BULLETIN OF THE
INSTITUTE OF MATHEMATICS
ACADEMIA SINICA
Volume 18, Number 1, March 1990

ONE DISCRETE TIME SERIES MODEL FOR FAT-TAILED INTEGER RANDOM VARIABLES: ZIPF PROCESS

BY

HSIAW-CHAN YEH(葉小蓁)

Abstract. A Markov chain Z.AR(1) with Zipf(III) distributed inputs analogous to those of Jacobs and Lewis (1978) and Yeh (1983) is defined and its properties are developed. It is shown by induction that the marginal stationary distributions are Zipf(III)variables. Also, we have derived the exact and asymptotic distributions of the extreme order statistics. Further, the Z.AR(1) process is shown to be closed under geometric minimization and maximization. Statistical inference for the Z.AR(1) process is developed. A stationary one-dependent moving average, Z.MA(1) process, is similarly constructed. It also has a Zipf(III)marginal distribution. Finally, a simple extension to mixed Z.ARMA(1,1) process and its correlational structure is briefly introduced.

1. Introduction. A wide variety of socio-economic integer variables such as the size of business firms, discrete income, have distributions which are fat upper tailed and reasonably well fitted by Zipf distributions. Additionally, many such variables are repeated observed over time and they form a disctete time series data.

The motivation behind the Zipf process is to provide a discrete time series model with which to analyze the fat tailed stationary sequence of dependent discrete random variables with Zipf marginal distribution.

Received by the editors August 25, 1988 and in revised form July 31, 1989.

AMS subject classification: 60C05

Key words: Markov chain, Zipf(III) variables, runs, transition matrix, autoregressive processes, extreme order statistics, Z.MA(1), Z.ARMA(1,1) processes.

The Zipf distribution is a discrete type of Pareto distribution (Arnold (1983)). In Yeh(1983), I have studied the Pareto processes for modelling continuous socio-economic time series with Pareto marginals. In this paper I develop and study an autoregressive Zipf process. It is a Markov chain. In addition I describe certain natural extensions to a one-dependent moving average and mixed Zipf ARMA model. Jacobs and Lewis (1978) described discrete time series models with geometric or Poisson marginals. The Zipf proecesses herein described are analogous to Jacobs and Lewis models with maximization playing the role played by addition in their models.

2. The Zipf autoregressive model Z.AR(1). In Arnold (1983) a random variable X is defined to be a Zipf(III) distribution and write $X \sim$ Zipf(III)(k_0, σ, γ) if its survival function is of the form

(2.1)
$$\bar{F}_X(k) = P(X \geq k) = [1 + (\frac{k - k_0}{\sigma})^{1/\gamma}]^{-1}, \text{ for } k = k_0, k_0 + 1, k_0 + 2, ...,$$

where $\sigma > 0$ is a scale-like parameter and $\gamma > 0$ is a shape parameter and k_0, the location parameter, is an integer. Let $\{A_n\}$ be the Zipf autoregressive model (Z.AR(1)) which is generated according to the probabilistic model

$$A_n = \max\{V_n A_{n-1}, (1 - V_n) Y_n\}, \tag{2.2}$$

where $\{V_n\}$ is a sequence of *i.i.d. Bernoulli*(ρ) random variables with $P(V_n = 1) = \rho$, $0 \leq \rho \leq 1$, a fixed constant for each $n = 1, 2, \ldots$. Also $\{Y_n\}$ is a sequence of *i.i.d.* Zipf(III)(k_0, σ, γ) random variables for each $n = 1, 2, \ldots$. Clearly, such sequence $\{A_n\}$ is a Markov chain with parameters k_0, σ, γ and ρ.

Property. *In equation* (2.2), *if* $A_0 \sim$ Zipf(III)(k_0, σ, γ) *then all the* A_n'*s in the process,* Z.AR(1), *have identical* Zipf(III)(k_0, σ, γ) *distribution.*

Proof. Suppose $A_{n-1} \sim$ Zipf(III)(k_0, σ, γ), then

$$\begin{aligned} \bar{F}_{A_n}(k) &= P(A_n \geq k) \\ &= P(\max\{V_n A_{n-1}, (1 - V_n) Y_n\} \geq k) \\ &= \rho P(A_{n-1} \geq k) + (1 - \rho) P(Y_n \geq k) \\ &= [1 + (\frac{k - k_0}{\sigma})^{1/\gamma}]^{-1} (\rho + 1 - \rho) \\ &= [1 + (\frac{k - k_0}{\sigma})^{1/\gamma}]^{-1}, \end{aligned}$$

i.e. $A_n \sim$ Zipf(III)(k_0, σ, γ) for all $n = 1, 2, \ldots$.

2.1 The autocorrelation structure of the Z.AR(1) process. Let $\{A_n\}$ be the sequence in the Z.AR(1) process. It follows from the previous property that all the $A_n \sim$ Zipf(III)(k_0, σ, γ) for all $n = 1, 2, \ldots$. For $j \geq 1$, j is an integer, the autocorrelation between A_n and A_{n+j} is calculated as follows:

From the representation (2.2) one finds

$$\begin{aligned} &E(A_n A_{n+j}) \\ &= \rho^j E(A_n^2) + \rho^{j-1}(1-\rho)E^2(A_n) + \rho^{j-2}(1-\rho)E^2(A_n) + \ldots \\ &\qquad + (1-\rho)E^2(A_n) \\ &= \rho^j E(A_n^2) + (1-\rho^j)E^2(A_n). \end{aligned}$$

It follows that

$$\mathrm{Cov}(A_n, A_{n+j}) = \rho^j \left\{E(A_n^2) - E^2(A_n)\right\} = \rho^j \mathrm{Var}(A_n).$$

Thus the autocorrelation of lag j $(j \geq 1)$ is

$$\mathrm{Corr}(A_n, A_{n+j}) = \rho^j. \tag{2.3}$$

It is geometrically decreasing as j increasing.

2.2 The transition matrix of the Z.AR(1) process. It follows from (2.2) that $\{A_n\}$ is a Markov chain with countable state space $E = \{k_0, k_0 + 1, \ldots\}$. Let P_{ij} represent the probability that the process $\{A_n\}$ will when in state i, next make a transition into state j. Then the matrix

$$P = (P_{ij})_\infty$$

is the transition matrix of one-step for the Z.AR(1) process.

For any two states $i, j \in E$, P_{ij} is evaluated as

$$\begin{aligned} P_{ij} &= P\{A_{n+1} = j | A_n = i\} \\ &= \frac{P\{A_n = i, \max\{V_{n+1}A_n, (1-V_{n+1})Y_{n+1}\} = j\}}{P(A_n = i)}. \end{aligned} \tag{2.4}$$

Note that A_n and Y_{n+1} are independently and identically distributed as Zipf(III)(k_0, σ, γ) with marginal pmf

$$\pi(j) = P(Y_{n+1} = j) \tag{2.5}$$
$$= [1 + (\frac{j - k_0}{\sigma})^{1/\gamma}]^{-1} - [1 + (\frac{j + 1 - k_0}{\sigma})^{1/\gamma}]^{-1} \quad \text{for } j \in E.$$

Then under some algebraic calculations, the transition probability is obtained as

$$P_{ij} = \begin{cases} (1-\rho)\pi(j), & \text{if } \; j \neq i, \\ \rho + (1-\rho)\pi(i), & \text{if } \; j = i, \end{cases} \tag{2.6}$$

for every pair $(i, j), i, j \in E = \{k_0, k_0 + 1, \dots\}$, where $\pi(i)$, $\pi(j)$ are defined as in equation (2.5).

2.3 The distribution of the runs. Let $A_1, A_2, \dots$ be a sequence in the Z.AR(1) process. Fix a state $i \in E$, the length of a run of $i's$ starting at time epoch one for $\{A_j\}$ is defined as

$$T_i = \inf\{j : A_j \neq i\} - 1.$$

The probability mass function (pmf) is derived as follows.

$$P(T_i = 0) = P(A_1 \neq i) = 1 - \pi(i),$$

where $\pi(i)$ is given in equation (2.5). Then for $l = 1, 2, 3, \dots$, by the Markov property of Z.AR(1), we have

$$\begin{aligned} P(T_i = l) &= P(A_1 = i, A_2 = i, \dots, A_l = i, A_{l+1} \neq i) \\ &= \sum_{k \in E - \{i\}} \pi(i) P_{ii}^{l-1} P_{ik} \\ &= (1-\rho)\pi(i)(1 - \pi(i)) P_{ii}^{l-1}, \end{aligned} \tag{2.7}$$

where P_{ii}, P_{ik} are given in (2.6). It is straightforward to check that

$$\sum_{l=0}^{\infty} P(T_i = l) = 1$$

and conclude that for any fixed state $i \in E$, T_i is a non-defective discrete random variable.

The survival function of T_i is directly calculated from the pmf of T_i, equation (2.7), as

$$P(T_i \geq n) = \begin{cases} 1, & \text{if } n = 0, \\ \pi(i)P_{ii}^{n-1}, & if \ n = 1, 2, \ldots \end{cases} \tag{2.8}$$

Property. *The expected run length for the* Z.AR(1) *process is always greater than or equal to the expected run length of an i.i.d. sequence of* Zipf(III) *random variables.*

Proof. From equation (2.8), the expected run length

$$E(T_i) = \sum_{n=1}^{\infty} P(T_i \geq n) = \sum_{n=1}^{\infty} \pi(i)P_{ii}^{n-1} = \frac{\pi(i)}{(1-\rho)\{1-\pi(i)\}}. \tag{2.9}$$

Clearly $E(T_i) \geq \frac{\pi(i)}{1-\pi(i)}$ for $0 \leq \rho \leq 1$. Refer to the equation (2.2), consider two particular cases

Case (i): $\rho = 0$, i.e., $V_n = 0$ with probability 1, then $A_n = Y_n$ for all $n = 1, 2, \ldots$, so the Z.AR(1) process is reduced to the sequence of *i.i.d.* Zipf(III)random variables and in this case $E(T_i) = \frac{\pi(i)}{1-\pi(i)}$.

Case (ii): $\rho = 1$, i.e., $V_n = 1$ with probability 1, then $A_n = A_{n+1}$ for all $n = 1, 2, \ldots$. Hence once $A_1 = i$, then $A_1 = A_2 = A_3 = \cdots = i$ so $E(T_i) = \infty$

3. Extreme order statistics.

3.1. Exact and asymptotic distribution of extremes. Let $A_1, A_2, \ldots, A_n$ be the first n observations from a Z.AR(1) process. Define

$$m_n = \min\{A_1, A_2, \ldots, A_n\}.$$

By the Markov property of $\{A_i\}$, the survival function of m_n can be calculated as

$$\begin{aligned} \bar{F}_{m_n}(i) &= P(m_n \geq i) = P(A_1 \geq i)\{P(A_2 \geq i | A_1 \geq i)\}^{n-1} \\ &= \frac{1}{1+(\frac{i-k_0}{\sigma})^{1/\gamma}} \left\{ \frac{1+\rho(\frac{i-k_0}{\sigma})^{1/\gamma}}{1+(\frac{i-k_0}{\sigma})^{1/\gamma}} \right\}^{n-1}, \end{aligned} \tag{3.1}$$

for any $i \in E$ and any integer $n \geq 1$.

The asymptotic distribution of T_n is readily deduced from (3.1). Assuming $0 < \rho < 1$, it follows that

$$\lim_{n\to\infty} P\{\frac{n^\gamma}{\sigma}(m_n - k_0) \geq x\} = e^{-x^{1/\gamma}(1-\rho)} \quad \text{for} \quad x > 0,$$

i.e., $\frac{[n(1-\rho)]^\gamma}{\sigma}(m_n - k_0) \xrightarrow{d} Weibull(\frac{1}{\gamma})$.

A similar analysis is for $M_n = \max\{A_1, A_2, \ldots, A_n\}$. The distribution function of M_n for any $k \in E$ is

$$F_{M_n}(k) = P(M_n < k) = P(A_1 < k)\{P(A_2 < k|A_1 < k)\}^{n-1}$$

$$(3.2) \qquad = \frac{1}{1+(\frac{k-k_0}{\sigma})^{1/\gamma}}\left\{\frac{1+\rho(\frac{k-k_0}{\sigma})^{-1/\gamma}}{1+(\frac{k-k_0}{\sigma})^{-1/\gamma}}\right\}^{n-1},$$

for any integer $n \geq 1$. If $0 < \rho < 1$, the asymptotic distribution of M_n is deduced from (3.2) which is

$$\lim_{n\to\infty} P\{\frac{1}{n^\gamma \sigma}(M_n - k_0) \leq t\} = e^{-(1-\rho)t^{1/\gamma}} \quad \text{for} \quad t > 0.$$

3.2. Geometric minima and maxima. Suppose that $A_1, A_2, \ldots$ is a sequence in the Z.AR(1) process and that N is a geometric random variable with probability function

$$P(N = n) = pq^{n-1} \quad \text{for} \quad n = 1, 2, \ldots, (q = 1 - p).$$

Assuming that N is independent of the $A_i's$, I define the geometric minimum as

$$m = \min\{A_1, A_2, \ldots, A_N\}.$$

Since N and $A_i's$ are independent,

$$P\{m \geq i|N = n\} = P\{m_n \geq i\} \quad \text{for} \quad i \in E,$$

where m_n is defined as in section 3.1. Thus

$$P(m \geq i) = \sum_{n=1}^{\infty} P\{m \geq i|N = n\}P(N = n)$$

$$(3.3) \qquad = [1 + [\frac{i - k_0}{\sigma(\frac{p}{1-q\rho})^\gamma}]^{1/\gamma}]^{-1},$$

i.e.,

$$m \sim \text{Zipf(III)}(k_0, \sigma(\frac{p}{1-q\rho})^{\gamma}, \gamma),$$

and observe that the Z.AR(1) process is closed under geometric minimization.

Analogously, the geometric maximum can be defined as

$$M = \max\{A_1, A_2, \ldots, A_N\}.$$

Using (3.2) and conditioning on N, can get

$$P(M \geq k) = \left\{1 + [\frac{i-k_0}{\sigma(\frac{1-q\rho}{p})^{\gamma}}]^{1/\gamma}\right\}^{-1}, \quad \text{for any } k \in E. \tag{3.4}$$

Thus

$$M \sim \text{Zipf(III)}(k_0, (\frac{1-q\rho}{p})^{\gamma}, \gamma),$$

and hence that the Z.AR(1) process is closed under geometric maximization.

To summarize the relation between (3.3) and (3.4), I may write

$$(\frac{p}{1-q\rho})^{-\gamma} m \stackrel{\text{d}}{=\!=} (\frac{p}{1-q\rho})^{\gamma} M \stackrel{\text{d}}{=\!=} A_1 \sim \text{Zipf(III)}(k_0, \sigma, \gamma).$$

This property is similar to that of Pareto processes. (Yeh, Arnold and Robertson (1988)).

4. Statistical inference for the Z.AR(1) process. Let $(a_1, a_2, \ldots, a_{n+1})$ be an observed sequence on the Z.AR(1) process. In this section I will introduce how to draw inferences about the parameters $\{\rho, \sigma, \gamma\}$ from the sample.

The likelihood of the observations $(a_1, a_2, \ldots, a_{n+1})$ is (Andersone and Goodman (1957))

$$L(a_1, a_2, \ldots, a_{n+1}) = P(a_1) \prod_{k=1}^{n} P(a_{k+1}|a_k),$$

where $P(a_{k+1}|a_k), k = 1, 2, \ldots, n$, are the transition probabilities defined in (2.4). The log-likelihood of $(a_1, a_2, \ldots, a_{n+1})$ is

$$\log L(a_1, a_2, \ldots, a_{n+1}) = \ln(P(a_1)) + \sum_{k=1}^{n} \ln P(a_{k+1}|a_k). \tag{4.1}$$

It is clear that the sample $(a_1, a_2, \ldots, a_{n+1})$ contains only one observation on the initial density

$$\pi(a_1) = P(A_1 = a_1) = P(a_1)$$
$$= \left[\frac{1}{1+(\frac{a_1-k_0}{\sigma})^{1/\gamma}}\right] - \left[\frac{1}{1+(\frac{a_1+1-k_0}{\sigma})^{1/\gamma}}\right]$$

and n observations on the set of transition densities

$$P(a_{k+1}|a_k), \qquad k = 1, 2, \ldots, n,$$

so that the information about $P(a_1)$ contained in $(a_1, a_2, \ldots, a_{n+1})$ does not increase with n. Thus in (4.1), the term $\ln P(a_1)$ is dominated by the sum of the other n terms, $\sum_{k=1}^{n} \ln P(a_{k+1}|a_k)$, as n becomes very large. Since the large-sample theory is usually applied to the development of statistical inference, so it will be convenient to drop the term, $\ln P(a_1)$ in (4.1) and take

$$\ln L = \sum_{k=1}^{n} \ln P(a_{k+1}|a_k)$$

as the log-likelihood function.

Note that all the values $a_1, a_2, \ldots, a_{n+1}$ are in the state space $E = \{k_0, k_0+1, \ldots\}$. For any $i \in \{a_1, a_2, \ldots, a_{n+1}\}$, let

$n_i =$ the number of $\{k:$ such that $a_k = i$ and $a_{k+1} = i$, for $1 \leq k \leq n\}$,

for any pair $(j,k), j, k \in \{a_1, a_2, \ldots, a_{n+1}\}$ with $j \neq k$,

$n_{jk} =$ the number of $\{k:$ such that $a_k = j$ and $a_{k+1} = j$, for $1 \leq k \leq n\}$.

Define two sets

$$D_1 = \{i | n_i > 0\},$$
$$D_2 = \{(j,k) | n_{jk} > 0\},$$

Clearly,

$$\sum_{i \in D_1} n_i + \sum_{(j,k) \in D_2} n_{jk} = n.$$

Then the log-likelihood of the given ordered set of sequence $(a_1, a_2, \ldots, a_{n+1})$ can be written as

$$\ln L = \sum_{i \in D_1} n_i \ln(P_{ii}) + \sum_{(j,k) \in D_2} n_{jk} \ln(P_{jk}).$$

Consider partial derivatives equations

$$(4.2)\qquad \frac{\partial \ln L}{\partial \rho} = \sum_{i\in D_1} n_i \frac{1}{P_{ii}} \frac{dP_{ii}}{d\rho} + \sum_{(j,k)\in D_2} n_{jk} \frac{1}{P_{jk}} \frac{dP_{jk}}{d\rho} = 0,$$

$$(4.3)\qquad \frac{\partial \ln L}{\partial \sigma} = \sum_{i\in D_1} n_i \frac{1}{P_{ii}} \frac{dP_{ii}}{d\sigma} + \sum_{(j,k)\in D_2} n_{jk} \frac{1}{P_{jk}} \frac{dP_{jk}}{d\sigma} = 0,$$

$$(4.4)\qquad \frac{\partial \ln L}{\partial \gamma} = \sum_{i\in D_1} n_i \frac{1}{P_{ii}} \frac{dP_{ii}}{d\gamma} + \sum_{(j,k)\in D_2} n_{jk} \frac{1}{P_{jk}} \frac{dP_{jk}}{d\gamma} = 0,$$

where P_{ii}, P_{jk} are given in (2.6) as

$$P_{ii} = \rho + (1-\rho)\pi(i) = \rho + (1-\rho)\left\{ \frac{1}{1+(\frac{i-k_0}{\sigma})^{1/\gamma}} - \frac{1}{1+(\frac{i+1-k_0}{\sigma})^{1/\gamma}} \right\},$$

$$P_{jk} = (1-\rho)\pi(k) = (1-\rho)\left\{ \frac{1}{1+(\frac{k-k_0}{\sigma})^{1/\gamma}} - \frac{1}{1+(\frac{k+1-k_0}{\sigma})^{1/\gamma}} \right\}.$$

Then in (4.2),

$$\frac{dP_{ii}}{d\rho} = 1 - \left\{ \frac{1}{1+(\frac{i-k_0}{\sigma})^{1/\gamma}} - \frac{1}{1+(\frac{i+1-k_0}{\sigma})^{1/\gamma}} \right\},$$

$$\frac{dP_{jk}}{d\rho} = -\left\{ \frac{1}{1+(\frac{k-k_0}{\sigma})^{1/\gamma}} - \frac{1}{1+(\frac{k+1-k_0}{\sigma})^{1/\gamma}} \right\},$$

and in (4.3)

$$\frac{dP_{ii}}{d\sigma} = (1-\rho)\left\{ \frac{1}{\gamma}\sigma^{\frac{1}{\gamma}-1}\left[\frac{(i-k_0)^{1/\gamma}}{(1+(\frac{i-k_0}{\sigma})^{1/\gamma})^2} - \frac{(i+1-k_0)^{1/\gamma}}{(1+(\frac{i+1-k_0}{\sigma})^{1/\gamma})^2}\right] \right\},$$

$$\frac{dP_{jk}}{d\sigma} = (1-\rho)\left\{ \frac{1}{\gamma}\sigma^{\frac{1}{\gamma}-1}\left[\frac{(k-k_0)^{1/\gamma}}{(1+(\frac{k-k_0}{\sigma})^{1/\gamma})^2} - \frac{(k+1-k_0)^{1/\gamma}}{(1+(\frac{k+1-k_0}{\sigma})^{1/\gamma})^2}\right] \right\},$$

and in (4.4)

$$\frac{dP_{ii}}{d\gamma} = (1-\rho)\left\{ \frac{1}{\gamma^2}\left[\frac{(\frac{i-k_0}{\sigma})^{1/\gamma}}{(1+(\frac{i-k_0}{\sigma})^{1/\gamma})^2} \ln(\frac{i-k_0}{\sigma}) - \frac{(\frac{i+1-k_0}{\sigma})^{1/\gamma}}{(1+(\frac{i+1-k_0}{\sigma})^{1/\gamma})^2} \ln(\frac{i+1-k_0}{\sigma})\right] \right\},$$

$$\frac{dP_{jk}}{d\gamma} = (1-\rho)\left\{ \frac{1}{\gamma^2}\left[\frac{(\frac{k-k_0}{\sigma})^{1/\gamma}}{(1+(\frac{k-k_0}{\sigma})^{1/\gamma})^2} \ln(\frac{k-k_0}{\sigma}) - \frac{(\frac{k+1-k_0}{\sigma})^{1/\gamma}}{(1+(\frac{k+1-k_0}{\sigma})^{1/\gamma})^2} \ln(\frac{k+1-k_0}{\sigma})\right] \right\},$$

The maximum likelihood estimates of ρ, σ, γ can be solved from equations (4.2), (4.3), (4.4) by Fletcher and Powell method (1963). The location parameter k_0 can be estimated by the minimum of $\{a_1, a_2, \ldots, a_{n+1}\}$, say $a_{(1)}$.

5. Extensions and related processes. Analogous to Yeh (1983)'s Pareto process model, two possible extensions of the Z.AR(1) process with Zipf(III) marginals to moving averages Z.MA(1) and mixed autoregressive-moving averages Z.ARMA(1,1) are discussed in this section.

5.1. The Zipf moving average process Z.MA(1). The Z.MA(1) process is defined is defined as

$$X_n = \max\{V_n Y_n, (1 - V_n) Y_{n-1}\}, \tag{5.1}$$

where $\{Y_n\}_0^\infty$ is a sequence of *i.i.d.* Zipf(III)(k_0, σ, γ) variables. Note that, unlike the Z.AR(1) process, the Z.MA(1) process is not a Markov chain in general.

I now claim that each X_i in the Z.MA(1) process has a Zipf(III)(k_0, σ, γ) marginal distribution. Let $\bar{F}_{X_i}(.)$ be the survival function of X_i. Then for any integer $k \geq k_0$,

$$\begin{aligned}
\bar{F}_{X_i}(k) = P(X_i \geq k) &= P\big(\max\{V_n Y_n, (1 - V_n) Y_{n-1}\} \geq k\big) \\
&= \rho P(Y_n \geq k) + (1 - \rho) P(Y_{n-1} \geq k) \\
&= \big[1 + (\frac{k - k_0}{\sigma})^{1/\gamma}\big]^{-1} (\rho + 1 - \rho) \\
&= \big[1 + (\frac{k - k_0}{\sigma})^{1/\gamma}\big]^{-1}.
\end{aligned}$$

Thus, $X_i \sim$ Zipf(III)(k_0, σ, γ) marginally for all $i = 1, 2, \ldots$.

The autocorrellation of $\{X_n\}$ is calculated as

$$\begin{aligned}
\mathrm{Cov}(X_n, X_{n+1}) = E(X_n X_{n+1}) - E(X_n) E(X_{n+1}) \\
= \rho^2 E(Y_n Y_{n+1}) + \rho(1 - \rho) E(Y_n^2) + \rho(1 - \rho) E(Y_{n-1} Y_{n+1}) \\
+ (1 - \rho)^2 E(Y_{n-1} Y_n) - E^2(X_n).
\end{aligned}$$

Note $E(X_n) = E(Y_n)$ and all the $\{Y_n\}$ are *i.i.d.* Zipf(III)(k_0, σ, γ) variables, hence

$$\begin{aligned}\mathrm{Cov}(X_n, X_{n+1}) &= \{\rho^2 + \rho(1-\rho) + (1-\rho)^2 - 1\}E^2(Y_n) + \rho(1-\rho)E(Y_n^2)\\ &= \rho(1-\rho)\{E(Y_n^2) - E^2(Y_n)\}\\ &= \rho(1-\rho)\mathrm{Var}(Y_n).\end{aligned}$$

Thus $\rho_1 = \mathrm{Corr}(X_n, X_{n+1}) = \rho(1-\rho)$. For

$$\begin{aligned}\mathrm{Cov}(X_n, X_{n+2}) &= E(X_n X_{n+2}) - E(X_n)E(X_{n+2})\\ &= \rho^2 E(Y_n Y_{n+2})\\ &\qquad + \rho(1-\rho)E(Y_n Y_{n+1}) + \rho(1-\rho)E(Y_{n-1}Y_{n+2})\\ &\qquad + (1-\rho)^2 E(Y_{n-1}Y_{n+1}) - E^2(Y_n)\\ &= \{[\rho + (1-\rho)]^2 - 1\}E^2(Y_n)\\ &= 0.\end{aligned}$$

In general, for $j \geq 3$, analogously, can get

$$\mathrm{Cov}(X_n, X_{n+j}) = 0.$$

Thus

$$\rho_j = 0 \qquad \text{for} \quad j \geq 2.$$

Moreover, from the generating procedure, equation (2.2),

$$\begin{aligned}X_n &= \max\{V_n Y_n, (1-V_n)Y_{n-1}\},\\ X_{n+j} &= \max\{V_{n+j}Y_{n+j}, (1-V_{n+j})Y_{n+j-1}\} \quad \text{for} \quad j \geq 2,\end{aligned}$$

it is clearly observed that X_n and X_{n+j} are independently distributed. Therefore, the Z.AR(1) is a one-dependent process and has a cut-off pattern of autocorrelation structure with autocorrelations $\rho_1 = (1-\rho)$ and $\rho_j = 0$ for any $j \geq 2$.

5.2. The mixed Zipf autoregressive-moving average process, Z.ARMA(1,1). In order to achieve parsimony it may be necessary to include both autoregressive and moving average terms and thus the mixed Z.ARMA(1,1) process is constructed as

$$\text{(5.2)}\qquad \begin{aligned} X_n &= \max\{U_nY_n, (1-U_n)A_{n-1}\}, \\ A_n &= \max\{V_nA_{n-1}, (1-V_n)Y_n\}, \end{aligned}$$

where $\{Y_n\}_1^\infty$ is a sequence of *i.i.d.* Zipf(III)(k_0, σ, γ) random variables and $\{U_n\}_1^\infty$, $\{V_n\}_1^\infty$ are independent sequences of independent Bernoulli random variables with

$$P\{U_n = 1\} = \beta \quad \text{and} \quad P\{V_n = 1\} = \rho,$$

for fixed $0 \leq \beta \leq 1$ and $0 \leq \rho \leq 1$.

Unless otherwise indicated, I will assume that A_0 has Zipf(III)(k_0, σ, γ) distribution and is independent of $\{Y_n\}$, $\{U_n\}$, and $\{V_n\}$ for any $n = 1, 2, \ldots$. I will claim that the $\{X_n\}_1^\infty$ forms a stationary sequence of dependent Zipf(III)(k_0, σ, γ) random variables. Note that from the Z.AR(1) process, I know $A_n's$ have Zipf(III) marginal distributions.

For $n = 1$, the survival function of X_1 is

$$\begin{aligned} \bar{F}_{X_1}(X) = P(X_1 \geq x) &= P\{\max\{U_1Y_1, (1-U_1)A_0\} \geq x\} \\ &= \beta P(Y_1 \geq x) + (1-\beta)P(A_0 \geq x) \\ &= [1 + (\frac{x-k_0}{\sigma})^{1/\gamma}]^{-1}, \end{aligned}$$

for any $x = k_0, k_0 + 1, \ldots$.

In general, for $n = 1, 2, \ldots$, the survival function of $\{X_n\}$ is

$$\begin{aligned} \bar{F}_{X_n}(X) = P(X_n > x) &= P\{\max\{U_nY_n, (1-U_n)A_{n-1}\} \geq x\} \\ &= \beta P(Y_n \geq x) + (1-\beta)P(A_{n-1} \geq x). \end{aligned}$$

Note that both $\{Y_n\}$ and $\{A_{n-1}\}$ are marginally distributed as Zipf(III) variables; moreover, they are independently distributed, hence

$$\begin{aligned} \bar{F}_{X_n}(x) &= \frac{1}{1 + (\frac{x-k_0}{\sigma})^{1/\gamma}}\{\beta + (1-\beta)\} \\ &= [1 + (\frac{x-k_0}{\sigma})^{1/\gamma}]^{-1}, \end{aligned}$$

i.e. $X_n \sim$ Zipf(III)(k_0, σ, γ).

Refer to (5.2), I find if $\beta = 0$ and $0 < \rho < 1$, then $X_n = A_{n-1}$ with probability one, so $\{X_n\}$ is just Z.AR(1) process, and if $\rho = 0$ and $0 < \beta < 1$, then $A_{n-1} = Y_{n-1}$ with probability one so $\{X_n\}$ is just Z.MA(1) process. In particular, if $\beta = 1$, and $0 \le \rho \le 1$, then $\{X_n\}$ is reduced to $\{Y_n\}$, an *i.i.d.* sequence of Zipf(III)variables. From this discussion, we know that the mixed Z.ARMA(1,1) is a process of great flexibility in modelling dependent Zipf(III)(k_0, σ, γ) variables.

According to (5.2), for $j \ge 1$,

$$X_{n+j} = \max\{U_{n+j}Y_{n+j}, (1 - U_{n+j})A_{n+j-1}\}.$$

The covariance between X_n and X_{n+j} is

$$\begin{aligned} &\mathrm{Cov}(X_{n+j}, X_n) \\ &= \beta^2\ \mathrm{Cov}(Y_{n+j}, Y_n) + \beta(1-\beta)\{\ \mathrm{Cov}(Y_{n+j}, A_{n-1}) +\ \mathrm{Cov}(Y_n, A_{n+j-1})\} \\ &\qquad + (1-\beta)^2\ \mathrm{Cov}(A_{n+j-1}, A_{n-1}). \end{aligned} \tag{5.3}$$

For $j \ge 1$, the terms

$$\mathrm{Cov}(Y_{n+j}, Y_n) = 0, \tag{5.4}$$

$$\mathrm{Cov}(Y_{n+j}, A_{n-1}) = 0, \tag{5.5}$$

also from equation (2.3), I get

$$\mathrm{Cov}(A_{n+j-1}, A_{n-1}) = \rho^j \mathrm{Var}(A_{n-1}). \tag{5.6}$$

In the third term $\mathrm{Cov}(Y_n, A_{n+j-1})$, note

$$A_{n+j-1} = \max\{V_{n+j-1}A_{n+j-1}, (1 - V_{n+j-1})Y_{n+j-1}\},$$

then

$$\begin{aligned} \mathrm{Cov}(Y_n, A_{n+j-1}) &= \rho\ \mathrm{Cov}(Y_n, A_{n+j-2}) + (1-\rho)\ \mathrm{Cov}(Y_n, Y_{n+j-1}) \\ &= \rho\ \mathrm{Cov}(Y_n, A_{n+j-2}) \\ &= \dots \\ &= \rho^{j-1}\ \mathrm{Cov}(Y_n, A_n), \end{aligned}$$

where

$$\begin{aligned} \text{Cov}(Y_n, A_n) &= \rho \ \text{Cov}(Y_n, A_{n-1}) + (1-\rho) \ \text{Cov}(Y_n, Y_n) \\ &= (1-\rho)\text{Var}(Y_n). \end{aligned} \tag{5.7}$$

Hence

$$\text{Cov}(Y_n, A_{n+j-1}) = \rho^{j-1}(1-\rho)\text{Var}(Y_n).$$

Putting together (5.4) through (5.7), then equation (5.3) is simplified as

$$\text{Cov}(X_{n+j}, X_n) = \beta(1-\beta)\rho^{j-1}(1-\rho)\text{Var}(Y_n)+(1-\beta)^2\rho^j\text{Var}(A_{n-1}). \tag{5.8}$$

Note the three sequences $\{X_n\}$, $\{Y_n\}$ and $\{A_n\}$ are marginally identically distributed as Zipf(III)variables, hence

$$\text{Var}(X_{n+j}) = \text{Var}(X_n) = \text{Var}(Y_n) = \text{Var}(A_{n-1}).$$

Then the autocorrelation of X_n and X_{n+j} for $j \geq 1$, is

$$\rho_X(j) = \text{Corr}(X_n, X_{n+j}) = \beta(1-\beta)\rho^{j-1}(1-\rho) + (1-\beta)^2\rho^j.$$

Note that $0 \leq \rho_X(j) \leq 1$ and for $j \geq 1$, $\rho_X(j)$ decreases geometrically with increasing j if $0 < \rho < 1$. Since $\rho_j(X)$ is independent of n, so the mixed Z.ARMA(1,1) process is second order covariance stationary.

6. Conclusions. In this article, I have presented an autoregressive Markov chain $\{A_n\}$ of Zipf(III) variables, Z.AR(1) process, in which A_n is an non-additive random nonlinear combination of the previous value, A_{n-1}, and an independent Zipf(III) random variable. The generating scheme of Z.AR(1) process is given in equation (2.1). Its autocorrelation structure, the distribution of runs, the extreme order statistics and the MLE for the parameters are all discussed detailedly in this paper. I also briefly studied the Zipf moving average process, Z.MA(1) and the mixed Zipf process, Z.ARMA(1,1).

In summary of this paper, I found

(i) All the Z.AR(1), Z.MA(1), and Z.ARMA(1,1) processes give a common Zipf(III)(k_0, σ, γ) marginal distribution.

(ii) The Zipf processes are analogous to the discrete-type Pareto processes (Yeh (1983)) and they have dependence character like that of the discrete time series with marginal geometric or Poisson distributions (Jacobs and Lewis (1978)).

(iii) The innovation variables $\{Y_i\}$ have *i.i.d.* Zipf(III)(k_0, σ, γ) distributions; this construction implies that the Zipf processes are easily simulated by a computer.

References

1. Anderson, T. W. and Goodman, L. A., *Statistical inference about Markov chains*, Annals of Mathematical Statistics, 28(1957), 89-109.

2. Arnold, B. C.,*Pareto Distribution*, International Cooperative Publishing House, Fairland, Maryland, 1983.

3. Chow, Y. S. and Teicher, H.,*Probability Theory*, Springer-Verlag, New York Inc. 1978.

4. Çinlar, E., *Introduction to Stochastic Processes*, Prentice Hall Inc., Englewood Cliffs, New Jersey. 1975.

5. Fletcher, R. and Powell, M. J. D., *A rapid descent method for minimization*, Computer Journal, 6, 2, (1963), 163-168.

6. Jacobs, P. A. and Lewis, P. A. W., *Discrete time series generated by mixtures, I: Correlational and runs properties*, J. R. Statist. Soc. B., 40(1978), 94-105.

7. Ross, S. M., *Introduction to Probability Models*, 3rd edition(1985), Academic Press Inc. United Kingdom.

8. Yeh, H. C., *Pareto Processes*, Unpublished Ph.D. dissertation (1983), University of California, Riverside.

9. Yeh, H. C., Arnold, B. C. and Robertson, C.,*Pareto Processes*, Journal of Applied Probability 25(1988), 291-301

Department of Finance, National Taiwan University, Taipei, Taiwan, R.O.C.