Laboratory Project 1:

Introduction to Random Processes

Random Signals Analysis (MVE136)

Mats Viberg and Lennart Svensson
Department of Signals and Systems
Chalmers University of Technology
412 96 Göteborg, Sweden

September 2013

1 Introduction

The purpose of this laboratory exercise is to illustrate some important concepts related to random processes. You will mostly use Matlab to compute and plot various quantities, and then contemplate on the result. To get a pass, you need to hand in a report with your results, and a carefully selected set of plots (there will be several!). The deadline for handing in the report is October 4th. You should work in groups of 2 students and send in the report as a pdf-file to `maryam.fatemi@chalmers.se`. Remember to explain/motivate the results that you observe and please do not forget to label the axes in all your figures!
The laboratory is mainly about random variables, conditional dependencies, ergodicity, auto-correlation and scatter plots. Having completed this laboratory you should be able to

- describe how the conditional distribution behaves if you know the correlation between two variables (at least if they are jointly Gaussian).
- visualize the typical realizations (say in a scatter plot) for pairs of variables for different correlations.
- estimate the auto-correlation function from data (at least in simple examples).

Random processes are frequently encountered in virtually all engineering applications. This is especially true when we are dealing with real data, collected by imperfect sensors and/or transmitted of random media. The key to a successful solution is then to understand random signals, or rather *models* of such processes. In this project we consider some basic concepts used to characterize random processes. Perhaps the most important is the nature of a random process itself. A short introduction is given in Section 4. More details regarding the background theory are given in [1], Chapters 5 and 8.

2 Pairs of Uncorrelated Random Variables

The goal of this exercise is to increase the understanding of what the joint distribution of two random variables means. In the first case, you will study uncorrelated variables, using the Gaussian and Uniform distributions respectively.

Task 2.1: Histogram Use Matlab to generate $N = 2000$ random variables x_k from the N(0,1) (`randn`) and u_k U(0,1) (`rand`) distributions[1] respectively, where k indicates "realization number k". Plot the histograms using the `hist` command. Note that you can control the number of "bins" in the histogram with a second argument.

Task 2.2: Joint Distribution Generate two vectors x and y of $N = 2000$ independent N(0,1) random variables. Illustrate the joint distribution by plotting y versus x using `plot(x,y,'.')` (this is called a scatterplot). What is the shape of the "equiprobability" lines (i.e. level curves in the joint distribution)? Next, continue with two vectors u and v from the U(-a,a) distribution, where a is selected to give variance one ($a = \sqrt{3}$). What is the shape of the scatterplot, and what can you say about the level curves now? Finally, generate a scatterplot of x versus u (Gaussian versus uniform) and explain the shape.

Task 2.3: Conditional Distribution An important situation in applications is that we know the outcome of one variable (e.g. from a measurement), and we are asked to make an estimate of an unobservable quantity. A crucial concept for this situation is the conditional distribution. The distribution of x when we know the value of y is denoted $p(x|y)$. It is a slice of the joint distribution $p(x, y)$, along the x-direction at some fixed $y = \hat{y}$, and then normalized to integrate to one: $p(x|y) = p(x, y)/p(y)$. Generate samples from this distribution by choosing a fixed value $\hat{y} = 0.5$ and a "tolerance" $\delta y = 0.1$. Go through the ordered pairs (x_k, y_k) (where $(\cdot)_k$ indicates the kth realization), and select those x_ks where $\hat{y} - \Delta y < y_k < \hat{y} + \Delta y$ (note that $\Delta y > 0$ is needed since we are dealing with a continuous-valued distribution). Plot the histogram of the so selected samples, i.e., from $p(x|y = \hat{y})$. Is it different from the marginal distribution $p(x)$? Why? Repeat the experiment with the pair (x_k, u_k), fixing $u_k = 0.5$. Explain the result.

3 Pairs of Correlated Random Variables

Task 3.1: Joint Distribution and Correlation In this task you will investigate how correlation (or, more generally, dependency) changes the joint distribution.

Start with the uncorrelated N(0,1) variables x and y from Task 2.1. Now, generate a new variable z according to[2]:

$$z = \alpha x + \sqrt{1 - \alpha^2}\, y\,,\ -1 \le \alpha \le 1\,.$$

Show (theoretically) that z is also N(0,1) and that the correlation between x and z is given by $r_{xz} = E[xz] = \alpha$. You shall now illustrate the meaning of the correlation by generating scatterplots for different values $\alpha \in \{0.5, -0.5, 0.9, -0.9\}$. Verify, for at least one of the cases, that z is also N(0,1), by plotting the histogram. After seeing the scatterplots, suggest a simple way to interpret correlation between two random variables!

Task 3.2: Conditional Distribution In Task 2.3, the conditional distribution is the same as the marginal, since the variables are uncorrelated. We will now investigate how correlation changes the picture. First, generate z as above, with $\alpha = 0.7$. Then generate samples from the distribution $p(x|z = 0.5)$ using the same procedure as in Task 2.3. Plot the histogram. Repeat for $p(x|z = -0.5)$. Then generate a new z with $\alpha = -0.7$ and plot the histogram for $p(x|z = 0.5)$. Summarize your observations, and explain how correlation means that one variable holds information about the other.

[1] Generally, the `typewriter` font is used for Matlab commands in this document.

[2] Note that the actual way the variable is generated is unimportant. Only its statistical properties matter. This is very fortunate for engineers, since the exact underlying physics/chemistry/biology etc. of a complicated engineering system is usually beyond comprehension. But most of the time we can get (sufficiently) good statistical models!

4 Random Processes: White Noise

A random process in discrete time is a sequence of random variables. Thus, a process $\{x[n]\}$ has actually two dimensions: the time variable n takes values from $..., -1, 0, 1, ...$, whereas the realization is chosen from the continuous "event space", according to the specified distribution. One of the simplest random processes is the white Gaussian noise (WGN) process, which is a sequence of uncorrelated random variables with a Gaussian distribution.
To illustrate the two-dimensional nature of a stochastic process, we can think of an $N \times K$ matrix X

$$X = \begin{bmatrix} x_1[1] & \dots & x_K[1] \\ \vdots & & \vdots \\ x_1[N] & \dots & x_K[N] \end{bmatrix} . \tag{1}$$

Thus, the nth row of X contains K different realizations of a sample $x[n]$, whereas column k is one realization of the whole sequence $\{x[1], \dots, x[N]\}$, indexed by the "realization number" k. We can therefore form the *Ensemble Average* $\mu_x[n]$ of a sample $x[n]$ as

$$\mu_x[n] = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} x_k[n],$$

and the *Time Average* for the kth realization is given by

$$\bar{x}_k = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} x_k[n] .$$

Note that we are using samples from the event space rather than treating it as continuous, mimicking the process of generating many observations of the same quantity. In most cases, the distribution is continuous, and the theoretical ensemble average is computed by an integral rather than a sum. We make no difference between these cases in the present exposition.
An important concept for stochastic processes is *ergodicity*, which means that ensemble and time averaging gives the same result. For this to be possible we must clearly have $\mu_x[n] = \mu_x$, independently of n, and $\bar{x}_k = \bar{x}$, independently of k. Thus, ergodicity is related to (wide-sense) stationarity. More details are given in Section 8.3 of [1].

Task 4.1: Ensemble and Time Averages Generate a matrix X as in (1), with $K = 256$ realizations of the stochastic process $x[n]$, $n = 1, \dots, N$. Use $N = 256$ and let each $x_k[n]$ be N(0,1) WGN. In Matlab, this is conveniently done as `X=randn(N,K)`. Plot the ensemble average $\hat{\mu}_x[n]$ as a function of n by averaging along the rows in X. Use `mean(X');`. Next, plot the time average as a function of realization (`mean(X);`). Does this process appear to be ergodic in the mean? Note that the averages are different from the theoretical values due to the finiteness of N and K.

Task 4.2: Joint Distribution and Correlation Illustrate the joint distribution of $x[n_1]$ and $x[n_2]$ by scatterplots, for a few values of n_1 and n_2. Do they look uncorrelated? Verify by computing the sample ensemble correlation

$$\hat{r}_x(n_1, n_2) = \frac{1}{K} \sum_{k=1}^{K} x_k[n_1] x_k[n_2]$$

(we call it "sample" because K is finite – the true ensemble average $r_x(n_1, n_2) = E[x(n_1)x(n_2)]$ would be obtained for $K \to \infty$).

5 Random Walk Process

In this part you shall study the simple random walk process, introduced in Example 8.5 on p. 282 in [1]. The process $x[n]$ is recursively generated as

$$x[n] = x[n-1] + w[n] , \tag{2}$$

where $w[n]$ is N(0,1) WGN. The filtering introduces a correlation between successive samples of $x[n]$. The first task is to theoretically characterize this correlation. Then you shall generate scatterplots to illustrate the theoretical results.

Task 5.1 (Theoretical): Mean Value Verify that $\mu_x[n] = 0$ for all n.

Task 5.2 (Theoretical): Variance Assume that the process is started with $x[0] = 0$, so that $x[1] = w[1]$ and $E[x^2[1]] = \sigma_w^2 = 1$. Derive a formula for recursively computing $P_x[n] = E[x^2[n]]$ in terms of $P_x[n-1]$. Exploit the fact that $x[n-l]$ is uncorrelated to $w[n]$ for any $l \geq 1$. Solve the resulting difference equation and give an explicit expression for $P_x[n]$. What happens as $n \to \infty$?

Task 5.3 (Theoretical): Auto-Correlation Compute the auto-correlation $r_x(n, n-1)$ by multiplying (2) with $x[n-1]$ and take expectation. Continue with $r_x(n, n-2)$ and generalize to $r_x(n, n-l)$ for any $l > 0$. Is the process wide-sense stationary ($r_x(n, n-l)$ independent of n)? Finally, give an explicit expression for the *Normalized Correlation Coefficient*

$$\rho_x(n, n-l) = \frac{r_x(n, n-l)}{\sqrt{P_x(n)P_x(n-l)}}$$

What happens as $n \to \infty$ (for fixed l)?

Task 5.4: Joint Distribution Generate a 256×256 matrix X similar to (1), where each column is generated as in (2). This can conveniently be done using the command `x=filter(1,[1 -1],w);` for each column. Plot all realizations in the same figure using `plot(X)`. Explain what you see! Is it consistent with your theoretical calculations? Next, generate scatterplots for pairs $(x[n_1], x[n_2])$ with $(n_1, n_2) \in \{(10, 9), (50, 49), (100, 99), (200, 199)\}$ and $(n_1, n_2) \in \{(50, 40), (100, 90), (200, 190)\}$. Comment on the plots in light of the theoretical computations!

Task 5.5: Sample Auto-Correlation Compute the sample ensemble auto-correlation $\hat{r}_x(n, n-1)$ as a function of n ($n = 2 : 256$). This is done by averaging the product $x[n]x[n-1]$ along the rows in the matrix X. Plot $\hat{r}_x(n, n-1)$ versus n, together with the theoretical values $r_x(n, n-1)$ in the same plot. Note the agreement with the theoretical values. In this experiment you used $K = 256$ realizations of the same process to compute the ensemble auto-correlation. Would it be possible to estimate the auto-correlation $r_x(n, n-1)$ from one realization only, for this process?

6 Damped Random Walk

As a final example you shall study a stationary random process. This time, let $x[n]$ be generated as

$$x[n] = 0.9x[n-1] + w[n]\,, \tag{3}$$

where $w[n]$ is N(0,1) WGN. Similar to the random walk, this is a first-order auto-regressive (AR) process. You shall repeat the above calculations for this model, which looks similar at first sight, but which has quite different behavior!

Task 6.1 (Theoretical): Variance Assume again that the process is started with $x[0] = 0$ so that $x[1] = w[1]$ and $E[x^2[1]] = \sigma_w^2 = 1$. Derive the recursion formula for $P_x[n]$ and give an explicit expression for $P_x[n]$. What is the limiting variance as $n \to \infty$?

Task 6.2 (Theoretical): Auto-Correlation Compute the auto-correlation function $r_x(n, n-l) = E[x[n]x[n-l]]$. Is this process wide-sense stationary? What if $n \to \infty$?

Task 6.3: Joint Distribution Generate a 256×256 matrix X similar to (1), where each column is generated as in (3) (Matlab: `x=filter(1,[1 -0.9],w);`). Plot all realizations in the same plot using `plot(X)` and comment on the results. Generate scatterplots for pairs $(x[n_1], x[n_2])$ with $(n_1, n_2) \in \{(10, 9), (50, 49), (100, 99), (200, 199)\}$ and $(n_1, n_2) \in \{(50, 40), (100, 90), (200, 190)\}$. Comment on the plots in light of the theoretical computations. Compare with Task 5.4.

Task 6.4: Sample Auto-Correlation Compute the sample ensemble auto-correlation $\hat{r}_x(n, n-1)$ versus n, as in Task 5.5, and plot $\hat{r}_x(n, n-1)$ and the theoretical value $r_x(n, n-1)$ versus n in the same figure. Is there any significant difference from Task 5.5? Since this process is stationary (for large n), we expect the same result also from a time average. Thus, the auto-correlation can be estimated from a single realization as

$$\hat{r}_x(l) = \frac{1}{N} \sum_{n=l}^{N} x[n]x[n-l]$$

Test this approach on one or two realizations, at least for $l = 1$, and compare to the theoretical auto-correlation and to the result obtained using ensemble averaging. Keep in mind that both K and N are finite, so you should not expect a perfect agreement.

References

[1] S. L. Miller and D. G. Childers, *Probability and Random Processes With Applications to Signal Processing and Communications.* Elsevier Academic Press, Burlington, MA, 2004.