SPECTRAL ESTIMATION UNDER NATURE MISSING DATA

Jui-Chung Hung[1], *Bor-Sen Chen*[2], *Wen-Sheng Hou, Li-Mei Chen*

[1] Ling-Tung College
[2]Department of Electrical Engineering
National Tsing-Hua University
Hsin-Chu 300, Taiwan.
Email: hong@moti.ee.nthu.edu.tw
TEL: 886-35-731155

ABSTRACT

This paper considers the problem of estimating the autoregressive moving average (ARMA) power spectral density when measurements are corrupted by noises and with missing data. The missing data model is based on a probabilistic structure with unknown. In this situation, the spectral estimation becomes a highly nonlinear optimization problem with many local minima. In this paper, we use the global search method of genetic algorithm (GA) to achieve a global optimal solution of this spectral estimation problem. From the simulation results, we have found that the performance is improved significantly if the probability of data missing is considered in the spectral estimation problem.

1. INTRODUCTION

The spectral estimation becomes a problem in parameter estimation based on the measured data. In most cases, it is assumed that the measurements always contain the signal. In fact, in practical situations there may be a nonzero probability that any measurement consists of noise alone, i.e., the measurements are not consecutive but contain missing data. The missing measurements are caused by a variety of reasons, e.g., a certain failure in the measurement, intermittent sensor failures, accidental loss of some collected data, or some of the data may be jammed, fading phenomena in propagation channels, and the effect of removing outliers 1-2]. Estimating the spectrum of stationary time series with missing data is more difficult than the spectral estimation problem for the case without missing data. The difficulty is that the standard definition of covariance in the statistical analysis of data does not directly apply if some of the measurements are unavailable [1]. Thus, many currently used parameter estimation algorithms do not apply to this situation. For example, standard techniques like the periodgram or the smoother periodgram will not apply to this situation, unless properly modified. This paper is concerned with the problem of spectral estimation when the data are corrupted with measurement noise and some data are missed. We assume the time points of missing data are unavailable and the probability of missing data is unknown. Since the covariance of corrupted noise and the probability of data missing also need to be estimated, the spectral estimation problem, based on ARMA modeling and the least square error criterion, become a highly nonlinear parameter estimation problem. The parameters of ARMA model and the probability of missing data are specified to minimize the mean square estimation error. There exist many local minima. In this situation, a GA based parameter estimation algorithm is proposed to achieve the global optimal solution of the spectral estimation problem.

Recently, the genetic algorithm has been introduced for optimization searching [3]. The genetic algorithm applies operators inspired by the mechanics of natural selection to a population of binary strings encoding the parameter space. It is a parallel global search technique that emulates natural genetic operators such as reproduction, crossover, and mutation. At each generation, it explores different areas of the parameter space, and then direct the search to the region where there is a high probability of finding improved per-

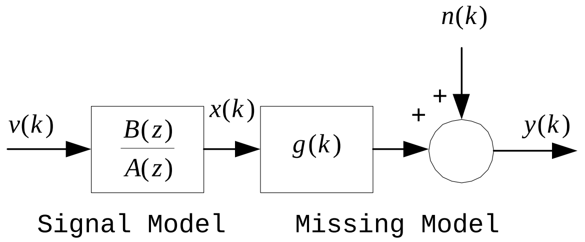

Figure 1: Signal model.

formance. Because the genetic algorithm simultaneously evaluates many points in parameter space, it is more likely to converge toward the global solution. In particular, it need not assume the search space being differentiable or continuous, and can also iterate several times on each datum received. Hence, it is very suitable to treat the global optimization problem of the nonlinear spectral estimation under the corrupted noises and missing data.

2. PROBLEM DESCRIPTION

Consider an ARMA time series $\{x(k)\}$ defined by the following stochastic process

$$x(k) = -\sum_{i=1}^{n} a_i x(k-i) + \sum_{i=0}^{q} b_i v(k-i) \qquad (1)$$

where $\{v(k)\}$ is a zero mean unit-variance white Gaussian noise. We assume that at random instant the signal $x(k)$ may be absent from the measurement. Let the sequence $\{x_m(k)\}$ be defined as

$$x_m(k) = g(k)x(k) \qquad (2)$$

where $g(k)$ is binary random variable such that

$$g(k) = \begin{cases} 1 & \text{if } x(k) \text{ is measured} \\ 0 & \text{if } x(k) \text{ is missing} \end{cases} \qquad (3)$$

Thus, $x_m(k)$ can be regarded as the measurement of $x(k)$ with missing data. Moreover, $x_m(k)$ is also assumed to be corrupted by zero mean measurement noise $n(k)$ with variance σ_n^2. Let $\{y(k)\}$ denote the observed output signal (see the Fig.1). Thus, the measurement equation is given by

$$y(k) = x_m(k) + n(k) = g(k)x(k) + n(k) \qquad (4)$$

The following assumptions are made:

(A1) The $v(k)$, $n(k)$, are mutually independent.

(A2) The sequence $g(k)$ is assumed to be asymptotically stationary and independent of $x(k)$. Furthermore, they are mutually independent. The probability P for the measurement $x_m(k)$ to be measured is assumed to be unknown and given by

$$E\left[g(k)\right] = \Pr\left[g(k) = 1\right] = P, \quad 0 < P \leq 1 \qquad (5)$$

where E denotes the expectation operator, and P is a fixed probability, independent of time. Thus the probability of missing measurement is $(1-P)$.

(A3) Each measurement has the fixed probability of being missed, and for different instants, the occurrences of missing data are mutually independent. Thus, by the assumptions, the power spectral density (PSD) of the ARMA time series $\{x(k)\}$ in (1) is

$$\begin{aligned} S_x(f) &= \frac{(b_0 + b_1 z^{-1} + \cdots + b_q z^{-q}).}{(1 + a_1 z^{-1} + \cdots + a_n z^{-n}).} \\ &\quad \frac{(b_0 + b_1 z^{1} + \cdots + b_q z^{q})}{(1 + a_1 z^{1} + \cdots + a_n z^{n})} \Big|_{z=e^{j2\pi f}} \\ &= \sum_{l=-\infty}^{\infty} r_x(l) z^{-l} \Big|_{z=e^{j2\pi f}} \end{aligned} \qquad (6)$$

where $r_x(l) = E\left[x(k)x(k-l)\right]$ is nonlinear function of the model parameters a_i for $i = 1, \ldots, n$, and $b_i$for $i = 0, \ldots, q$.

From equation (4-6), we can obtain the PSD of the received data is

$$\begin{aligned} S_y(f) &= \sum_{l=-\infty}^{\infty} r_y(l) z^{-l} \Big|_{z=e^{j2\pi f}} \\ &= P^2.S_x(f) + \left[\sigma_n^2 + P.(r_x(0) - P.r_x(0))\right] \\ &= P^2 \sum_{l=-\infty}^{\infty} r_x(|l|) e^{-j2\pi f l} + K \end{aligned} \qquad (7)$$

where $K = \sigma_n^2 + P.(r_x(0) - P.r_x(0))$. Therefore, if the system parameters $a_1, \cdots, a_n$, $b_0, \cdots, b_q$ are given, the PSD of system can be computed from (8), Based on the analysis above, the power spectral estimation problem under corrupted noise and missing data is to estimate the system parameters $a_1, \cdots, a_n$, $b_0, \cdots, b_q$ from the missing data sequence $\{y(k)\}$ in (4). In this spectral estimation problem, not only the system parameters $a_1, \cdots, a_n,\ b_0, \cdots, b_q$ are unknown but also the probability P and the noise's variance σ_n^2 are also needed to be estimated.

Inspection the equation(10), the series as $\sum_{l=-\infty}^{\infty} e^{-j2\pi fl}$ as a power series has the property of completeness, we can obtain

$$r_y(l) = \begin{cases} P^2\, r_x(0) + K & \text{for } l = 0 \\ P^2\, r_x(l) & \text{for } l \neq 0 \end{cases} \tag{8}$$

From the received data sequence $\{y(k)\}$, let us define the sample covariance by

$$\widehat{r}_y(l) = \frac{1}{N-l} \sum_{k=l+1}^{N} y(k)y(k-l) \tag{9}$$

where N is number of received data length. Note that the sample covariances $\{\widehat{r}_y(l)\}$ are unbiased estimates of the true covariances $\{r_y(l)\}$. The main idea is to search for the polynomials $\sum_{l=-\infty}^{\infty} r_y(l)z^{-l}$ such that the corresponding sequence $\{\widehat{r}_y(0), \cdots, \widehat{r}_y(M)\}$, from (10) and (12), a reasonable criterion is to minimize the mean square error as

$$\begin{aligned} & \min J(a_1, \cdots, a_n, b_0, \cdots, b_q, P, \sigma_n^2) && (10) \\ = \quad & \min \left\{ \sum_{l=0}^{M} (r_y(l) - \widehat{r}_y(l))^2 \right\} \\ = \quad & \min \left\{ (r_y(0) - \widehat{r}_y(0))^2 + \sum_{l=1}^{M} (r_y(l) - \widehat{r}_y(l))^2 \right\} \end{aligned}$$

In general, J is a very highly nonlinear function of the probability P, noise variance σ_n^2, and the coefficients a_i for $i = 1, \ldots, n$, and $b_i$for $i = 0, \ldots, q$. There may exist many local minima. It is very difficult to find the global minimum of J in (13) by the conventional methods. Genetic algorithms are optimization and machine learning algorithms, initially inspired from the processes of natural selection and evolution genetics. Therefore, in this study, genetic algorithms will be employed to specify the coefficients of P, σ_n^2, a_i for $i = 1, \ldots, n$, and b_i for $i = 0, \ldots, q$ to solve the spectral estimation problem under corrupted noise and missing data in (13).

3. GENETIC ALGORITHM IN SPECTRAL ESTIMATION UNDER CORRUPTED NOISE AND MISSING DATA

3.1. Simple Genetic Algorithm

Genetic algorithms are stochastic optimization algorithms. Their initial mechanisms are originally motivated by natural selection and evolutionary genetics. The underlying principles of genetic algorithm were first published by Holland in 1962 [3]. The mathematical framework was developed in the late 1960's, and has been presented in Holland's pioneering books[4].

In this paper, a simple genetic algorithm is used. It is an iterative procedure which maintains a constant size population Γ of candidate solutions. In each generation, the genetic algorithm is composed of three operators: (1)reproduction, (2) crossover, and (3)mutation [3-4]. These operators are implemented by performing the basic tasks of coping strings, exchanging portion of strings, and changing the state of bit from $1's$ to $0's$. These operators ensure that the best members of the population will survive, and their information contents are preserved and combined to generate better population(offsprings). That is to improve the performance of the next generation. It is shown in Schema theorem [4] that the genetic search algorithm will converge exponentially from the view point of schema. With the above descriptions, the procedure of a simple genetic algorithm is given as follows.

(1) Generate randomly a population of binary strings.

(2) Calculate the fitness for each string in the population.

(3) Create offspring strings by three operators (reproduction, crossover, and mutation).

(4) Evaluate the new strings and calculate the fitness for each string.

(5) If the search aim is achieved, stop and return; else go to (3).

3.2. Design Procedure

Based on the above analysis, the design procedure of spectral estimation of time series with noises and missing data is divided into the following steps.

Step 0: Given the received data $y(k)$. Compute the $\left\{\hat{r}_y\ (0), \cdots, \hat{r}_y\ (M)\right\}$..

Step 1: Generate random population of T chromosomes.

Step 2: Find the impulse response of the function $\frac{(b_0+b_1z^{-1}+\cdots+b_qz^{-q}).(b_0+b_1z^1+\cdots+b_qz^q)}{(1+a_1z^{-1}+\cdots+a_nz^{-n}).(1+a_1z^1+\cdots+a_nz^n)}$ up to M.

Step 3: Compute the minimum mean square error

$$\left\{\sum_{l=0}^{M}(r_y(l) - \hat{r}_y\ (l))^2\right\}$$

Step 4: Compute the corresponding fitness value

Step 5: Remain the best chromosome intact into next generation.

Step 6: Use genetic operators (reproduction, crossover, and mutation) to generate new chromosomes into next generation.

Then, repeat the procedure from step 2 to step 6 until a suitable parameter set is obtained.

4. DESIGN EXAMPLES

In this section, an example is given to illustrate both the design procedure and the performance of the proposed method. The number of evaluations was taken as equal to 100000 for the genetic algorithm. In the case of the genetic algorithm, as population was 1000, the generation number was equal to 100 for all functions. The simulative results are obtained by averaging 20 independent Monte Carlo (MC) runs.

Consider a system in Fig.1 with missing received signal corrupted by colored noise such that

$$\begin{aligned}\frac{B(z)}{A(z)} &= \frac{1}{1+1.57z^{-1}+2.1z^{-2}+1.4386z^{-3}+0.8409z^{-4}},\\ \sigma_n^2 &= 0.1\end{aligned}$$

The results are summarized in Fig. 2. The results indicate that the method behaves well in the cases with 40% of missing samples.

5. REFERENCES

[1] B. Porat and B. Friedlander, “ARMA estimation of time series with missing observations,” *IEEE Trans. Information Theory*, vol. IT-30, pp. 823-831, 1984.

[2] Rosen and B. Porat, “Optimal ARMA parameter estimation based on the sample covariances for data with missing observations,” *IEEE Trans. Information Theory*, vol. 35, pp. 342-349, 1989.

[3] Holland, “Outline for a logical theory of adaptive systems,” *J. ACM*, vol. 3, pp. 297-314, 1962.

[4] D. E. Goldberg, *Genetic algorithms in search, optimization, and machine learning, reading*, MA: Addition Wesley, 1989.

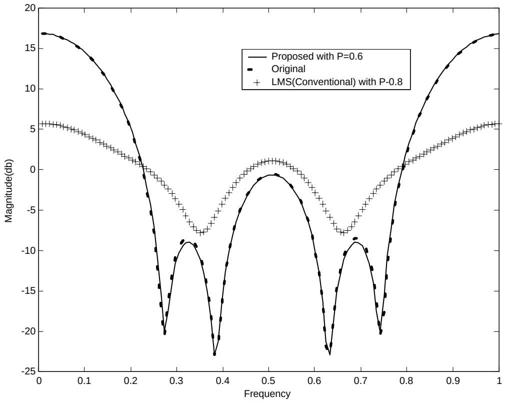

Figure 2: The power spectrum density. The solid line represents original. The dash line represents proposed method with $P = 0.6$. The + represents LMS method with $P = 0.8$.