Homework 6

Statistical Tests on Time Series

Name: Jinhan Zhang

Class: ECE 251A Digital Signal Processing I

Date: 02/28/2017

Statistical Tests on Time Series

• Objective

Use Matlab to do statistical tests on different time series, use statistical methods to test the properties of time series such as stationarity and if it is Gaussian distribution by Runs test and Chi-square goodness of fit test.

• Background

Statistical tests are usually used to make quantitative decisions concerning a process or processes. This is completed by determining whether there is enough evidence to reject a hypothesis about the process, which is called the null hypothesis. So in this assignment, we will create time series and use statistical tests to determine if they are stationary and if they are Gaussian distribution.

• Approach

Firstly, we will create four 1024-point time series. The first one is a Gaussian white noise with zero-mean and unit variance. The second one is a sinusoid ($\omega = \frac{\pi}{16}$) with $\sqrt{2}$ amplitude plus the Gaussian white noise as the first one. The third one is a sinusoid ($\omega = \frac{\pi}{16}$) with $\sqrt{2 \times 10^{0.9}}$ amplitude plus the Gaussian white noise as the first one. The fourth one is just a concatenation with first one half zero-mean unit-variance Gaussian white noise and the second half zero-mean Gaussian white noise but variance equals to 4.

Then for each time series, we firstly plot it in time domain and calculate its estimated mean, variance and standard deviation. Then we plot its histogram to see its probability density function estimate. Finally, we do power spectrum estimation using Welch's method, in which we choose averaging time K=15 and M=128 length periodograms with 50% overlap.

Before we test whether each time series is Gaussian distribution or not, we do Runs test on each time series to see if it's stationary. And then we only do chi-square test on those time series which we can't reject its stationarity.

• Results

Fig. 1-4 show the time series we build, and all their mean, variance and standard deviation are shown on the caption. Fig. 1-3 have the same mean of -0.036049, but with covariance to be 1.0238, 1.9989 and 8.09018 respectively. This is due to the introduction of sinusoid and their different amplitudes. A4 has a mean of 0.0087181 and variance of 2.5871.

Fig. 5-8 show their corresponding histograms, and we also superimpose a normal density with their corresponding estimated mean and variance on the histogram. We can see from the plot that the histogram of A1 fits the normal density curve very well, while A2 has some artifacts in the histogram compared with normal density curve, A3 is squeezed in the middle part and A4 looks like being pulled up in the middle part.

Fig. 9-12 show the smoothed estimation of the power spectrum for four time series. Fig. 9 just fluctuates around 0 dB, it's very reasonable due to that A1 is a zero-mean unit-variance Gaussian white noise. In Fig. 10 we can see 2 peaks at ± 0.03125 cycle/sample with amplitude about 16.6 dB, which corresponds to the part of sinusoid we add. Fig.11 has 2 peaks at ± 0.03125 cycle/sample with amplitude about 25.7 dB, which is due to higher amplitude of the sinusoid. Fig. 12 fluctuates around about 4 dB, which is due to that A4 has a half to be zero-mean four-variance Gaussian white noise.

Then we do runs test on these time series. We compute $\hat{E}_i[x^2(n)] = \frac{1}{10}\sum_{m=0}^{9} x^2(m+10i)$, $\mathrm{i} = 0, \ldots, 99$ and assign $+/-$ by comparing with their median, then count how many runs we get. We choose $\alpha = 0.05$ and have that $\mathrm{N} = 100$ and runs test is a two-sided test. From

Table A.6 $r_{50;1-\alpha/2} = 40$ and $r_{50;\alpha/2} = 61$, so we have our confidence interval for this runs test as $[40,61]$. Fig. 13 shows the result of sample mean-square estimates for series A1, we can't reject the hypothesis at $\alpha = 0.05$ level of significance since $r_1 = 55$ falls within the range between 40 and 61. Fig. 14 shows the result of sample mean-square estimates for series A2, we can't reject the hypothesis at $\alpha = 0.05$ level of significance since $r_2 = 59$ falls within the range between 40 and 61. For time series A3, Fig. 15 shows the result of sample mean-square estimates, and we reject the hypothesis at $\alpha = 0.05$ level of significance since $r_3 = 75$ falls outside the range between 40 and 61. Fig. 16 is the sample mean-square estimates for series A4, we can also reject the hypothesis at $\alpha = 0.05$ level of significance because $r_4 = 16$ falls outside the range between 40 and 61. All the hypothesis above refers to that "The process is stationary".

Since we only can't reject the stationarity hypothesis for time series A1 and A2, we will do chi-square test for normality of them in this part. We firstly use "norminv(k,0,1)" to generate boundaries of histogram for zero-mean unit-variance Gaussian distribution so that the probability of each bin is equal, where k equals to "[0:1/30:1]" in Matlab. Then we get 31 boundary points, multiply them with the estimated standard deviation and then plus the estimated mean. And we use these 31 boundary points to plot histograms for time series, then use "histcounts" function to count the number of points in each bin. Then for every number of points, we subtract the expected number of points (i.e. 1000/30 in this case) and then divide the square of them by the expected number of points and sum the result we will get the χ^2 value. Fig 17 and 18 is the histogram used in chi-square test for time series A1 and A2 respectively. We draw a table for the process as Table 1 and 2 below.

Table 1. Calculation for chi-square goodness of fit test A1

Interval Number	Upper Limit of Interval			f	F	$\lvert F-f \rvert$	$\frac{(F-f)^2}{F}$
	α	z_α	$x = sz + \bar{x}$				
1	0.9667	-1.83	-1.89	33	33.33	0.33	0.0033
2	0.9333	-1.50	-1.55	34	33.33	0.67	0.0135
3	0.9000	-1.28	-1.33	35	33.33	1.67	0.0837
4	0.8667	-1.11	-1.16	30	33.33	3.33	0.3327
5	0.8333	-0.97	-1.02	27	33.33	6.33	1.2022
6	0.8000	-0.84	-0.89	34	33.33	0.67	0.0135
7	0.7667	-0.73	-0.77	34	33.33	0.67	0.0135
8	0.7333	-0.62	-0.66	33	33.33	0.33	0.0033
9	0.7000	-0.53	-0.57	34	33.33	0.67	0.0135
10	0.6667	-0.43	-0.47	36	33.33	2.67	0.2139
11	0.6333	-0.34	-0.38	28	33.33	5.33	0.8524
12	0.6000	-0.25	-0.29	35	33.33	1.67	0.0837
13	0.5667	-0.17	-0.21	37	33.33	3.67	0.4041
14	0.5333	-0.08	-0.12	27	33.33	6.33	1.2022
15	0.5000	0	-0.04	39	33.33	5.67	0.9646
16	0.4667	0.08	0.04	41	33.33	7.67	1.7650
17	0.4333	0.17	0.14	40	33.33	6.67	1.3348
18	0.4000	0.25	0.22	38	33.33	4.67	0.6543
19	0.3667	0.34	0.31	33	33.33	0.33	0.0033
20	0.3333	0.43	0.40	28	33.33	5.33	0.8524
21	0.3000	0.53	0.50	25	33.33	8.33	2.0819
22	0.2667	0.62	0.60	33	33.33	0.33	0.0033
23	0.2333	0.73	0.70	37	33.33	3.67	0.4041
24	0.2000	0.84	0.81	29	33.33	4.33	0.5625
25	0.1667	0.97	0.95	40	33.33	6.67	1.3348
26	0.1333	1.11	1.09	34	33.33	0.67	0.0135
27	0.1000	1.28	1.26	25	33.33	8.33	2.0819
28	0.0667	1.50	1.48	32	33.33	1.33	0.0531
29	0.0333	1.83	1.82	29	33.33	4.33	0.5625
30	0			40	33.33	6.67	1.3348
				Sum=1000	Sum=1000		Sum=18.442
N=1000		$\bar{x}$ =-0.0360		s=1.0118		n=27	$\chi^2 = 18.442$

Table 2. Calculation for chi-square goodness of fit test A2

Interval Number	Upper Limit of Interval			f	F	$\lvert F-f \rvert$	$\frac{(F-f)^2}{F}$
	α	z_α	$x = sz + \bar{x}$				
1	0.9667	-1.83	-2.62	22	33.33	11.33	3.8515
2	0.9333	-1.50	-2.16	43	33.33	9.67	2.8055
3	0.9000	-1.28	-1.85	33	33.33	0.33	0.0033
4	0.8667	-1.11	-1.61	42	33.33	8.67	2.2553
5	0.8333	-0.97	-1.41	40	33.33	6.67	1.3348

6	0.8000	-0.84	-1.22	42	33.33	8.67	2.2553
7	0.7667	-0.73	-1.07	32	33.33	1.33	0.0531
8	0.7333	-0.62	-0.91	28	33.33	5.33	0.8524
9	0.7000	-0.53	-0.79	25	33.33	8.33	2.0819
10	0.6667	-0.43	-0.64	32	33.33	1.33	0.0531
11	0.6333	-0.34	-0.52	36	33.33	2.67	0.2139
12	0.6000	-0.25	-0.39	35	33.33	1.67	0.0837
13	0.5667	-0.17	-0.28	21	33.33	12.33	4.5613
14	0.5333	-0.08	-0.15	27	33.33	6.33	1.2022
15	0.5000	0	-0.04	35	33.33	1.67	0.0837
16	0.4667	0.08	0.08	29	33.33	4.33	0.5625
17	0.4333	0.17	0.20	35	33.33	1.67	0.0837
18	0.4000	0.25	0.32	29	33.33	4.33	0.5625
19	0.3667	0.34	0.44	26	33.33	7.33	1.6120
20	0.3333	0.43	0.57	41	33.33	7.67	1.7650
21	0.3000	0.53	0.71	42	33.33	8.67	2.2553
22	0.2667	0.62	0.84	36	33.33	2.67	0.2139
23	0.2333	0.73	1.00	22	33.33	11.33	3.8515
24	0.2000	0.84	1.15	26	33.33	7.33	1.6120
25	0.1667	0.97	1.34	32	33.33	1.33	0.0531
26	0.1333	1.11	1.53	46	33.33	12.67	4.8163
27	0.1000	1.28	1.77	43	33.33	9.67	2.8055
28	0.0667	1.50	2.08	41	33.33	7.67	1.7650
29	0.0333	1.83	2.55	27	33.33	6.33	1.2022
30	0			32	33.33	1.33	0.0531
				Sum=1000	Sum=1000		Sum=44.905
N=1000		$\bar{x}$ =-0.0360		s=1.4138		n=27	$\chi^2 = 44.905$

We choose $\alpha = 0.05$ and we have $n = 30 - 3 = 27$ in this case. Since chi-square test is one-side test, so we have $X_{27;0.05} = 40.11$. From Table 1 we can get $\chi^2 = 18.442 < X_{27;0.05}$ for time series A1, so we can't reject the hypothesis that "A1 is Gaussian distribution" at the $\alpha = 0.05$ level of significance in this test. And from Table 2 $\chi^2 = 44.905 > X_{27;0.05}$ for time series A2, so we can reject the hypothesis that "A2 is Gaussian distribution" at $\alpha = 0.05$ level of significance in this test. And also I find that A2 is not frequently rejected when I run this program for many times, this is mainly because chi-square is not a very strong test for normality, when I use Kolmogorov-Smirnov test for A2 the hypothesis gets rejected almost every time.

- **Summary**

In this assignment, we do runs test on 4 time series we generate to test their stationarities and then do chi-square goodness of fit test for normality to those time series which the runs test indicates are stationary. We get to know the process of runs test and chi-square goodness of fit test for normality, and the difference between them when we try to pick up the threshold values from their corresponding tables since runs test is two-sided test and chi-square test is one-side test. We can see that statistical test could not always pick up what is definitely right, but make decision by determining whether there is enough evidence to reject a hypothesis at some level of significance. And also, chi-square goodness of fit test is not strong enough for the normality test but it is effective to some degree.

- **Plots**

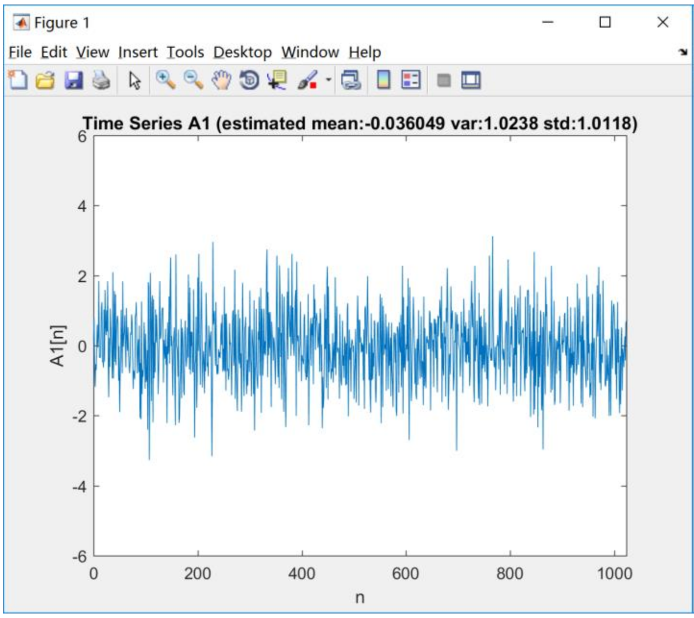

Fig. 1

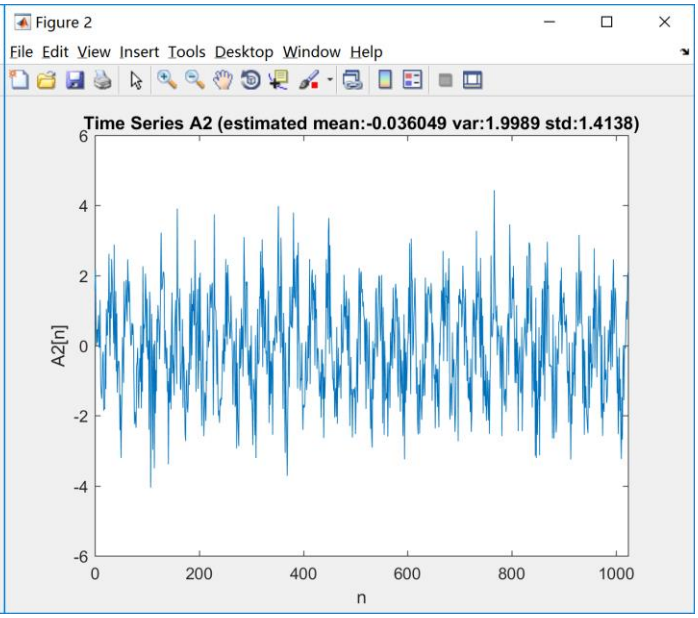

Fig. 2

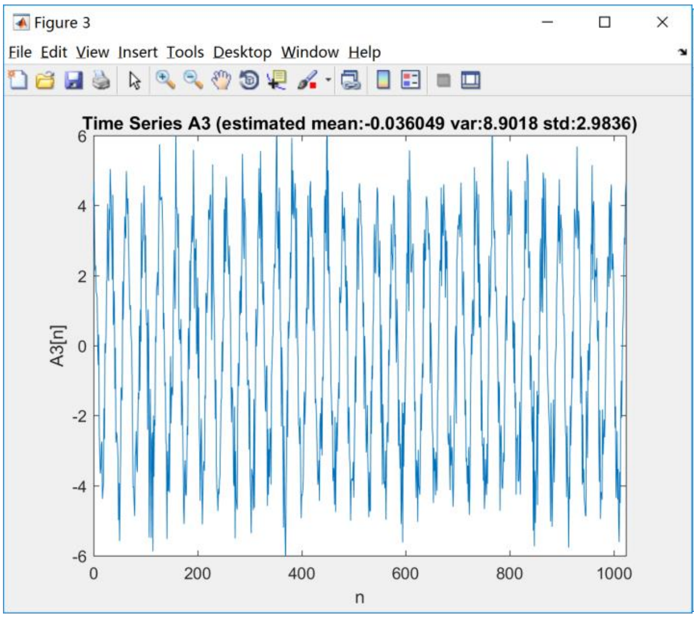

Fig. 3

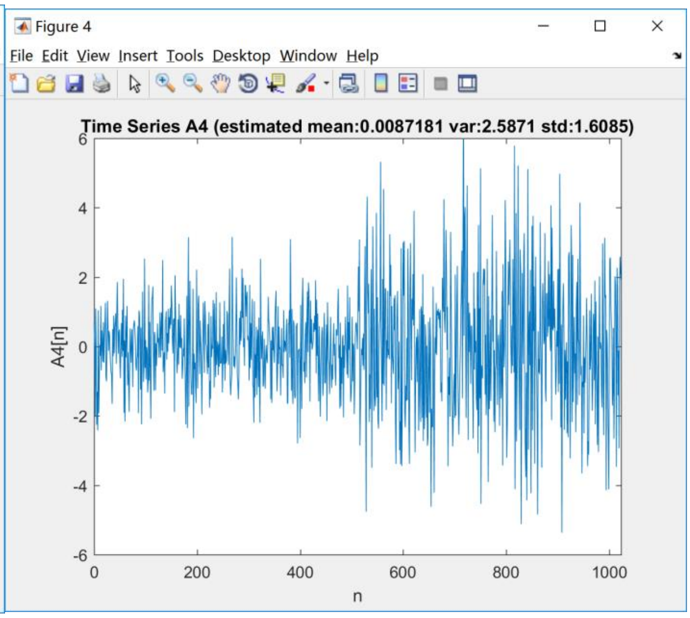

Fig. 4

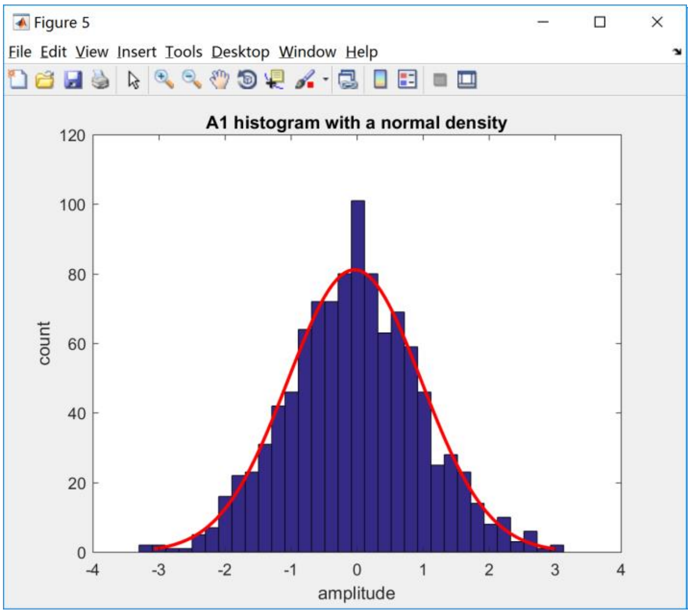

Fig. 5

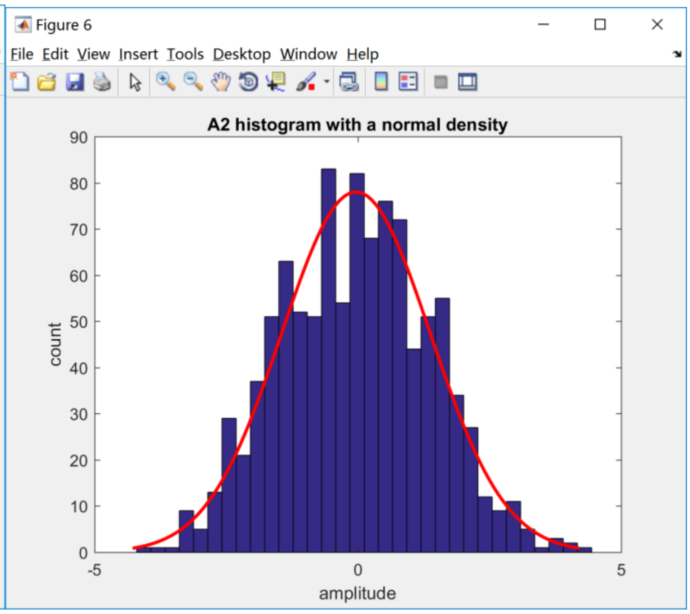

Fig. 6

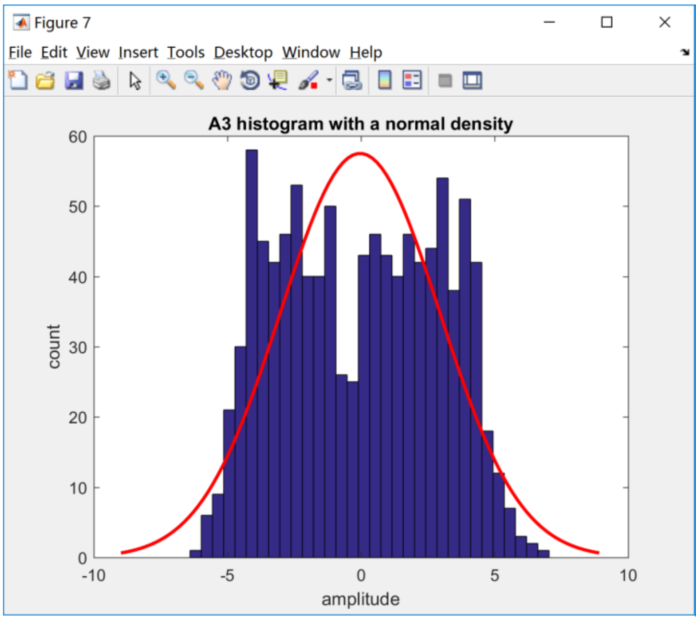

Fig. 7

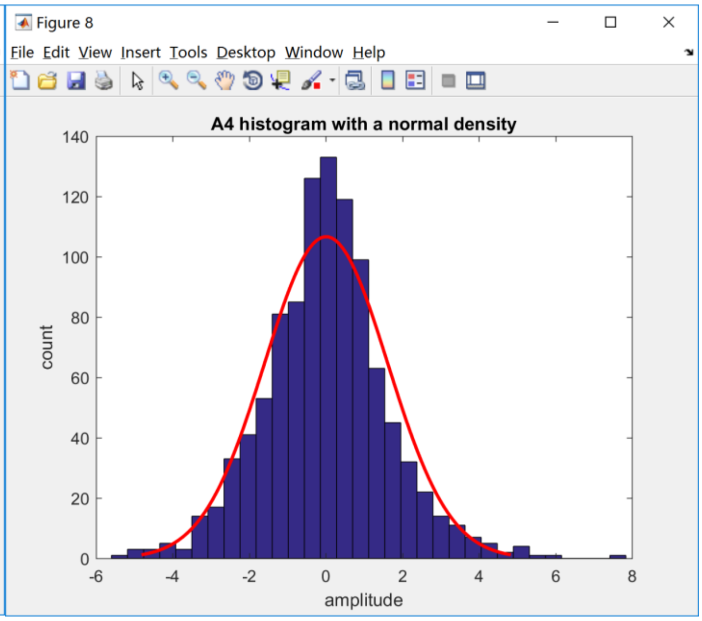

Fig. 8

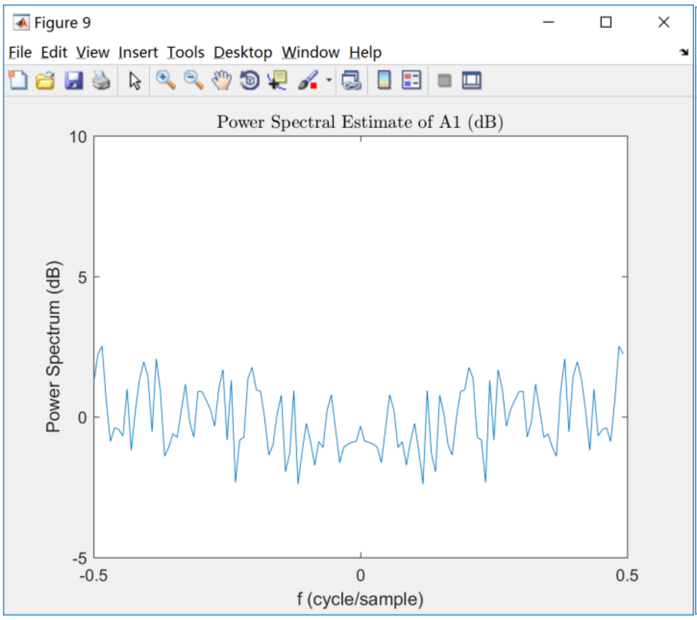

Fig. 9

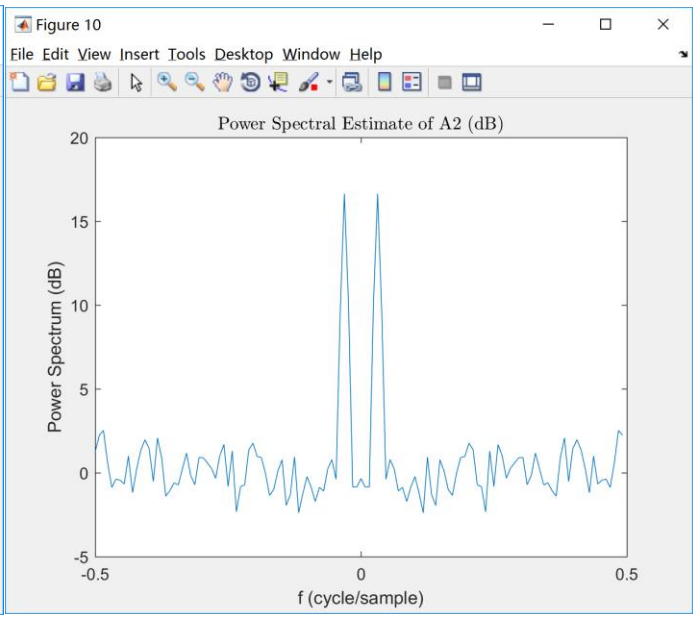

Fig. 10

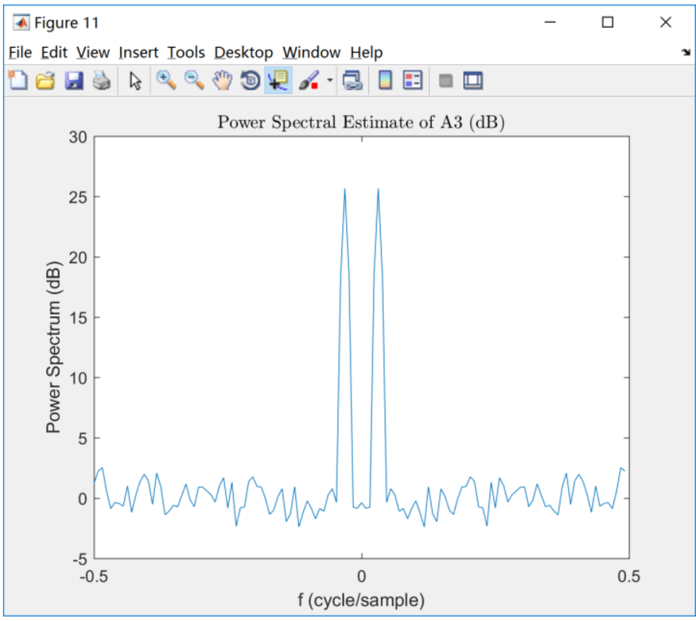

Fig. 11

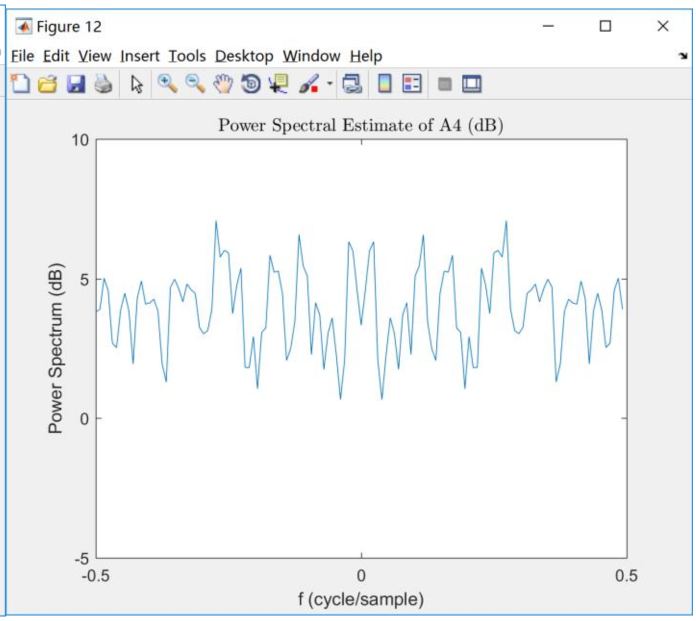

Fig. 12

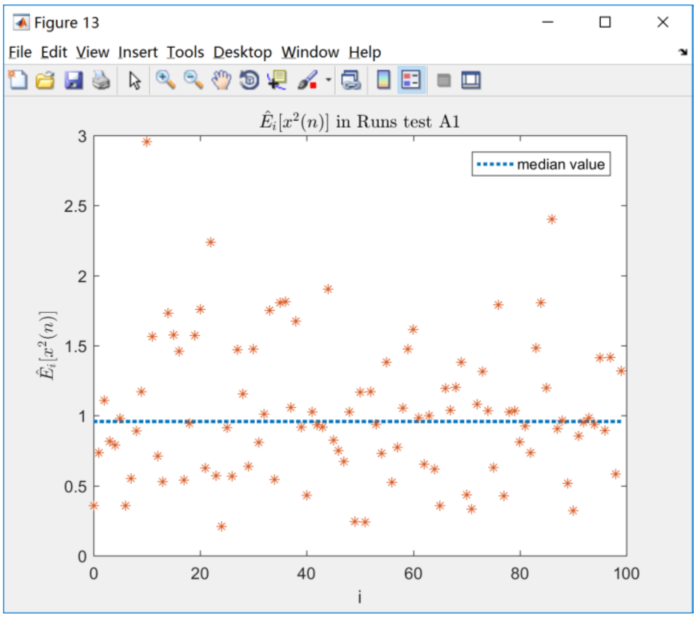

Fig. 13

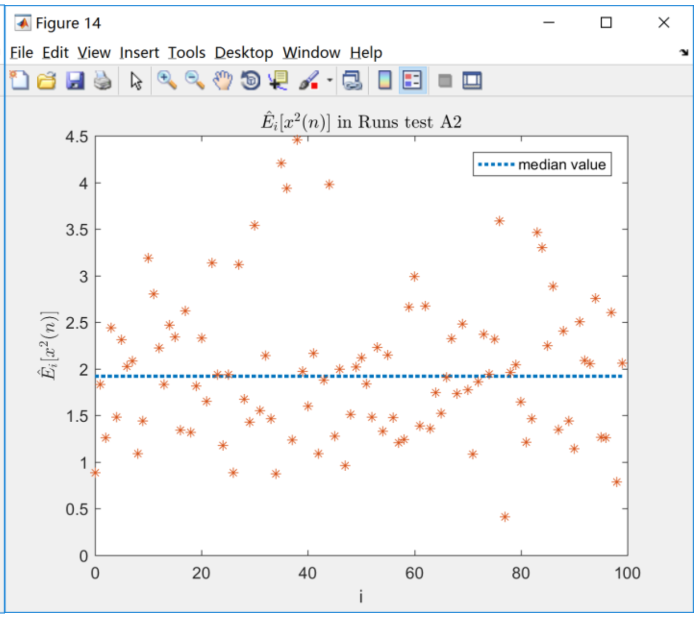

Fig. 14

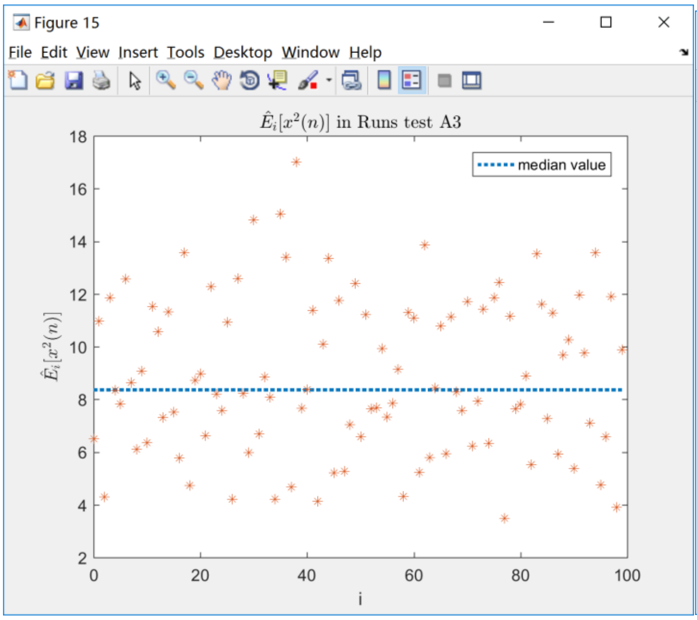

Fig. 15

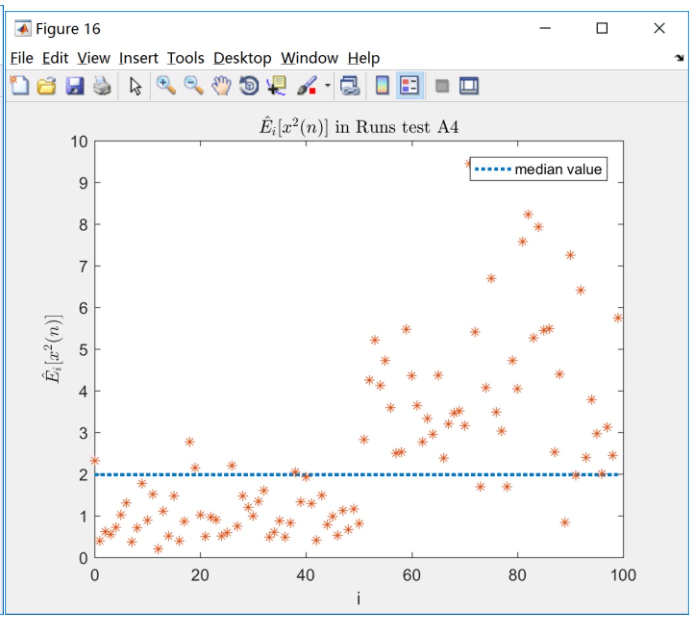

Fig. 16

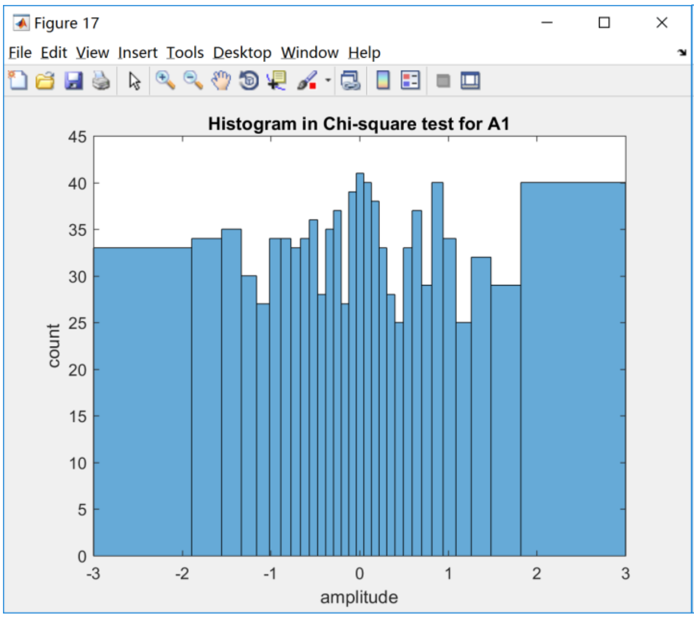

Fig. 17

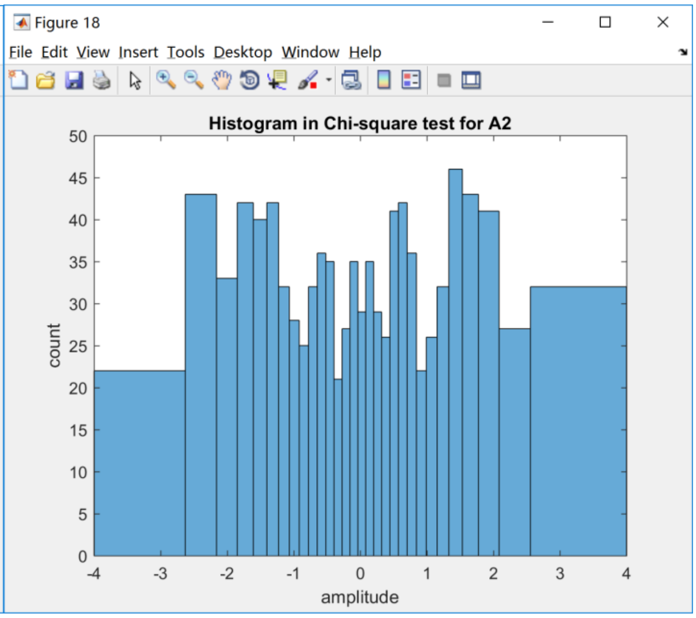

Fig. 18

- Appendix

Script:

```
clear;clc;close all;
%% Part A
rng(669);   %% random seed
N=1024;
n=0:N-1;
A1=randn([1,1024]);
A2=A1+sqrt(2)*cos(pi/16*n);
A3=A1+sqrt((10^(0.9))*2)*cos(pi/16*n);
A4_1=randn([1,512]);
A4_2=2*randn([1,512]);
A4=[A4_1 A4_2];
%% Part B1
mean1=mean(A1);
var1=var(A1);
std1=std(A1);
figure(1);
plot(n,A1);axis([0,N-1,-6,6]);
title(['Time Series A1 (estimated mean:',num2str(mean1),' var:',...
    num2str(var1),' std:',num2str(std1),')']);
xlabel('n');ylabel('A1[n]');

mean2=mean(A2);
var2=var(A2);
std2=std(A2);
figure(2);
plot(n,A2);axis([0,N-1,-6,6]);
title(['Time Series A2 (estimated mean:',num2str(mean2),' var:',...
    num2str(var2),' std:',num2str(std2),')']);
xlabel('n');ylabel('A2[n]');

mean3=mean(A3);
var3=var(A3);
std3=std(A3);
figure(3);
plot(n,A3);axis([0,N-1,-6,6]);
title(['Time Series A3 (estimated mean:',num2str(mean3),' var:',...
    num2str(var3),' std:',num2str(std3),')']);
xlabel('n');ylabel('A3[n]');

mean4=mean(A4);
var4=var(A4);
std4=std(A4);
figure(4);
plot(n,A4);axis([0,N-1,-6,6]);
title(['Time Series A4 (estimated mean:',num2str(mean4),' var:',...
```

```
    num2str(var4),' std:',num2str(std4),')']);
xlabel('n');ylabel('A4[n]');
%% Part B2
figure(5);
histfit(A1);
title('A1 histogram with a normal density');
xlabel('amplitude');ylabel('count');

figure(6);
histfit(A2);
title('A2 histogram with a normal density');
xlabel('amplitude');ylabel('count');

figure(7);
histfit(A3);
title('A3 histogram with a normal density');
xlabel('amplitude');ylabel('count');

figure(8);
histfit(A4);
title('A4 histogram with a normal density');
xlabel('amplitude');ylabel('count');
%% Part B3
NN=128;
figure(9);
S1 = fftshift(pwelch(A1,128,64,128,1,'twosided','psd'));
plot([-0.5:1/NN:0.5-1/NN],10*log10(S1));
xlabel('f (cycle/sample)');ylabel('Power Spectrum (dB)');
title('Power Spectral Estimate of A1 (dB)','Interpreter','latex');
axis([-0.5,0.5,-5,10]);

figure(10);
S2 = fftshift(pwelch(A2,128,64,128,1,'twosided','psd'));
plot([-0.5:1/NN:0.5-1/NN],10*log10(S2));
xlabel('f (cycle/sample)');ylabel('Power Spectrum (dB)');
title('Power Spectral Estimate of A2 (dB)','Interpreter','latex');

figure(11);
S3 = fftshift(pwelch(A3,128,64,128,1,'twosided','psd'));
plot([-0.5:1/NN:0.5-1/NN],10*log10(S3));
xlabel('f (cycle/sample)');ylabel('Power Spectrum (dB)');
title('Power Spectral Estimate of A3 (dB)','Interpreter','latex');

figure(12);
S4 = fftshift(pwelch(A4,128,64,128,1,'twosided','psd'));
plot([-0.5:1/NN:0.5-1/NN],10*log10(S4));
xlabel('f (cycle/sample)');ylabel('Power Spectrum (dB)');
title('Power Spectral Estimate of A4 (dB)','Interpreter','latex');
axis([-0.5,0.5,-5,10]);
```

```
%% C1
f=0:99;
[ ms1 , med1 ] = run_test( A1 );
aa=ms1-med1;
runs1=sum(xor((aa(1:end-1) > 0),(aa(2:end) > 0)));
figure(13);
plot(f,med1*ones(1,size(ms1,2)),':','LineWidth',2);hold on;
plot(f,ms1,'*');title('$\hat{E}_i[x^2(n)]$ in Runs test
A1','Interpreter','latex');
legend('median value');
xlabel('i');ylabel('$\hat{E}_i[x^2(n)]$','Interpreter','latex');

[ ms2 , med2 ] = run_test( A2 );
aa=ms2-med2;
runs2=sum(xor((aa(1:end-1) > 0),(aa(2:end) > 0)));
figure(14);
plot(f,med2*ones(1,size(ms2,2)),':','LineWidth',2);hold on;
plot(f,ms2,'*');title('$\hat{E}_i[x^2(n)]$ in Runs test
A2','Interpreter','latex');
legend('median value');
xlabel('i');ylabel('$\hat{E}_i[x^2(n)]$','Interpreter','latex');

[ ms3 , med3 ] = run_test( A3 );
aa=ms3-med3;
runs3=sum(xor((aa(1:end-1) > 0),(aa(2:end) > 0)));
figure(15);
plot(f,med3*ones(1,size(ms3,2)),':','LineWidth',2);hold on;
plot(f,ms3,'*');title('$\hat{E}_i[x^2(n)]$ in Runs test
A3','Interpreter','latex');
legend('median value');
xlabel('i');ylabel('$\hat{E}_i[x^2(n)]$','Interpreter','latex');

[ ms4 , med4 ] = run_test( A4 );
aa=ms4-med4;
runs4=sum(xor((aa(1:end-1) > 0),(aa(2:end) > 0)));
figure(16);
plot(f,med4*ones(1,size(ms4,2)),':','LineWidth',2);hold on;
plot(f,ms4,'*');title('$\hat{E}_i[x^2(n)]$ in Runs test
A4','Interpreter','latex');
legend('median value');
xlabel('i');ylabel('$\hat{E}_i[x^2(n)]$','Interpreter','latex');
%% C2
NC=1000;
k=[0:1/30:1];
interval=norminv(k,0,1);

interval1=interval*std1+mean1;
figure(17);
histogram(A1(1:NC),interval1);
```

```
cnt1 = histcounts(A1(1:NC),interval1);
cnt1 = ((cnt1 - 1000/30).^2 )/ (1000/30);
chis1=sum(cnt1);
title('Histogram in Chi-square test for A1');
xlabel('amplitude');ylabel('count');

interval2=interval*std2+mean2;
figure(18);
histogram(A2(1:NC),interval2);
cnt2 = histcounts(A2(1:NC),interval2);
cnt2 = ((cnt2 - 1000/30).^2 )/ (1000/30);
chis2=sum(cnt2);
title('Histogram in Chi-square test for A2');
xlabel('amplitude');ylabel('count');
```

Function:

```
function [ ms , med ] = run_test( A )

ms=zeros(1,100);
for i=0:99
    sub=A(i*10+1:i*10+10);
    sub=sub.^2;
    ms(i+1)=sum(sub)/10;
end;
med=median(ms);

end
```