Early Detection of Epidemic Outbreaks and Financial Bubbles Using Autoregressive Models with Structural Changes

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

ABSTRACT

This paper is a further development of our previous work presented at SUGI 2006, NESUG 2007 and SGF 2008. Our basic model is AR(1) (first-order autoregression). This model emerges as a natural approximation of a classic susceptible-infectious-recovered (SIR) model. It inherits first principles of SIR models and can be used in both epidemiological and financial applications for early detection of epidemics and financial bubbles. We consider epidemic outbreaks or financial bubbles as structural changes in the autoregressive coefficient in AR(1) models. We propose two methods of estimation of the autoregressive parameter: least-squares and median-ratio based methods, discuss the questions of bias correction and confidence intervals construction. The value of our first-order autoregressive coefficient less than one corresponds to a stationary, no-epidemic/no-bubble regime. If the parameter is greater than one, we have an explosive case (outbreak of epidemic or bubble). When the coefficient is equal to one, we have a unit root case. Under some conditions, least-squares estimates and confidence intervals, based on the observed data in a chosen time window, allow us to decide which case is more appropriate. We propose two alternative strategies for early detection of structural changes. In both strategies we use some generalizations of the Fisher's F-test: so-called supremum F-tests (essentially equivalent to supremum Likelihood-Ratio tests), and end-of-sample breakpoint S-tests. These tests can be used in rather general situations and usually have a better power performance. Also we provide simulation results.

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

INTRODUCTION

Early recognition of either bioterrorist attacks or emerging epidemics is of greatest importance nowadays. Also in financial markets, it is very important to be able to detect as early as possible the emergence of stock market bubbles. It can seem unusual that these two seemingly unrelated phenomena are 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 emergency 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 approximation. 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), and Shtatland & Shtatland (2008). 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 context.

With such highly infectious and rapidly spreading diseases as SARS, influenza, measles or smallpox it is a challenge to build adequate models for early detection of these exponentially developing processes. According to Mohtashemi, Kleinman and Yih (2007), temporal anomaly detection is a key component of real-time syndromic surveillance. Although space-time detection
methods are sometimes more powerful at anomaly detection than purely temporal ones, spatial data are typically scarce. In addition, many health-care institutions routinely collect data in the form of time series. That is why effective purely temporal surveillance techniques are still needed. 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. 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 (2008) 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 were based solely on the parsimony principle. In Shtatland (2007) and Shtatland & Shtatland (2008), we proposed strong theoretical grounds for this. In this paper, we provide further development of our approach. In doing so, we are guided by Mohtashemi et al. (2006) in which SIR models were successfully used for early detection of respiratory infection outbreaks. 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. Here, we need to estimate only one parameter, the first-order autoregressive coefficient. This parameter and its least squares estimate have a very simple epidemiological meaning. It is interesting that in our approach detection thresholds are defined naturally in terms of this coefficient. In all papers cited above, thresholds are defined more or less arbitrarily. We propose two estimates of the first-order autoregressive coefficient: the classic ordinary least squares (OLS) estimate and the median-unbiased ratio-based estimate, discuss OLS bias and confidence intervals for OLS estimates. Also we discuss strategies for early detection of outbreaks: a naïve strategy based on confidence intervals for the autoregressive coefficient built on a full sample (we developed this naïve approach in Shtatland (2007) and Shtatland & Shtatland (2008)) 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.

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 Pedersen (2002). Although the presence of bubbles and herding behavior in financial markets is widely accepted, there are still few theoretical models which generate such phenomena. The model proposed in this paper is one of these few. It is based on the premises that (1) a speculative bubble is inherently social-psychological and (2) it is a short-term rather than long-term, fundamental phenomenon.

SIR MODELS IN EARLY DETECTION

A SIR model can be described by the following first-order nonlinear system of difference equations (we use the notations of Mohtashemi et al. (2006)):

\[
\begin{align*}
S_{n+1} &= S_n - \beta S_n I_n, \\
I_{n+1} &= I_n + \beta S_n I_n - \delta I_n, \\
R_{n+1} &= R_n + \delta I_n
\end{align*}
\]  

(1)

where \( S_n, I_n \) and \( R_n \) represent the respective numbers of susceptible, infected and recovered individuals correspondingly on day \( n \); \( \beta \) is the infection transmission rate and \( \delta \) is the average rate of recovery from infection; \( 1/\delta \) can be considered the mean duration of the infectivity (in days). Below, we will use another transmission parameter, alternative to \( \beta \). Note that Mohtashemi et al. (2006) use only the first two equations in (1) with the motivation that there is not enough time for temporary removal of the recovered population to be of significance to the dynamics. Note also that in the context of their article (the emergency department of a large, academic pediatric...
hospital), variables $S_n$, $R_n$ cannot be observed or measured systematically. Only $I_n$ can be obtained indirectly, through some calculations, by using the observable daily number of patients $v_j$ presenting to the emergency department on day $j$. Mohtashemi et al. (2006) propose the following approximate formula for $I_n$:

$$I_n = \sum_{j=n-d+1}^{n} v_j$$

(2)

where $d$ is an average number of days of infectivity per patient, i.e. $d = 1/\delta$. Thus Mohtashemi et al. (2006) assume that the overall number of infected on day $n$ can be approximated by the sum of the number of visits to the emergency department during the past $d$ days. This is a reasonable approximation if parameter $d$ is adequately chosen or estimated. Also Mohtashemi et al. (2006) suppose that the mean duration of infectivity $d = 1/\delta = 7$, which is a clinically realistic assumption for influenza. And even though the assertion $d = 7$ is rather approximate (according to Wearing, Rohani and Keeling (2005), influenza lasts for 3 to 5 days), it is very convenient because it allows to compensate for the day of the week effect in $v_j$ variability. According to Mohtashemi et al. (2006), the goodness of such approximation can potentially impact the accuracy of early detection results and is open to discussion. Summarizing, we can see an urgent need in a closed equation for $I_n$ alone. Fortunately, such an equation can be obtained from the first two equations in (1) by eliminating unobservable variable $S_n$ (see equation (2) in Mohtashemi et al. (2006)):

$$I_{n+2} = I_{n+1} / I_n - \beta I_{n+1} (I_{n+1} - (1 - \delta)I_n)$$

(3)

Equation (3) is a strongly nonlinear second-order difference equation, which is not easy to interpret. In (3) we have only one parameter to estimate, $\beta$. Mohtashemi et al. (2006) use this equation for estimating $\beta$ through least-squares regression for each time window of a chosen length. The authors proceed from the premise that the transmission rate $\beta$ is seldom unchanging and may be considered approximately constant only over short time periods. For each day in a year, they choose a time window of length $T$ (from the present day to the past), form $T - 2$ equations and estimate $\beta$ by using least-squares regression. In their paper, the results for $T = 7$ are reported. Again, this choice of $T$ seems reasonable since it compensates for the day of the week variation in data. In addition, it is made in the spirit of early detection requirements. At the same time the choice $T = 7$ results in a very high everyday variability of the $\beta$ estimates, comparable to the variability of the time series $I_n$ itself. It looks unusual since $\beta$ is the only fundamental parameter of our model. We can conjecture that the time window length of 14 or 28 would provide smoother estimates of $\beta$ and in some cases serve the needs of early detection of outbreaks at the same level as $T = 7$ (note that there is a 7-day periodicity in both applications of interest: early detection of epidemic outbreaks and stock market bubbles). Though, sometimes it is necessary to work with $T < 7$ also. Unfortunately, there is no explicit formula for the least-squares estimates of $\beta$ and it is not easy to investigate their properties, in particular sensitivity to the time-window length $T$. This disadvantage will be overcome within our autoregression modeling approach. After getting the estimate of $\beta$ for each current day, Mohtashemi et al. (2006) compare it with the mean infection rate for the same day of year in their historical data (they have 7 training years). If the current estimate exceeds the threshold based on the training data (in terms of the mean and standard deviation of $\beta$), their detection algorithm generates an alarm. Mohtashemi et al. (2006) test the performance of their detection system by using simulated outbreaks. But what to do if historical data are either not available at all or are unrepresentative of the current behavior for a variety of reasons: changes in treatments, coding, and reporting practices, appearance of new influenza strains, etc. We will see below that our approach with autoregressive approximation of SIR models can help us in this situation.

**AUTOREGRESSIVE MODELS IN EARLY DETECTION**

In trying to build autoregressive models for early detection of epidemic outbreaks, we use SIR model (1) and equation (3). The 2nd order of equation (3) suggests that we can limit ourselves to AR(2) models. Thus, regrouping (3) we arrive at the equation of the form
\[ l_{n+2} = a_1 l_{n+1} + a_2 l_n \]  \hfill (4)

where
\[ a_1 = \frac{l_{n+1}}{l_n} - \beta (l_{n+1} - l_0) \]  \hfill (5)
\[ a_2 = -\beta \delta l_{n+1} \]  \hfill (6)

Equation (4) is a nonlinear difference equation with time- and state-dependent coefficients. This class of equations has been described, for example, in Kato and Ozaki (2002). Also, (4) can be considered an equation with slowly varying coefficients. Indeed, in the context of Mohtashemi et al. (2006) we can think that \( \beta \approx 2/10000, \delta \approx 1/7, l_{n+1} \approx 350, \) and \( (l_{n+1} - l_0) \approx 20 \) or less at the early phase of epidemics. Thus, \( a_2 \approx 0.01, \) this coefficient can be neglected and our 2nd order difference equation is reduced to the 1st order equation. Further, it can be shown that term \( \beta (l_{n+1} - l_0) \) in (5) is usually less than 0.004 and can also be neglected. Thus, \( l_{n+1} / l_n \) can be considered the main part of parameter \( a_1, \) and consequently this ratio \( l_{n+1} / l_n \) should be a good estimate of parameter \( a_1 \) (see Hurwicz (1950), Zielenski (1999) and Luger (2005)). Below we will return to the ratio-based estimates.

Of course, both of the dropped terms can produce an effect on the course of epidemics during much longer time than the starting period of the epidemic in which we are interested. As shown, equation (4) can be reduced to the following 1st order linear equation (we change the time index: \( n + 2 \rightarrow n + 1 \)):

\[ l_{n+1} \approx a_1 l_n . \]

This equation can be considered an indication of an exponential growth. Also, at the early phase of an outbreak, we can assume \( S_n \approx S_0 \approx N \) (where \( S_0 \) is the initial number of susceptible at the beginning of the epidemic (if it ever happens) and \( N \) is the total number of people in our closed population: \( N = S + I + R, \) with no birth, no death). With this assumption we arrive at the equation

\[ l_{n+1} \approx (1 + \beta N - \delta )l_n \]  \hfill (7)

Now we introduce an alternative transmission rate \( \beta^* = \beta N. \) Both \( \beta \) and \( \beta^* \) are popular in the epidemiological community. For more information about various forms of transmission rates see, for example, McCallum, Barlow and Hone (2001). According to Begon et al. (2002), transmission rates \( \beta \) and \( \beta^* \) discussed here are likely to remain benchmarks against which actual transmission dynamics are judged, and to remain key elements in most mathematical models of transmission. Note that \( \beta \) and \( \beta^* \) have different dimensions. With the alternative infection transmission rate, our difference equation takes a very simple form

\[ l_{n+1} \approx (1 + \beta^* - \delta )l_n \]  \hfill (8)

Finally, note that we have performed a number of approximations in developing linear equation (8) from strongly nonlinear equation (3). Each of these approximations results in some small error, and it is not easy to take those errors into account individually and altogether. In addition, we can have some stochastic variation in \( \beta \) and \( \delta. \) Combining all these sources of uncertainty in one, we arrive at our final AR(1) model

\[ l_{n+1} = (1 + \beta^* - \delta )l_n + e_n \]  \hfill (9)

where error term \( e_n \) is an additive white Gaussian noise. At this moment, we are not interested in specifics of this noise. Note that we can derive the equations (7), (8) and (9) not from (4), but directly from the second equation in (1):
\[ l_{n+1} = l_n + \beta S_n l_n - \delta l_n \]

By using the same assumption \( S_n \approx S_0 \approx N \) as before, we immediately get linear models (7), (8) and (9). This is a result of linearization: our originally nonlinear system becomes (to a very good approximation) linear because the assumption \( S \approx N \) turns our nonlinear transmission term \( \beta SI \) into a linear one, \( \beta^* I \). The sequence of approximations (4) - (9) discussed above can be symbolically expressed as \( \text{SIR} \rightarrow \text{AR}(2) \rightarrow \text{AR}(1) \). We adopt model (9) as our basic model for pre-epidemic and emerging epidemic processes. Note that almost all ARMA / ARIMA models for syndromic surveillance for respiratory diseases (influenza, SARS) use a 1st order autoregressive component (see Shtatland (2007)). That is why using AR(1) to describe a pre-epidemic process does not seem unusual. However, we are not interested in the most adequate description of a quiet, stationary period before the epidemic. Our goal is early detection of an outbreak, the take-off of the epidemic, and model (9) seems to be adequate for this purpose. When \( \beta^* - \delta < 0 \), i.e. the rate of recovery is greater than the transmission rate, we have a non-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. It is very important because of 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). It is important to emphasize that our AR(1) model (9), being based on the SIR model, inherits its first principles.

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

Statistical inference about model (9) (confidence intervals, hypotheses testing, etc.) is based on estimates of the sole autoregressive parameter \( a_1 = 1 + \beta^* - \delta \). For simplicity, we drop subscript index 1 here and to the rest of the paper. Thus, equation (9) can be written as

\[ l_{n+1} = a l_n + e_n. \quad (9') \]

It is safe to assume that \( a \) is always greater than 0 in our context. There exists a well-developed theory of estimating the AR(1) parameter. It includes ordinary least-squares, 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 the ordinary least-squares (OLS) method that provides the following estimate of parameter \( a \) based on the time series \( l_1, l_2, \ldots, l_T \):

\[ \hat{a}(T) = \frac{\sum_{n=2}^{T} l_n l_{n-1}}{\sum_{n=1}^{T-1} l_n^2} \quad (10) \]

\( T \) can be considered a baseline of historical data used for estimating parameters of the model and making decisions. Here and further in the paper, a typical value of \( T \) is 7, though larger values (e.g., \( T = 14 \)) and especially smaller values (from 2 to 6) can be useful. Note that for general ARMA processes, a nonlinear iterative least-squares procedure must be used for estimating ARMA parameters. It is well known that 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 (10), which can be easily used and interpreted. The properties of estimate (10) are well known. It is consistent, i.e. \( \hat{a}(T) \rightarrow a \) (the real value of the autoregressive parameter) as \( T \rightarrow \infty \). At the same time \( \hat{a}(T) \) is a biased estimator. The bias \( b_T(a) \) 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

\[ b_T(a) = -2a / T \quad (11) \]

and if \( a > 1 \) (an explosive, outbreak case), then
$b_T(a) \approx - C T^\kappa / a^T \quad (12)$

where $C$ is a known constant which depends on $a$. In Le Breton and Pham (1989), there is a more complicated formula for $b_T(a)$ in the unit root case ($a = 1$) that is similar to (11) with a different constant. This case is not as important to us as the two previous ones. In practice we are never in a position of knowing the true mean of the process of interest. If the mean is unknown, we have to use a model with intercept. In this case the bias becomes even larger than (11), it is given by following formula

$b_T(a) \approx - (1 + 3a) / T \quad (11')$

(see Kendall (1954), Sawa (1978), Orcutt and Winokur (1969)). It is interesting that the bias $b_T(a)$ is always negative which means that the ordinary least-squares estimator always underestimates a. The bias can be rather substantial when the unknown parameter $a$ is in the vicinity of 1. For example, according to (12) an outbreak value $a = 1.1$ might be estimated as 0.96 (the bias equals 0.14 or 13% of the real value) with $T = 7$. Let us remind that $T = 7$ is used in Reis, Pagano and Mandl (2003) and Mohtashemi et al. (2006). The example above shows how easy it is to misspecify the real regime if we do not take the bias into account. It is important to remember that formulas (11), (11') and (12) are asymptotic and ideally are meant to be used for large enough $T$. So using them for $T = 7$ is rather problematic. Instead of using formulas (11), (11') and (12) for bias correction we can also use a modified least-squares estimator defined by formula (10')

$\hat{a}(T)= \sum_{n=2}^{T} I_n I_{n-1} / \sum_{n=2}^{T-1} I_n^2 \quad (10')$

which differs from (10) 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. According to Provost and Sanjel (2005), this simple correction can be surprisingly effective, including cases with small values of $T$.

**MEDIAN-UNBIASED ESTIMATES OF AR(1) PARAMETERS**

Another interesting and more general approach to unbiased estimating of autoregressive parameter $a = 1 + \beta \ast - \delta$ in (9), was initiated by Hurwicz (1950) and developed further by Zielinski (1999) and Luger (2005). In particular, Hurwicz (1950) observed that every ratio $I_{n+1} / I_n$, $n = 1, 2, \ldots, T-1$, is a median-unbiased estimator of $a$ (though inefficient in statistical terms, since these ratios have a Cauchy distribution). Hurwicz (1950) hypothesized that the median of the ratios mentioned above would be a more efficient estimate of $a$ and perhaps an unbiased one. This hypothesis has been proved by Zielinski (1999) who has showed that the Hurwicz estimator

$\hat{a}^{\text{Hur}}(T)=\text{median}(I_2 / I_1, I_3 / I_2, \ldots, I_T / I_{T-1}) \quad (13)$

is median-unbiased, robust against any deviation from Gaussian distribution, including heavy tails as well as contamination with outliers. Moreover, it has been proved that innovations $w_n$ in (9) are not necessarily identically distributed. Luger (2005) has showed that the results mentioned above remain true under more general distributional assumptions, without assuming statistical independence. Thus, assumptions basic for the least squares theory which are very difficult to verify in practice or to prove theoretically, are not necessary for the median-unbiased estimator approach. The fact that ratio $I_{n+1} / I_n$ is the main part of coefficient $a$ in (5) (see above) and properties of this ratio as an unbiased estimator of a will allow us to use $I_{n+1} / I_n$ as the most local estimate which is based only on the two consecutive terms of time series. This locality will be important for us in the light of early detection requirements (epidemics and explosive behaviors in finance).
TIMELINESS OF DETECTION

Mohtashemi et al. (2006) propose the following classification of timeliness of detection (with 7-day time window and simulating outbreaks lasting 7 days): “early detection” is defined as detection during the first 3 days of outbreak, “intermediate detection” represents detection during the 4th and 5th days of outbreak, and “late detection” corresponds to detection in the last 2 days of outbreak. Using the SIR model with detection time window $T = 7$ and simulated outbreaks, Mohtashemi et al. (2006) report cumulative sensitivity of 10% on the very 1st day of the outbreak, 30% - on the 2nd day, 50% - on the 3rd day, 63% - on the 4th day, 75% - on the 5th day, and 78% and 87.5% - on the 6th and 7th days correspondingly. Thus, according to Mohtashemi et al. (2006) their SIR model combined with Least-Squares estimation provides 50% sensitivity at the early stage of the epidemics (first 3 days) and 75% sensitivity at the end of the intermediate detection stage. Of course, real outbreaks may last well beyond 7 and even 14 days. Mohtashemi et al. (2006) and Reis, Pagano and Mandl (2003) focused on the first few days because useful detection systems should be able to recognize outbreaks within that time frame.

COMPARISONS OF OLS AND MEDIAN-UNBIASED ESTIMATES

Combining the information above with theoretical results regarding formulas (10) – (13), the simulation results from Marriott and Pope (1954), Orcutt and Winokur (1969), Sawar (1978), Andrews (1993), and Provost and Sanjel (2005), and the outcomes of our simulations, we can conclude:

- As our AR(1) model is a linear approximation (and a good one) of the SIR model in Mohtashemi et al. (2006), we can expect a comparable performance for such short-term dynamics as flu-like epidemics (with seven-day simulated outbreaks). To achieve such a performance it is recommended to use a combination of estimates (10), (10') and (13) with bias-corrections (11), (13) and (11') not only for $T = 7$ but also for shorter time windows including $T = 2$. According to Reis, Pagano and Mandl (2003), time window $T$ should be at most 7 days, i.e. the authors assume the possibility of $T < 7$. Also, to have a non-zero sensitivity on the 1st day, we must use the ratio $I_{n+1}/I_n$ among other statistics. Note that for $T = 2$, both OLS estimator (10) and Hurwicz estimator (13) are equal to the ratio $I_{n+1}/I_n$. This is the only case when estimators (10) and (13) are equal and both represent the instant change on the recent day comparatively to the previous one. This is why the ratio $I_{n+1}/I_n$ should be an important component of our statistical toolbox. We will return to using shorter time windows later in the hypotheses testing section.

- We have compared the OLS estimate (10) with $\hat{a}_{Hur}$ estimate (13) using simulations. Our simulations results with 10,000 simulations for $T = 4, 5, 6, 7$ and 14, and $a = 0.7, 0.8, 0.9, 1.0, 1.1, 1.2$ and 1.47 (the last value is taken as a parameter of exponential growth for simulating outbreaks in Mohtashemi et al. (2006)) show that $\hat{a}_{Hur}(T)$ is much less biased and much less variable across the values of $T$ than its OLS counterpart (10) for all combinations of $T$ and $a$. Bias of all our estimates decreases with the increase in $a$ for $a > 1$. When we average our OLS estimates across simulations, using median rather than mean, both bias and variability dramatically decrease. Also, the results depend on the fact whether we apply some induction period with discarding some terms (in our case 50 terms) of simulated time series. We find that discarding can significantly improve both bias and variability. See also Anderson (1979).

- Actually, statistics (10), (10'), (11), (11'), (12) and (13) 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. In particular, statistic $\hat{a}_{Hur}(T)$ (with $T = 7$) being a robust and stable estimate is ideal for either stationary, pre-epidemic state or purely epidemic regime, but not for the periods including both pre-epidemic and epidemic days. For example, with $T = 7$ the estimate $\hat{a}_{Hur}(T)$ will likely not react to the first days of the epidemic (from the 1st to the 3rd day), so we can miss the early detection at all. At the same time, OLS estimate can potentially react to the very 1st day, though probably in a rather weak form. To have this reaction stronger we have to use smaller time window, including $T = 3$ and even $T = 2$. As noted above, $I_{n+1}/I_n$ is the only estimate capable of detecting change-point at the very first day.
Finally, when estimating parameter $a = 1 + \beta^* - \delta$ we actually estimate $\beta^* - \delta$, the difference between the transmission rate and recovery rate. This difference is a threshold parameter, closely related to another famous threshold parameter, $R_0 = \beta^* / \delta$, which is known in epidemiology as the basic reproductive ratio. Parameter $\beta^* - \delta$ can be called “an engine of the epidemic”. Also note that here we use $\beta^*$ as the transmission rate parameter, and not $\beta$. Another interesting observation is that the type of behavior (no-epidemic vs. an outbreak regime) is determined not by $\beta^*$ and $\delta$ separately, but by their difference $\beta^* - \delta$. In our approach we do not need to estimate or make some assumptions about $\beta^*$ and $\delta$ separately (for example, assume that $\delta = 1/7$ and then estimate $\beta^*$). We need to estimate only the difference $\beta^* - \delta$.

Also, it should be noticed that all the results on OLS and median-unbiased estimates of the first-order autoregressive parameter $a$ cannot be used alone for early detection purposes. But our further inference (confidence intervals and hypothesis testing) is heavily based on these estimates.

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. 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”. Also, we consider here only OLS estimates though generalizations are possible to other classes of estimates, including median-unbiased ones (in the spirit of Andrews (1993a) and Luger (2005)). We will return to confidence intervals for median-unbiased estimators elsewhere.

Even with bias correction discussed above, the point estimate of the form (10) 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:

- If $a < 1$, then the probability distribution of
  \[(\hat{a}(T) - a) \frac{T^{1/2}}{(1 - \hat{a}^2(T))^{1/2}}\]
is close to the Gaussian distribution $N(0,1)$ when the time window $T$ is large enough (see Giraitis and Phillips (2006) and Phillips and Han (2006)).

- If $a > 1$, then the probability distribution of
  \[(\hat{a}(T) - a) \frac{\hat{a}(T)}{(\hat{a}(T) - 1)}\]
is close to the standard Cauchy distribution $C$ when the time window $T$ is large enough (see Phillips, Wu and Yu (2007) and Phillips and Magdalinos (2007)).

As a result, we have the following two families of the approximate two-sided confidence intervals for parameter $a$:

- If $a < 1$ (a stationary case), the 100(1 - $\alpha$)% confidence interval for $a$ is given by the formula
  \[\left(\hat{a}(T) - ((1 - \hat{a}^2(T))^{1/2} / T^{1/2}) N_{\alpha}, \quad \hat{a}(T) + ((1 - \hat{a}^2(T))^{1/2} / T^{1/2}) N_{\alpha}\right)\]

(14)
where $N_\alpha$ is the two-tailed $\alpha$ percentile critical value of the standard Gaussian distribution. For 90, 95 and 99 percent confidence intervals, these critical values are as follows: $N_{0.10} = 1.645$, $N_{0.05} = 1.96$, $N_{0.01} = 2.576$. Note that for small sample sizes (e.g., $T = 7$), using confidence limits based on the Student’s $t$ distribution rather than Gaussian one, is more appropriate.

- If $a > 1$ (an explosive case), the 100(1 - $\alpha$)% confidence interval for $a$ is given by the formula

$$(\hat{a}(T) - (|\hat{a}(T)| - 1)) / \hat{a}(T) C_\alpha, \ \hat{a}(T) + (|\hat{a}(T)| - 1) / \hat{a}(T) C_\alpha)$$

(15)

where $C_\alpha$ is the two-tailed $\alpha$ percentile critical value of the standard Cauchy distribution. For 90, 95 and 99 percent confidence intervals, these critical values are as follows: $C_{0.10} = 6.315$, $C_{0.05} = 12.7$, $C_{0.01} = 63.657$ (see Phillips, Wu and Yu (2007)). Gaussian critical values are much smaller than the corresponding Cauchy ones, which is not surprising since the Cauchy distribution has heavier tails.

In the following hypothetical examples we demonstrate a very typical problem: confidence intervals in (14) are often too wide to be practical for the time window $T = 7$ adopted in Reis, Pagano and Mandl (2003) and Mohtashemi et al. (2006). For example, if $\hat{a}(T) = 1.25$ (with bias correction already taken into account and $\delta \approx 1/7$) which corresponds to a high value of the basic reproductive ratio: $R_0 \approx 2.75$, then the 90% two-sided confidence interval based on the Cauchy distribution is of the form: (1.25 - 0.36, 1.25 + 0.36). This interval contains the borderline value $a = 1$ and therefore we are in a “Grey Zone” and fail to detect a rather strong signal as a take-off of the epidemic. Now let us consider an example with $\hat{a}(T) = 1.5$ (with bias correction already taken into account and $\delta = 1/7$) which corresponds to $\beta^* \approx 0.643$ and $R_0 \approx 4.5$. It is an important example, since $\hat{a}(T) = 1.5$ is very close to the parameter of exponential growth for simulated outbreaks ($A = 1.47$) in Mohtashemi et al. (2006). Using 90 percent confidence interval (14) with $C_{0.10} = 6.315$, we obtain the interval (1.5 - 0.46, 1.5 + 0.46) which does not contain $a = 1$ and thus we have to reject stationarity or a no-outbreak regime. But this rejection (with 10% significance level) is almost borderline. To improve the situation we can build a one-sided confidence interval, instead of the typically recommended two-sided one. As a result, we arrive at the 95 % confidence interval (1.5 - 0.46, + $\infty$) that does not cover $a = 1$. Of course, changing significance level from $\alpha = 10\%$ to $\alpha = 5\%$ makes us psychologically more ready to reject stationarity in favor of an outbreak. But what we did can be considered a kind of statistical trick to get a more or less tolerable (in terms of the level of confidence) confidence interval. We can add to this that early detection in syndromic surveillance is inherently multiple testing problem since estimating, testing and making decisions are performed daily. Thus, significance levels used in testing procedures cannot be taken at face value. Moreover, they could be absolutely misleading. Note that most probably our early detection problem with $T = 7$ (or even less) should not be considered in terms of statistical significance only, practical significance can be even more important.

### 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 premises 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 bias in this situation. To avoid it we propose two alternatives: mid-sample and end-of-sample tests, and their combination.

**Mid-sample tests.** These tests also known as supremum tests. See Andrews (1993b) for a general case, and Diebold and Chen (1996) who focus on dynamics models, in particular AR(1) model, and references therein. To serve both applications of interest – epidemiological and financial, we are choosing more “neutral” notations for our basic model

$$Y_{t+1} = aY_t + e_t \quad t = 1,2,..., T \quad (9'')$$
Here, $Y_t$ is used instead of $I_t$, and index $t$ – instead of $n$, parameter $a$ is not necessarily equal to $1 + \beta^* - \delta$. Time $T$ (which is equal to our sample size) corresponds to the most recent observation and time $t = 1$ corresponds to the earliest available observation. Every day, our time window is shifted to the right. Each day, based on the sample available, we have to decide whether our sample is homogeneous or we have a breakpoint in the time window. It is natural to assume that $T = 7$ or $T = 14$ for the epidemic outbreak setting, and $T = 14$ or $T = 28$ in the financial context. Let us remind that we have a 7-day periodicity in both epidemiological and financial applications.

The null hypothesis of structural stability (parameter $a$ is unchanged in the time window) is expressed by equation (9''). The alternative hypothesis of a one-time structural break is

$$
Y_{t+1} = a_1 Y_t + e_t \quad t = 1, \ldots, T^* , \\
Y_{t+1} = a_2 Y_t + e_t \quad t = T^* + 1, \ldots, T ,
$$

(16)

where $1 < T^* < T$, i.e. a hypothetical breakpoint is separated from the beginning and end of the time window. In the case of early detection of epidemic outbreaks and stock market bubbles, it is reasonable to assume that $a_1 < 1$ and $a_2 > 1$. Also, it is assumed typically that $T^*$ is separated from both ends of the time window by 5% to 15% of the sample size (that is why the term "mid-sample"). In our case with 5% separation and $T = 7$ or 14, it means that $2 \leq T^* \leq 6$ or $2 \leq T^* \leq 13$ correspondingly. Our basic test statistics are the "supremum" statistics of Andrews (1993b), $Sup-W$ and $Sup-LR$:

$$
Sup-W = \max_T \left[ T \left( \sum_{t=1}^T \hat{e}_t^2 - \sum_{t=T+1}^{T^*} \hat{e}_t^2 - \sum_{t=T+1}^{T^*} \hat{e}_{2t}^2 \right) / \left( \sum_{t=1}^{T^*} \hat{e}^2_t + \sum_{t=T+1}^{T} \hat{e}_{2t}^2 \right) \right]
$$

$$
Sup-LR = \max_T \left[ T \left( \sum_{t=1}^T \hat{e}_t^2 \right) / \left( \sum_{t=1}^{T^*} \hat{e}^2_t + \sum_{t=T+1}^{T} \hat{e}_{2t}^2 \right) \right]
$$

(17)

where $\hat{e}_t$, $\hat{e}_1$, and $\hat{e}_{2t}$ are OLS residuals for the whole time window $1, \ldots, T$ and subsamples $1, \ldots, T^*$ and $T^* + 1, \ldots, T$ correspondingly. These residuals are calculated by using OLS estimates of the first-order autoregressive parameters $a$, $a_1$ and $a_2$ or their median-unbiased alternatives.

Taking into account a very simple relationship between Wald statistic $W$ and likelihood ratio statistic $LR$

$$
LR = T \log(1 + W / T),
$$

(18)

we can conclude that both supremum tests are equivalent. Note also that some variant of the Sup-LR test underlies the space-time detection approach by Kulldorff et al. (2005).

Andrews (1993b) used an asymptotic approach based on the assumption that subsample sizes before and after the breakpoint are both very large (theoretically tend to infinity). Of course, this assumption is unacceptable in our situation with $T = 7$, 14 or even 28. Focusing on smaller samples, Diebold and Chen (1996) proposed the bootstrap approximation to the finite-sample distribution which appeared consistently accurate, in contrast to the Andrews asymptotic approximation. In this paper we follow the bootstrap approach of Diebold and Chen (1996).

**Bootstrap approximation.** The bootstrap formalizes a very intuitive idea. Using observed data, we calculate our test statistic ($Sup-W$ or $Sup-LR$). Then we generate many pseudo-data samples using parameter $a$ estimated under the null hypothesis of structural stability (this is either OLS estimate $\hat{a}(T)$ or median-unbiased estimate $\hat{a}_{Hur}(T)$) and pseudo-disturbances drawn with replacement from estimated model’s residuals, and then calculate the statistic for each sample of
pseudo-data. By using the “bootstrap distribution” built in this way, the approximate p-value can be obtained. This procedure is called a “nonparametric bootstrap”. No distributional assumptions are made in it. According to Diebold and Chen (1996), the bootstrap approximation outperforms the asymptotic approximation for sample sizes as small as $T = 5$ and 10. As shown earlier, OLS estimates are downward biased, especially for small $T$, and median-unbiased estimates $\hat{a}_{\text{Hur}}$ can be very useful. Note that according to formulas (17) in calculating Sup-$W$ and Sup-$LR$ statistics we inevitably have to work with estimates $\hat{a}_{\text{OLS}}$ and $\hat{a}_{\text{Hur}}$ for very small subsamples of time window $T$.

**End-of-sample or S-tests.** As mentioned above mid-sample tests are designed to check whether the breakpoint of interest happened in the middle of the time window, i.e. before the most recent day $T$, and consequently we can miss the very first day of the epidemic or the emerging bubble. To detect the very beginning of the regime switch, we can use the so-called end-of-sample tests, introduced by Andrews (2003) in the econometric context. Unlike the Andrews (1993b) approach, the basic test statistic does not compare parameters estimate before and after the breakpoint. Instead, it compares full-sample estimates of the residual variance to the size of the residual at the very end of the sample:

$$ S = T\hat{\theta}_{T}^{2} / \left( \sum_{t=1}^{T} \hat{\theta}_{t}^{2} \right) $$

In formula (19), the end of the sample is considered as one, the most recent day. If the end of the sample includes more than one day which is appropriate for larger time windows (for example, $T = 28$), the formula for the S-test is more complicated and is not given here. To calculate p-values we can use a bootstrap approach exactly as for mid-sample tests.

Finally, some remarks about mid-sample and end-of-sample tests should be mentioned:

- The reader can easily recognize some variants of the classic F-test in mid-sample and end-of-sample tests.
- Since mid-sample and end-of-sample tests are applied to different parts of the time window, it is advisable to use a combination of both (see flow-chart below).
- OLS and / or median-unbiased estimates of the first-order autoregressive parameter discussed above are integral parts (though in hidden form) of test statistics (17) and (19).
- Implementation of mid-sample and end-of-sample tests requires each day multiple using of PROC ARIMA with ESTIMATE statement which provides the estimates of the first-order autoregressive parameter $a$ and residuals $\hat{\theta}_{t}$, $\hat{\theta}_{\text{Hur}}$ and $\hat{\theta}_{2t}$ needed for calculating test statistics (17) and (19). For mid-sample tests, all calculations can be performed for each day in a loop with index $T^* (1 < T^* < T)$. SAS code for these tests is currently in preparation.

**APPLICATIONS TO ECONOMIC AND FINANCIAL TIME SERIES: SIR, AR(1) MODELS, STOCK MARKET BUBBLES AND MICRO-BUBBLES**

Although the presence of bubbles and herding behavior in financial markets is widely accepted there are still few theoretical models which generate such phenomena. For example, the paper by Phillips, Wu and Yu (2007) being a fundamental research article on modeling bubble formation, uses AR(1) as an ad-hoc, purely empirical model, without any explanation of why AR models, in particular AR(1), should be used. In this connection, it is interesting to mention the paper by Shive (2006) in which the author used a SIR model as an epidemic model of investor behavior that results in building a stock market bubble. Shive (2006) reports periods of significant epidemic behavior in buying and selling that correspond to the dramatic price movements in some stocks in Finland from 12/1994 to 1/2004. The author refers to Robert Shiller (2005) in which a speculative bubble is defined “as a situation in which news of price increases spurs investor enthusiasm, which spreads by psychological contagion from person to person”. We can add to this that Shiller
is world famous by predicting the year 2000 stock market crash. Also, he is one of the first econometricians to describe a new bubble in the making – the housing bubble.

As in Shive (2006), global financial bubbles (like NASDAQ bubbles, industry bubbles, etc.) are not so important for us. They are mostly of academic, post factum or posterior interest. Our interest is more practical and limited to micro-bubbles which are related to individual stocks and for shorter time span. Typically, time is measured in months / quarters for global bubbles and in days or even intra-day units for micro-bubbles. Global bubbles usually belong to low-frequency phenomena, and micro-bubbles are high-frequency processes. As additional reading, see also Abreu and Brunnermeier (2003) for a comprehensive paper on bubbles and crashes, Brunnermeier and Pedersen (2002) for an interesting work on predatory trading, and Brunnermeier and Nagel (2004) for a very influential paper on hedge funds and technology bubbles. In Brunnermeier and Pedersen (2002), the authors focus more on higher frequency phenomena, including micro-bubbles rather than global bubbles.

With psychological and physiological contagion as a common and natural justification for using SIR models and their linear approximations, AR(1) models (both) in epidemiology and finance, both applications share model equations (9) or (9'), least-squares estimates (10) or (10'), median-unbiased estimate (13), confidence intervals (14) or (15), and mid-sample and end-of-sample tests (17) and (19). The real strength of the approach with using SIR and then AR(1) models in financial applications is the straightforward, almost clinical way in which it explains how and why financial bubbles happen as they do. Still, we can conclude that using autoregressive models in describing explosive phenomena seems to be in an emerging state.

**WORKFLOW FOR EARLY DETECTING EXPLOSIVE BEHAVIORS**

We suggest the following workflow for early detecting explosive behaviors:

1) Each day, the ordinary least-square (OLS) estimate (10) and median-unbiased estimate (13) of the only autoregression parameter \(a\) are iteratively calculated, based on the past \(T\) values of the time series \(l_i\) or \(Y_t\) (the present date value is included). The changes in the time index are obvious. Note that we have to calculate estimate not only for the whole time window, but for all subsamples \((1,...,T^*)\) and \((T^*+1,..., T)\) where \(2 \leq T^* \leq T - 1\), as it is required in mid-sample tests (17). Also, we calculate all residuals \(\hat{e}_t, \hat{e}_t^{fr}, \hat{e}_t^{mr}\) and eventually test statistics \(\text{Sup-W} or \text{Sup-LR}\). Day \(T^*\) that provides maximum to the test statistic is a breakpoint candidate and is subject to further scrutiny. Mid-sample tests (17) allow us to check whether the outbreak has happened in the middle of the time window and we have already missed the beginning.

↓

2) Using bootstrap we calculate the empirical \(p\)-value for the mid-sample test. If this bootstrap \(p\)-value is smaller than a nominal test size \(\alpha\) (usually \(\alpha = 1\% or 5\% or 10\%), and is subject to our choice) then we reject the null hypothesis of stability in the middle of the time window and declare day \(T^*\) (which maximizes the test statistic) as a change-point with significance level of \(\alpha\). Of course, we can scrutinize this conclusion.

↓

3) If the null hypothesis cannot be rejected, then we can either apply the confidence interval approach discussed above (see formulas (14) and (15)) or use the end-of-sample test (19). In both cases, bias correction is advisable. Bias correction is performed by using either formulas (11), (11') and (12) or a modified least-squares estimator given by (10'). Comparing OLS and median-unbiased estimates (13) can
be useful here. (The) end-of-sample test (19) helps us to check whether the most recent day is the beginning of (the) outbreak.

↓

4) If statistical tests (17) and (19) are not convincing in terms of statistical significance ($p$-value) and / or practical significance (break size), we can still turn to the empirical rules described in Shtatland and Shtatland (2008). Also, using higher-order moment estimates (Huzii (1981)) and a partial correlation approach (Anderson (1990)) is a possibility.

SAS code to implement steps 1 - 4 is work in progress.

SUMMARY, CONCLUSIONS, AND FUTURE WORK

In this paper, we propose a simple but powerful and practical model for early detection of epidemic outbreaks and stock market bubbles for individual stocks. As mentioned above, it seems unusual that these two seemingly unrelated phenomena can be combined under one umbrella – SIR models. 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 SIR models and AR(1) processes as their natural local approximation. First, based on a SIR model, a workhorse of epidemiology, we show that in the autoregressive approximation we can limit ourselves with the models of the second or even first order. Then, by using methods of linearization, we conclude that the most appropriate local model is AR(1). The autoregression coefficient of this model has a very simple epidemiological meaning and can be used as a natural threshold parameter in epidemic detection. To the best of our knowledge, this paper and Shtatland (2007), Shtatland and Shtatland (2008) are the first publications that report this threshold parameter for early detection in biosurveillance applications. We can consider our AR(1) model as a paradigm-model for early detection of epidemic outbreaks and financial bubbles. Note that AR(1) is already widely used as a paradigm-model in the unit root theory. The proposed detection approach is based on a combination of the least-squares and median-unbiased estimators of the autoregressive coefficient, approximate two-sided and one-sided confidence intervals (built by using the Gaussian and Cauchy limit distributions), and mid-sample and end-of-sample statistical tests. Note that typically ARMA processes are considered synonyms to the stationary models, and the explosive case is ignored as a nuisance. Here AR(1) models with the autoregressive coefficient greater than one play a major role. Note that there is an important difference between our approach and that of Mohtashemi et al. (2006), which is worth mentioning. Mohtashemi et al. (2006) heavily use historical data (more than 4389 days), and we do not ($T = 7,14$ days or less). Our approach can be preferred when historical data are either not available at all or are unrepresentative of the current behavior for a variety of reasons: changes in treatments, coding, and reporting practices, appearance of new influenza strains, etc. (See more about problems with using historical data in Burkom, Murphy and Shmueli (2007)). Also, our approach can be used not only in epidemiological and financial time series but in any application where we are interested in detection of abrupt switches from the normal, stationary regime to the exponential growth. Models of the form (9), (9') and (9'') seem well suited to capturing the essential features of epidemic, economic, financial and other processes that can undergo mildly explosive behavior ( stock market bubbles for individual stocks is a good example).

In our future work we are planning to investigate relationships between our approach to early detection of structural changes and intervention analysis / interrupted time series. In this paper, we work with endogenous variables only: either the numbers of infected individuals for epidemic outbreaks or the numbers of investors, transactions (and related volume, stock prices as well), etc., in the stock market context. In intervention analysis, some exogenous information can be used through input time series containing discrete-value variables that flag the occurrence of an
event affecting our main response series. This event is an interruption of the normal evolution of the response time series, which in absence of interruption is usually assumed to be a pure ARMA process. Intervention models are typically used to analyze the effect of the intervention. Intervention effects are usually associated with some new regulations, laws, provisions, etc. Our interest in intervention analysis is different: we would like to combine our endogenous information with exogenous knowledge provided by some input time series. It could be, for example, information about vaccination in the epidemiological context or some insider information in the sock market bubble setting, etc. This usage of additional, exogenous information can be performed within a Bayesian approach. It can be particularly important in making decisions in borderline situations which we called “Grey Zone” in Shtatland and Shtatland (2008). Also, it will be of obvious interest in future work to compare the power of the bootstrapped versions of $\text{Sup-W}$ or $\text{Sup-LR}$ tests with particular attention paid to location of the break and “distance” of the alternative from the null.

REFERENCES

Le Breton, A. and Pham, D. T. (1989). On the bias of the least squares estimator for the first


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
tel: (781) 662-9578
email: eshtatland@yahoo.com