification of methods used for outbreak detection. In general, outbreak methods use threshold values or threshold intervals to signal an alert, all based on historical data.

There are methods based on linear regression with model selection using criteria like AIC or BIC. However, outbreak detection is made under uncertainty, as noise is present in early signals of influenza surveillance [12]. Statistical methods that ignore this uncertainty may result in overconfident predictions. Bayesian methods provide a way to account for uncertainty in both data and model selection [13]. In this paper we introduce a Bayesian outbreak detection using regression models with model selection based on Bayes factors; see Hoeting et al. [13] for a review. Examples of Bayesian model comparison in linear models are [14, 15]. Smith and Spiegelhalter [16] present a review of selection criteria for linear models in terms of Bayes factors. Guo and Speckman [17] examine consistency of Bayes factors in the comparison problem for linear models. One key difference from most other methods is that the method introduced in this paper is not based on historical data alone, but rather on the exponential nature of an epidemic outbreak. For the purposes of this paper, prior information regarding influenza-like illnesses was used to build prior distributions which in turn are useful to estimate the Bayes factors for model selection.

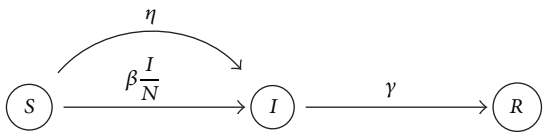

Figure 1: *SIR* epidemic model. Parameter β is the contact rate, γ is the recovery rate, and η accounts for infections due to imported cases. No births or deaths are taken into account given the time frame of an epidemic outbreak (few months for ILI).

The paper is organized as follows. Section 2 describes the results that lead to the outbreak detection method proposed in this paper. Section 3 applies the proposed method to synthetic and real data sets. Section 3.3 discusses the feasibility of our approach. Finally, Section 4 summarizes our findings and offers some perspectives.

2. Materials and Methods

Let us consider the epidemic process outlined in Figure 1. Let $S(t)$, $I(t)$, and $R(t)$ denote the number of susceptible, infected, and recovered individuals at time t and the population size $N(t) = S(t)+I(t)+R(t)$. The deterministic *SIR* model, without imported infections, that is, $\eta = 0$, is defined through the following ODE system [18]:

$$\begin{aligned}\frac{dS(t)}{dt} &= -\beta\frac{I}{N}S,\\ \frac{dI(t)}{dt} &= \beta\frac{I}{N}S-\gamma I,\\ \frac{dR(t)}{dt} &= \gamma I.\end{aligned} \tag{1}$$

β is the per capita contact rate between susceptible and infected individuals and γ is the infection recovery rate. At the onset of an epidemic outbreak the number of infected individuals is small (relative to N); that is, $I(t_0) = I_0 \approx 0$ and $R(t_0) = 0$ at initial time t_0. Therefore $S(t_0) \approx N$ and

$$\frac{dI}{dt} = \beta\frac{I}{N}S-\gamma I \approx (\beta-\gamma)I \tag{2}$$

for $t \approx t_0$; consequently

$$I(t) \approx I_0\exp(\beta-\gamma t) = I_0\exp(\gamma(R_0-1)t). \tag{3}$$

Here $R_0 = \beta/\gamma$ is the basic reproductive number, which is defined as the expected number of secondary infections caused by an infectious individual in a totally susceptible population during the time the individual spends in the infectious compartment. An epidemic may occur if R_0 is greater than one, while a basic reproductive number smaller than one will not sustain an epidemic; see [18]. Of note, the basic reproductive number does not change if $\eta \neq 0$.

In the remainder of the paper we write $\Delta R_0 = R_0 - 1$. Thus $I(t) \approx I_0\exp(\gamma\Delta R_0 t)$ and therefore

$$\log(I(t)) \approx \log(I_0) + \gamma\Delta R_0 t. \tag{4}$$

That is, the logarithm of the number of infected individuals is linear in t during an epidemic outbreak.

On the other hand, outside epidemic outbreaks we expect that the number of infected reported cases varies around a background level, either around zero or an average number of reports as it is the case in influenza-like illnesses (ILI reports, examples to be analyzed in Section 3). By chance, the number of infected persons reports may vary around the average, with temporary runs going up (or down). In such a case we may fit a linear model in the original scale; namely,

$$I(t) \approx a + bt. \tag{5}$$

The basis for our approach is to compare models (4) and (5), with a short run of reports, using the machinery of Bayesian model selection (see Section 2.1). If the exponential (i.e., linear in log scale) model is selected, it will signal the possible start of an epidemic outbreak. It will be crucial to properly code in the prior distribution for ΔR_0 and b a clear separation between the two models, since for small values of ΔR_0 both models may be quite similar (since $e^x \approx 1+x$, for small x). We explain the model selection and prior selection in the following sections.

2.1. Bayesian Model Comparison. Given a data set of reported cases $I(t_i)$, $i = 1, 2, \ldots, k$ at times t_i, we consider a sliding window of n consecutive reports $I(t_i)$ to compare the statistical models defined by expressions (4) and (5). Before the outbreak, a linear model explains better the reported cases. On the other hand, during the early phase of the epidemic outbreak the number of infected individuals grows exponentially; thus the exponential model should be selected by the Bayes factors and the onset of the outbreak detected. Next we present an outline of Bayes factors and Bayesian model comparison and the basis for our approach.

Given two hypotheses H_1 and H_2 corresponding to the alternative models M_1 and M_2 for data D and parameters θ_1 and θ_2, the posterior distribution in each case is $f(\theta_j \mid D) = p(\theta \mid H_i)p(D \mid \theta, H_i)/p(D \mid H_i)$, $j = 1, 2$. Here $p(\theta_j \mid H_i)$ and $p(y \mid \theta_j, H_i)$ are the prior and likelihood for model i and

$$p(D \mid H_i) = \int p(\theta \mid H_i)\, p(y \mid \theta, H_i)\, d\theta \tag{6}$$

is the normalization constant in each case. The basis of Bayesian model selection is that we can calculate the posterior distribution that each model, or each hypothesis, H_i, is true. Namely, from Bayes's theorem we have

$$p(H_i \mid D) = \frac{p(D \mid H_i)\, p(H_i)}{p(D \mid H_1)\, p(H_1) + p(D \mid H_2)\, p(H_2)}, \tag{7}$$

where $p(H_i)$ is the prior probability assigned for model i. The Bayes factor ($B_{1,2}$) comparing these two models is given by the odds ratio of model M_1 versus model M_2; that is,

$$B_{1,2} = \frac{p(H_1 \mid D)}{p(H_2 \mid D)} = \frac{p(D \mid H_1)\, p(H_1)}{p(D \mid H_2)\, p(H_2)}. \tag{8}$$

Table 1: Model parameters summary of parameters used for both synthetic data generation and outbreak detection method.

Parameter	Value	Dimension	Description
η	100	Days	Infection importation rate
γ	7	Days	Infection recovery time
n	3	Reporting interval	Length of the window used to compare the models
p	2		Parameter index

Intuitively, the Bayes factor provides a measure of whether data D have increased or decreased the odds on H_1 versus H_2. Thus $B_{1,2} > 1$ signifies that H_1 (or M_1) is relatively more probable than H_2 (or M_2) given D [19]. The optimal decision is therefore to choose the model with the highest posterior probability, that is, model 1 if $B_{1,2} > 1$ and model 2 otherwise.

Note that Bayes factors do not make sense when using improper priors (due to unspecified constants) and are sensitive to vague or default a priori distributions; see [20]. However, in this paper we use strong and informative (and indeed proper) priors aimed at distinguishing both models. Therefore the mentioned issues, thoroughly discussed in the Bayesian literature, should be of no concern in the current setting.

Let us denote by M_1 the exponential model in (4) and M_2 the linear model given in (5). Let D be the data at hand, either $I(t_i)$ for model 1 or $\log I(t_i)$ for model 2, $i = 1, 2, \ldots, k$. Then we assume

$$D \sim N_n\left(X\theta, \sigma^2 I_n\right). \tag{9}$$

Thats is, $D \in \mathbb{R}^n$ follows a normal distribution with mean $X\theta$ and covariance matrix $\sigma^2 I_n$, where I_n is the identity matrix; $X \in \mathbb{R}^{n\times 2}$ and $\theta \in \mathbb{R}^2$ are the design matrix and the parameter vector, respectively. We will require a different design matrix X and prior distributions, for each model M_i.

To perform a standard conjugate Bayesian analysis on this linear model [19, 21, 22] we proceed as follows; please see Appendix A for more details. We use the Normal-Inverse Gamma (NIG) prior distribution:

$$\theta, \sigma^2 \sim \mathrm{NIG}\left(\theta_0, \Sigma_0, \alpha_0, \beta_0\right); \tag{10}$$

θ_0 corresponds to the location parameter, Σ_0 is the covariance matrix (for $\theta \mid \sigma^2 \sim N_2(\theta_0, \sigma^2\Sigma_0)$), and α_0 and β_0 denote the parameters of the Inverse-Gamma distribution (for $\sigma^2 \sim \mathrm{InvGa}(\alpha_0, \beta_0)$), in the usual way. The posterior distribution results in a $\mathrm{NIG}(\theta_n, \Sigma_n, \alpha_n, \beta_n)$, where

$$\theta_n = \left(\Sigma_0^{-1} + X^T X\right)^{-1}\left(\Sigma_0^{-1}\theta_0 + X^T D\right), \tag{11a}$$

$$\Sigma_n = \left(\Sigma_0^{-1} + X^T X\right)^{-1}, \tag{11b}$$

$$\alpha_n = \alpha_0 + \frac{n}{2}, \tag{11c}$$

$$\beta_n = \beta_0 + \frac{1}{2}\left[\theta_0^T\Sigma_0^{-1}\theta_0 + D^T D - \theta_n^T\Sigma_n^{-1}\theta_n\right]. \tag{11d}$$

The normalization constant in (6), required by the Bayes factor, is

$$\begin{aligned} p(D) &= \iint p\left(D \mid \theta, \sigma^2\right) p\left(\theta, \sigma^2\right) d\theta\, d\sigma^2 \\ &= \frac{(2\pi)^{-(n+p)/2}\,\Gamma(\alpha_n)}{(2\pi)^{-p/2}\,\Gamma(\alpha_0)} \frac{\beta_n^{-\alpha_n}\sqrt{|\Sigma_n|}}{\beta_0^{-\alpha_0}\sqrt{|\Sigma_0|}} \end{aligned} \tag{12}$$

(see Appendix A for more details).

From (4) and (5) it is clear that the design matrices X are

$$\begin{pmatrix} 1 & 0 \\ 1 & \gamma \\ 1 & 2\gamma \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \tag{13}$$

for the log-linear (exponential) and linear models, with θ^T equal to $(\log(I_0), \Delta R_0)$ and (a, b), respectively.

Other relevant parameters are explained and set in Table 1. In the following section we discuss and establish prior distributions for each model, setting the hyperparameters of the prior NIG distribution.

2.2. Prior Distributions. As mentioned in Section 2.1, it is crucial to separate both models through a prior distribution that distinguishes clearly the exponential growth from a linear fluctuation. The basic reproduction number R_0 plays a central role in the prior information. Here, prior information of our approach is set for influenza-like illnesses; other prior specifications could be attempted for another type of epidemic outbreaks. It is known that for seasonal influenza R_0 is approximately 1.5 [23]; therefore prior expectation for ΔR_0 will be centered at 0.5. Moreover, in calibrating our models we have found that the bigger the population size N the sharper the prior needed, where the prior variance should decrease as $1/N$. This rule is in agreement with standard hypothesis in physics; in a well mixed system the amplitude of fluctuations scales like the square root of the system size [24].

For each data window, we first subtract its corresponding mean, for either the logged or the original data, and center the prior linear model around 0. Consequently, the hyperparameters θ_0 and Σ_0 for the NIG prior are set to

$$\begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix}, \quad \begin{pmatrix} \log(10)^2 & 0 \\ 0 & \frac{1}{N} \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 10^2 & 0 \\ 0 & 2 \end{pmatrix}, \tag{14}$$

for the log-linear (exponential) and linear models, respectively. The outbreak detection method introduced here is

robust to other reasonable settings for these hyperparameters. The only critical value is the variance for ΔR_0, which, as mentioned above, needs to be adjusted with the population size as $1/N$.

The remaining hyperparameters are set to $\alpha_0 = (1/2)(n - p)$ and $\beta_0 = (1/2)(n-p)\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ is the observed variance in the data window, for either the logged or the original data. Thus, the prior variance is centered near the observed variance for each model.

Indeed, in a pure inference scenario it is questionable to use data driven prior distributions. However, in the current setting it is desired to distinguish between the linear and exponential models and not in fact the estimation of the regression parameters themselves, which are regarded as nuisance. By subtracting the mean and centering the prior of θ_1 (either to ΔR_0 or to b) to 0 and by setting a priori $E(\sigma^2) \approx \widehat{\sigma}^2$ we are helping the inference of the regression parameters in each case (and equally for both models). This is a key feature of the proposed approach, since we will use a small window of three consecutive reports, and uncentered priors would blur the relative weight of each model, rendering the model comparison useless. Overall, the prior distribution selection at this stage should be regarded as a pragmatic approach to making the outbreak detection procedure work.

Once the outbreak is detected we may then try to estimate R_0 using the data window at hand. Again, since the data set is very small, we will use a noninformative prior (see [19]) and use the marginal posterior for the regression parameters of the log-linear (exponential) model to estimate R_0. The corresponding marginal posterior for the whole $\theta = (\log(I_0), \Delta R_0)^T$ parameter is $\mathrm{St}_p(\widehat{\theta}, 0.5(X^TX)(n-2)\widehat{\beta}_n^{-1}, n-2)$, where $\widehat{\theta}_n = (X^TX)^{-1}X^TD$ and $\widehat{\beta}_n = 0.5(I - X\widehat{\theta}_n)^TD$ (indeed, D is the logged data). The marginal distribution of any one of the entries of θ is a univariate Student t distribution. We are interested in θ_2 (corresponding to ΔR_0); thus $\theta_2 \sim \mathrm{St}((\widehat{\theta}_n)_2, s^2(X^TX)_{22}, n - p)$. We will use the posterior expected value, $\widehat{\theta}_2 = (\widehat{\theta}_n)_2$, of this posterior marginal to estimate R_0; namely, $\widehat{R}_0 = \widehat{\theta}_2 + 1$. Also, since γ is fixed an estimator for β can be produced with $\widehat{\beta} = (\widehat{\theta}_2 + 1)\gamma$.

In Section 3 we compute B_{12} over a moving window of four consecutive data points, that is, $N = 4$, to decide whether changes are due to data oscillations (linear model is selected and $B_{12} < 1$), or the onset of exponential growth occurs (the exponential model is selected and $B_{12} > 1$) and an epidemic outbreak is expected.

3. Results

We have tested the predictive capacity of the outbreak detection method proposed in this paper with real and synthetic data sets. The real data sets used are from the Spanish influenza outbreak in San Francisco, USA, in 1918 (see [25]) and data of the acute respiratory illnesses (ARI) from San Luis Potosí, México (see Noyola and Arteaga-Domínguez [26]).

Outbreak information and model relevant parameters like the infection rate (β), the basic reproductive number (R_0), and the week of outbreak were estimated. In each figure, red dots indicate three consecutive points in which the exponential model is selected over the linear model; that is, $B_{12} > 1$. Grey points indicate one single four-point window in which $B_{12} > 1$. As explained in the previous section, once the outbreak is detected we use the log-linear model, with a noninformative prior, to produce estimators for both R_0 and β.

Table 2: Estimates obtained for the detected outbreak.

N	$\widehat{R}_0$	Week of outbreak	$\widehat{\beta}$
5000	1.23	2	0.17
10000	1.36	7	0.19
500000	1.91	8	0.27
1000000	1.35	14	0.19

3.1. Synthetic Data Analysis. To create synthetic data we have avoided committing an "Inverse Crime" [27]. Synthetic data was produced with a closely related but different model to the one assumed in (4) or (5) to be producing the infectious reports. Namely, we use the Gillespie algorithm to make a realization of the *SIR* epidemic model with demographic stochasticity [28]. Initially all individuals are susceptible and the epidemic outbreak is due to imported cases. The frequency of imported cases is controlled with parameter η; see Figure 1. Of note, the deterministic model (1) is the mean field equation of this stochastic *SIR* model. Moreover, in a real scenario data is accumulated over the reporting time frame (daily, weekly, etc., reports for infected persons). We then accumulate the simulated data over the reported time frame to produce the synthetic infectious reports $I(t_i)$. Also, a linear autoregressive process is added to the synthetic data to simulate a background of diseases caused by other agents, as it is the case of influenza-like illness. Simulations have $R_0 = 1.5$, $\gamma = 1/7$ (days); the rate of imported cases is $\eta \in [10^{-7}, 10^{-4}]$ depending on the population size N. Reports are accumulated weekly. Some examples are presented in Figure 2 and the estimates for R_0 and γ are presented in Table 2.

3.2. Real Data Analysis. Real surveillance data sets account for medical consultation cases. These numbers represent infected persons seeking medical attention at health centers. For influenza, it is estimated that as low as 17% of the infected population seek medical consultation and approximately 75% of people with seasonal or pandemic influenza do not exhibit symptoms [29]. However, under normal circumstances reports are proportional to the actual number of infected people and exponential growth in the number of infected people will be shown as such in the reported cases. In the following examples we do not explicitly model subreporting, obtaining good results in all cases.

The Spanish influenza of 1917-18 was a pandemic considered among the most devastating ones in history [30, 31]. Figure 3 shows a data set corresponding to San Francisco, USA, spanning from September 24th to November 24th.

Our detection method identifies an outbreak on October 10th. The estimated parameters associated with this epidemic are $\widehat{\beta} = 0.53$ and $\widehat{R}_0 = 3.7$. Both the estimated R_0 and

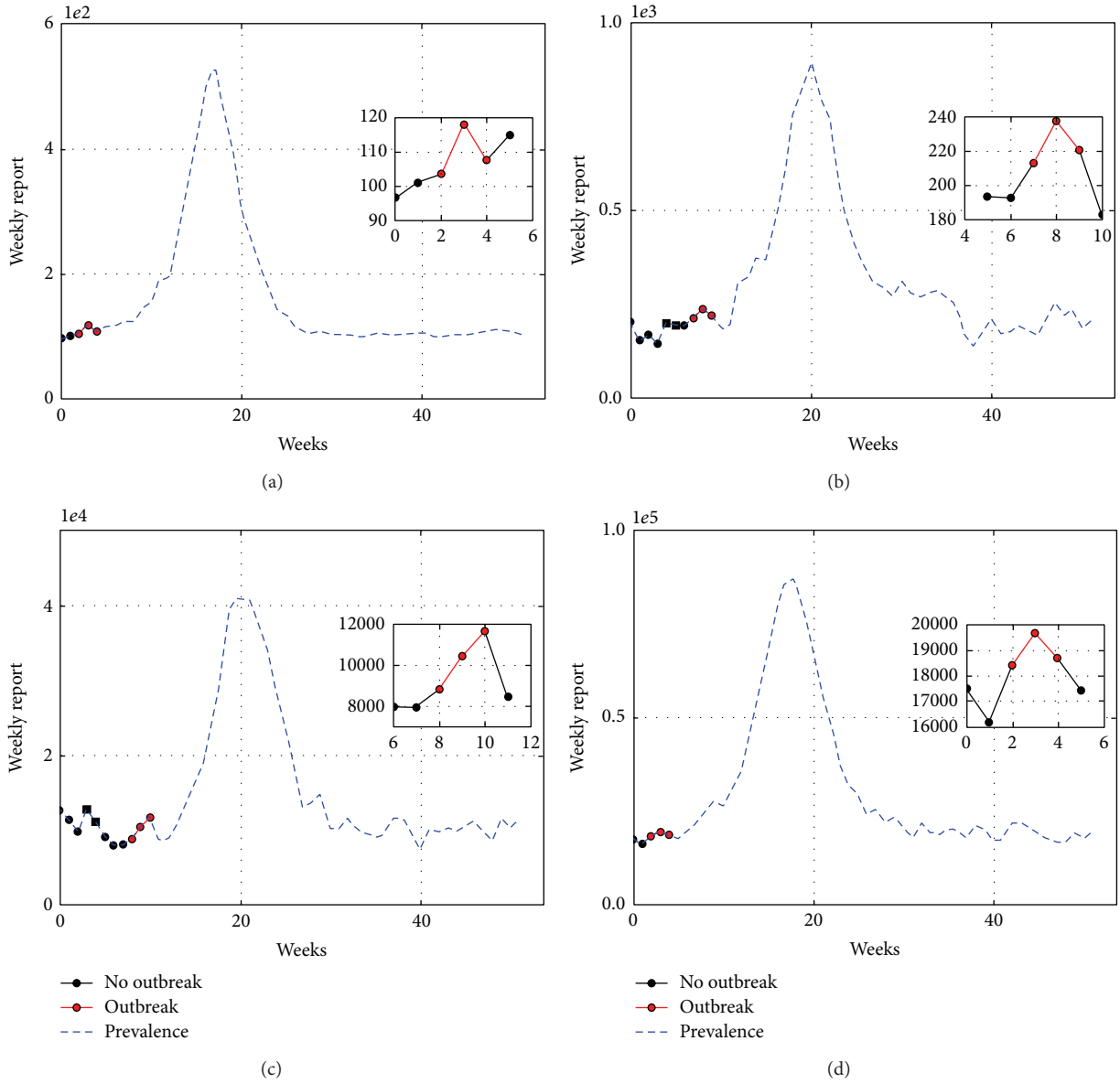

FIGURE 2: Outbreak detection for population sizes of (a) $N = 5{,}000$, (b) $N = 10{,}000$, (c) $N = 500{,}000$, and (d) $N = 1{,}000{,}000$. Data was generated with a realization of a *SIR* model with demographic stochasticity and imported cases. Outbreaks detection improves as the population size grows.

outbreak day are comparable with the values calculated by Chowell et al. [23].

Data of acute respiratory infections (ARI) in San Luis Potosí, México, are available in Noyola and Arteaga-Domínguez [26]. Here, we analyze ARI weekly reports from the winter seasons of 2000 to 2008. Reports refer to epidemiological weeks, for which week 1 is week 25 of the calendar year (i.e., mid June). Data for 2002-2003 and 2003-2004 winter seasons are plotted in Figure 4 along with outbreak detection results. In this series of data sets the seasonal outbreak is consistently detected between epidemiological weeks 13 and 15 with R_0 between 1.3 and 2.5; see Table 3.

Of note, other questions from ARI surveillance may be addressed; for instance, when do the weekly reports of ARI exceed the historical mean? However, in this paper we limit ourselves to the introduction of the detection method and leave other questions of disease surveillance for future research.

3.3. Discussion. We have introduced an outbreak detection method based on Bayesian linear regression and Bayes factors. Our method performs correctly in real and synthetic examples. Undoubtedly a key component of this method is the structure of the prior information used to distinguish the exponential from the linear model. In the above examples we have focused on influenza-like illness (ILI) or acute respiratory infection (ARI). Consequently, the prior expectation for R_0 was set equal to 1.5. We anticipate that other diseases may

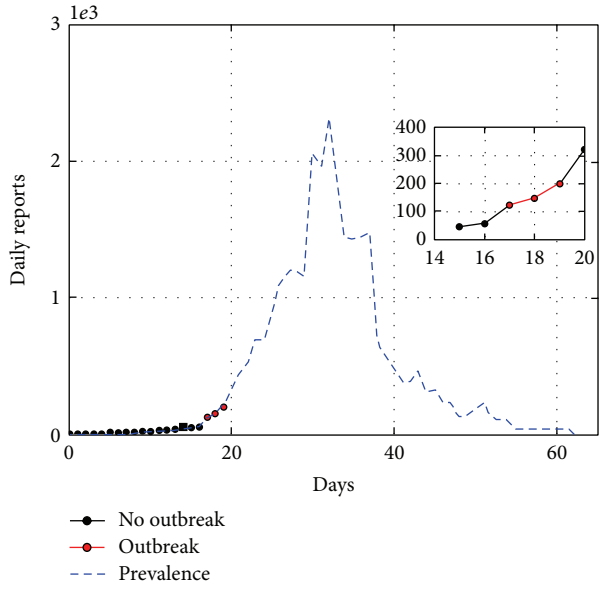

FIGURE 3: Spanish influenza in San Francisco, USA, 1918. Population 550000. Outbreak spanned from September 24th to November 24th. Method detected outbreak on the 17th day of the outbreak (October 10th). Estimated parameters are $\beta = 0.53$ and $R_0 = 3.7$.

TABLE 3: Parameters of acute respiratory infection records from San Luis Potosí 2000–2009. Population is approximately 2,000,000.

Year	$\widehat{R}_0$	Week of outbreak	$\widehat{\beta}$
2000-2001	1.57	8	0.22
2001-2002	1.29	7	0.18
2002-2003	1.34	7	0.19
2003-2004	1.37	8	0.19
2004-2005	1.59	8	0.23
2005-2006	1.32	8	0.19
2006-2007	1.42	8	0.20
2007-2008	2.5	11	0.36

be modeled correctly using previous reports of the expected value of the basic reproductive number. We have learned that the prior variance for ΔR_0 needs to reduce as $1/N$, where N is the population size. This choice may be justified recalling that in a well mixed physical system fluctuations scale like the square root of the system size.

In the examples presented above the outbreak is detected in the presence of underreporting. The good performance of the method is explained considering the fact that the method is based on detecting a qualitative feature of the surveillance data instead of a quantitative threshold. Methods based on historical thresholds may have difficulties in detecting an outbreak happening within or below average historical report levels. Of note, our method uses historical data to calibrate prior distributions; for example, historical data is used to model how much we allow surveillance data to oscillate while in the autoregressive regime. Moreover, the method introduced in this paper allows us to estimate important parameters like infection rate (β) and the basic reproductive number (R_0) which provide valuable information regarding outbreak behavior. The estimation of these quantities was made using a sliding window of three consecutive reports.

Bayesian outbreak detection was applied to two types of real data sets. It consistently succeeded in making an early detection and the estimated R_0 and β values were in agreement with values reported in the literature.

A Python-Scipy implementation of our approach may be downloaded from http://www.cimat.mx/jac/software; a user friendly interphase is available at request from the authors.

4. Conclusions

Outbreak detection is an important problem in surveillance of infectious diseases. The development of robust methods of early outbreak detection remains an active research area.

In this paper we use Bayes factors to detect a breakpoint that characterizes the onset of an epidemic outbreak in influenza-like illness surveillance data. The breakpoint characterizes the change from an autoregressive regime to exponential behavior of reported cases at the beginning of an epidemic outbreak. The detection method was successfully used on synthetic and real data sets. The resulting algorithm is straightforwardly implemented. The mathematical methods behind the algorithm are simple but contrast with other proposed methods which are based on calculating thresholds and control charts. Of note, our approach has no free parameters to tune.

The prior distributions used arise from coding information available for influenza-like illness. It is apparent that the method may be applied to surveillance data of other infectious diseases, for example, acute diarrheal diseases, provided enough prior information about the disease of interest is available.

Certainly, it is important to detect outbreaks before they have fully developed, that is, when the number of cases is still low. Our outbreak detection method seems to be able to achieve an early detection of influenza-like illness outbreaks, when synthetic and real data are analyzed. Furthermore, it allows us to make quantitative estimations for important parameters regarding the epidemic. The estimated parameters in the data sets analyzed are in agreement with previously published values.

Some features like the optimal number of reports required to identify an outbreak, optimal number of consecutive Bayes factors required to call an outbreak, and so forth are left as subject of further research.

Appendix

A. Details on the Prior and Posterior Distributions and Obtaining the Normalizing Constants

Let us denote by M_1 the linear model $I(t) = a + bt$, modeling the background data, and M_2 the exponential model given by $\log(I(t)) = \log(I_0) + \gamma \Delta R_0 t$, modeling the early outbreak. Let

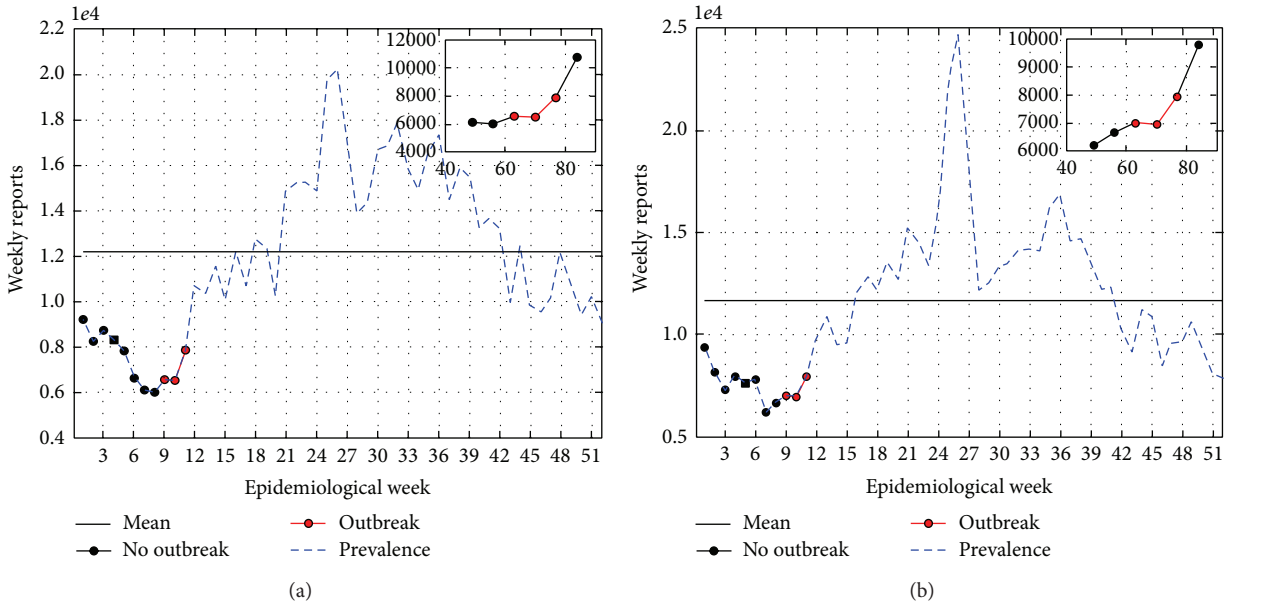

FIGURE 4: ARI reports from SLP, Mexico, winter seasons of (a) 2000-2001, $\widehat{\beta} = 0.22$, $\widehat{R}_0 = 1.57$, outbreak detected at epidemiological week 8, and (b) 2003-2004, $\widehat{\beta} = 0.19$, $\widehat{R}_0 = 1.37$, outbreak detected at epidemiological week 8.

D be the data, either $I(t_i)$ for model 1 or $\log(I(t_i))$ for model 2. Then, we assume

$$D \sim N_n\left(X\theta, \sigma^2 I_n\right); \tag{A.1}$$

that is, $D \in \mathbf{R}^n$, follows a normal distribution with mean $X\theta$ and covariance matrix $\sigma^2 I_n$, where I_n is the identity matrix, and $X \in \mathbf{R}^n$ and $\theta \in \mathbf{R}^2$ are the design matrix and the parameter vector, respectively. The following details may also be found in [22].

A.1. The NIG Prior. To perform a standard conjugate Bayesian analysis on this linear model, we use the Normal-Inverse Gamma (NIG) prior distribution as follows:

$$\theta, \sigma^2 \sim \mathrm{NIG}\left(\theta_0, \Sigma_0, \alpha_0, \beta_0\right). \tag{A.2}$$

This two-dimensional NIG distribution signifies that

$$\theta \mid \sigma^2 \sim N_2\left(\theta_0, \sigma^2\Sigma_0\right), \tag{A.3}$$

where θ_0 correspond to the a priori location parameter and Σ_0 the a priori covariance matrix for θ and α_0 and β_0 denote the hyperparameters for the a priori Inverse-Gamma distribution for σ^2; consider

$$\sigma^2 \sim \mathrm{IG}\left(\alpha_0, \beta_0\right). \tag{A.4}$$

The functional form of this prior distribution is given by

$$\begin{aligned} p\left(\theta, \sigma^2\right) &= p\left(\theta \mid \sigma^2\right) p\left(\sigma^2\right) = N_2\left(\theta_0, \sigma^2\Sigma_0\right) \times \mathrm{IG}\left(\alpha_0, \beta_0\right) \\ &= \frac{\beta_0^{\alpha_0}}{(2\pi)^{p/2}\left|\Sigma_0\right|^{1/2}\Gamma\left(\alpha_0\right)}\left(\frac{1}{\sigma^2}\right)^{\alpha_0+p/2+1} \\ &\quad \times \exp\left[-\frac{1}{\sigma^2}\left\{\beta_0 + \frac{1}{2}\left(\theta-\theta_0\right)^T\Sigma_0^{-1}\left(\theta-\theta_0\right)\right\}\right] \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\alpha_0+p/2+1} \\ &\quad \times \exp\left[-\frac{1}{\sigma^2}\left\{\beta_0 + \frac{1}{2}\left(\theta-\theta_0\right)^T\Sigma_0^{-1}\left(\theta-\theta_0\right)\right\}\right], \end{aligned} \tag{A.5}$$

where $\Gamma(\cdot)$ represents the Gamma function and the $\mathrm{IG}(\alpha_0, \beta_0)$ prior density for σ^2 is given by

$$\begin{aligned} p\left(\sigma^2\right) &= \frac{\beta_0^{\alpha_0}}{\Gamma\left(\alpha_0\right)}\left(\frac{1}{\sigma^2}\right)^{\alpha_0+1}\exp\left(-\frac{\beta_0}{\sigma^2}\right), \\ &\sigma^2 > 0, \quad \beta_0 > 0, \quad \alpha_0 > 0. \end{aligned} \tag{A.6}$$

A.2. The Likelihood. The likelihood function for each model is defined as the joint probability of observing the data viewed as a function of the parameters; consequently

$$\begin{aligned} &P\left(D \mid \theta, \sigma^2\right) \\ &\quad = N\left(X\theta, \sigma^2 I_n\right) \\ &\quad = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \times \exp\left\{-\frac{1}{2\sigma^2}\left(D - X\theta\right)^T\left(D - X\theta\right)\right\} \end{aligned} \tag{A.7}$$

viewed as a function of θ and σ^2 and fixing D.

A.3. The Posterior NIG Distribution. The posterior distribution is defined as $p(\theta, \sigma^2 \mid D) = p(\theta, \sigma^2)p(D \mid \theta, \sigma^2)/p(D)$, where $p(D) = \int p(\theta, \sigma^2)p(D \mid \theta, \sigma^2)d\theta\, d\sigma^2$ is the marginal distribution of the data.

We have that

$$
\begin{aligned}
&p\left(\theta,\sigma^2 \mid D\right)\\
&\quad=\frac{\mathrm{NIG}\left(\theta_0,\Sigma_0,\alpha_0,\beta_0\right)\times N\left(X\theta,\sigma^2 I_n\right)}{p\left(D\right)}\\
&\quad\propto\frac{\beta_0^{\alpha_0}}{(2\pi)^{p/2}\left|\Sigma_0\right|^{1/2}\Gamma\left(\alpha_0\right)}\left(\frac{1}{\sigma^2}\right)^{\alpha_0+p/2+1}\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\\
&\qquad\times\exp\left\{-\frac{1}{\sigma^2}\left[\beta_0+\frac{1}{2}\left(\theta-\theta_0\right)^T\Sigma_0^{-1}\left(\theta-\theta_0\right)\right]\right\}\\
&\qquad-\frac{1}{2\sigma^2}\left\{\left(D-X\theta\right)^T\left(D-X\theta\right)\right\}\\
&\quad\propto\left(\frac{1}{\sigma^2}\right)^{\alpha_0+(p+n)/2+1}\\
&\qquad\cdot\exp\left\{\frac{-1}{\sigma^2}\left[\beta_0+\frac{1}{2}\left(\theta-\theta_0\right)^T\Sigma_0^{-1}\left(\theta-\theta_0\right)\right.\right.\\
&\qquad\qquad\left.\left.+\left(D-X\theta\right)^T\left(D-X\theta\right)\right]\right\}.
\end{aligned}
\tag{A.8}
$$

Using the identity

$$
u^TAu-2\alpha^Tu=\left(u-A^{-1}\alpha\right)^TA\left(u-A^{-1}\alpha\right)-\alpha^TA^{-1}\alpha
\tag{A.9}
$$

we may write

$$
\begin{aligned}
&\frac{1}{\sigma^2}\left[\beta_0+\frac{1}{2}\left(\theta-\theta_0\right)^T\Sigma_0^{-1}\left(\theta-\theta_0\right)+\left(D-X\theta\right)^T\left(D-X\theta\right)\right]\\
&\quad=\frac{1}{\sigma^2}\left[\beta_n+\frac{1}{2}\left(\theta-\theta_n^*\right)^T\Sigma_n^{-1}\left(\theta-\theta_n^*\right)\right],
\end{aligned}
\tag{A.10}
$$

where

$$
\begin{aligned}
\theta_n&=\left(\Sigma_0^{-1}+X^TX\right)^{-1}\left(\Sigma_0^{-1}\theta_0+X^TD\right),\\
\Sigma_n&=\left(\Sigma_0^{-1}+X^TX\right)^{-1},\\
\alpha_n&=\alpha_0+\frac{n}{2},\\
\beta_n&=\beta_0+\frac{1}{2}\left[\theta_0^T\Sigma_0^{-1}\theta_0+D^TD-\theta_n^T\Sigma_n^{-1}\theta_n\right].
\end{aligned}
\tag{A.11}
$$

Therefore,

$$
\begin{aligned}
&p\left(\theta,\sigma^2 \mid D\right)\\
&\quad=\left(\frac{1}{\sigma^2}\right)^{\alpha_0+(n+p)/2+1}\\
&\qquad\times\exp\left\{-\frac{1}{\sigma^2}\left[\beta_n+\frac{1}{2}\left(\theta-\theta_n\right)^T\Sigma_n^{-1}\left(\theta-\theta_n\right)\right]\right\},\\
&p\left(\theta,\sigma^2 \mid D\right)\propto\mathrm{NIG}\left(\theta_n,\Sigma_n,\alpha_n,\beta_n\right).
\end{aligned}
\tag{A.12}
$$

A.4. The Normalization Constant. This is the constant required by the Bayes factor. We need to compute the distribution $p(D \mid \sigma^2)$ by integrating out β and subsequently integrate out σ^2 to obtain $p(D)$. Accordingly,

$$
\begin{aligned}
&p\left(D \mid \sigma^2\right)\\
&\quad=\int p\left(D \mid \theta,\sigma^2\right)p\left(\theta \mid \sigma^2\right)d\theta\\
&\quad=\int N\left(X\theta,\sigma^2 I_n\right)\times N\left(\beta_0,\sigma^2\Sigma_0\right)d\theta\\
&\quad=\frac{1}{\left(2\pi\sigma^2\right)^{(n+p)/2}\left|\Sigma_0\right|^{1/2}}\\
&\qquad\cdot\int\exp\left[-\frac{1}{2\sigma^2}\left\{\left(D-X\theta\right)^T\left(D-X\theta\right)\right.\right.\\
&\qquad\qquad\left.\left.+\left(\theta-\theta_0\right)^T\Sigma_0^{-1}\left(\theta-\theta_0\right)\right\}\right]d\theta\\
&\quad=\frac{1}{\left(2\pi\sigma^2\right)^{(n+p)/2}\left|\Sigma_0\right|^{1/2}}\\
&\qquad\times\int\exp\left[-\frac{1}{2\sigma^2}\left\{\left(D-X\theta_0\right)^T\left(I+X\Sigma_0X^T\right)^{-1}\left(D-X\theta_0\right)\right\}\right.\\
&\qquad\qquad\left.+\left(\theta-\theta_n\right)^T\Sigma_n^{-1}\left(\theta-\theta_n\right)\right]d\theta\\
&\quad=\frac{1}{\left(2\pi\sigma^2\right)^{(n+p)/2}\left|\Sigma_0\right|^{1/2}}\\
&\qquad\cdot\exp\left\{-\frac{1}{2\sigma^2}\left(D-X\theta_0\right)^T\left(I+X\Sigma_0X^T\right)^{-1}\left(D-X\theta_0\right)\right\}\\
&\qquad\times\int\exp\left[-\frac{1}{2\sigma^2}\left\{\left(\theta-\theta_n\right)^T\Sigma_n^{-1}\left(\theta-\theta_n\right)\right\}\right]d\theta\\
&\quad=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\left(\frac{\left|\Sigma_n\right|}{\left|\Sigma_0\right|}\right)^{1/2}\\
&\qquad\cdot\exp\left\{-\frac{1}{2\sigma^2}\left(D-X\theta_0\right)^T\left(I+X\Sigma_0X^T\right)^{-1}\left(D-X\theta_0\right)\right\}\\
&\quad=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}\left|I+X\Sigma_0X^T\right|^{1/2}}\\
&\qquad\cdot\exp\left\{-\frac{1}{2\sigma^2}\left(D-X\theta_0\right)^T\left(I+X\Sigma_0X^T\right)^{-1}\left(D-X\theta_0\right)\right\}\\
&\quad=N\left(X\theta_0,\sigma^2\left(I+X\Sigma_0X^T\right)\right).
\end{aligned}
\tag{A.13}
$$

Here the matrix identity $|A+BDC| = |A||D||D^{-1}+CA^{-1}B|$ was applied to obtain

$$
\left|I_n+X\Sigma_0X^T\right|=\left|\Sigma_0\right|\left|\Sigma_0^{-1}+X^TX\right|=\left(\frac{\left|\Sigma_0\right|}{\left|\Sigma_n\right|}\right).
\tag{A.14}
$$

Now, the marginal distribution of $p(D)$ is obtained as follows:

$$\begin{aligned} p(D) &= \int p\left(D \mid \theta, \sigma^2\right) p\left(\theta, \sigma^2\right) d\theta\, d\sigma^2 \\ &= \int N\left(X\theta, \sigma^2 I_n\right) \times \mathrm{NIG}\left(\theta_0, \Sigma_0, \alpha_0, \beta_0\right) d\theta\, d\sigma^2 \\ &= \mathrm{MVSt}_{2\alpha_0}\left(X\theta, \frac{\beta_0}{\alpha_0}\left(I + X\Sigma_0 X^T\right)\right). \end{aligned} \tag{A.15}$$

In more detail, we have

$$\begin{aligned} & p(D) \\ &= \iint p\left(D \mid \theta, \sigma^2\right) p\left(\theta, \sigma^2\right) d\theta\, d\sigma^2 \\ &= \iint N\left(D \mid X\theta, \sigma^2 I_n\right) \times \mathrm{NIG}\left(\theta, \sigma^2 \mid \theta_0, \Sigma_0, \alpha_0, \beta_0\right) d\theta\, d\sigma^2 \\ &= \frac{\beta_0^{\alpha_0}}{(2\pi)^{p/2}\left|\Sigma_0\right|^{1/2}\Gamma(\alpha_0)} \\ &\quad \cdot \iint \left(\frac{1}{\sigma^2}\right)^{\alpha_n + p/2 + 1} \\ &\qquad \cdot \exp\left\{-\frac{1}{\sigma^2}\left[\beta_n + \frac{1}{2}\left(\theta - \theta_n\right)^t \Sigma_n^{-1}\left(\theta - \theta_n\right)\right]\right\} d\theta\, d\sigma^2 \\ &= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)(2\pi)^{(n+p)/2}\sqrt{|\Sigma_0|}} \frac{\Gamma(\alpha_n)(2\pi)^{p/2}\sqrt{|\Sigma_n|}}{\beta_n^{\alpha_n}} \\ &= \frac{(2\pi)^{-(n+p)/2}\Gamma(\alpha_n)}{(2\pi)^{-p/2}\Gamma(\alpha_0)} \frac{\beta_n^{-\alpha_n}\sqrt{|\Sigma_n|}}{\beta_0^{-\alpha_0}\sqrt{|\Sigma_0|}}. \end{aligned} \tag{A.16}$$

Thus, the posterior distribution is

$$\begin{aligned} p\left(\theta, \sigma^2 \mid D\right) &= \frac{p\left(\theta, \sigma^2\right) \times p\left(D \mid \beta, \sigma^2\right)}{p(D)} \\ &= \frac{\mathrm{NIG}\left(\theta_0, \Sigma_0, \alpha_0, \beta_0\right) \times N\left(X\theta, \sigma^2 I\right)}{\mathrm{MVSt}_{2\alpha_0}\left(X\theta, \left(\beta_0/\alpha_0\right)\left(I + X\Sigma_0 X^T\right)\right)}, \end{aligned} \tag{A.17}$$

which indeed reduces (after some algebraic manipulation) to the $\mathrm{NIG}(\theta_n, \Sigma_n, \alpha_n, \beta_n)$ density.

The marginal distribution of any one of the entries of θ_n is a univariate Student t distribution. This is used and the correct parameters are described in Section 2.2 to estimate R_0 and infection rate (β).

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] W. W. Thompson, L. Comanor, and D. K. Shay, "Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease," *Journal of Infectious Diseases*, vol. 194, supplement 2, pp. S82–S91, 2006.

[2] L. Simonsen, "The global impact of influenza on morbidity and mortality," *Vaccine*, vol. 17, no. 1, pp. S3–S10, 1999.

[3] H. Zhou, W. W. Thompson, C. G. Viboud et al., "Hospitalizations associated with influenza and respiratory syncytial virus in the United States, 1993–2008," *Clinical Infectious Diseases*, vol. 54, no. 10, pp. 1427–1436, 2012.

[4] S. L. Murphy, J. Xu, and K. D. Kochanek, "Deaths: final data for 2010," *National Vital Statistics Reports*, vol. 61, no. 4, pp. 1–117, 2013.

[5] Centers for Disease Control and Prevention, *Seasonal Influenza (Flu)*, Centers for Disease Control and Prevention, 2014, http://www.cdc.gov/flu/weekly/summary.htm.

[6] M. A. Martínez-Beneito, D. Conesa, A. López-Quílez, and A. López-Maside, "Bayesian Markov switching models for the early detection of influenza epidemics," *Statistics in Medicine*, vol. 27, no. 22, pp. 4455–4468, 2008.

[7] C. Pelat, P.-Y. Boëlle, B. J. Cowling et al., "Online detection and quantification of epidemics," *BMC Medical Informatics and Decision Making*, vol. 7, no. 1, article 29, 2007.

[8] S. Unkel, C. P. Farrington, P. H. Garthwaite, C. Robertson, and N. Andrews, "Statistical methods for the prospective detection of infectious disease outbreaks: a review," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 175, no. 1, pp. 49–82, 2012.

[9] C. Sonesson and D. Bock, "A review and discussion of prospective statistical surveillance in public health," *Journal of the Royal Statistical Society. Series A. Statistics in Society*, vol. 166, no. 1, pp. 5–21, 2003.

[10] R. Brookmeyer and D. F. Stroup, *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*, Oxford University Press, 2004.

[11] R. E. Watkins, S. Eagleson, R. G. Hall, L. Dailey, and A. J. Plant, "Approaches to the evaluation of outbreak detection methods," *BMC Public Health*, vol. 6, no. 1, article 263, 2006.

[12] G. F. Cooper, D. H. Dash, J. D. Levander, W.-K. Wong, W. R. Hogan, and M. M. Wagner, "Bayesian biosurveillance of disease outbreaks," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 94–103, AUAI Press, 2004.

[13] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–417, 1999.

[14] J. M. Dickey, "The weighted likelihood ratio, linear hypotheses on normal location parameters," *Annals of Mathematical Statistics*, vol. 42, pp. 204–223, 1971.

[15] D. J. Spiegelhalter and A. F. Smith, "Bayes factors for linear and Log-Linear models with vague prior information," *Journal of the Royal Statistical Society Series B (Methodological)*, vol. 44, no. 3, pp. 377–387, 1982.

[16] A. F. Smith and D. J. Spiegelhalter, "Bayes factors and choice criteria for linear models," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 42, no. 2, pp. 213–220, 1980.

[17] R. Guo and P. L. Speckman, "Bayes factor consistency in linear models," in *Proceedings of the International Workshop on Objective Bayes Methodology*, Valencia, Spain, June 2009.

[18] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Review*, vol. 42, no. 4, pp. 599–653, 2000.

[19] J. M. Bernardo and A. F. Smith, *Bayesian Theory*, vol. 405, John Wiley & Sons, 2009.

[20] J. O. Berger and L. R. Pericchi, "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 109–122, 1996.

[21] S. Kunz, "The bayesian linear model with unknown variance," Tech. Rep., Mimeo, New York, NY, USA, 2009.

[22] S. Banerjee, Bayesian linear model: Gory details 1 the nig conjugate prior family, 2014, http://www.biostat.umn.edu/ ph7440/pubh7440/BayesianLinearModelGoryDetails.pdf.

[23] G. Chowell, H. Nishiura, and L. M. A. Bettencourt, "Comparative estimation of the reproduction number for pandemic influenza from daily case notification data," *Journal of the Royal Society Interface*, vol. 4, no. 12, pp. 155–166, 2007.

[24] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry, Volume 1*, Elsevier, 1992.

[25] M. C. J. Bootsma and N. M. Ferguson, "The effect of public health measures on the 1918 influenza pandemic in U.S. cities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, pp. 7588–7593, 2007.

[26] D. E. Noyola and G. Arteaga-Domínguez, "Contribution of respiratory syncytial virus, influenza and parainfluenza viruses to acute respiratory infections in San Luis Potosí, Mexico," *Pediatric Infectious Disease Journal*, vol. 24, no. 12, pp. 1049–1052, 2005.

[27] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, vol. 160, Springer Science & Business Media, 2006.

[28] M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*, Princeton University Press, Princeton, NJ, USA, 2008.

[29] A. C. Hayward, E. B. Fragaszy, A. Bermingham et al., "Comparative community burden and severity of seasonal and pandemic influenza: results of the Flu Watch Cohort Study," *The Lancet Respiratory Medicine*, vol. 2, no. 6, pp. 445–454, 2014.

[30] J. K. Taubenberger and D. M. Morens, "1918 influenza: the mother of all pandemics," *Revista Biomédica*, vol. 17, pp. 69–79, 2006.

[31] A. W. Crosby, *America's Forgotten Pandemic: The Influenza of 1918*, Cambridge University Press, Cambridge, UK, 2003.