Application of an Integer-Valued Autoregressive Model to Hit Phenomena

Yasuko Kawahata
The University of Tokyo:
Graduate School of Information Science and Technology
Hongo, Bunkyo-ku, Tokyo, Japan
Tottori University:
Graduate School Faculty of Engineering
Koyama, Tottori, Tottori, Japan
Email: purplemukadesan@gmail.com

Tamio KOYAMA
Shiga University:
The Center for Data Science Edutation and Research
Banba, Hikone, Shiga, Japan
Email: tamio-koyama@biwako.shiga-u.ac.jp

***Abstract*—We propose a new model for hit phenomena. Our model is based on the Integer-Valued autoregressive model in form of a stochastic difference equation, and it describes behaviors of count data sequences. Utilizing our model, we give a theoretical formulation of the concept "hit", and a systematic method deciding whether given time series count data contains "hit".**

I. Introduction

The Intentions of humans living in society can be measured by using social network systems. In this paper, we focus our attention on the decay of intention in societies that appears in the time variation of the number of blog or twitter posts per day. In psychology, it is well-known that a forgetting curve has exponential form [1]. On the other hand, Crane and Sornette found that the forgetting curve jas a power-law form from other experiments [2]. For Japanese social media, Sano et al. found that power laws generally approximate the functional forms of growth and decay with various exponents values between -0.1 and -2.5 [3], [4]. Contrarily, in formulating a mathematical model for hit phenomena [5], Ishii et al. indicated that the decay of the reputation of a movie is exponential using the observed data concerning the reputation of 25 movies on blogs. This exponential decay is built into the model. In [6], some count-series data concerned with social scandals scandals were found to have neither an exponential form nor a power law. In order to analyze such data, a model combining both functions was introduced.

All of these models for hit phenomena are continuous and described by ordinary differential equations. We are interested in a discrete analog of these models. In this paper, we focus on the case of the exponential decay and apply the integer-valued autoregressive (INAR) model to hit phenomena. This model is a basic tool for time series analysis. A standard autoregressive model has the form of a stochastic difference equation. Each element in a sequence of the standard autoregressive model takes a real number. The INAR model is an analogy of the model for count-data sequences; thus, any element in a sequence of this model takes a non-negative integer value. For the basic properties of INAR models, see [7] and [8].

The concept of 'hit' can be considered to be an explosive growth of an index that express intentions of humans in society. Counts of posting of blogs and sales of certain goods are examples of such indices. Our model is a stochastic difference equation that describes a low which such phenomena follows. We utilize this model to mathematically formulate the concept of a "hit".

The organization of this paper is as follows. In Section II, we review basic results concerning the INAR model. In Section III, we introduce a new model for hit phenomena based upon the INAR model. Estimation of parameters in the model is also discussed. Utilizing this model, we obtain a mathematical formulation of the notion of a hit. After a theoretical discussion, we give an example with fictitious data generated by a computer. In Section IV, we apply our method to data in the real world.

II. Review of the INAR Model

In this section, we review the basic properties of the INAR model. For the standard AR model, basic results include the theorem for the stationarity condition and the Yule–Walker equation. Analogies of these results hold in the case of the INAR model.

A. Definition

Before explaining the INAR model, let us recall the well-known AR model. The AR model of order p is a sequence of random variables $(X_t)_{t\in\mathbf{Z}}$ that satisfy a stochastic difference equation:

$$X_t = \sum_{i=1}^{p} c_i X_{t-i} + \varepsilon_t \tag{1}$$

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

where c_i is a real number and ε_t is a random variable with the standard normal distribution. [1]

The INAR model is an analogy of the AR model for the count-data sequences introduced in [7]. Since each element of a count-data sequence takes a non-negative integer value, the stochastic difference equation (1) does not work well. In fact, a linear combination of the elements of a count data sequence with real number coefficients can easily take a non-integer value. For this reason, the INAR model does not adopt a linear combination of random variables in the past history of the sequence.

Let $r \in \mathbf{R}^p$ be parameters. We assume that $0 \leq r_i \leq 1\,(i = 1, \ldots, m)$, $\sum_{i=1}^{p} r_i \leq 1$. For a non-negative integer-valued random variable X, let $(r_1 * X_t, r_2 * X_t, \ldots, r_p * X_t)$ be a vector of random variables such that

$$\mathbf{P}\left(r_i * X_i = k_i, 1 \leq i \leq p | X = x\right) = \frac{x!}{k_1! \cdots k_p!} r_1^{k_1} \cdots r_p^{k_p}$$

where $k_1 + \cdots + k_p = x$ and $k_i \in \mathbf{Z}_{\geq 0}\,(1 \leq i \leq p)$. Note that $r_i * X$ denotes a new random variable rather than the product of r_i and X. The *INAR model of order* p is a sequence of random variables X_t satisfying the following stochastic difference equation:

$$X_t := Y_t + \sum_{i=1}^{p} r_i * X_{t-i} \quad (t \in \mathbf{Z}) \tag{2}$$

where $\{Y_t\}_{t \in \mathbf{Z}}$ is a sequence of non-negative integer-valued random variables [2]

B. Yule–Walker Equation

In the AR model, the method of least squares for the estimation of the coefficient in (1) derives the Yule–Walker equation. An alogous argument is successful in the case of the INAR model. Here we shortly review a discussion of the Yule-Walker equation for the INAR model.

To estimate the parameters of the INAR model, we minimize the following sum of squared errors:

$$\sum_{t=p+1}^{N} \left(X_t - E\left(X_t | X_s = x_s, t-1 \geq s \geq t-p\right)\right)^2. \tag{3}$$

When we have an outcome $X_t = x_t\,(t = 1, \ldots, N)$, the sum of the squared errors (3) can be written as

$$L(\mu_y, r_1, \ldots, r_p) := \sum_{t=p+1}^{N} \left(x_t - \mu_y - \sum_{i=1}^{p} r_i x_{t-i}\right)^2$$

where we set $\mu_y := E\left(Y_t\right)$. Since the derivatives of L can be written as

$$\frac{\partial L}{\partial \mu_y} = -2(N-p)\left(\bar{x}_0 - \mu_y - \sum_{j=1}^{p} \bar{x}_j r_j\right),$$

$$\frac{\partial L}{\partial r_i} = -2(N-p)\left(\gamma_{i0} - \bar{x}_i \mu_y - \sum_{j=1}^{p} \gamma_{ij} r_j\right),$$

the minimum point of L satisfies the equation

$$\begin{pmatrix} \bar{x}_0 \\ \gamma_{i0} \\ \vdots \\ \gamma_{p0} \end{pmatrix} . = \begin{pmatrix} 1 & \bar{x}_1 & \cdots & \bar{x}_p \\ \bar{x}_1 & \gamma_{11} & \cdots & \gamma_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_p & \gamma_{p1} & \cdots & \gamma_{pp} \end{pmatrix} \begin{pmatrix} \mu_y \\ r_1 \\ \vdots \\ r_p \end{pmatrix} \tag{4}$$

Here we substitute

$$\bar{x}_i := \frac{1}{N-p} \sum_{t=p+1}^{N} x_{t-i},$$

$$\gamma_{ij} := \frac{1}{N-p} \sum_{t=p+1}^{N} x_{t-i} x_{t-j}.$$

Equation (4) is the Yule–Walker equation for the INAR model.

III. Model for Hit Phenomena

A. Model

In this section, we apply the INAR model to hit phenomena. Since power laws are very important in the analysis of social data, we assume that the increments of the time series obey the Pareto distribution. Recall the definition of the INAR model (2). According to this definition, we only assume that Y_t is a non-negative integer-valued random variable. Hence, we can take any distribution with support $\mathbf{Z}_{\geq 0}$ for the distribution of Y_t. In order to describe the power law in the increment of count-data sequances, we assume that

$$\mathbf{P}\left(Y_t = k\right) = \int_0^{\infty} \frac{\lambda_t^k}{k!} e^{-\lambda_t} \frac{\alpha}{(\lambda_t + 1)^{\alpha+1}} d\lambda_t, \tag{5}$$

where $\alpha > 0$ is a parameter. Form view point of Bayesian statistics, the variable Y_t is a Poisson random variable with parameter λ_t, and the prior distribution of the parameter λ_t is Pareto with parameter α. Hence, our INAR model for hit phenomena can be written as the following stochastic difference equation:

$$X_t := Y_t + \sum_{i=1}^{m} \beta_i * X_{t-i} \quad (t \in \mathbf{Z}).$$

Here the distribution of Y_t is defined by (5), and α and $\beta = (\beta_1, \ldots, \beta_m)^{\top}$ are the parameters of this model.

[1] We also need the following technical assumption: the random variable ε_t is independent from the σ-algebra $\sigma\left(X_s,\, s < t\right)$ generated by the random variables $X_s (s < t)$. For an explanation of σ-algebra and other basic information concerning probability theory, see [9].

[2] We also assume that $r_i * X_{t-i}$ is independent from $\sigma\left(X_s,\, s < t - i\right)$ and that Y_t is independent from $\sigma\left(X_s,\, s < t\right)$.

B. Estimation

Suppose that the outcome $X_t = x_t$ and the parameter $\alpha \in \mathbf{R}$, $\beta \in \mathbf{R}^p$ are given. Solving the Yule–Walker equation (4), we can estimate $\mu_y = \mathbf{E}(Y_t)$. Since the parameter α can be written as

$$\alpha = 1 + \frac{1}{\mathbf{E}(Y_t)},$$

we can estimate parameters α and β.

In order to estimate the value of each Y_t, we utilize the Bayesian statistics. We assume that the random variable λ_t with Pareto distribution is the parameter of the Poisson random variable Y_t. From the view point of Bayesian statistics, the prior probability of λ_t is $P(\lambda_t) = \alpha/(\lambda_t + 1)^{\alpha+1}$ and the posterior distribution of λ_t is

$$\begin{aligned} &P(\lambda_t | x_t, \ldots, x_{t-m}) \\ &= \frac{P(x_t | \lambda_t, x_{t-1}, \ldots, x_{t-m}) P(\lambda_t | x_{t-1}, \ldots, x_{t-m})}{P(x_t | x_{t-1}, \ldots, x_{t-m})}. \end{aligned}$$

Here we substitute

$$\begin{aligned} &P(x_t | \lambda_t, x_{t-1}, \ldots, x_{t-m}) \\ &= \sum_{(k_0,\ldots,k_m)\in K} \frac{\lambda_t^{k_0}}{k_0!} e^{-\lambda_t} \prod_{i=1}^{p} \binom{x_{t-i}}{k_i} r_i^{k_i} (1-r_i)^{x_{t-i}-k_i} \\ &\quad \left(K := \left\{ (k_0, \ldots, k_m) \middle| \begin{array}{l} 0 \le k_i \le x_{t-i}, \\ k_0 + \cdots + k_m = x_t \end{array} \right\} \right), \\ &P(\lambda_t | x_{t-1}, \ldots, x_{t-m}) = \frac{\alpha}{(\lambda_t + 1)^{\alpha+1}}, \\ &P(x_t | x_{t-1}, \ldots, x_{t-m}) \\ &= \int_0^\infty P(x_t | \lambda_t, x_{t-1}, \ldots, x_{t-m}) \frac{\alpha}{(\lambda_t + 1)^{\alpha+1}} d\lambda_t. \end{aligned}$$

In Section IV, we demonstrate that applying the Bayesian method obtains good accuracy, however, it requires a long computation time. Hence, we consider an alternative method with low computational complexity. For the estimation of λ_t, we utilize the maximum likelihood method. Since the binomial distribution can be approximated by a Poisson distribution, we can approximate the distribution of X_t by the Poisson distribution of the mean $\lambda_t := a_t + \sum_{i=1}^{p} \beta_i E(X_{t-i})$. Hence, the likelihood function can be written as

$$\prod_t \frac{\lambda_t^{x_t}}{x_t!} e^{-\lambda_t} \frac{\alpha}{(a_t - 1)^{\alpha+1}}$$

Since the derivative of its logarithm with respect to a_t is

$$\frac{x_t}{\lambda_t} - 1 - \frac{\alpha + 1}{a_t - 1},$$

we can easily find which the a_t value that maximizes the likelihood.

C. Formulation of a "hit"

In our model, we assume that the increment of count-data sequences follows a Pareto distribution. By the method described in the previous subsection, the parameter α of the Pareto distribution and the parameter λ_t of the Poisson random variable Y_t can be estimated. We denote by λ_t' the estimated value of λ_t. The probability that the parameter λ_t is greater than the estimated value λ_t' can be estimated as

$$\mathbf{P}(\lambda_t > \lambda_t') = \frac{1}{(1 + \lambda_t')^{\alpha}}.$$

When the probability is less than $q\,(0 \le q \le 1)$, we say that *a hit of level q occurred at t.* This is our formulation of the notion of a hit.

D. Example

In order to describe how our model and method work, we generate a fictitious count-data sequence with a computer and apply our method to the data.

We generate a fictitious count-data sequence $(x_1, \ldots, x_N)$ with length $N = 10'000$ in the case where $m = 2$ and

$$\alpha = 2.001, \qquad \begin{pmatrix} r_1 & r_2 \end{pmatrix}^\top = \begin{pmatrix} 0.8 & 0.1 \end{pmatrix}^\top. \tag{6}$$

In Figure 1, we show the values of the count-data sequence form $t = 1$ to $t = 500$. The black line shows the values of x_t and the red line shows those of y_t, which is an outcome of Y_t.

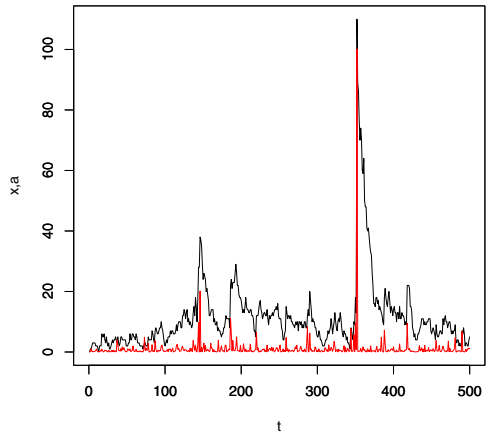


Figure 1. An example of Poisson AR model

This fictitious count-data sequence determines each coefficient of the Yule–Walker equation (4), and we solve it in the cases where $N = 100, 200, \ldots, 10000$. Figure 2 shows the result. The left graph in Figure 2 shows the estimated values of α, and the right shows those of β. The black and red lines show the values of β_1 and β_2 respectively. As the length of the count-data sequence increases, the estimated

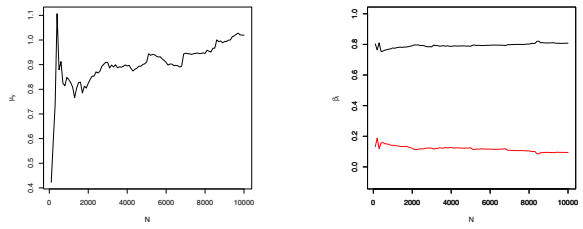


Figure 2. Parameter estimation for α and β

values of parameters converge to the true value. Figure 3 shows the estimation result of a_t. The black line shows the expectation value of the posterior distribution of a_t and the red line shows the true value of a_t.

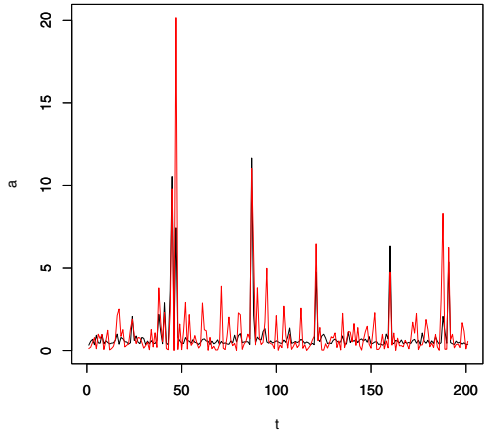


Figure 3. Estimation for a_t

IV. Application of The INAR Model to Real Data

In this section, we apply our model and method to data in the real world. In particular, we use the audience-rating data provided by Video Research Ltd. (a Japanese market-research company) for 42 prefectures in Japan, omitting only 5 provinces for reasons that will be described later). These data include audience survey of television programs as well as media research, including listening-rate surveys of the radio programs. By cooperation from this company, this data is views of the outflow of per minute in the TV news of the day, have taken up continues, the inflow number. Figure 4 shows count sequences of the data. The data comprises of four items (all,keep,in, and out), and the items have the following relations at each time k:

$$\begin{aligned} \mathrm{all}[k] &= \mathrm{keep}[k] + \mathrm{in}[k], \\ \mathrm{keep}[k] &= \mathrm{all}[k-1] - \mathrm{out}[k]. \end{aligned}$$

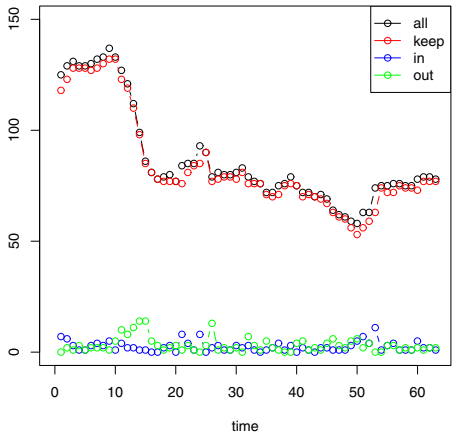


Figure 4. Count data sequences of tvw.csv

Solving the Yule–Walker equation (4) in the case where $1 \leq m \leq 3$, we can estimate the parameters α and β. Table I shows this result.

p	α	$\mathbf{E}(Y_t)$	β_1	β_2	β_3
1	1.375735	2.661453	0.960710	-	-
2	1.295989	3.378499	1.355874	-0.401148	-
3	1.254385	3.931049	1.282080	-0.208174	-0.125711
4	1.230873	4.331384	1.257210	-0.254269	0.127991

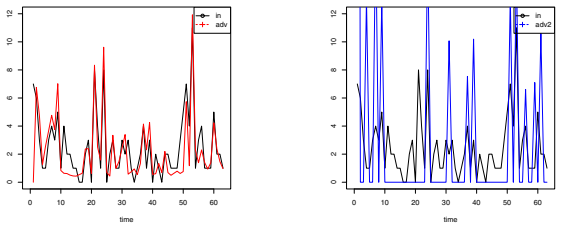
Table I
ESTIMATED VALUES OF α AND β

Since we assumed that $0 \leq \beta_i \leq 1$, the estimated values are not suitable except in the case where $p = 1$. On the other hand, the mean of the item in is 2.507937, and the estimated value of $\mathbf{E}(Y_t)$ should be near to this value. We confirm that the estimated value is nearest to the mean when $p = 1$.

In order to estimate the "advertisement effect," we apply the Bayesian analysis discussed in Subsection III-B. Figure 5 (a) shows the result of the Bayesian analysis with $p = 1$, $\alpha = 1.375735$ and $\beta = (0.960710)$, which are the results of the Yule–Walker equation. Figure 5(b) shows the maximal points of the approximation of the likelihood function. The black lines in both figures show the count series of the item "in." The expectations of posterior distribution for a_t estimate the value of "in" better than the approximated likelihood function.

References

[1] E. Eddinghaus, *Memory: A Contribution to Experimental Psychology*. New York: Teachers College, Columbia University, 1913.



(a) (b)

Figure 5. Estimation of the "advertisement effect"

[2] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," in *Proceedings of the National Academy of Sciences*, V. I. Keilis-Borok, Ed., vol. 105, 2008, pp. 15 649–15 653.

[3] Y. Sano, K. Yamada, H. Watanabe, H. Takayasu, and M. Takayasu, "Empirical analysis of collective human behavior for extraordinary events in the blogosphere," *Physical Review E*, vol. 87, p. 012805, January 2013.

[4] Y. Sano, "Empirical analysis and modeling of word frequency time series in social media," Ph.D. dissertation, Tokyo Institute of Technology, 2013.

[5] A. Ishii, H. Arakaki, N. Matsuda, S. Umemura, T. Urushidani, N. Yamagata, and N. Yoshida, "The 'hit' phenomenon: a mathematical model of human dynamics interactions as s stochastic process," *New Journal of Physics*, vol. 14, p. 063018, June 2012.

[6] A. Ishii and T. Koyabu, "Analysis of behavior of attenuation of social memories on movie and social scandal using sociophysics approach," in *Proceedings of the 47th ISCIE International Symposium on Stochastic Systems Theory and Its Applications*, December 2015, pp. 204–209.

[7] E. McKenzie, "Some arma models for dependent sequences of poisson counts," *Advances in Applied Probability*, vol. 20, no. 4, pp. 822–835, December 1988.

[8] A. A. Alzaid and M. Al-Osh, "An integer-valued pth-order autoregressive structure (inar(p)) process," *Journal of Applied Probability*, vol. 27, no. 2, pp. 314–324, January 1990.

[9] D. Williams, *Probability with Martingales*. Yew York: Cambridge University Press, 1991.