Detecting Influenza Epidemics Using Hidden Markov Models With Bayesian Approach

Seyed Yousef Mousavinasab

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Science at

Concordia University

Montreal, Quebec, Canada

August 2007

Your file Votre référence
ISBN: 978-0-494-34758-4
Our file Notre référence
ISBN: 978-0-494-34758-4

ABSTRACT

Detecting Influenza Epidemics Using Hidden Markov Models with Bayesian Approach

Seyed Yousef Mousavinasab

In this thesis, we present a statistical method for detecting influenza epidemics. First, we use a hidden Markov model with Bayesian approach to partition the influenza data into two groups, one group for the epidemic states and another one for the non-epidemic states. Then, we detect the start of the epidemic phase of the disease through introducing a warning threshold. This warning threshold is efficient in increasing the detection rates while decreasing the false alarm rates. Finally, we compare the established hidden Markov model with the traditional seasonal ARIMA model.

Acknowledgements

I would like to express my deepest gratitude to my thesis supervisors, Dr. W. Sun and Dr. Z. Khalil, for their guidance, encouragement and patience throughout the preparation for this thesis. I also thank them for their kind support during the period of my graduate study at Concordia University.

I am grateful to the faculty and staff of the Department of Mathematics and Statistics, especially Dr. Y.P. Chaubey, for their support and encouragement. Thanks also go to my colleagues and friends for helpful discussions.

Finally, thanks to my parents and my sister, Firuzeh, for everything they have done for me, in all manner of ways, during my studies.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

Infectious diseases continue to be a major cause of human suffering and death around the world. They are responsible for more than 10 million deaths every year, which represent one third of annual mortality for the entire planet. In particular, they are directly responsible for about 40% of deaths in the developing countries. Infectious diseases continually threat all persons, regardless of age, sex, lifestyle, ethnic background and socioeconomic status. They have imposed a very heavy financial burden on the society. Despite that a great deal of efforts has been made to prevent and control the infectious diseases, they continue to be a common and significant problem of modern medical science [1-4].

In order to obtain a better understanding of the transmission mechanisms of infectious diseases and to predict their spread, it is essential to apply statistical modeling. In recent years, many successful statistical models have been proposed to interpret and predict the dynamics of infectious diseases. They play an important role in helping make public health decisions on the control of infectious diseases. These models are adapted to a variety of different practical circumstances, depending on the disease involved, the community being studied and the kind of epidemiological or administrative questions being asked.

For successful modeling, it is crucial to understand the course of infection within an individual and the patterns of infection within communities of people. Aspects of the model that require careful choice include the number of population variables and equations needed for a sensible characterization of the system, the typical relationships among the various rate parameters, and the form of the mathematical expression that captures the essence of the transmission process. In this thesis, we present a new statistical method for detecting influenza epidemics, which uses hidden Markov model (HMM) with Bayesian approach. In the following two sections, we will give a brief introduction to HMM and Bayesian approach, respectively.

1.2 Hidden Markov Model (HMM)

Hidden Markov Model (HMM) is a powerful statistical tool for modeling random systems that can be characterized by an underlying process generating an observable sequence. The well-known Russian mathematician, Andrey Andreyevich Markov, gave his name to the mathematical theory of Markov processes in the early twentieth century; however, the theory of HMMs was systematically developed by Baum and his colleagues in the 1960s [5].

An HMM, as the name suggests, is a Markov model in which the states cannot be observed (i.e., hidden) but symbols that are consumed or produced by transition are observable. In recent years HMM and its simple extensions are widely used in a variety of fields including computer vision, natural language understanding, speech recognition and synthesis, Biostatistics (to name a few). Often HMMs are a usual way of modeling a system and in other cases they are "force-fit" to a problem to which they are not quite ideal. The vast popularity of HMMs is that very fast and linear time algorithms exist for some of the most important HMM problems [6].

In an HMM, the unobserved sequence of states follows a Markov chain, which determines the probabilities of the observed states. For instance, given a binary time series, each event might be generated by one of two Bernoulli distributions. The process switches from one distribution to another one according to the state of the hidden Markov chain, in this manner generating state dependence.

Let T be a realization of an integer-valued random variable with a complex distribution. An HMM is then defined as a set of T discrete scalar variables $Q_{1:T}$ and T other variables $X_{1:T}$. $Q_{1:T}$ and $X_{1:T}$ may be either discrete or continuous (and either scalar- or vector-valued). Further, for an HMM we have that $\{Q_{t:T}, X_{t:T}\}$ are independent of $\{Q_{1:t-2}, X_{1:t-1}\} | Q_{t-1}$ and X_t is independent of $\{Q_{(1:T)\setminus t}, X_{(1:T)\setminus t}\} | Q_t$, for all $t \in 1:T$. Let Θ be the *state space* with cardinality $|\Theta| < \infty$. Then each Q_t takes its values in the finite set Θ.

The above definition of HMM is very flexible in the sense that it does not limit the number of states in the Markov chain or the implementation of the dependencies (e.g.,

general regression, probability table, etc.). The only condition for a joint probability distribution (over an appropriately typed set of random variables) to form an HMM is that it follows the above-mentioned conditional independence rules.

From this definition, we can deduce the following properties for an HMM:

1) Since $\{Q_{t:T}, X_{t:T}\}$ are independent of $\{Q_{1:t-2}, X_{1:t-1}\} | Q_{t-1}$ and X_t is independent of $\{Q_{(1:T)\setminus t}, X_{(1:T)\setminus t}\} | Q_t$, the future is conditionally independent of the past given the present. In particular, Q_t is independent of $Q_{1:t-2} | Q_{t-1}$, which means that the variables $Q_{1:T}$ form a discrete-time, discrete-valued, first-order Markov chain.

2) The independency of $\{Q_{t:T}, X_{t:T}\}$ and $\{Q_{1:t-2}, X_{1:t-1}\} | Q_{t-1}$ implies that Q_t is independent of $\{Q_{1:t-2}, X_{1:t-1}\} | Q_{t-1}$. This means that for $1 \le t' < t \le T$, given Q_{t-1}, $X_{t'}$ is unable to affect Q_t. However, given Q_{t-1}, Q_t is not necessarily unaffected by the future variables.

3) The random variables $X_{1:T}$ form a general discrete time stochastic process. Since X_t is independent of $\{Q_{(1:T)\setminus t}, X_{(1:T)\setminus t}\} | Q_t$, assigning a value to Q_t, makes the distribution of X_t independent of all $X_{t'}$ for $t' \neq t$, i.e., all the variables in both the future and the past. One implication is that X_t *is independent of* $X_{t+1} | \{Q_t, Q_{t+1}\}$, which follows since X_t is independent of $\{X_{t+1}, Q_{t+1}\} | Q_t$ and X_t is independent of $X_{t+1} | Q_{t+1}$.

The above conditional independence properties imply that, for a given T, the joint distribution over all the variables may be expanded as follows:

$$p(x_{1:T}, q_{1:T}) = p(q_1) \prod_{t=2}^{T} p(q_t | q_{t-1}) \prod_{t=1}^{T} p(x_t | q_t). \quad (1.1)$$

From (1.1) we conclude that the following quantities are required to parameterize an HMM: the distribution over the initial chain variable $p(q_1)$, the conditional "transition" distributions for the first-order Markov chain $p(q_t | q_{t-1})$, and the conditional distribution for the other variables $p(x_t | q_t)$. These quantities correspond to the classical definition of HMM [6].

Let us label the initial distribution of Q as π, where π is a vector of length $|\Theta|$. Then, $p(Q_1 = i) = \pi_i$, where π_i is the i^{th} element of π. The observation probability distributions are denoted by $b_j(x) = p(X_t = x \mid Q_t = j)$ and the associated parameters depend on $b_j(x)$'s family of distributions. The Markov chain is typically assumed to be time-homogeneous, with stochastic matrix A, where $(A)_{ij} = p(Q_t = j \mid Q_{t-1} = i)$ for all t. We represent the collection of all HMM parameters as $\lambda \stackrel{\Delta}{=} (\pi, A, B)$, where B represents the parameters corresponding to all the observation distributions.

In most applications, the Markov chain $Q_{1:T}$ is hidden and the variables $X_{1:T}$ are observed. Therefore the Markov model is referred to as a hidden Markov model. Note that for some values of t, X_t might be also missing or hidden. Also note that in some cases $X_{1:T}$ is hidden and $Q_{1:T}$ is observed, which again leads to an HMM. Hereafter without loss of generality, we call $Q_{1:T}$ the hidden variables and $X_{1:T}$ the observations.

With the above definition, an HMM can be simultaneously viewed as a generator and a stochastic acceptor. Like any random variable, say "Y", one may obtain a sample from that random variable (e.g., flip a coin), or given a sample, say y, one may compute the probability of that sample $p(Y = y)$ (e.g., the probability of heads). To sample from an HMM, we first take a set of samples $q_{1:T}$ from the Markov chain $Q_{1:T}$. For this we first generate q_1, then proceed by generating q_2 using the conditional distribution of Q_2 given Q_1. After generating $q_{1:T}$ at each time point t we produce a sample of X_t using $p(X_t \mid q_t)$, the observation distribution according to the hidden variable value at time t. This is explained in [7] by an example where we first choose a sequence of urns and then a sequence of balls from each urn. To sample just from $X_{1:T}$, we follow the same procedure, i.e., although we do not need samples $q_{1:T}$ we have to generate them anyways.

Note that to generate each sample of $X_{1:T}$, we need a new and different sample of $Q_{1:T}$. Generating different samples of $X_{1:T}$ based on a fixed observation $q_{1:T}$ and using marginal distribution $p(X_{1:T} \mid q_{1:t})$ does not reflect the HMM properties. This is because

in HMM two different observation samples typically originate from two different state assignments to the hidden Markov chain. In other words, to reflect the HMM properties each sample $x_{1:T}$ must be generated according the marginal distribution $p(X_{1:T}) = \sum_{q_{1:T}} p(X_{1:T}, q_{1:T})$ and not from the conditional distribution $p(X_{1:T} \mid q_{1:t})$ for some fixed hidden variable assignment $q_{1:T}$.

Although a collection of values $x_{1:T}$ have presumably been produced according to some specific (but unknown) assignment to the hidden variables, a given $x_{1:T}$, could have been produced from one of many different realizations $q_{1:T}$. The probability $p(x_{1:T})$ can be computed by marginalizing away all possible assignments to $Q_{1:T}$ as follows:

$$p(x_{1:T}) = \sum_{q_{1:T}} p(x_{1:T}, q_{1:T})$$

$$= \sum_{q_{1:T}} p(q_1) \prod_{t=2}^{T} p(q_t \mid q_{t-1}) \prod_{t=1}^{T} p(x_t \mid q_t).$$

We will use two ways to graphically depict an HMM. The first way is to use a directed state-transition graph as in Figure 1.2A.

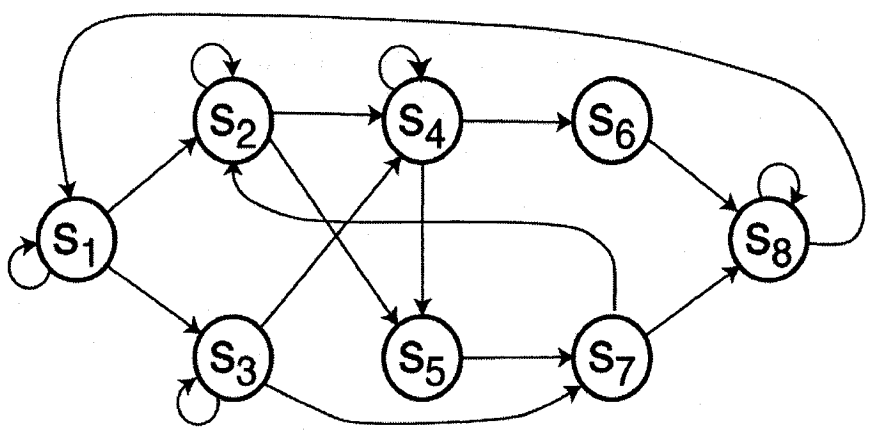

Figure 1.2A: Stochastic finite state automaton [9] view of an HMM. In this case, only the possible (i.e., non-zero probability) hidden Markov chain state transitions are shown [8].

In this illustration each node represents one of the states in Θ and an edge connects node i to node j if and only if $a_{ij} > 0$. A directed state-transition graph does not show HMM's output distributions or the conditional independence properties. Only the allowable transitions in the HMM's underlying Markov chain are illustrated by this graph. The transition matrix associated with Figure 1.2A is as follows [8]:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 & 0 & 0 & 0 \\ 0 & a_{22} & 0 & a_{24} & a_{25} & 0 & 0 & 0 \\ 0 & 0 & a_{33} & a_{34} & 0 & 0 & a_{37} & 0 \\ 0 & 0 & 0 & a_{44} & a_{45} & a_{46} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_{57} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{68} \\ 0 & a_{72} & 0 & 0 & 0 & 0 & 0 & a_{78} \\ a_{81} & 0 & 0 & 0 & 0 & 0 & 0 & a_{88} \end{pmatrix},$$

where the explicitly mentioned a_{ij}'s have non-zero values. Starting from a particular state j at a certain time, one can produce an observation sample from the distribution $b_j(x)$, which is the observation distribution corresponding to state j, and then advance to the next state according to the non-zero transitions.

The second way to graphically depict HMMs is illustrated in Figure 1.2B.

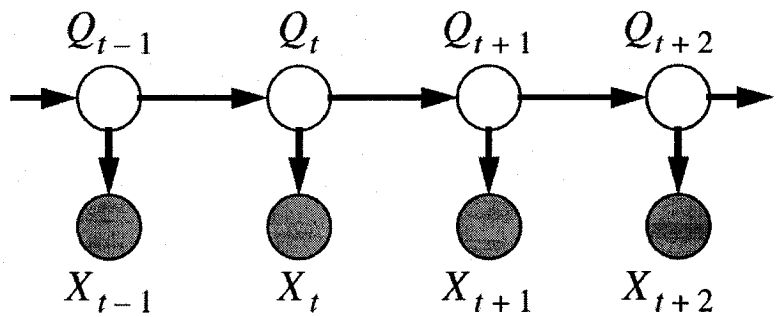

Figure 1.2B: A hidden Markov model

This method shows how HMMs are one instance of a directed graphical model (DGM), also called Bayesian network [10]. In this case, the HMM conditional independence

properties are shown but the hidden Markov-chain topology is unspecified. That is, using any of the equivalent schemas such as the directed local Markov property, the conditional independence properties implied by Figure 1.2B are identical to those expressed in the definition of HMM. For example, the variable X_t does not depend on any of X_t's non-descendants $(\{Q_{(1:T)\setminus t}, X_{(1:T)\setminus t}\})$ given X_t's parent Q_t. When one considers the HMM statistical dependencies, this directed graphical model is preferred to the directed state-transition model. On the other hand, when one analyzes the underlying hidden Markov chain topology, the directed state-transition model is more useful.

1.3 Bayesian Approach

Bayesian data analysis is based on Bayes' theorem (Bayes' rule). Let us first understand the Bayes' rule. Suppose we want to make inference about a random variable θ through observing another random variable, say "y". Let $p(\theta)$ be the distribution of θ. If θ and y are statistically independent, then $p(y,\theta) = p(y)p(\theta)$. However, if θ and y are dependent, then the dependence of y on θ is displayed by $p(y,\theta) = p(y \mid \theta)p(\theta)$. By the definition of conditional probability, $p(\theta \mid y) = \frac{p(y,\theta)}{p(y)}$. Then, we obtain Bayes' theorem:

$$p(\theta \mid y) = \frac{P(y \mid \theta)p(\theta)}{p(y)}.$$

$p(\theta \mid y)$ is called the *posterior distribution* of θ given the observed data y while $p(\theta)$ is called *the prior distribution.*

In the case of n possible outcomes $(\theta_1,\ldots,\theta_n)$, we have

$$p(\theta_j \mid y) = \frac{p(y \mid \theta_j)p(\theta_j)}{p(y)} = \frac{p(y \mid \theta_j)}{\sum_{i=1}^{n} p(\theta_i)p(y \mid \theta_i)}.$$

Set $\Theta = (\theta_1,\ldots,\theta_n)$. Then, the continuous multivariate form of Bayes' theorem is

$$p(\Theta \mid y) = \frac{p(y \mid \Theta)p(\Theta)}{p(y)} = \frac{p(y \mid \Theta)p(\Theta)}{\int p(y \mid \Theta)d\Theta}.$$

Similar to the univariate case, the prior distribution of the unknown parameters is assumed to be $p(\Theta)$, while $p(\Theta \mid y)$ is the posterior distribution given the prior $p(\Theta)$ and the data y.

The methods of maximum likelihood (ML) and Bayesian analysis are closely related. Let $l(\Theta \mid X)$ be the assumed likelihood function. According to the ML estimation, we compute the maximum of l (as a function of Θ given the data X) and would use the local curvature to construct confidence intervals. Hypothesis testing follows using the likelihood-ratio (LR) statistics. The strengths of ML estimation rely on its *large-sample* properties, namely that when the sample size is sufficiently large, we can assume both normality of the test statistics about its mean and that LR tests follow the χ^2 distributions. These good features do not necessarily hold for small samples.

An alternative way to proceed is to start with some initial knowledge/guess about the distribution of the unknown parameters, $p(\Theta)$. By Bayes' theorem, we have that

$$\begin{aligned} p(\Theta \mid X) &= \frac{1}{p(X)} \cdot p(X \mid \Theta) \cdot p(\Theta) \\ &= (\text{normalizing constant}) \cdot p(X \mid \Theta) \cdot p(\Theta) \\ &= \text{Constant} \cdot \text{Likelihood} \cdot \text{Prior}, \end{aligned}$$

where $p(X \mid \Theta) = l(\Theta \mid X)$ is just the likelihood function. Since our concern is to make inference about Θ, $\frac{1}{p(X)}$ is a constant with respect to Θ. Because of this, the posterior distribution is often written as $p(\Theta \mid X) \propto l(\Theta \mid X)p(\Theta)$, where the symbol $\propto$ means "proportional to". Note that the constant $p(X)$ normalizes $p(X \mid \Theta) \cdot p(\Theta)$ to one, and hence can be obtained by integration, $p(X) = \int_{\Theta} p(X \mid \Theta) \cdot p(\Theta)d\Theta$. By taking logs and ignoring the constant, the effect of the data can be seen:

$$\textit{log (posterior)} = \textit{log (likelihood)} + \textit{log (prior)}.$$

A critical feature of Bayesian analysis is the choice of the prior. The role of the prior depends on the amount of data that we have in use. The basic limit theorem of Bayesian probability theory tells us that the posterior probability converges to the actual distribution as more data arrives. This also means that the influence of the prior on the posterior distribution diminishes as more data arrives. The point here is that when the data contain significant information, even a bad prior does not greatly influence the posterior. If the posterior is highly dependent on the prior, then the likelihood function may not contain sufficient information. However, if the posterior is relatively stable over a choice of priors, then the data indeed contain significant information.

Often, only a subset of the unknown parameters is really of concern to us, the rest being nuisance parameters that are really not of concern to us. A very strong feature of Bayesian analysis is that we can remove the effect of the nuisance parameters by simply integrating them out of the posterior distribution to generate a marginal posterior distribution for the parameters of interest. We write the vector of unknown parameters as $\Theta = (\Theta_1, \Theta_n)$, where Θ_n is the vector of nuisance parameters. Integrating over Θ_n gives the desires marginal as

$$p(\Theta_1 \mid y) = \int_{\Theta_n} p(\Theta_1, \Theta_n \mid y) d\Theta_n .$$

In Bayesian data analysis the posterior probability distribution characterizes the uncertainty in the model estimated from a given set of data. One way to explore the posterior and hence characterize parameter uncertainty is to employ the Markov Chain Monte Carlo (MCMC) technique, which effectively generates a random sequence of model realizations. There are several sources for description of MCMC algorithms for HMMs. Readers are referred to [11] for an excellent overview. For further information on Bayesian statistical modeling, [12, 13 and 14] are suggested.

2 Analyzing Influenza Data Using HMMs with Bayesian Approach

In this chapter, we will use HMM with Bayesian approach to analyze the influenza data. The influenza virus causes one of the world's most serious breathing illnesses. It is potentially fatal and infects humans in tens of millions every year. The influenza virus changes or mutates every year. These changes cause seasonal epidemics of respiratory infections which are sometimes life threatening. Although many people regard the flu as a seasonal annoyance, influenza can lead to serious complications such as bacterial pneumonia and, for the elderly patients, influenza can be fatal [15].

Since early detection of influenza epidemics could have a serious impact on the number of lives saved, we place great emphasis on influenza surveillance by monitoring indicators of the spread of epidemics. In the following, we will analyze a monthly incidence rates data by applying HMM with Bayesian approach to estimate model parameters as well as introducing a warning threshold for detecting the start of the epidemic phase of the disease. We start with the data.

2.1 Influenza Data

The data set was obtained from the Sentinel Network of 500 general practitioners in France. Herein, we use the data collected between January 1991 and December 2005 (available at www.b3e.jussieu.fr/sentiweb/). Figure 2.1A depicts the time series of the monthly incidence rates data. It shows a clear mixture of two dynamics: the non-epidemic dynamic with incidence rates that vary according to a seasonal pattern (low-level dynamic) and the epidemic dynamic, in which the incidence rate increases sharply at irregular intervals.

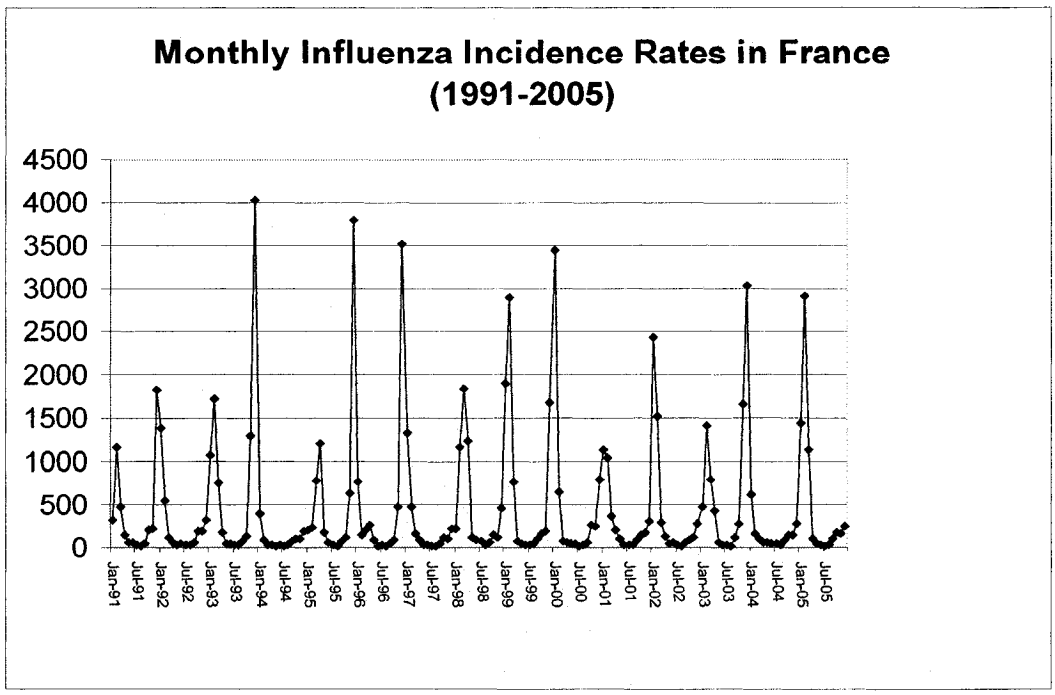

Figure 2.1A: Monthly influenza incident rates, France, 1991-2005

The current approach to influenza surveillance implemented by the Center for Disease Control (CDC) [15] is based on Serfling's method [16], which models the incidence rates due to influenza by the cyclic regression equation:

$$\mu_t = \gamma + \beta t + \delta \cos(\frac{2\pi t}{r}) + \varepsilon \sin(\frac{2\pi t}{r}) + e_t, \qquad (2.1)$$

where μ_t is the number of susceptible to influenza in month t. The parameter γ is the baseline monthly incident rates, without seasonal and secular trends, and e_t is noise, which is assumed to have mean zero and variance σ^2. The component βt describes secular trend, and the sine-wave component $\delta \cos(\frac{2\pi t}{r}) + \varepsilon \sin(\frac{2\pi t}{r})$ models period of influenza epidemics. Assuming that the errors are uncorrelated and normally distributed, the standard least squares method is used to estimate the model parameters $\gamma, \beta, \delta, \varepsilon, \sigma^2$ from non-epidemic data, and to compute confidence intervals about the predicted values. The predicted values are given by

$$\hat{\mu}_{j,t} = \hat{\gamma}_j + \hat{\beta}_j t + \hat{\delta}_j \cos(\frac{2\pi t}{r}) + \hat{\varepsilon}_j \sin(\frac{2\pi t}{r}). \qquad (2.2)$$

The confidence bounds are computed as $\hat{\mu}_t \pm t_{\alpha/2} SE(\hat{\mu}_t)$, where $SE(\hat{\mu}_t)$ is the estimated standard error of the prediction, and $t_{\alpha/2}$ is the $(1-\alpha/2)$ percentile of a Student's t distribution. The confidence bounds are used to define a time varying epidemic threshold that is adjusted for trend and seasonal effects [15].

Because of delays in detection due to the nature of the data, Serfling's method has been adapted to monitor either the proportion of influenza disease or hospital visit data for influenza [17, 18]. Although applied to data that should provide earlier indications of epidemics, the detection is still based on cyclic regression and suffers from two failures: the need for non-epidemic data to model the baseline distribution, and the fact that observations are treated as independent and identically distributed. The need for non-epidemic data is a fundamental obstacle toward the development of an automated surveillance system for influenza and it is overcome by the use of HMMs to segment the time series of influenza indicators into epidemic and non-epidemic phases. L. Start and F. Carrat [19] introduced this model.

We briefly review the following sort of HMMs, which is employed in our work.

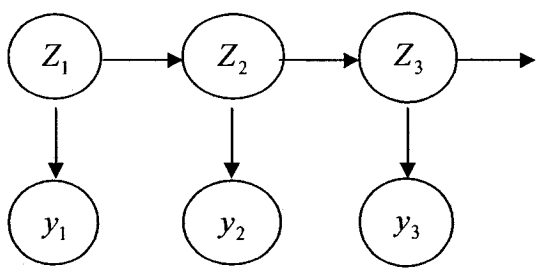

Figure 2.1 B: Standard hidden Markov model

Figure 2.1B depicts the graphical model for the standard HMM. Each pair of vertices represents a mixture model that is a pair $(z_t, y_t), t = 1,\ldots,n$, with $z_t \in \{1,\ldots,K\}$ and $y_t \mid z_t \sim f_{z_t}(y_t)$, where K is the cardinality of the state space. Generally the y_t's are observed and the sequence $\{z_t\}_{t=1}^{n}$ called the state sequence, is assumed to be unobserved

or hidden. For $t \in \{1,\ldots,n\}, y_t$ is conditionally independent of all the remaining variables conditional on z_t. That means:

$$P(y_t \mid y_1,\ldots,y_{t-1},z_1,\ldots,z_t) = P(y_t \mid z_t), \qquad \text{(I)}$$

$$P(z_t \mid y_1,\ldots,y_{t-1},z_1,\ldots,z_{t-1}) = P(z_t \mid z_{t-1}). \qquad \text{(II)}$$

Equation (II) follows a Markov chain of order 1 with transition probability matrix $P = (p_{ij})$, where $p_{ij} = P(z_t = j \mid z_{t-1} = i)$, $i, j = 1,\ldots,m$; $t = 2,\ldots n$.

We assign Gaussian distributions to the 2-state HMM $(m = 2)$, non-epidemic and epidemic rates, by using Serfling's cyclic regression method. The model is implemented by Bayesian approach using the software WinBugs. The main idea is to learn the model from the data; that is, to estimate the model parameters form the data. This can be achieved by considering prior distributions for parameters via WinBUGS, which is described in the following section.

2.2 HMM for Incidence Rates Series

As Figure 2.1A shows, the time series of influenza incidence rates often reveal trend and seasonality. We apply Serfling's method to model the data. This requires prior distributions [20] for the parameters which are introduced as follows for $j = 1,\ldots,K$ (focused on the case $K = 2$ in our work).

$$\gamma_j \sim dnorm(0.0, 1.0E-6),$$

$$\beta_j \sim dnorm(0.0, 1.0E-6),$$

$$\delta_j \sim dnorm(0.0, 1.0E-6),$$

$$\varepsilon_j \sim dnorm(0.0, 1.0E-6),$$

$$\sigma_j \sim dgamma(0.001, 0.001),$$

$$p_{j,1:K} \sim dbeta(1,1),$$

where "dnorm(μ,σ)" represents normal distribution with mean μ and variance σ^2, "dgamma(α,β)" represents gamma distribution with mean $\frac{\alpha}{\beta}$ and variance $\frac{\beta^2}{\alpha}$,

"dbeta(α,β)" represents beta distribution with mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2+(\alpha+\beta+1)}$ and $p_{j,1:K}$ denotes the vector of K transition probabilities from state $j, j=1,...,K$. The model has been implemented in the WinBUGS package.

"WinBUGS" (the MS Windows operating system version of BUGS: Bayesian Analysis Using Gibbs Sampling) is a versatile package that has been designed to carry out MCMC computations for a wide variety of Bayesian models [21]. The software is currently distributed electronically free-of-charge from the BUGS Project website: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml. WinBUGS assumes a Bayesian model in which all quantities are treated as random variables. The model consists of a defined joint distribution over all unobserved data (parameters and missing data) and observed quantities (the data); so we need to condition on the data in order to obtain posterior distribution over the parameters and unobserved data [21].

By running the WinBUGS program with 10000 iterations (the source code is available in Appendix), the parameters in Equation (2.1) are summarized in Table 2.1.

State	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\varepsilon}$	Transition probabilities	MSE	R-squared
Non-Epidemic	109.2	0.049	86.67	33.31	$\begin{pmatrix} 0.864 & 0.136 \\ 0.286 & 0.714 \end{pmatrix}$	2.65E3	0.728
Epidemic	478.3	1.224	897.4	323.3			

Table 2.1: Estimated parameters in cyclic regression and transition probabilities in HMM with Bayesian approach

Thus, Equation (2.2) for epidemic and non-epidemic states is given by

$$\mu_{ne,t} = 109.2 + 0.049t + 86.67\cos(\frac{\pi t}{6}) + 33.31\sin(\frac{\pi t}{6}),$$

$$\mu_{e,t} = 478.3 + 1.224t + 897.4\cos(\frac{\pi t}{6}) + 323.3\sin(\frac{\pi t}{6}).$$

Therefore, the model for the surveillance data is presented by

$$\mu_t = (1 - Z)\mu_{ne,t} + Z\mu_{e,t},$$

where $Z=0$ for non-epidemic state and $Z=1$ for epidemic state. This model follows a mixture Gaussian distribution, with its parameters depending on the epidemic and non-epidemic states.

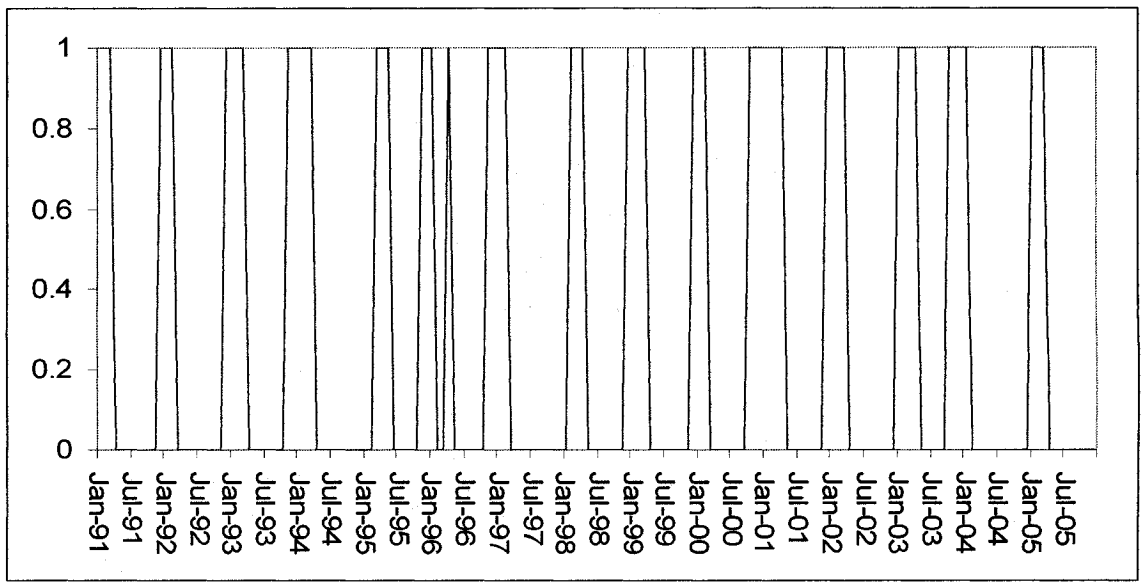

Figure 2.2A: Sequence of epidemic and non-epidemic states

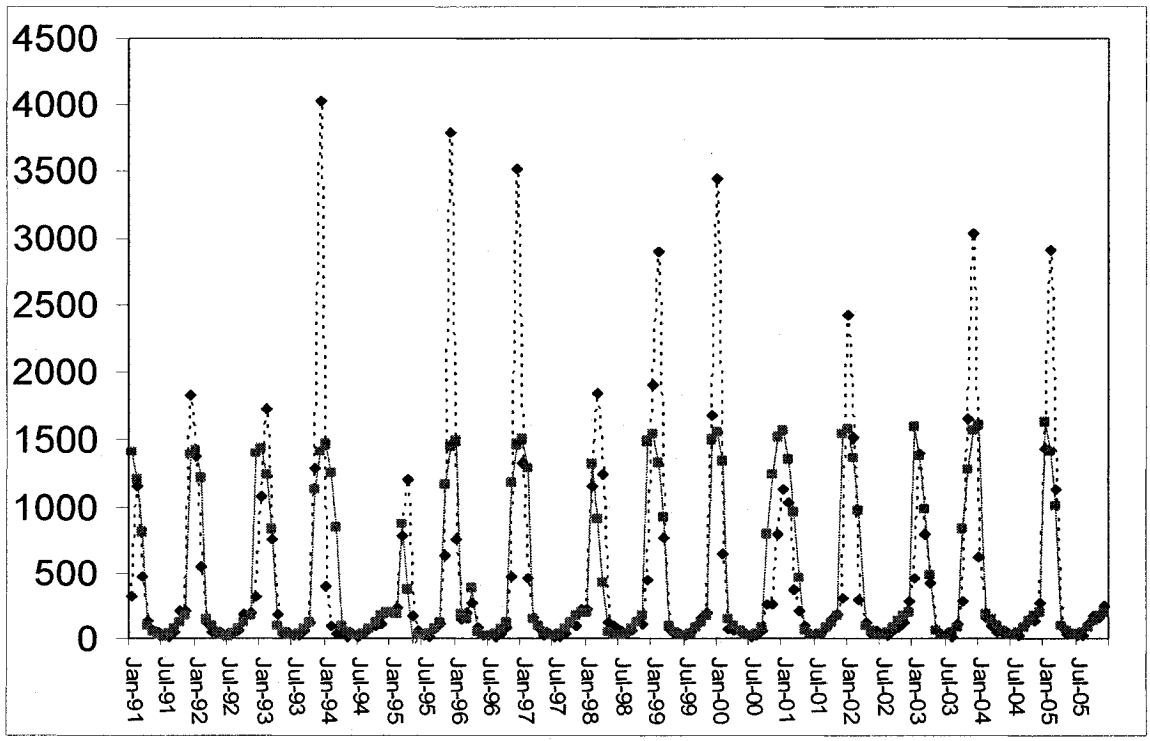

Figure 2.2B: Fitted model based on HMM with Bayesian approach

Through visual judgment, we can see that the model describes the data well. The curve fluctuates with observed data, which shows robustness of the model for detecting epidemic and non-epidemic phases as well as forecasting. More precise comparison of the established HMM with the traditional SARIMA model using statistical criterions will be given in Chapter 3.

The term epidemic refers to the uncommon sudden increase of the occurrence of a disease and is defined as the occurrence of the incidence that exceeds the expected one. This definition thus assumes a mixture of two phases, non-epidemic phase for the expected incidences and epidemic phase for the excess incidences, which is shown in Figure 2.2C.

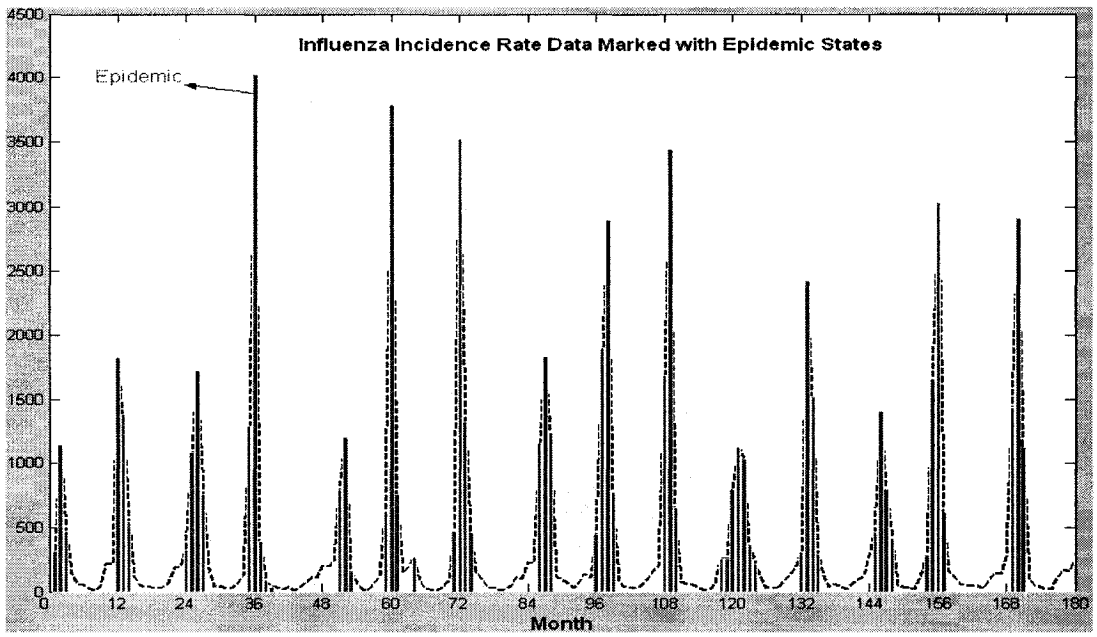

Figure 2.2C: Normal/normal model with Bayesian approach

2.3 Warning Threshold for Epidemic Detection

In this section, we will introduce a statistical method for detecting the starting point of the epidemic phase of the influenza disease, which can be applied to other infectious diseases where detection of the epidemic phase is concerned. To this end, we conduct the receiver operating characteristic (ROC) analysis, which is briefly introduced as follows.

The ROC curve analysis [22, 23] is an effective method for evaluating the performance and accuracy of diagnostic tests by considering decision thresholds. The ROC methodology is appropriate in situations with binary classifications, for example, epidemic/non-epidemic, diseased/normal cases, which has a wide range of applications in medicine. Graphically, an ROC curve is a plot of the relationship between sensitivity (y-axis) and its 1-specificity (x-axis) for the binary classifier system. Each point on the graph represents a possible diagnostic cutoff value (threshold) and the empirical ROC curve is formed by all possible decision thresholds. Let us consider a two-group diagnostic test for deciding whether a certain disease, in a particular area, is in epidemic

or non-epidemic phase. In practice, the distributions of the test results will overlap as shown in Figure 2.3A, which means that we seldom observe a complete separation between the two groups [24].

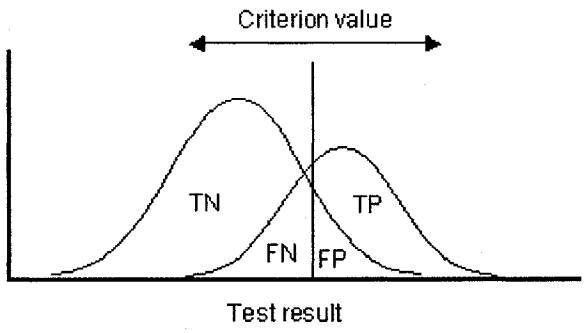

Figure 2.3A: Overlap of distributions

There are four possible outcomes from the binary classification (positive or negative epidemic) for every possible cut-off point or criterion value: 1) There will be some presence of epidemic cases correctly detected as positive which is called *true positive* (TP). 2) Some presence of epidemic cases will be detected as negative (non-epidemic), then it is called *false negative* (FN). 3) Some non-epidemic cases will be correctly detected as negative which is called *True Negative* (TN). 4) Some non-epidemic cases will be detected as positive (epidemic) which is named *false positive* (FP). The outcomes for the test are summarized in the following table:

Epidemic					
Test	**Present**	**N**	**Absent**	**n**	**Total**
Positive	True Positive	a	False Positive	c	$a + c$
Negative	False Negative	b	True Negative	d	$b + d$
Total		$a + b$		$c + d$	
Sensitivity	$= \frac{a}{a+b}$		**Specificity**	$= \frac{d}{c+d}$	

Table 2.2 Chart outcomes of a test

Herein *sensitivity* is the probability that a test result will be positive when the epidemic is present (true positive rate) and *specificity* is the probability that a test result will be negative when the epidemic is not present (true negative rate). Thus the ROC curve can be formed by plotting the fraction of *true positive* (sensitivity) versus fraction of *false positive* (1- specificity).

In Section 2.2 we have considered two Gaussian distributions with different parameters for uncovering epidemic and non-epidemic states of the data. In order to obtain the empirical distributions for the epidemic and non-epidemic states, we partitioned the data into two separate groups by using HMM and considering the trend and seasonality for the two Gaussian distributions. Bayesian approach has been used to estimate the parameters.

Usually there is an overlap between the two distributions for the epidemic and non-epidemic states. Hence, the warning threshold for detecting epidemic (WTDE) should be a point in the common region of the two distributions. In order to balance the sensitivity and specificity for detection, we conduct the ROC analysis (see Figure 2.3B), which has been introduced in the above. The introduced WTDE is a valuable indicator for public health managers to make decisions. If a health surveillance system finds that the disease incidences go beyond the WTDE, then it would be the time for taking special and more preventive actions in involved areas.

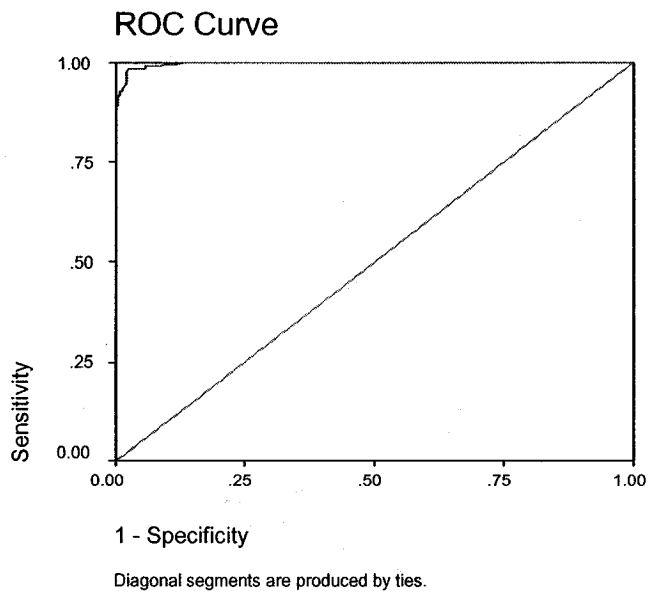

Figure 2.3B: Receiver operating characteristic (ROC)

After using an HMM to uncover the hidden states (epidemic and non-epidemic states), we are able to fit a proper empirical (Gaussian) distribution for each group. Since there is an overlap between the two distributions (see Figure 2.3C), the WTDE should be a point in the common region of the two distributions. We set the lower bound of the common region to be the lower bound of the confidence interval for the epidemic distribution, and set the upper bound of the common region to be the upper bound of the confidence interval for the non-epidemic distribution. Then, we search the best choice of the WTDE by conducting the ROC analysis. Finally, the WTDE is defined to be the point with the best-balanced sensitivity and specificity, see Table 2.3.

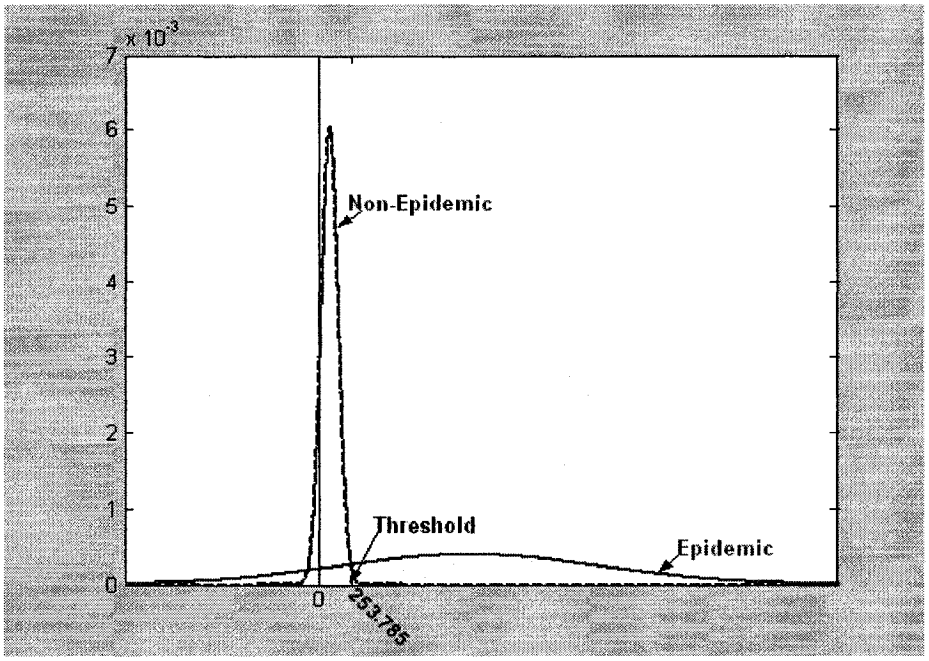

Figure 2.3C: Warning threshold for detecting epidemic phase of influenza disease

Cutoff point	Sensitivity	1-Specificity	Misclassification error 1[a]	Misclassification error 2[b]
253.785	0.931	0.16	0.055	0.179

Table 2.3: Cutoff point for warning epidemic states

All computations have been performed using WinBUGS and Matlab as well as SPSS for some statistical analyses.

To compare the agreement between HMM and WTDE for detecting epidemic and non-epidemic states, we consider the Kappa Statistics [25]. Kappa statistics is a common tool for comparing the agreement against that might be expected by chance. Kappa statistics can be thought of as the chance-corrected proportional agreement, and possible values range from positive-one (perfect agreement) to zero (no agreement above that expected by chance) to negative-one (complete disagreement), which is calculated as follows:

Kappa statistics= (Observed agreement - Chance agreement) / (1 - Chance agreement)

		HMM		*Total*
		Non-epidemic	*Epidemic*	
		Count (percent)	count (percent)	count (percent)
WTDE	*Non-epidemic*	**122 (67.8%)**	**1 (0.56%)**	**123 (68.36%)**
	Epidemic	**0 (0)**	**57 (31.66%)**	**57 (31.64%)**
Total		**122(67.8%)**	**58(32.22%)**	**180**

Table 2.4: Comparison of HMM and WTDE results for influenza data

Observed agreement = (122+57) / 180 = 0.983,
Chance agreement = 0.678 * 0.683 + 0.322 * 0.316 = 0.564,
Kappa statistics = 0.96.

For these counts, we find that the Kappa statistics is equal to 0.96, which indicates that there is almost perfect agreement between the result specified by the two detection methods (p-value < 0000...), where 58 and 57 months are detected as epidemic states by HMM and WTDE methods, respectively.

3 A Comparison of HMM with SARIMA Model

3.1 SARIMA Model for Incidence Rates Series

The Seasonal Auto Regressive Integrated Moving Average (SARIMA) model is a traditional model for incidence rates data, which is conventionally described using the Box-Jenkins notation system: $(p,d,q)(P,D,Q)_s$. The (p,d,q) term contains the parameters of the non-seasonal part of the model and the $(P,D,Q)_s$ term contains those of the seasonal component, where s is the length of the seasonal period (=12 in our case). The process for SARIMA modeling involves identification of the initial values of p,d and q using the auto-correlation/partial auto-correlation methods and estimation of the p (auto-regressive) and q (moving average) components to see if they contribute significantly to the model or one or the other should be dropped [26]. Readers may refer to [27] for details on SARIMA modeling.

Due to the nature of the data, the time series has significant seasonal variations (see Figure 3.1A). To identify the appropriate SARIMA model to the monthly influenza data, we start with making the series stationary by removing the gross features of seasonality, see Figures 3.1B and 3.1C.

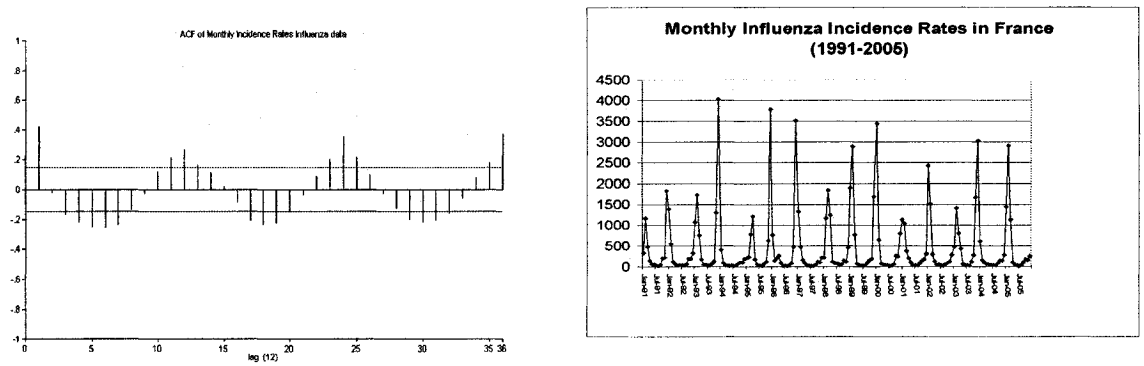

Figure 3.1A: Time series of influenza incidence rates and its ACF

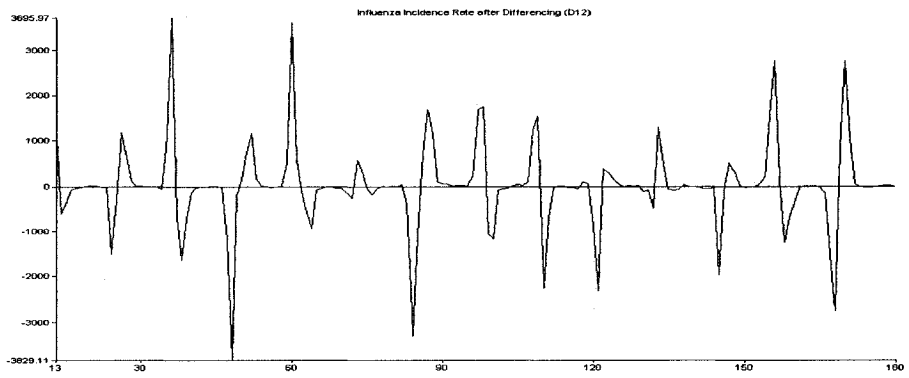

Figure 3.1B: Time Series of influenza data after differencing

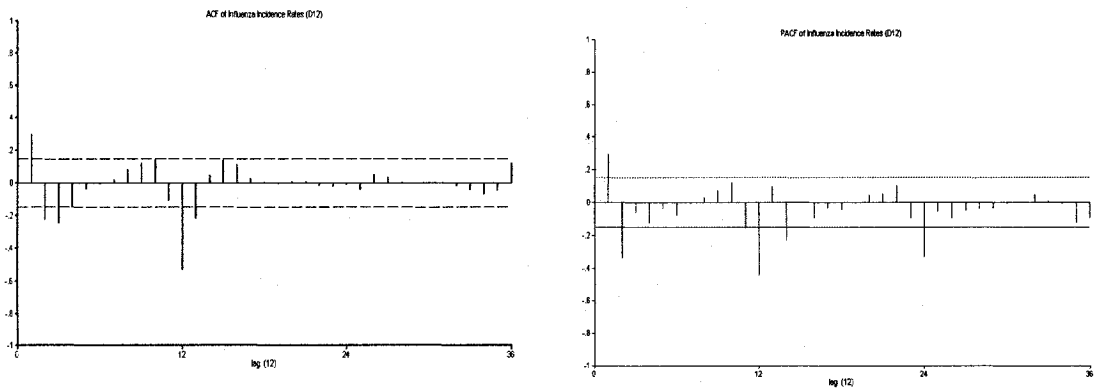

Figure 3.1C: ACF, PACF plots for time series of influenza disease

The seasonally differenced data is now stationary; however, the autocorrelation function (ACF) has the significant values at lag 12 and 13.

To determine the possible persistence structure in the monthly influenza data, the ACF and the partial ACF (PACF) are used. The closeness of fits is checked using the Portmanteau test (Q-statistics) [28]. The procedures is repeated until an appropriate model is found. A combination of Akaike information criterion (AIC) and Bayesian information criterion (BIC) [29] is used to select the best SARIMA model among the possible models that fulfill most of the diagnostic checks for the time series of influenza disease. The best fitted SARIMA model for the data is

ARIMA (1,0,1) $(1,0,1)_{12}$ *process with equation*:

$$(1+0.02B)\,(1-0.97B^{12})\,X_t = (1+0.45B)\,(1-1.04B^{12})\,W_t,$$

where B is the backward shift operator.

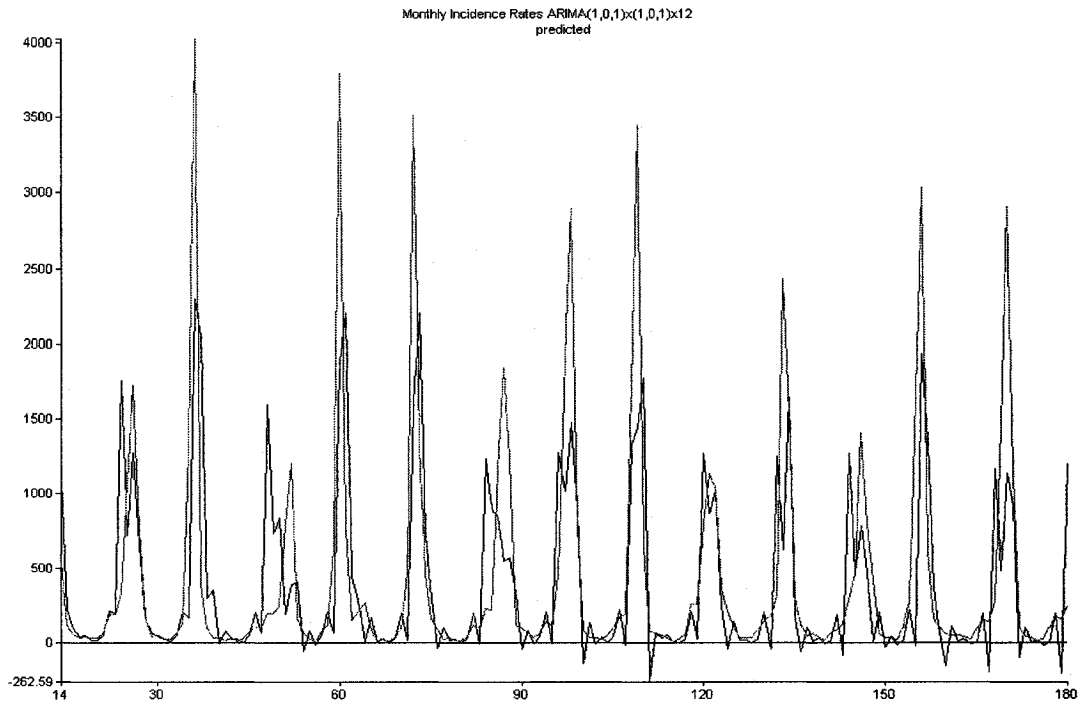

Figure 3.1D: Fitted SARIMA model for the observed series

(The plot which assigns negative values to the influenza data is the fitted SARIMA model)

The computation required for fitting the SARIMA models to our data is carried out using the ASTSA and SPSS.*V*12 computer packages.

Sometimes it is useful to describe the properties of the time series in a frequency domain. The Fourier transform is a mathematical technique for transferring a domain function into a frequency spectrum. The Fast Fourier Transform is a method for performing this process more efficiently. The spectrum is defined as:

$$f(\lambda) = \frac{\sigma^2}{2\pi}\sum_h \gamma(h)\exp(-i\lambda h), \qquad \textit{where } \gamma(h) \textit{ is the autocoriance function.}$$

From Figure 3.1E and the sample ACF, we except that the influenza incidence rates in winter, specifically in January, would be more vulnerable to seasonal fluctuations. Using frequency domain analysis, we attempt to verify that this is indeed true. By trying three

smoothing constants (L = 1, 2, 3), we expect that the peak of the spectrum would be around $\lambda = 0.0833$, which corresponds to the period of twelve months.

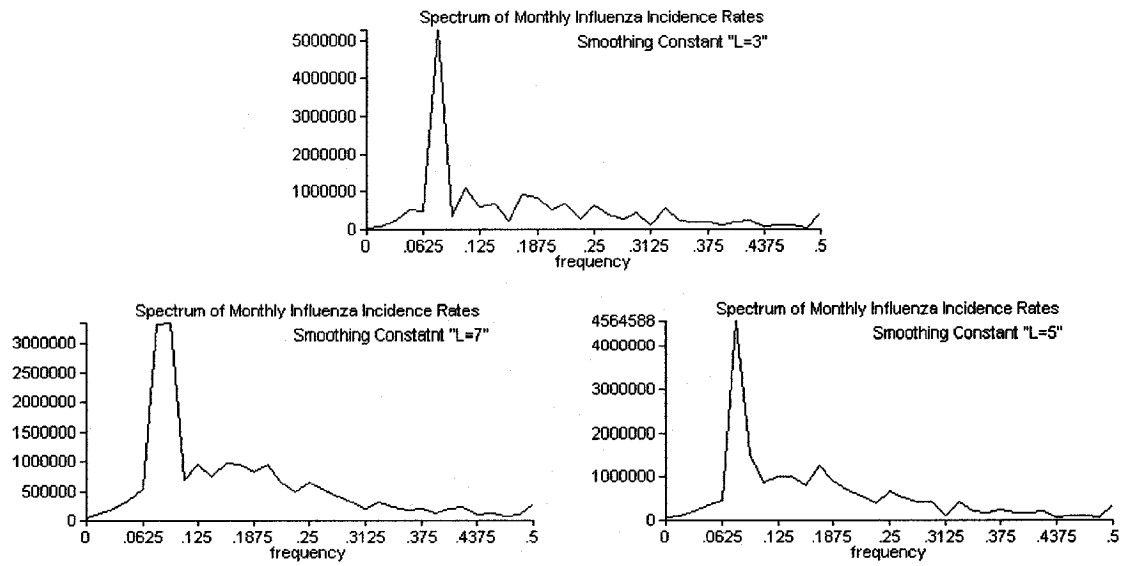

Figure 3.1E: Spectrum of monthly influenza incidence rates

It is nearly apparent that all of the powers are concentrated at the frequency of 0.0833, which corresponds to the period of twelve months (see Figure 3.1E). Therefore, we are confirmed by the spectrum analysis that there is a seasonal trend of twelve months for the influenza data.

3.2 Model Comparison

In this section, the HMM established for the data in Chapter 2 will be compared with the SARIMA model using various statistical criterions, such as the Mean Square Error (MSE) and the R^2, as well as predicted values for the following five consecutive months. The results are displayed in Table 3.1.

Model	Approach	MSE	R-squared	Month				
				181	182	183	184	185
				Predicted-Values				
ARIMA	Non-Bayesian	3.57E3	0.645	389	1124	70	-199	18
HMM	**Bayesian**	**2.65E3**	**0.728**	**411**	862	**571**	**319**	**72**
				Observed-Values				
				403	1459	856	274	60

Table 3.1: Comparison of SARIMA and HMM using MSE, R^2 and predicted-values criterions

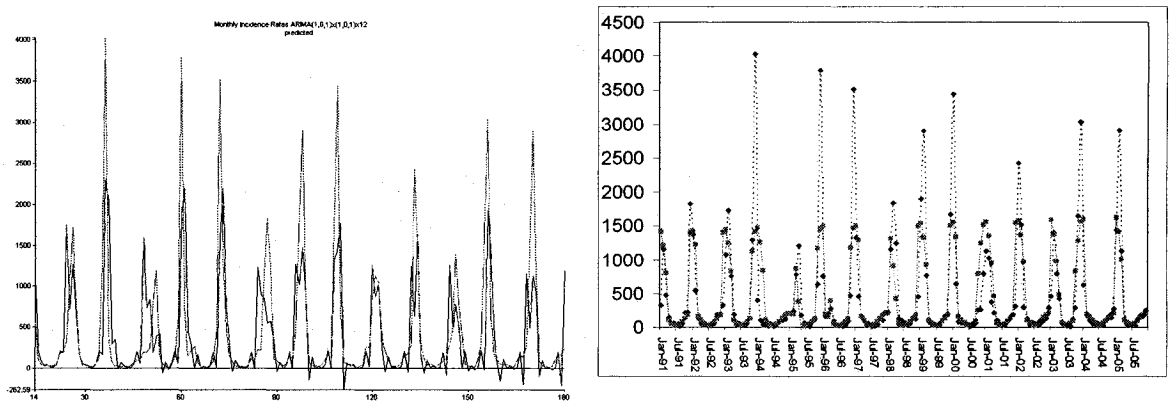

Figure 3.2A: Comparison of fitted SARIMA and HMM Bayesian models

(The plot which assigns negative values to the influenza data is the fitted SARIMA model)

As expected, the HMM with cyclic regression model and parameters estimated by the Bayesian approach, has lower MSE and higher R^2 than the SARIMA model. Also the HMM makes generally more accurate predictions than the SARIMA model that assigns negative values for some periods, which is not realistic. However, at month "182", February, the SARIMA model predicts a value that is closer to the corresponding observation. Since the months of January and February are characterized by high rates of

the reported disease (see Figure 2.2A), they are understood to be the epidemic phase for the SARIMA model and the HMM as well. As long as predicted rates are drastically above a given cutoff point during this period, the difference between the two predictions, SARIMA and HMM, in relation to the observed value is not significant. Therefore, the values predicted for these months are not representative of the overall fit of the model for decision making. Consequently, the better performance of the SARIMA at month "182" cannot be used to infer an advantage over its counterpart.

Furthermore, the HMM Bayesian model can be justified by using another statistical criterion: Deviance Information Criterion (DIC). The DIC is defined in analogy with the AIC. The model with the smallest DIC is regarded as the model that would best predict a replicate dataset [30]. A brief introduction to DIC is as follows.

Let measurement of fit via deviance equal: $D(data,\theta) = -2\log L(data \mid \theta)$. We need two quantities, which can be approximated using the MCMC sampler output in WinBUGS; First, D averaged over the posterior distribution of parameter, $E_{\theta|y}[D]$ and Second, D evaluated at the posterior mean of parameter, $D(E_{\theta|y}[\theta])$. Then complexity measured by estimate of the "effective number of parameters" is

$$PD = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]$$

$$= \overline{D} - D(\overline{\theta}).$$

And finally the DIC is defined as

$DIC = D(\overline{\theta}) + 2PD$ (These quantities are easy to compute in an MCMC run.)

$= \overline{D} + PD$ or equal to "*goodness of the fit + complexity*".

Models with smaller value of DIC are better supported by the data. The DIC is built into WinBUGS and can be monitored by running the software. The calculated DIC for our model, HMM Bayesian model, and SARIMA model, are 263 and 284, respectively. Therefore, the value of the DIC criterion also confirms the appropriate selection of the HMM Bayesian model for the data.

3.3 Discussion

The HMM with Bayesian approach (MCMC) provides a highly flexible framework for model exploration. The model established by using this approach has demonstrated its efficiency through our data set in comparison with the traditional SARIMA model. Also, an improved statistical method has been employed for developing the warning thresholds for the influenza disease in France. The warning threshold is an effective and capable indicator that can be used in health surveillance systems for strategic decision-making.

Influence of different prior distributions for the hidden Markov states is posed as a further question for the future work. In addition, the lack of user-friendly computer programs for utilizing HMMs makes it especially inconvenient when working with mixed distributions. We expect to work with programmers to develop enhanced statistical packages using, for example SPSS and S-Plus, in order to facilitate the further implementation of HMMs.

Appendix

WinBUGS Source Code for Analyzing Influenza Incidence Rates Data with Bayesian Approach

```
model;
{
  z0 ~ dbern(pInit)
  pInit ~ dbeta(1,1)
  z[1] ~ dbern(p[1])
  p[1] <- ((1-z0)*p01) + (z0*p11)
mean[1] <- ((1-z[1])*(mu0 + (beta0 * 1) + (delta0 * cos(2 * pi * 1/12)) + (epsilon0 * sin(2 * pi * 1/12)))) + (z[1]*(mu1 + (beta1
* 1)+ (delta1 * cos(2 * pi * 1/12)) + (epsilon1 * sin(2 * pi * 1/12))))

# Z= 0 means Non-Epidemic State & Z = 1 means Epidemic State

for( i in 2 : n ) {
  z[i] ~ dbern(p[i])
  p[i] <- ((1-z[i-1])*p01) + (z[i-1]*p11)
  y[i] ~ dnorm(mean[i],tau[i])
  tau[i] <-((1-z[i])*tau0) + (z[i]*tau1)
  mean[i] <- ((1-z[i])*(mu0 + (beta0 * i) + (delta0 * cos(2 * pi * i/12)) + (epsilon0 * sin(2 * pi * i/12)))) + (z[i]*(mu1 + (beta1 *
i)+ (delta1 * cos(2 * pi * i/12)) + (epsilon1 * sin(2 * pi * i/12))))
}
  tau0 ~ dgamma(0.001,0.001)
  tau1 ~ dgamma(0.001,0.001)
  p01 ~ dbeta(1,1)
  p11 ~ dbeta(1,1)
  mu0 ~ dnorm( 0.0,1.0E-6)
  mu1 ~ dnorm( 0.0,1.0E-6)
  beta0 ~ dnorm( 0.0,1.0E-6)
  beta1 ~ dnorm( 0.0,1.0E-6)
 delta0 ~ dnorm( 0.0,1.0E-6)
  delta1 ~ dnorm( 0.0,1.0E-6)
  epsilon0 ~ dnorm( 0.0,1.0E-6)
  epsilon1 ~ dnorm( 0.0,1.0E-6)
}


list(n=180,pi=3.14159,y=c(data))


list(mu0=0,mu1=0,p01=0.5,p11=0.5,pInit=0.5,beta0=0,beta1=0,delta0=0,delta1=0,epsilon0=0,epsilon1=0,tau0=1,tau1=)
```

References

[1] Institute of Medicine. Emerging infections: microbial threats to health in the United States. Washington, DC: National Academy Press, 1992.

[2] W.J. Martone, W.R. Jarvis, D.H. Culver and R.W. Haley. Incidence and nature of endemic and epidemic nosocomial infections. In J.V. Bennett and P.S. Brachman eds. Hospital infections. Third edition. Boston, MA: Little, Brown and Company, 1992.

[3] CDC. Preventing emerging infectious diseases: a strategy for the 21st century. Atlanta, GA: CDC, in press.

[4] Source: Institute Pasteur; see http://www.pasteur.fr/

[5] L. Baum et. al. A maximization technique occurring in the statistical analysis of probabilistic function of Markov chains. Annals of Mathematical Statistics, 41:164-171, 1970.

[6] L. Rabiner. A tutorial on Hidden Markov models and selected applications in speech recognition. Proceeding of IEEE, 1989.

[7] L.R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series, 1993.

[8] J. Bilmes. *What HMMs Can do*? E TRANSACTIONS on Information and Systems. Vol. E89-D, No.3, pp.869-891, 2002.

[9] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[10] F.V. Jensen. *An Introduction to Bayesian Networks*. Springer, 1996.

[11] S. Scott. Bayesian method for hidden Markov model: recursive computing in the 21st century, 2002.

[12] E. T. Jaynes. Bayesian methods - An introductory tutorial. In J. H. Justice, editor, *Maximum Entropy and Bayesian Methods in Applied Statistics*. Cambridge University Press, Cambridge, 1986.

[13] P. Congdon. Applied Bayesian Modeling, Wiley, UK, 2003

[14] A. Gelman, J. Carlin, H. Stern and D. Rubin. *Bayesian Data Analysis* 2nd ed., Chapman & Hall/CRC, 2003.

[15] CDC. Department of Health and Human Services of Centers for Disease Control and Prevention (http://www.cdc.gov).

[16] R. E. Serfling. Methods for current statistical analysis of excess pneumonia-Influenza deaths. *Public Health Reports*, 78:494–506, 1963.

[17] L. Simenson, K. Fukuda, L. B. Schonberg and N. J. Cox. The impact of influenza epidemics on hospitalizations. *The Journal of Infectious Diseases*, 181:831–837, 2000.

[18] F.C. Tsui, M. M.Wagner, V. Dato, and C. C. H. Chang. Value ICD-9-Coded chief complaints for detection of epidemics. In *Proceedings of the Annual AMIA Fall Symposium*, 2001.

[19] Y. LeStrat and F. Carrat. Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine*, 18 (24):3463–78, 1999.

[20] R. Winkler. The assessment of prior distributions in Bayesian analysis. Journal of the American Statistical Association, 62:776--800. 57, 1967.

[21] D.G. Fryback, N.K. Stout and M.A. Rosenberg. An elementary introduction to Bayesian computing using WinBUGS, *International Journal of Technology Assessment in Health Care*, Vol. 17, No. 1, pp. 98–113, winter 2001.

[22] C.E. Metz. Basic principles of ROC analysis. 8:283-298. Semin. Nucl. Med. 1978.

[23] M. H. Zweig and G. Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinicalmedicine. *Clinical Chemistry, 39*, 561–577, 1993.

[24] T.G. Tape. *Interpreting diagnostic tests.* University of Nebraska Medical Center, http://gim.unmc.edu/dxtests/Default.htm.

[25] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), pp. 159-174, 1977.

[26] G.E.P. Box and G. Jenkins. *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco 1976.

[27] P. J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting,* 2nd Edition, Springer, 2002.

[28] G.M. Ljung and G.E.P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65 (2), pp. 297-303, 1978.

[29] H. Akaike. A new look at statistical model identification. *IEEE Trans. Automat. Contr. AC*, 19, pp. 716-23, 1974.

[30] D. J Spiegelhalter, N.G Best, B.P. Carlin and A. van der Linde. *Bayesian Measures of Model Complexity and Fit*, 2002.