Why We Need a Bayesian Approach to Early Detection of Epidemic Outbreaks and Financial Bubbles Using First-Order Autoregressive Models with Structural Changes

Ernest S. Shtatland, eStatConsulting, Stoneham, MA
Timur Shtatland, eStatConsulting, Stoneham, MA

ABSTRACT

We propose a Bayesian alternative to the frequentist approach developed in our NESUG 2007/2008 and SGF 2008 presentations. This approach includes first-order autoregression modeling, estimating autoregression parameters, building confidence intervals and hypothesis testing. For the purpose of early detection, this statistical inference is performed daily using short samples of previous observations. The results obtained within the frequentist approach are important, however, this approach suffers from well-known downward bias and related asymmetric distribution of classic least square estimates of autoregression parameters, so that standard methods for constructing confidence intervals based on symmetry are inaccurate. In our Bayesian alternative, we can avoid these disadvantages. Additionally, everyday testing results in a multiplicity problem and requires some adjustment, which is not always effective within the frequentist approach. In contrast, the current consensus suggests that correct adjusting is automatic within the Bayesian paradigm and that Bayesian testing of many hypotheses does not pose problems different than testing a single hypothesis, so no adjustment is needed. In addition, no resampling testing is necessary, which makes Bayesian detection procedures faster than frequentist counterparts. The Bayesian framework allows us to incorporate prior and any other type of exogenous information, which can compensate for the shortness of everyday baseline samples. Bayesian methods provide naturally interpretable results: they output the posterior probability that an outbreak has occurred. We can sound the alarm whenever posterior probability of an outbreak exceeds some threshold.

The intended audience: SAS users of all levels who work with SAS®/STAT and SAS®/ETS.

INTRODUCTION AND OVERVIEW OF OUR FREQUENTIST APPROACH

Early recognition of either bioterrorist attacks or emerging epidemics is of greatest importance nowadays, especially in the light of a new epidemic of so-called Mexican Flu aka Swine Flu aka A(H1N1) and finally A/California/04/2009(H1N1). The last name, A/California/04/2009(H1N1), is more accurate because the name A(H1N1) defines only the subtype of the virus A not the individual strain. Note that the notoriously known Spanish Influenza (1918-1919) also belongs to this subtype. This epidemic started in April 2009 in Mexico. In a short time WHO declared phase 5 - out of a possible 6 - meaning that a global outbreak was "imminent". Eventually, on June 11, 2009, WHO elevated the level to 6, that is Swine Flu was announced pandemic.

Also in financial markets, it is very important to be able to detect as early as possible the emergence of market bubbles and anti-bubbles. 2009 is a year of a global unfolding financial and economic crisis: the accumulation of several bubbles and their interplay and mutual reinforcement has led to an illusion of a "perpetual money machine". This crisis, initially focused on mortgage backed securities, has been cascading into a global economic recession, whose increasing severity and uncertain duration has led and is continuing to lead to massive losses and damage for billions of people. The cumulative effect of the real estate and financial derivative bubbles and crashes resulted in the deep loss of trust, not only in Wall Street, but more importantly in Main Street.

It seems rather unusual that these two seemingly unrelated phenomena (epidemics and financial bubbles) can be combined under one umbrella. The reason is that both phenomena have a common cause – contagion: physiological contagion in case of epidemics and psychological
contagion in case of stock market bubbles. As a result these phenomena can be described by using susceptible-infected-recovered (SIR) models and AR(1) processes as their natural local approximations. For the definition of SIR models and more details see the next section. Of course, using classic SIR models in epidemiology is not new. Actually, SIR is considered a workhorse in modern epidemiology. But using SIR models in finance, with contagion as a motivation, is quite new. See for example, Shive (2006), Shiller (2005), Shtatland (2007), Shtatland & Shtatland (2008a and 2008b). Note that early detection of epidemic outbreaks and financial bubbles can be formulated as a more general problem: detection of structural changes. Below in the paper, we will mostly use epidemiological context, keeping in mind that almost everything (models, estimates, confidence intervals and hypotheses testing, etc.) is applicable to the financial setting. In epidemiological context we will be interested basically in respiratory infection outbreaks on a background of influenza. Some results on the recent swine flu (A/California/04/2009(H1N1)) epidemic will be mentioned.

Most health-care (and financial) institutions routinely collect data in the form of time series. That is why effective purely temporal surveillance techniques are so important. Autoregressive moving average (ARMA) processes are among the most popular and frequently used temporal models. The literature on using ARMA models in biosurveillance is abundant, not to mention econometrics in which time series analysis plays a major role. Here we mention only the most important for us sources: Reis and Mandl (2003), Reis, Pagano and Mandl (2003), Earnest et al. (2005), and Lai (2005). Also see Shtatland, Kleinman and Cain (2006), Shtatland (2007), and Shtatland & Shtatland (2008a and 2008b) for numerous additional references. In spite of their popularity, AR and ARMA models in biosurveillance share a common disadvantage: they are used as purely empirical models, with no specific substance matter meaning for coefficients. Our suggestions in Shtatland, Kleinman and Cain (2006) on using low-order ARMA processes are motivated solely by the parsimony principle. In Shtatland (2007) and Shtatland & Shtatland (2008a and 2008b), we propose strong theoretical grounds for this. In doing so, we are based on SIR models. In the course of linearization of nonlinear SIR difference equations, which is justified by the requirements of early detection of the initial phase of the epidemic, we arrive at a linear AR(1) model of epidemics that inherits first principles of SIR. As a result, we have to estimate only one basic parameter, the first-order autoregressive coefficient. This parameter and its classic ordinary least squares (OLS) estimate have a very simple epidemiological meaning. It is important that in our approach, detection thresholds are defined naturally in terms of this coefficient. In Reis and Mandl (2003), Reis, Pagano and Mandl (2003) and Mohtashemi et al. (2006), thresholds are defined more or less arbitrarily, and they are heavily based on extensive historical data (daily visits during more than 10 years). In Shtatland & Shtatland (2008a) we propose in addition to OLS estimate a median-unbiased ratio-based estimate, compare both estimates, discuss OLS bias and confidence intervals for OLS estimates. In Shtatland & Shtatland (2008b) we compare various strategies for early detection of outbreaks: a strategy based on confidence intervals for the autoregressive coefficient built on a full sample (we call it naive) and two alternative strategies based on so-called supremum F-tests (essentially equivalent to supremum Likelihood-Ratio tests), and end-of-sample breakpoint S-tests. The supremum F-test is known also as a mid-sample test.

Note also that among mentioned above sources, Earnest et al. (2005) and Lai (2005) are devoted to SARS epidemics in mainland China and Singapore (2002-2003); Reis and Mandl (2003) and Reis, Pagano and Mandl (2003) work with respiratory diseases visits in general in the emergency department of Boston Children’s Hospital, an urban, tertiary care pediatric teaching facility. In Shtatland (2007) and Shtatland & Shtatland (2008a and 2008b), we are “influenza-oriented” (we are driven by data on influenza or influenza-like-illnesses (ILI) following Mohtashemi et al. (2006).

As mentioned above our results can also be used in describing explosive behaviors of economic and financial time series (e.g., stock market bubbles and micro-bubbles, related to individual stocks). See for example Phillips, Wu and Yu (2007), Shive (2006), Shiller (2005), Abreu and Brunnermeier (2003), and Brunnermeier and Nagel (2004). Although the presence of bubbles and
herding behavior in financial markets is widely accepted, there are still few theoretical models, which try to explain such phenomena. The model proposed in this paper is one of these few.

AUTOREGRESSION MODELS IN EARLY DETECTION

We are starting with the SIR models which are usually described by the first-order nonlinear system of difference equations with variables $S_n$, $I_n$ and $R_n$ that represent the respective numbers of susceptible, infected and recovered individuals correspondingly on day $n$. That is why the abbreviation — SIR. Typically, variables $S_n$, $R_n$ cannot be observed or measured systematically and our attention is concentrated on $I_n$. We can derive a closed second-order nonlinear equation for $I_n$, only. But this equation is not easy to interpret. The system of difference equations for variables $S_n$, $I_n$ and $R_n$ and the closed second-order nonlinear equation for $I_n$ alone are described in Shtatland & Shtatland (2008a and 2008b). For simplicity, here we will limit ourselves to some approximations. Since our goal is just early detection of structural changes such as emerging of epidemics or economics bubbles and not an adequate description of the process of interest as a whole, we can perform some approximations which result in the first-order autoregressive model (AR(1)) for variable $I_n$ (the number of infected or investors on day $n$)

$$I_{n+1} = (1 + \beta - \delta)I_n + e_n$$

(1)

where $\beta$ is the infection transmission rate, $\delta$ is the average rate of recovery (so that $d = 1/\delta$ is the mean duration of infectivity in days) and error term $e_n$ is an additive Gaussian noise. At this moment, we are not interested in specifics of this noise. See the details of the sequence of equations leading from the deterministic nonlinear SIR model to the stochastic but linear AR(1) model (1) in Shtatland & Shtatland (2008a and 2008b).

It is important to emphasize that our resulting AR(1) model (1) is not an ad hoc model. Being based on the SIR model, it inherits its first principles. With our objective to detect an outbreak, a take-off of the epidemic as early as possible, model (1) seems to be most adequate: it is simple, flexible, robust and easily interpretable. Really, when $\beta - \delta < 0$, i.e. the rate of recovery is greater than the transmission rate, we have a no-epidemic, stationary regime. When $\beta - \delta > 0$, there is an exponential growth, outbreak of epidemic. If $\beta - \delta = 0$, we have the so-called unit root case. Thus, there is a natural threshold in terms of $\beta - \delta$ to discriminate between these regimes. This difference is closely related to another famous threshold parameter, $R_0 = \beta / \delta$, which is known in epidemiology as the basic reproductive ratio or basic reproductive number. $R_0$ is a key quantity used to estimate transmissibility of infectious diseases. $R_0$ can be defined also as the average number of secondary cases generated by a single primary case during its entire period of infectiousness in a completely susceptible population.

Since we are especially interested in flu epidemics nowadays, here are some estimates of $R_0$ for some influenza or influenza-like outbreaks. It is known that for seasonal flu, the basic reproductive number for seasonal flu seems to range from 1.5 to 3.0. First estimates of parameter $R_0$ for the swine flu are around 1.5. Thus, swine flu (at least its first wave) looks less "transmittable" than typical seasonal flu. And yet, it should be remembered that the first wave of the Spanish Flu also had the basic reproductive rate as low as 1.49 whereas the second wave developed $R_0$ between 3.0 and 4.0 and resulted in worldwide mortality ranging from 50 to 100 million deaths (the worst epidemic in recent history). See for example Chowell et al. (2006). It is clear that the ability to quickly detect and institute control efforts at the early stage of an influenza pandemic is directly linked to the final levels of morbidity and mortality in the population.

Although $R_0$ is the most popular parameter for SIR models, for their approximation, AR(1), more appropriate and convenient is parameter $\beta - \delta$ which can be called “an engine of the epidemic” or just the rate of growth. The right side of equation (1) justifies this terminology. It is interesting that the type of behavior (a no-epidemic vs. an outbreak regime) is determined not by $\beta$ and $\delta$ separately, but through their difference $\beta - \delta$. Again, we would like to emphasize the lack of such
a natural threshold in all previous works on time series modeling for biosurveillance. See, for example, Reis, Pagano and Mandl (2003) and Mohtashemi et al. (2006), with some additional references in Shtatland & Shtatland (2008b). In these papers thresholds are determined by using massive (up to several years) historical data. But what to do if we have no historical data at all as in the case of a new emerging epidemic of swine flu of 2009?

STATISTICAL INFERENCE FOR AR(1) MODELS: OLS ESTIMATES OF AR(1) PARAMETERS

Statistical inference about model (1) (confidence intervals, hypotheses testing, etc.) is based on estimates of the sole autoregressive parameter \( a = 1 + \beta - \delta \). For simplicity, equation (1) can be written as

\[
I_{n+1} = aI_n + e_n. \tag{1'}
\]

It is safe to assume that \( a \) is always greater than 0 in our context. Moreover, typically we can assume that \( a > 0.5 \). There exists a well-developed theory of estimating the AR(1) parameter. It includes ordinary least-squares (OLS), Yule-Walker, Burg, and various modified least-squares estimators (see, for example, Provost and Sanjel (2005) and references therein). The most widely used estimator is OLS method that provides the following estimate of parameter \( a \) based on the time series \( I_1, I_2, \ldots, I_T \):

\[
\hat{a}(T) = \frac{\sum_{n=2}^{T} I_n I_{n-1}}{\sum_{n=1}^{T-1} I_n^2} \tag{2}
\]

\( T \) can be considered a baseline (time window) of data used for estimating parameters of the model and everyday decision-making. In the present paper like in Shtatland & Shtatland (2008a and 2008b), a typical value of \( T \) is 7 though larger values (e.g., \( T = 14 \)) can be useful. This choice of \( T \) seems reasonable since it compensates for the day of the week variation in data. It is well-known that there is strong 7-day periodicity in both applications of interest: early detection of epidemic outbreaks and stock market bubbles.

Note that for general ARMA processes, a nonlinear iterative least-squares procedure must be used for estimating ARMA parameters. This procedure does not always converge successfully for a given set of data, particularly if the starting values of parameters are far from the resulting least-squares estimates (SAS/ETS® User’s Guide (1993), pp 140-141). In case of AR(1), we have a very simple, explicit formula (2), which can be easily used and interpreted. The properties of estimate (2) are well known. It is consistent, i.e. \( \hat{a}(T) \to a \) (the real value of the autoregressive parameter) as \( T \to \infty \). At the same time \( \hat{a}(T) \) is a biased estimator.

The bias \( b_1(a) \) depends on the real value of the parameter to be estimated and the time window \( T \). According to Le Breton and Pham (1989), if \( 0 < a < 1 \) (a stationary case), then

\[
b_1(a) = -2a/T \tag{3}
\]

and if \( a > 1 \) (an explosive, outbreak case), then

\[
b_1(a) \approx -C T^{\beta} / a^{T} \tag{4}
\]

where \( C \) is a known constant. It is interesting that bias is always negative – towards stationarity. Note also that formulas (3) and (4) are not exact. They are asymptotical and as such should be used with care for finite \( T \), especially for small values.

As mentioned in Introduction, in Shtatland & Shtatland (2008a and 2008b) we also discussed median-unbiased (MU) estimates of the autoregressive parameter \( a \), compared them with OLS
estimates (2) and found them very useful. Here we will limit ourselves with estimate (2) and its slight modification

$$\hat{a}(T)=\sum_{n=2}^{T} l_n / \sum_{n=2}^{T-1} l_n$$  \hspace{1cm} (2')

which differs from (2) only in the denominator (now summation is performed from $n = 2$ to $n = T - 1$, rather than from $n = 1$ to $n = T - 1$). The idea behind this correction is very simple: dropping a positive term in the denominator results in overall estimate increase – to compensate our bias. According to Provost and Sanjel (2005), this simple correction can be surprisingly effective, including cases with small values of $T$.

CONFIDENCE INTERVALS FOR AUTOREGRESSIVE PARAMETER IN AR(1) MODELS: A NAÏVE APPROACH

As in the previous section, the theory below is developed only for stable regimes in which parameter $a$ is considered unchanged (but unknown). Thus properly speaking, it cannot be very useful in early detection situation, when the first days of the time window $T$ are non-epidemic and then the epidemic begins. That is why we call this theory “naïve”.

Even with bias correction discussed above, the point estimate of the form (2) cannot serve as the sole basis to distinguish the case $a < 1$ (stationarity) from $a > 1$ (outbreak). The point estimate can be considered rather as the “best guess” for an unknown parameter. To choose between these possibilities, we have to use either confidence intervals, or hypotheses testing, or both. Although a statistician will recognize the typical hypothesis-testing problem in biosurveillance (with Type I and Type II errors), in this section we prefer to use an equivalent confidence-interval language following Tukey (1991) who argues strongly that confidence intervals are more informative and “more honest” than p-values. Confidence interval results are based on the following statements:

1. If $a < 1$, then the probability distribution of $$(\hat{a}(T) - a) \frac{T^{\frac{1}{2}}}{(1 - \hat{a}^2(T)^{\frac{1}{2}})}$$ converges to the Gaussian distribution $N(0,1)$ when $T \to \infty$ (see Giraitis and Phillips (2006) and Phillips and Han (2006)).
2. If $a > 1$, then the probability distribution of $$(\hat{a}(T) - a) \frac{\hat{a}(T)}{(\hat{a}^2(T) - 1)}$$ converges to the standard Cauchy distribution when $T \to \infty$ (see Phillips, Wu and Yu (2007) and Phillips and Magdalinos (2007)).

Based on these distributions: Normal and Cauchy, we can build approximate confidence intervals for parameter $a$. There are 2 problems with these confidence intervals though. First, they are asymptotic, and can hardly be used for such short samples as $T = 7$ or 14. Second, the short sample distribution of $\hat{a}(T)$ is asymmetric and so strongly skewed to the left that the standard methods for constructing confidence intervals based on symmetry are likely to be extremely inaccurate. Most applied work on time series econometrics / biosurveillance relies on asymptotics in order to justify using OLS and traditional (symmetric) confidence intervals, but this is just a way of ignoring the finite sample distribution.

We studied the bias and asymmetry problems with the short sample distribution of $\hat{a}(T)$ by using simulations. Our simulations results for $T = 7, 14$ and $a = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2$ and 1.5 invariably demonstrate both effects. It is interesting that OLS bias (negative) and skewness
(to the left) are closely related. Also important is the following observation: OLS bias increases when \( a \) increases from \( a = 0.5 \) to \( a = 1 \); then with farther increase of \( a \) from \( a = 1 \) to \( a = 1.5 \) bias has a steep decrease. At the same time, the variance of \( \hat{a}(T) \) also drops sharply for \( a > 1 \). Similar behavior of \( \hat{a}(T) \) and its variance is described in Provost and Sanjel (2005) and Jarocinski and Marcet (2009). Such a simultaneous change in behavior of OLS estimate \( \hat{a}(T) \) and its variance can be used as an additional indication of the take-off of an epidemic / bubble.

Actually, all statistics mentioned above are developed for steady-state regimes in which parameter \( a \) is unchanged in time window \( T \) and we have either a stationary regime or an exponential growth. In early detection, at some moments we are dealing with a mix of these regimes, so strictly speaking our statistics are not applicable. As a result, statistic \( \hat{a}(T) \) can be a good estimate for either stationary, pre-epidemic state or purely epidemic regime, but completely misleading for the periods containing both pre-epidemic and epidemic days.

**HYPOTHESES TESTING FOR AR(1) MODELS: MID-SAMPLE TESTS VS. END-OF-SAMPLE TESTS**

As mentioned above, our naïve approach to confidence intervals is based on using the whole time window of size \( T \) and the assumption that during this time window the regime is unchanged, that is, we are either in the stationary state or epidemic regime. But what if a regime switch happens exactly in the time window? This situation is of special interest in early detection. Inevitably, we will have an incorrect estimate in this situation. To avoid it we proposed in Shtatland and Shtatland (2008b) two alternatives: mid-sample and end-of-sample tests, and their combination. These tests are called so because they are designed to locate a breakpoint in the middle or in the end of the sample (time window) correspondingly. They are based on OLS estimate \( \hat{a}(T) \) though this estimate does not appear in the formulas explicitly.

Implementation of mid-sample and end-of-sample tests requires massive computations: each day multiple using of PROC ARIMA (SAS/ETS®) with ESTIMATE statement, which provides the estimates of the first-order autoregressive parameter \( a \) and residuals needed for calculating test statistics. We can add to this that \( p \)-values for these tests can be calculated only through bootstrapping. For more details about these tests including formulas for test statistics etc., see Shtatland and Shtatland (2008b) and references therein. Here we note only that end-of-sample tests are designed to answer the question whether the most recent day is the beginning of the epidemic. At the same time, when using mid-sample tests we are assuming that likely we have already overlooked the epidemic takeoff which has happened somewhere in the middle of the recent time window.

**SUMMARY OF THE PROBLEMS WITH THE FREQUENTIST APPROACH**

Within our frequentist approach we can obtain important results, but there are serious drawbacks in this approach. Some of them are listed below.

- A substantial downward bias of \( \hat{a}(T) \) – based on \( \hat{a}(T) \) only we can be late with sounding alarm.

- A finite sample distribution of \( \hat{a}(T) \) is asymmetric (strongly skewed to the left) so standard confidence intervals based on symmetry are likely to be extremely inaccurate.

- Also, we rely on asymptotics in order to justify using OLS and traditional (symmetric) confidence intervals, but this is just a way of ignoring the finite sample distribution.

- Everyday decision making requires correction for multiple testing (\( p \)-value correction) which is nontrivial and not always effective within the frequentist approach.
The only practical and natural way to evaluate $p$-value seems to be estimation of $p$-values through resampling which results in heavy computation.

**BAYESIAN APPROACH: WHY WE NEED IT**

In our Bayesian alternative, we can avoid or at least soften disadvantages mentioned above. In particular, it can be shown that Bayesian analysis does not explicitly adjust for multiplicity of tests, because a correct adjustment is *automatic* within the Bayesian paradigm. In addition, no resampling testing is necessary which makes Bayesian detection procedures faster than their frequentist counterparts.

The Bayesian framework allows us easily to incorporate prior and any other type of exogenous information. It helps us to compensate for the shortness of everyday baseline samples ($T = 7$ or $14$). Bayesian methods provide naturally interpretable results: they output the posterior probability that an outbreak has occurred. We can sound the alarm whenever posterior probability of an outbreak exceeds some reasonable threshold.

A Bayesian approach to our model (1')

$$I_{n+1} = a I_n + e_n.$$  

where $a$ is our autoregressive parameter and $(e_n)$ is Gaussian white noise ($e_n \sim N(0, \sigma^2)$), should start with a prior distribution for parameters $a$ and $\sigma^2$. We treat these parameters differently. Parameter $a$ is considered basic: this parameter alone determines whether an epidemic or a bubble is already in progress. As to parameter $\sigma^2$, it is not as important for us. We consider $\sigma^2$ as a nuisance parameter and will estimate it beyond the Bayesian approach, within the frequentist framework. Thus, our attention will be limited to the priors for $a$ only. At the very beginning of surveillance, without any specific information about the process of interest, it is natural to assume some uninformative, in particular flat or uniform prior for $a$.

Sims and Uhlig (1991) noted a simple but very interesting fact that under the uniform prior for $a$, its posterior (assuming known $\sigma^2$) is Gaussian or Normal:

$$p(a | I_1, I_2, \ldots, I_T) = N(\hat{a}(T), \sigma^2 / \sum_{T-1}^T I_n^2)$$  

where $\hat{a}(T)$ is our familiar OLS estimate (2). The very appearance of $\hat{a}(T)$ in the Bayesian context is rather surprising. Even more surprising is the fact that a strongly skewed frequentist distribution of $\hat{a}(T)$ turns into a symmetric Normal posterior. As explained in Sims (1988), the downward bias of the OLS estimator is exactly compensated by the increasing precision of OLS for higher values of $a$, and as a result, the posterior is symmetric around the OLS estimator. The variance of distribution (7) depends on the particular observed sample $I_1, I_2, \ldots, I_T$. Note that we can use either formula (7) with $\sigma^2$ considered known, or apply an alternative formula

$$p(a | I_1, I_2, \ldots, I_T) = N(\hat{a}(T), S^2 / \sum_{T-1}^T I_n^2)$$  

where $S^2$ is the least squares estimate of $\sigma^2$

$$S^2 = \sum_{T-1}^T (I_{n+1} - \hat{a}(T) I_n)^2 / (T - 1)$$

Notice also that term $\sum_{T-1}^T \hat{I}_n^2$ in the denominator in formula (7) has a very simple and important meaning. It is an estimate of the so-called *Fisher's Information* on parameter $a$ ($FI(a)$) supplied by the sample $I_1, I_2, \ldots, I_T$. In our simple case (assuming $I_0 = 0$ and $\sigma^2 = 1$), this information can be defined as

$$FI(a) = E(\sum_{T-1}^T \hat{I}_n^2)$$
Here “E” stands for expectation or statistical averaging (see for example, White (1958)). Roughly speaking, the more variability we have in the sample, the more information about the parameter can be extracted and, as a result, the more precision in our estimate \( \hat{a}(T) \).

Thus, starting from the uninformative (uniform) prior for unknown \( a \), we arrived at the Gaussian or Normal posterior distribution with the parameters which can be estimated as \( \hat{a}(T) \) and \( \sigma^2 / \sum_{T=1}^{T} \sigma_n \). Since the process of early detecting of epidemic outbreaks or financial bubbles or any other structural changes is sequential by nature, our today’s posterior will serve as tomorrow’s prior so we should apply the Bayesian procedure routinely day by day. The effectiveness of the Bayesian routine is determined by the effectiveness of computing posterior distributions from prior distributions and likelihood. Of course, we could follow the Bayesian procedure strictly and work with two-dimensional prior-posterior distributions for both parameters \( a \) and \( \sigma^2 \) by using the popular Normal-Gamma (or Normal-Wishart in the case of multidimensional autoregressive model) prior / posterior technique (Uhlig (1994)). As a result, the marginal posterior distribution of \( a \) becomes a noncentral Student’s \( t \)-distribution with 3 parameters that are not easy to calculate anew while updating information.

Instead, we prefer to work in a Bayesian way with parameter \( a \) only, by conditioning on \( \sigma^2 \), that is considering either \( \sigma^2 = \text{constant} \) or estimating it using formula (8)). This approach can be called quasi-Bayesian or combined Bayesian-frequentist. Being not purely Bayesian, the approach allows to work with the one-dimensional Normal distribution only and to use simple formulas for transforming priors into posteriors. It can be explained as follows. Suppose that at our present day \( T \), we have a Normal prior with the mean \( \delta_{\text{prior}} \), variance \( \nu_{\text{prior}} \) and precision \( \nu_{\text{prior}} = 1 / \nu_{\text{prior}} \). Here it appears more convenient to work with precision, not variances. The observations available to us at this moment are \( I_1, I_2, \ldots, I_T \). Then a posterior distribution of the parameter \( a \) is again Normal with the mean \( \delta_{\text{post}} \), the variance \( \nu_{\text{post}} \) and precision \( \nu_{\text{post}} = 1 / \nu_{\text{post}} \) calculated as follows

\[
H_{\text{post}} = H_{\text{prior}} + \sum_{n=1}^{T} \sigma_n \quad (10)
\]

\[
\delta_{\text{post}} = \left( \frac{H_{\text{prior}} / \sum_{n=1}^{T} \sigma_n \delta_{\text{prior}} + \sum_{n=1}^{T} \sigma_n}{H_{\text{prior}} + \sum_{n=1}^{T} \sigma_n} \right) \delta_{\text{prior}} + \left( \frac{\sum_{n=1}^{T} \sigma_n}{H_{\text{prior}} + \sum_{n=1}^{T} \sigma_n} \right) \delta_{\text{post}} (11)
\]

In other words, the posterior mean of the autoregressive parameter is a weighted sum of the prior mean of this parameter and its OLS estimate based on the recent information in \( I_1, I_2, \ldots, I_T \). It is interesting that that averaging of the prior and sample means is performed with the weights equal to their corresponding precisions. This pattern occurs over and over again in Bayesian analysis. Also interesting that if we do not have any prior information (formally it means: \( \delta_{\text{prior}} = 0 \) and \( \nu_{\text{prior}} = 0 \)) then formula (11) reduces to our familiar OLS estimate \( \hat{a}(T) \). Thus, we can conclude that the Bayesian approach allows us to use any kind of additional information, whereas in the frequentist approach we are limited to the most recent sample \( I_1, I_2, \ldots, I_T \) with \( T = 7 \) or 14. Thus, formula (11) can compensate for shortness of our everyday baseline sample (time window). This equation exemplifies the optimal Bayesian way of carrying information from one time period to the next. One more remark regarding (11). It is easy to see a very close relationship between equation (11) and the two very popular in biosurveillance, econometrics and statistical process control techniques: the Kalman filter and the exponentially weighted moving averaging (EWMA).

**BAYESIAN HYPOTHESIS TESTING**

Bayesian hypothesis testing is less formal than non-Bayesian one. In fact, Bayesian researchers typically summarize the posterior distribution without applying a rigid decision process. If one wanted to apply a formal process, Bayesian decision theory is the way to go because it is possible to get a probability distribution over the parameter space and one can make expected utility calculations based on the costs and benefits of different outcomes. Considerable energy has been given, however, in trying to map Bayesian statistical models into the null hypothesis testing framework, with mixed results at best.
In our particular situation, once the posterior Normal distribution with parameters $\hat{\alpha}_{\text{post}}$ and $V_{\text{post}}$ (or $H_{\text{post}}$) is obtained we can perform a Bayesian test for

$$H_0: \alpha > 1 \text{ vs. } H : \alpha < 1$$

(12)

We calculate posterior masses for those regions, and based on these posterior probabilities, we accept or reject $H_0$. In doing so, we should take into account the cost of type I and type II errors. Thus, the Bayesian method, based on (11) and (12) provides us with easily interpretable results: it outputs the posterior probability that an outbreak has occurred. Hence, for a Bayesian, the degree of belief is naturally quantified. This makes it easy for a user (e.g., a public health official or an individual/institutional investor) to decide whether to investigate each potential outbreak based on the costs of false positives and false negatives. We can "sound the alarm", whenever the posterior probability of an outbreak exceeds some threshold.

In testing hypothesis like (12), it is especially important to have information as specific as possible about the range of probable values of the parameter $\alpha$ during some particular epidemic. Currently in the light of Swine Flu, we are especially interested in influenza outbreaks. That is why we will return to the published estimates of the parameters $R_0$ briefly mentioned earlier in the section AUTOREGRESSION MODELS IN EARLY DETECTION. For more details, see Chowell et al. (2006), Wearing, Rohani and Keeling (2005), and Mohtashemi et al. (2006). Let us remind that:

- $R_0 \approx 1.5$ for Swine Flu or A/California/04/2009(H1N1), which is a preliminary estimate.
- $1.5 \leq R_0 \leq 3.0$ for seasonal flu.
- $R_0 \approx 3.0$ for SARS (Severe Acute Respiratory Syndrome).
- $3.0 \leq R_0 \leq 4.0$ for Spanish Influenza, the second wave (1918-1919).

The fact that basic reproductive number for A/California/04/2009(H1N1) is located at the lower end of the corresponding interval for seasonal flu sounds quite optimistically. But as we mentioned before, the first wave of Spanish Influenza had also $R_0 \approx 1.5$ and (which is more important) both Swine Flu and Spanish Influenza belong to the same subtype A/(H1N1).

We have to transform the information above into estimates for our autoregression parameter $\alpha$. In doing so, we use formulas $R_0 = \beta / \delta$ and $\alpha = 1 + \beta - \delta$. Parameter $\delta$ (the rate of recovery) varies less than the infection transmission rate $\beta$. Nevertheless, it is easier to interpret the inverse to $\delta$: $1/\delta$ is the mean duration of infectivity in days. We have to add to this that according to Chowell et al. (2006) and Wearing, Rohani and Keeling (2005), infected individuals can pass the flu virus for 3-7 days following symptom onset which means that $1/7 \leq \delta \leq 1/3$. Thus for influenza, the shortest duration of infectivity is 3 days, the longest infectious period is 7 days and the median infectious time is 5 days. We will work with these 3 scenarios. Note that Mohtashemi et al. (2006) assume only the case with $1/\delta = 7$. Most calculations in Shtatland (2007) and Shtatland & Shtatland (2008a and 2008b) are also performed under this assumption. Using the above mentioned formulas with parameters $a$, $\beta$, $\delta$ and $R_0$ in combination with estimates of $R_0$ for various types of influenza we obtain the following estimates of autoregression parameter $a$ for 3 different variants of the duration of infectivity:

- For Swine Flu or A/California/04/2009(H1N1)
  - If $1/\delta = 3$, then $a = 1.167$;
  - If $1/\delta = 5$, then $a = 1.1$;
  - If $1/\delta = 7$, then $a = 1.07$;
- For seasonal flu
  - If $1/\delta = 3$, then $1.167 \leq a \leq 1.667$;
  - If $1/\delta = 5$, then $1.1 \leq a \leq 1.4$;
  - If $1/\delta = 7$, then $1.07 \leq a \leq 1.29$;
• For Spanish Influenza, the second wave (1918-1919)
  
  If $1/\delta = 3$, then $1.667 \leq a \leq 2.0$;
  
  If $1/\delta = 5$, then $1.4 \leq a \leq 1.6$;
  
  If $1/\delta = 7$, then $1.29 \leq a \leq 1.43$;
  
From these estimates we can conclude that the minimal value of $a$ is 1.07 (Swine Flu with $1/\delta = 7$), then follow $a = 1.1$ (Swine Flu with $1/\delta = 5$) and $a = 1.17$ (the same Swine Flu with $1/\delta = 3$). This implies the growth rate of 7%, 10% and 17% (per day) correspondingly. Naturally, these estimates are realistic only at the beginning stage of the emerging epidemic when the whole population is susceptible (not immune). For comparison, the NASDAQ Bubble in the 1990s showed a rate of 4% per month (see Phillips, Wu and Yu (2007)). The maximal value of parameter $a$ for influenza-like illnesses (ILI) is 2.0 (Spanish Influenza), which means that the number of infected is doubling every day. This rate can be sustained only at the early phases of epidemics.

The simplest way of performing test (12) is to check each day whether the posterior mean of the autoregressive parameter, $\tilde{a}_\text{post}$ exceeds the lower level of the intervals mentioned above. Note that when we apply test (12) every day, it can potentially result in a multiplicity problem and requires some adjustment, which is nontrivial and not always effective within the frequentist approach. In contrast, Bayesian analysis does not explicitly adjust for multiplicity of tests, the argument being that a correct adjustment is automatic within the Bayesian paradigm. In other words, Bayesian testing of many hypotheses does not pose problems different than testing a single hypothesis, and no adjustment is needed. For details, see for example, Berry and Hochberg (1999), Scott and Berger (2003), Bayarri and Berger (2004), and Farcomeni (2004). In addition, no resampling testing is necessary which makes Bayesian detection procedures faster than their frequentist counterparts.

**CONCLUSION**

In this paper we propose a Bayesian alternative to the frequentist approach developed in our NESUG 2007/2008 and SGF 2008 presentations for early detection of epidemic outbreaks and financial bubbles and anti-bubbles. We demonstrate the advantages of the Bayesian approach over the frequentist counterparts. In doing so, we use the data of the most recent influenza epidemics.

**REFERENCES**


SAS/STAT® and SAS/ETS® are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

**CONTACT INFORMATION**

Ernest S. Shtatland
eStatConsulting
200 Park Terrace Drive, Suite 234
Stoneham, MA 02180
email: eshtatland@yahoo.com