

# Automated Detection of Influenza Epidemics with Hidden Markov Models

Toni M. Rath<sup>1</sup>, Maximo Carreras<sup>2</sup>, and Paola Sebastiani<sup>2</sup>

<sup>1</sup> Department of Computer Science  
University of Massachusetts at Amherst, MA 01003 USA  
trath@cs.umass.edu

<sup>2</sup> Department of Mathematics and Statistics  
University of Massachusetts at Amherst, MA 01003 USA  
{carreras, sebas}@math.umass.edu

**Abstract.** We present a method for automated detection of influenza epidemics. The method uses Hidden Markov Models with an Exponential-Gaussian mixture to characterize the non-epidemic and epidemic dynamics in a time series of influenza-like illness incidence rates. Our evaluation on real data shows a reduction in the number of false detections compared to previous approaches and increased robustness to variations in the data.

**Keywords:** Cyclic regression, Hidden Markov Models, Influenza Surveillance, Time Series.

## 1 Introduction

Influenza is a contagious disease that affects between 10% and 20% of U.S. residents every year. Although most people regard the flu as a seasonal annoyance, influenza can lead to serious complications such as bacterial pneumonia and, for the elderly or immunocompromised patients, influenza can be fatal — it is estimated that an average of about 20,000 people die from influenza in the U.S. every year, and 114,000 per year have to be admitted to the hospital as a result of influenza. Because early detection of influenza epidemics could have a serious impact on the number of lives saved, great emphasis has been placed on influenza surveillance by monitoring indicators of the spread of epidemics.

The current approach to influenza surveillance is based on Serfling’s method [1]. The method uses cyclic regression to model the weekly proportion of deaths from pneumonia and influenza and to define an epidemic threshold that is adjusted for seasonal effects. Because of delays in detection caused by the nature of the data, several authors have proposed to use other indicators of influenza epidemics, such as the proportion of patient visits for influenza like illness (ILI) [2, 3] that should provide earlier indications of an influenza epidemic. Although applied to different data, the detection is still based on cyclic regression and suffers from several shortcomings, such as the need for non-epidemic data to model the baseline distribution, and the fact that observations are treated as independent and identically distributed. Although the second issue can be overcome by proper modeling of the time series data and examples are discussed in [4,

5], the first issue is a fundamental obstacle toward the development of an automated surveillance system for influenza.

Recent work in [6] suggested the use of Hidden Markov Models [7] to segment time series of influenza indicators into epidemic and non-epidemic phases. There are two advantages in this approach. The first advantage is that the method can be applied to historical data without the need for distinguishing between epidemic and non-epidemic periods in the data, thus opening the way toward an automated surveillance system for influenza. The second advantage is that the observations are supposed to be independent given knowledge about the epidemic, whereas Serfling’s method assumes marginal independence of the data.

For detection of influenza epidemics, the authors in [6] suggest using a mixture of Gaussian distributions to model the ILI rates and demonstrate the detection accuracy using a data set of historical data collected between January 1985 and December 1996 from a sentinel network of 500 general practitioners in France. We analyze the same data set and show that better detection accuracy can be achieved by modeling the data using a mixture of Exponential and Gaussian distributions. The change of the underlying distributions removes the need for explicit modeling of trend and seasonal effects that can introduce systematic bias in the detection accuracy.

The remainder of this paper is structured as follows. The next section describes the traditional approach to detection of influenza epidemics based on Serfling’s method. Section 3 reviews the basic concepts of Hidden Markov models. In section 4 we describe the approach proposed by [6] and introduce our model for epidemic detection. The evaluation, which compares these two methods, is described in Section 5. Conclusions and suggestions for further work are in Section 6.

## 2 Serfling’s Method for Epidemic Detection

The current approach to influenza surveillance implemented by the Centers for Disease Control in the U.S. is based on Serfling’s method [1] and the objective is to determine an epidemic threshold. The method models the weekly number of deaths due to pneumonia and influenza by the cyclic regression equation

$$y_t = \mu_0 + \theta t + \alpha \sin(2\pi t/52) + \beta \cos(2\pi t/52) + \epsilon_t \quad (1)$$

where  $y_t$  is the number of susceptible to deaths from pneumonia and influenza in week  $t$ , when there is no epidemic. The parameter  $\mu$  is the baseline weekly number of deaths, without seasonal and secular trends, and  $\epsilon_t$  is noise which is assumed to have mean 0 and variance  $\sigma^2$ . The component  $\theta t$  describes secular trend, and the sine-wave component  $\alpha \sin(2\pi t/52) + \beta \cos(2\pi t/52)$  models the annual recurrence of influenza epidemics. Assuming that the errors are uncorrelated and normally distributed, the standard least squares method is used to estimate the model parameters  $\mu, \theta, \alpha, \beta, \sigma^2$  from non-epidemic data, and to compute confidence bounds about the predicted values. The predicted values are given by

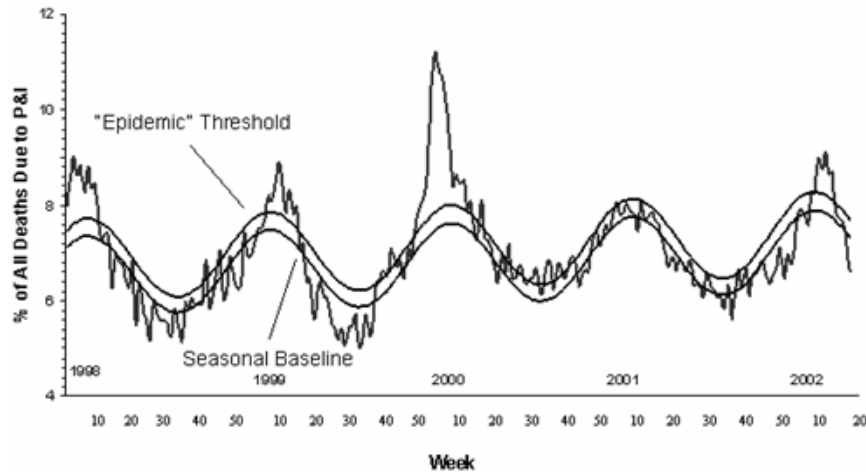
$$\hat{y}_t = \hat{\mu}_0 + \hat{\theta}t + \hat{\alpha} \sin(2\pi t/52) + \hat{\beta} \cos(2\pi t/52)$$

The confidence bounds are then computed as

$$\hat{y}_t \pm t_{\alpha/2} SE(\hat{y}_t),$$

where  $SE(\hat{y}_t)$  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.

As an example, Figure 1 describes the epidemic threshold computed by using Serfling's method on the proportion of deaths for pneumonia and influenza recorded in the USA from 122 sentinel cities between January 1998 and May 2002. The Figure is reproduced from the web site of the Centers for Disease Control and Prevention (CDC).



**Fig. 1.** Weekly proportion of deaths from pneumonia and influenza in 122 sentinel cities in the USA between January 1998 and May 2002 after normalization with the number of deaths for all causes. Source: CDC, *Morbidity and Mortality Weekly Report* (<http://wonder.cdc.gov/mmwr/mmwrmort.asp>).

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 like illness or hospital visit data for influenza [2, 3]. Although applied to data that should provide earlier indications of epidemics, the detection is still based on cyclic regression and suffers of two shortcomings: 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 Hidden Markov Models to segment time series of influenza indicators into epidemic and non-epidemic phases. This approach was introduced in [6] and is reviewed in the next section.

### 3 Hidden Markov Models

Hidden Markov Models (HMMs) are a convenient statistical tool for explaining serial dependency in data. They can be described by a sequence of pairs of random variables  $(Y_t, S_t)$ ,  $t = 1, \dots, n$ , satisfying the following conditional independence assumptions:

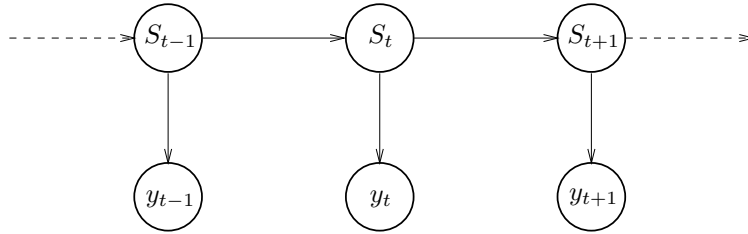
$$P(y_t | y_1, \dots, y_{t-1}, s_1, \dots, s_t) = P(y_t | s_t) \quad (2)$$

$$P(s_t | y_1, \dots, y_{t-1}, s_1, \dots, s_{t-1}) = P(s_t | s_{t-1}), \quad (3)$$

where  $y_t$  and  $s_t$  are, respectively, the realized values of  $Y_t$  and  $S_t$  and  $P(\cdot)$  denotes a generic probability function. The variable  $S_t$  is discrete and takes on one of  $m$  values  $1, \dots, m$ . Figure 2 shows a graphical representation of an HMM in which the directed arrows define the conditional dependencies. The sequence  $\{S_t\}_{t=1}^n$ , called the state sequence, is assumed to be unobserved or *hidden* and, from Equation (3), it follows a Markov Chain of order 1 with transition probability matrix  $P = (p_{ij})$ , where

$$p_{ij} = P(S_t = j | S_{t-1} = i) \quad i, j = 1, \dots, m; \quad t = 2, \dots, n$$

and initial probability distribution  $\pi = (\pi_1, \dots, \pi_m)'$ , with  $\pi_i = P(S_1 = i)$  for an  $m$ -state HMM.



**Fig. 2.** Illustration of the conditional dependencies in an HMM model.

The conditional distribution of  $Y_t$  given  $S_t = i$  has a parametric form  $f_{it}(y_t; \theta_i)$ , where  $\theta_i$  is a vector of unknown parameters. The notation  $f_{it}$  suggests that the conditional density could change with the state as well as with the time. An example of the former is the model in this work: we describe the non-epidemic rates with an Exponential distribution and the epidemic rates with a Gaussian distribution. An example of the latter is the trend and seasonality model for the distribution means in [6].

When applying the model to the data, there are two main objectives:

1. Learning the model form the data; that is, estimating the model parameters from the data. This is achieved by maximum likelihood techniques using the EM-algorithm, known in the HMM literature as the Baum-Welch algorithm [8].

2. Decoding the most likely sequence of hidden states that generated the data. This is done by means of the Viterbi algorithm [9].

The EM-algorithm is the standard technique for computing maximum likelihood estimates of the model parameters from an incomplete data set, see [10]. The method works by alternating an E-step to an M-step until convergence is reached. In the E-step, missing data are replaced by their expected values given the current parameter estimates; in the M-step, maximum likelihood estimates of the parameters are computed using the data completed in the E-step.

## 4 Epidemic Detection Using Hidden Markov Models

Le Strat & Carrat [6] proposed to detect epidemic and non-epidemic phases of influenza by modeling ILI rates with HMMs using a mixture of Gaussian distributions. They focused on a 2-state model and parameterized the means using Serfling's approach, so that the rates of both the epidemic and non-epidemic phases are modeled as in Equation (1). There are two main shortcomings to this approach:

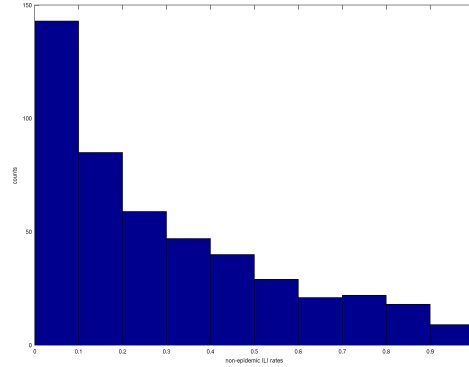
- The use of a Gaussian distribution for modeling the non-epidemic rates is inappropriate, because it assigns non-zero probability to the occurrence of negative incidence rates.
- The model for the means, although rich in parameters, is quite rigid; for instance, it does not allow for variation in the amplitude and frequency of the sinusoidal pattern. This lack of flexibility raises the question of how much data should be used in the model estimation. For example, modeling data from influenza mortality surveillance has shown that more accurate forecasts are based on a few years historical data, and the CDC influenza surveillance practice is to use five years historical data [11–13]. It is clear that using too little data will lead to an unreliable fit that would not extend well to future data. On the other hand, a large amount of data will yield bias and under/overestimation problems.

To address these problems, we propose to use a 2-state HMM, where non-epidemic rates are modeled with an Exponential distribution, and epidemic rates with a Gaussian distribution. Unlike Le Strat & Carrat's model, the means of the densities do not have any seasonal dependency. Non-epidemic rates are always positive and approximately exponentially distributed. As an example, Figure 3 shows a histogram of non-epidemic rates, using a visually determined threshold. The rates are skewed to the right, and follow, approximately, an exponential distribution.

Our model is simple and describes the data well, without the need for modeling seasonality and trend effects. It allows us to capture out-of-season, non-epidemic peaks, which tend to cause false detections in Le Strat & Carrat's periodic model.

## 5 Evaluation

Le Strat & Carrat [6] tested their model (referred to as *periodic model* in the following) for epidemic detection on a time series of ILI incidence rates in France from 1985 to



**Fig. 3.** Histogram of non-epidemic ILI incidence rates using an empirical threshold.

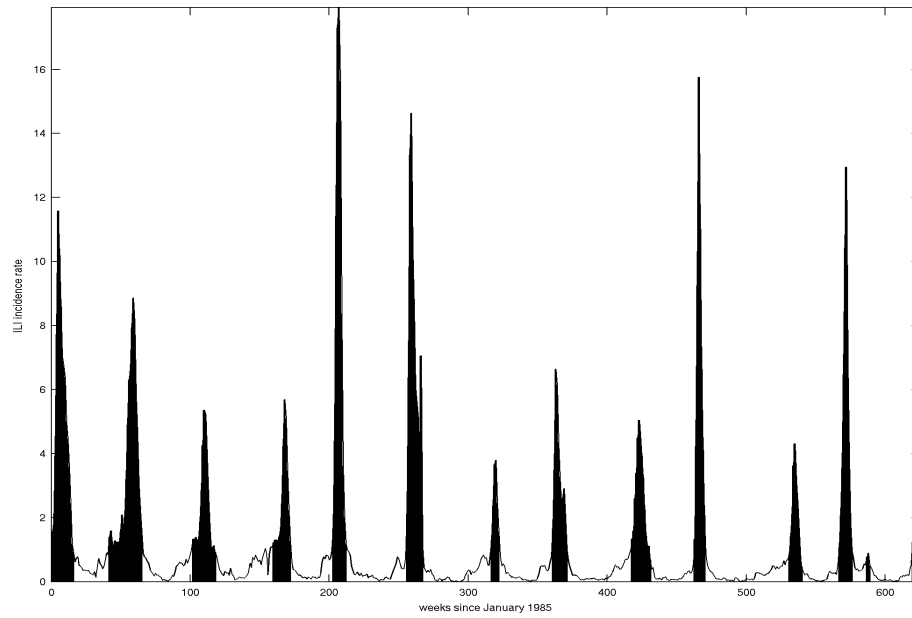
1996, collected by a network of sentinel physicians. For reasons of comparability, we performed the evaluation of our model (referred to as *Exponential*) on the same time series. Both models were implemented in *Matlab*<sup>3</sup> and the parameters estimated from the time series using the Baum-Welch/EM estimation algorithm [8]. The updated parameters are then passed to the Viterbi algorithm [9], which extracts the most likely state sequence that generated the observed time series. Table 5 summarizes the classification induced by the two models, and Figure 4 shows the ILI time series as classified by the two models, with the epidemic periods shaded dark.

Model	Number of epidemic weeks	Number of non-epidemic weeks
Periodic	159	465
Exponential	124	500

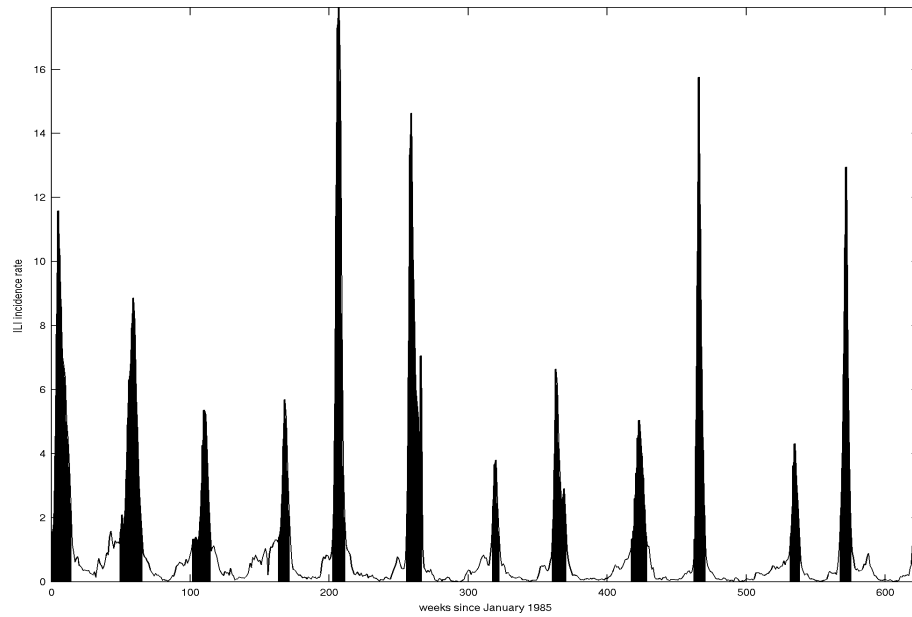
**Table 1.** Summary of the classification of the two models

An ideal classification separates the data into two parts: the non-epidemic, seasonal background pattern with low incidence rates and the epidemic periods with the pronounced peaks. Visual inspection shows that the Exponential model proposed here provides a more cautious classification than the periodic model proposed in [6] and it classifies a smaller number of weeks as epidemic weeks. Furthermore, the periodic model (Figure 4(a)) often tends to detect epidemic periods too early and non-epidemic periods too late, and such mistakes are very minimal with the Exponential model (Figure 4(b)): on average, our approach classified 10.40 weeks per year as epidemic, which is more realistic than 13.31 (over 3 months) obtained with the periodic model. For example, the CDC reports that, in the last 6 years, only during the influenza season 1996-1997 the

<sup>3</sup> Source code available from <http://ciir.cs.umass.edu/~trath/prj/epdet.html>



(a) normal/normal model with periodic means by Le Strat & Carrat,

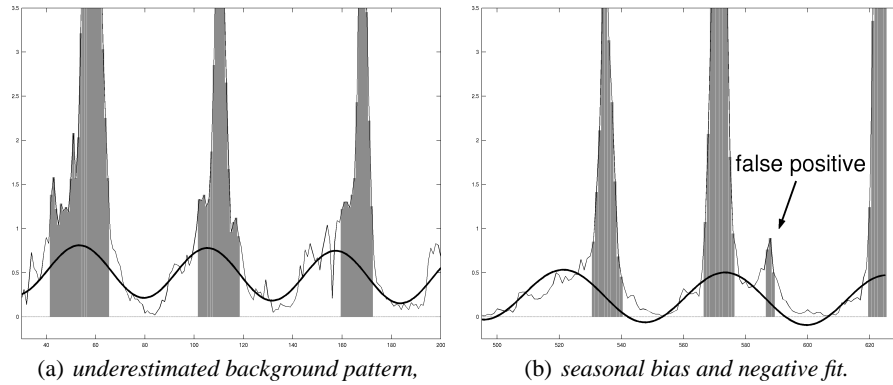


(b) exp/normal model proposed in this work

**Fig. 4.** Time series of weekly ILI rates in France from 1985 to 1996. Dark shaded areas indicate periods that were identified as epidemic by the respective models.

epidemic threshold was exceeded for more than 13 weeks, and the average number of epidemic weeks per year was about 10, see for example [14]. This number is consistent with our results.

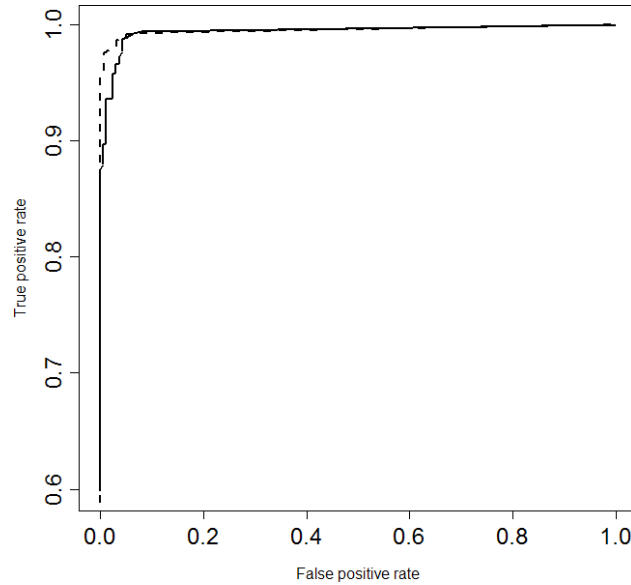
Another advantage of the Exponential model is its classification robustness: in the periodic model, small peaks occurring outside the influenza season trigger classifications (e.g. early classification of the second peak and one extra classification between the last two epidemic seasons in the periodic model). The Exponential model performs much better in these situations.



**Fig. 5.** Magnifications of the classification results with the periodic model. The shaded regions indicate the detected epidemic seasons, the sinusoidal pattern represents the estimated mean of the normal distribution which models the non-epidemic seasons.

We attribute these differences in classification performance to the rigid modeling of seasonality in the periodic model, which does not adapt to variations in the data. The effects can be seen in Figure 5, which shows magnifications of the classification results with the periodic model. In Figure 5(a), the beginning of the time series is magnified, where the non-epidemic background pattern varies in a larger range than in the rest of the series. Since the amplitude of the periodic, non-epidemic pattern is a constant that is estimated from the whole time series, the amplitude is underestimated for this part of the data. The consequence is that some non-epidemic incidence rates appear high compared to the estimated seasonal mean, and the corresponding weeks are therefore classified as epidemic. Furthermore, the seasonal model is unable to adapt to bias in the data: Figure 5(b) shows a part of the time series, where a small peak that occurs outside of the influenza season causes a false detection (marked with an arrow). This figure also shows a situation where the non-epidemic seasonal fit falls below zero, which is unrealistic.

Using an Exponential distribution to model non-epidemic incidence rates can lessen these problems. The Exponential distribution is a more adequate model, since it does not allow for negative incidence rates (as in the periodic model) and it is able to explain small out-of-season peaks in the non-epidemic portion of the data. This model prevents

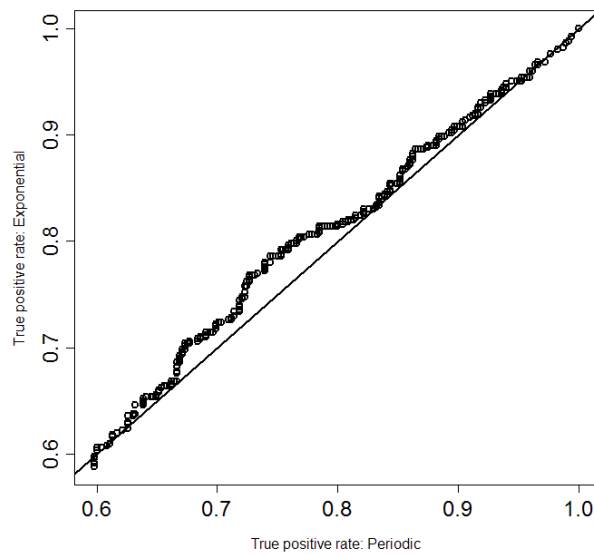


**Fig. 6.** Receiver Operating Characteristic (ROC) curve plotting the true positive rate against the false positive rate for the different epidemic thresholds. Solid line: ROC curve induced by the periodic model. Dashed line: ROC curve induced by the exponential model.

the false positive detections that occur in regions where the periodic model does not adequately adjust to the seasonal incidence rate pattern.

Because the original data were not labelled into epidemic and non-epidemic weeks, we further compared the classification determined by the two models with the classification that would be induced by an adjusted Serfling's method trained on the non-epidemic weeks. The periodic model labels 465 weeks as non-epidemic weeks and 159 as epidemic weeks. We used the non-epidemic weeks to estimate the parameters in Equation 1 (constrained to avoid negative estimates of the incidence rates) and then used the fitted model to detect epidemic and non epidemic weeks on the whole time series, using different epidemic thresholds. To measure the detection sensitivity and specificity of the periodic model, we then labeled as false positive the weeks that were detected as epidemic weeks by the periodic model and were below the epidemic threshold induced by Serfling's method. We also labeled as false negative the weeks that were detected as non-epidemic weeks by the periodic model and were above the epidemic threshold induced by Serfling's method. For example, when the threshold was set to have  $\alpha = 0.05$ , 4 of the weeks detected as epidemic weeks by the periodic model were below the epidemic threshold, whereas 24 of the weeks detected as non-epidemic weeks were above the threshold, resulting in a proportion of false positive  $4/159=0.02$  and a proportion of false negative  $24/465=0.0516$ . Varying the epidemic threshold with  $\alpha > 0.5$  results in different false positive and false negative rates that are plotted in the Receiver Operating Characteristic (ROC) curve in Figure 6 (solid line).

The Exponential model labels 500 weeks as non-epidemic weeks and 124 as epidemic weeks. The non-epidemic weeks were used to estimate the parameters of Equation 1. As above, we then used the fitted model to detect epidemic and non-epidemic weeks on the whole time series and to define the false positive and negative rates for different epidemic thresholds. For example, when the threshold was set to have  $\alpha = 0.05$ , none of the weeks detected as epidemic weeks by the exponential model was below the epidemic threshold, whereas 25 of the weeks detected as non-epidemic weeks were above the threshold, resulting in a false positive rate  $0/124=0$  and a false negative rate  $25/500=0.0500$ . The dashed line in Figure 6 depicts the tradeoff between false positive and negative rates and shows the slightly superiority of the exponential model with a ROC curve that is closer to the  $y$ -axis compared to the ROC curve associated with the periodic model. Figure 7 plots the true positive rates for the Exponential ( $y$ -axis) versus the periodic model ( $x$ -axis) for  $\alpha > 0.5$  and shows the larger true positive rate of the Exponential model.



**Fig. 7.** True positive rates of the Exponential ( $y$ -axis) versus the periodic model ( $x$ -axis).

## 6 Conclusions

We have proposed a new model for the detection of influenza epidemics in time series of ILI incidence rates using Hidden Markov Models. Unlike previous approaches, we do

not model explicitly the periodicity in the data to detect epidemics and non-epidemic weeks. We use a hidden variable to describe the unobservable epidemic status and, conditional on the absence of epidemics, we use an Exponential distribution to model the incidence rates of ILI. ILI incidence rates observed during influenza epidemics are modeled as Gaussian distributions. The exponential distribution provides a better adjustment to the variations in the data and removes the need for constrained parameter estimation that would be needed with Gaussian models. An evaluation of the proposed model demonstrated the improved performance over previously reported techniques.

## Acknowledgments

This work was supported by the Alfred P. Sloan Foundation (Grant 2002-12-1) and by the National Science Foundation (Grant IIS-0113496).

## References

1. R. E. Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports*, 78:494–506, 1963.
2. 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.
3. 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.
4. L. Wang, P. Sebastiani, J. Tsimikas, and K. Mandl. Automated surveillance for influenza pandemics. Technical report, Department of Mathematics and Statistics. University of Massachusetts at Amherst., 2003. Submitted.
5. P. Sebastiani and K. Mandl. Biosurveillance and outbreak detection using syndromic data. In *Data Mining: Next Generation Challenges and Future Directions*. AAAI/MIT Press, 2003. In press.
6. Y. LeStrat and F. Carrat. Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine*, 18 (24):3463–78, 1999.
7. L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–285, 1989.
8. L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.
9. G. D. Forney. The Viterbi algorithm. *Proceedings of the IEEE*, 61:268–278, 1973.
10. A. P. Dempster, D. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38, 1977.
11. C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society A*, 159:547–63, 1996.
12. L. C. Hutwagner, E. K. Maloney, N. H. Bean, L. Slutsker, and S. M. Martin. Using laboratory-based surveillance data for prevention: An algorithm for detecting salmonella outbreaks. *Emerging Infectious Diseases*, 3:395–400, 1997.
13. L. Stern and D. Lightfoot. Automated outbreak detection: a quantitative retrospective analysis. *Epidemiology and Infection*, 122:103–110, 1999.
14. Centers for Disease Control. Morbidity and mortality tables, 2002. <http://wonder.cdc.gov/mmwr/mmwmort.asp>.