CHAPTER 2

PRELIMINARIES

2.0 Introduction

In this chapter, we present the terms, theorems and concepts that will be used in the thesis.

2.1 Parameter Estimation

In literature, many parameter estimation techniques have been discussed, for example, the method of moments, M-estimation, minimum divergence estimations and so on. In this section, we present an overview of the most commonly used estimation methods which include maximum likelihood estimation and minimum Hellinger estimation. These estimations are asymptotically consistent, asymptotically efficient, unbiased and asymptotically normally distributed under large sample.

2.1.1 Maximum Likelihood Estimation

Let $S = \{x_1, x_2, \ldots, x_N\}$ be a sample containing N independent and identically distributed (iid) observations with unknown probability function (pf), $pr(x_i, \boldsymbol{\theta}),\ i = 1,\ \ldots,\ N$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ is the vector of unknown parameters belong to a set Θ. Maximum likelihood estimation (MLE) provides the estimation of the unknown parameters $\boldsymbol{\theta}$ in a model by maximizing the sample likelihood function. The log-likelihood of $\boldsymbol{\theta}$ is defined as

$$\ln L(\boldsymbol{\theta}|S) = \prod_{i=1}^{N} pr(x_i, \boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \Theta$$

where it can be maximized by solving the first order condition

$$\frac{\partial \ln L(\boldsymbol{\theta}; x)}{\partial \theta_i} = \sum_{i=1}^{N} \frac{\partial \ln pr(x_i; \boldsymbol{\theta})}{\partial \theta_i} = 0 \tag{2.1}$$

The solutions of the equations (2.1) are denoted by $\overset{\wedge}{\theta}_1,\ \overset{\wedge}{\theta}_2,\ ...,\ \overset{\wedge}{\theta}_q$.

Succinctly, MLE estimates $\hat{\boldsymbol{\theta}} = \arg \max_{\theta} \left[\ln L(\boldsymbol{\theta}|S) \right]$ or formally, the MLE can be defined as $\log L(\hat{\boldsymbol{\theta}}|S) \geq \log L(\boldsymbol{\theta}|S)$ for all $\boldsymbol{\theta}$.

Equation (2.1) requires additional regularity conditions which are listed below:

Let denote the true value of $\boldsymbol{\theta}$.

(R1) The pfs are distinct; that is, $\boldsymbol{\theta} \neq \boldsymbol{\theta}' \Rightarrow pr(x_i; \boldsymbol{\theta}) \neq pr(x_i; \boldsymbol{\theta}')$

(R2) The pfs have common support for all $\boldsymbol{\theta}$.

(R3) The point is an interior point in Θ .

The first assumption states that the parameters identify the pfs. The second assumption implies that the support of x_i does not depend on $\boldsymbol{\theta}$*.* Additional information about the regularity conditions can be found, for instance, in Rao (1973) or Lehman and Casella (1998).

Although MLE is a powerful method, there are limitations. Firstly, the ML equations cannot be solved directly if they are complicated. Secondly, ML estimates are badly biased for misspecified models where the data can be seriously affected by outliers. The application of numerical optimization method is essential to overcome the complexity of MLE and this approach has lead to the development of the effective algorithms like the Expectation-Maximization (EM) algorithm (Dempster and Laird).

Minimum divergence methods which are robust to existence of outliers, such as the minimum Hellinger distance estimation, have also been proposed.

2.1.2 Minimum Hellinger Distance Estimation

Minimum Hellinger distance estimation (MHDE) is an attractive alternative to MLE as it is robust under model misspecification and at the same time, it retains the desirable properties of MLE. Extensive research has been done on MHDE. Beran (1977) introduced the density-based minimum Hellinger divergence estimator with robustness properties. Then, Stather (1981), Tamura and Boos (1986) and Simpson (1989) continued to contribute in this area of research. It is shown that under regularity conditions, MHDE can simultaneously enjoy the property of robustness and asymptotic efficiency of the first order.

For count data, the MHDE is defined as follows. Suppose that x_1, x_2, …, x_N are samples from a discrete parametric model with pf, $f_{\boldsymbol{\theta}}(x)$ where $\boldsymbol{\theta}$ may be a vector of parameters, which is, $\boldsymbol{\theta} \in \Theta \subset R^k$. Also, let $d(x)$ denote the empirical density function and π_x be the frequency of the observation x, where

$$d(x) = \pi_x / N, \qquad x = 0,1,2,\ .\ .\ .\ .$$

The MHDE minimizes the measure of discrepancy between a nonparametric density estimate, ($d(x)$, obtained from the data) and the $f_{\boldsymbol{\theta}}(x)$ from the parametric model, i. e., the minimum Hellinger distance estimator of $\boldsymbol{\theta}$ minimizes the quantity

$$\sum_x (d^{1/2}(x) - f_{\boldsymbol{\theta}}^{1/2}(x))^2 \tag{2.2}$$

From (2.2), it is easily seen that the numerical optimization for the MHDE is rather involved if $f_{\boldsymbol{\theta}}(x)$ is complex.

2.1.3 Simulated Annealing

The method of simulated annealing (SA) was introduced by Metropolis *et al.* (1953) as a local searching tool to minimize a criterion function on a large finite set. Also, it can be applied to an optimization on a continuous set and for simulation; see Kirkpatrick *et al.* (1983), Aldous (1987, 1900) and Neal (1993, 1995).

The term "simulated annealing" aims to draw on the analogy with the cooling process which a metal ingot experiences after it has been formed in a heating process. SA lowers the temperature slowly until the system achieves its minimal potential energy configuration and "freezes". SA operates iteratively by choosing an element *y*, which is accepted or rejected as the new configuration, from the neighbourhood of the present configuration *x*. The temperatures *T*, a choice of terminology made by analogy to the cooling process described above, controls $pr(x,y,T)$ which is the probability of accepting a move of y, given the present configuration *x*. A common choice for $pr(x,y,T)$ is the Metropolis acceptance probability for $T > 0$ as shown below:

$$pr(x,y,T) = \begin{cases} 1 \quad , & \text{if } f(y) < f(x) \\ \exp\left(\dfrac{f(y)-f(x)}{T}\right), & \text{if} f(y) < f(x) \end{cases}$$

where *f* is the objection function. Note that there are possibilities for SA to accept y even if *y* is worse than *x* and this allows the algorithm to avoid from getting stuck in local maxima.

At high temperatures (with sufficient energy), the system accepts moves almost randomly regardless whether they are uphill or downhill. Conversely, at low temperatures, the probability of making uphill moves will drop. When no further moves are accepted, the system may reach an extremum state which is in a locally or globally minimum state.

Algorithm 2.1: Simulated annealing

Purpose:

Minimize the objective function, F_*

Input:

```
Choose a configuration $x_0$;
Select the initial temperature $T_0$;
Set $x \leftarrow x_0$ and $T \leftarrow T_0$;
```

Output:

$\hat{F}_*$ as an estimate of F_* and LT denoting the point x at which $\hat{F}_*$ occurs.

Method:

```
Repeat:
    Repeat:
    Choose a new configuration $y$ from the neighborhood of $x$;
    If $f(y) \geq f(x)$
    Then $x \leftarrow y$;
     Otherwise,
        Randomly generate a $U$ from $u(0,1)$;
        If $U < \exp\left[\frac{f(y) - f(x)}{T}\right]$, $x \leftarrow y$;
        Else $x \leftarrow x$
Until iteration count $= n_{iter}$;
Decrease $T$ according to the temperature schedule;
Until stopping criterion = true;
```

X is the approximation to the optimal configuration;

End.

In short, the fundamental idea of simulated annealing is that a change of scale called temperature, allows for faster moves on the surface of the function *f* to be minimized. Ideally, SA is suitable for many optimization problems as it is powerful in searching a global extremum of a function that has many local extrema and may not be smooth. Most importantly, the method does not require calculation of derivatives as in the case with most of the optimization techniques used.

2.2 Hypothesis testing

In this section, a brief review of the chi-square goodness-of-fit test, Rao's score test and likelihood ratio test will be provided.

2.2.1 Chi-Square Goodness-of-Fit Test

The chi-square goodness-of-fit (gof) test is a popular non-parametric test introduced by Karl Pearson (1990) and is widely in use because of its simple and straightforward computation. This test judges whether the discrepancy between the probability model and the observed data is acceptable. Under the null hypothesis, it is claimed that there is no significance difference between the proposed model and the observed data. While for the alternative hypothesis, significance difference is assumed. Intuitively, a large test statistic value implies poor fitting as the observed and the expected values are not close to each other.

Let the cell probabilities be $pr(x_i, \boldsymbol{\theta}),\ i = 1,\ ...,\ k$ which involve q unknown parameters $\boldsymbol{\theta} = (\theta_1,\ ...,\ \theta_q)$. These parameters are estimated by $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1,\ ...,\ \hat{\theta}_q)$ and the estimated cell probabilities are $\hat{p}r(x_i, \boldsymbol{\theta}),\ i = 1,\ ...,\ k$.

The following formula is used to compute the chi-square test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(\pi_i - N\hat{p}r_i)^2}{N\hat{p}r_i} = \sum_{i=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where π_i is the *i*th cell of observed frequency and $N = \sum_{i=1}^{k} \pi_i$. The χ^2 test statistic is asymptotically chi square distributed with $k - 1 - q$ degrees of freedom.

For count frequency data, cells are normally combined until the expected frequencies exceed a minimum value. However, for the bivariate situation, the grouping is less straightforward as there is no obvious procedure for grouping the frequencies. The general procedure is to group several expected frequencies together.

2.2.2 Rao’s Score Test and Likelihood Ratio Test

The Rao’s score (RS) test and the likelihood ratio (LR) test converge to the same limiting chi-square distribution when the null hypothesis is true. Moreover, they share the same asymptotic power characteristic and are monotonic functions of each other. The RS test is constructed only from the restricted estimates of parameters from the null hypothesis. Meanwhile, the LR test requires both of the restricted and unrestricted estimation under the null and alternative hypothesis. When the sample size is large, the RS and LR tests are equivalent; see Cox and Hinkley, 1974.

Suppose that $S = \{x_1, x_2, \ldots, x_N\}$ is a set of sample as defined in the previous section, having probability functions $pr(x_i, \boldsymbol{\theta}_j)$, $j = 1, \ldots, q$. We consider a test of a composite hypothesis as follows

$$H_0 : R_1(\boldsymbol{\theta}) = \ldots = R_r(\boldsymbol{\theta}) = 0$$

$$H_1 : \boldsymbol{\theta} \in \Theta\text{-}\Theta_0$$

where $\boldsymbol{\theta}$ belong to a subset $\Theta_0 \quad \Theta$.

I. Rao's Score Test

The score test statistic for the H_0 given above is

$$T = \left(V(\hat{\boldsymbol{\theta}})' I(\hat{\boldsymbol{\theta}})^{-1} V(\hat{\boldsymbol{\theta}}) \right) \sim \chi^2(r) \tag{2.3}$$

where $\hat{\boldsymbol{\theta}}$ is the MLE under the restrictions of the composite hypothesis, i.e. $\hat{\boldsymbol{\theta}} = \arg \max\limits_{R_1(\boldsymbol{\theta}) = \ldots = R_r(\boldsymbol{\theta}) = 0} \ln L(\boldsymbol{\theta} | S)$.

$V(\hat{\boldsymbol{\theta}})$ denotes the *i*th efficient score vector function given as

$$\begin{aligned} V(\hat{\boldsymbol{\theta}}) &= \left(\frac{\partial \ln L(\hat{\boldsymbol{\theta}} | S)}{\partial \theta_i} \right) \\ &= \left(\frac{1}{pr(x_1, \hat{\boldsymbol{\theta}})} \frac{\partial pr(x_1, \hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_i} + \ldots + \frac{1}{pr(x_n, \hat{\boldsymbol{\theta}})} \frac{\partial pr(x_n, \hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_i} \right), \quad i = 1, \ldots, q \end{aligned}$$

with *N* being the sample size.

If $I(\boldsymbol{\theta})$ is the information matrix evaluated at the MLE $\hat{\boldsymbol{\theta}}$, we write

$$I(\hat{\boldsymbol{\theta}}) = E\left\{ -\left[\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}} | S)}{\partial \theta_m \partial \theta_n} \right] \right\}, \; m, n = 1, \ldots, q \; .$$

II. Likelihood Ratio Test

The LR test compares the maximized log-likelihood values for the null and alternatives models. Hence we need to compute the LR criterion for the composite hypothesis which is expressed in term of the likelihood function in order to apply the LR test. The LR statistic is

$$\lambda = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta} | S)}{\sup_{\boldsymbol{\theta} \in \Theta - \Theta_0} L(\boldsymbol{\theta} | S)}$$

where the numerator is the constrained maximization of the likelihood function under the null hypothesis and the denominator is obtained through unconstrained maximization of the likelihood function under the alternative hypothesis.

Under certain regularity conditions, the test statistic is expressed as

$$LR = -2\left(\ln L(\hat{\boldsymbol{\theta}} | S) - \ln L(\tilde{\boldsymbol{\theta}} | S)\right) \sim \chi^2(r)$$

where

$\hat{\boldsymbol{\theta}} = \arg \max_{R_1(\boldsymbol{\theta})=\ldots=R_r(\boldsymbol{\theta})=0} \left[\ln L(\boldsymbol{\theta} | S)\right]$, $\tilde{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta - \Theta_0} \ln L(\boldsymbol{\theta} | S)$ and r = number of independent restrictions.

The calculation of LR test statistic requires fitting of two models compared to the RS test. Simulation studies suggest that the LR test may appear to be better when the given sample size is small.

2.3 Statistical Power Analysis

The power of a statistical test is the probability which leads to rejection of the null hypothesis when it is untrue, that is, the probability of making the correct decision. A specific statistical power analysis is built according to the characteristics of the null hypothesis and it is greatly affected by the choice of significance level, sample size and effect size. The ideal power value suggested by Cohen (1988, page 56) to detect a reasonable departure from the null hypothesis is 0.80.

In this section, we briefly discuss the significance level, the reliability of the sample and sample size and also the effect size.

2.3.1 Significance Level

The significance level determines the critical region of rejecting the null hypothesis. Mathematically, it is termed the error of the first kind, the Type I error, which is the probability of rejecting the null hypothesis when it is true. Since it is the rate of accepting the alternatives hypothesis, it should be taken as a relatively small value. Similarly, a more stringent standard of rejection implies a smaller value of the significance level. Typical values for the significance level are 0.01, 0.05 and 0.10.

2.3.2 Reliability of Sample and Sample Size

It is impractical and difficult for a researcher to collect data from the entire population and normally a sample is obtained from the population for inference. Hence, the use of sampling and the reliability of the chosen sample are important considerations as decisions are frequently made based on the sample data. As mentioned by Cohen (1988, p.8) “Whatever else sample reliability may be dependent upon, it always depends upon the size of the sample”. If the sample size is too small, the precision of the sample which can be expected to approximate the population of interest is extremely low. If the sample size is larger, the margin error will be reduced and of course the results will provide greater reliability. Clearly, sample size is the invariant feature of the sample precision and it is one of the factors which we use to determine the power of the test.

2.3.3 The Effect Size

The effect size is an index of degree of departure from the null hypothesis. When the null hypothesis is true, the effect size takes the value zero. Increase in the effect size (with significance level and sample size fixed) is the same as the increase of the departure from the null hypothesis and this implies an increase in the power of the statistical test.

2.4 Construction of Bivariate Discrete Distribution

Various methods have been introduced to construct bivariate distributions, such as the construction of copulas, mixing and compounding, trivariate reduction, mixtures of convolution, transformation and so on; see Balakrishnan and Lai (2009), Chapter 5. In this section, we discuss two most commonly used methods which are the mixtures of convolution and trivariate reduction method.

2.4.1 Mixtures of Convolution

Let $\varphi_1(t_1,t_2)$ and $\varphi_2(t_1,t_2)$ be the probability generating functions (pgf) of two independent discrete bivariate distributions and their joint probability mass functions be $f_1(x_1,x_2)$ and $g_1(x_1,x_2)$, respectively. Through the mixture of the convolution of the pgfs, a new bivariate distribution defined as $\varphi(t_1,t_2)=\varphi_1(t_1,t_2)\varphi_2(t_1,t_2)$ can be formed. The joint probability mass function of this mixture of convolution can be determined as follows:

Let

$$(\varphi_1\cdot\varphi_2)^{(m,n)}=\sum_{k=0}^{m}\sum_{\ell=0}^{n}\binom{m}{k}\binom{n}{\ell}\varphi_1^{(m-k,n-\ell)}\varphi_2^{(k,\ell)} \tag{2.4}$$

where $\varphi_1^{(i,j)}(t_1,t_2)=\dfrac{\partial^{i+j}}{\partial^i t_1 \partial^j t_2}\varphi_1(t_1,t_2)$.

By applying

$$(\varphi_1 \cdot \varphi_2)^n = \sum_{k=0}^{n}\binom{n}{k}\varphi_1{}^k \varphi_2{}^{n-k}$$

$$\varphi(t_1,t_2)^{(x_1,x_2)} = \sum_{k=0}^{x_1}\sum_{\ell=0}^{x_2}\binom{x_1}{k}\binom{x_2}{\ell}\varphi_1{}^{(x_1-k,x_2-\ell)}\varphi_2{}^{(k,\ell)}$$

$$\varphi(t_1,t_2)^{(x_1,x_2)}\,|_{t_1=t_2=0} = \sum_{k=0}^{x_1}\sum_{\ell=0}^{x_2}\binom{x_1}{k}\binom{x_2}{\ell}\varphi_1{}^{(x_1-k,x_2-\ell)}\varphi_2{}^{(k,\ell)}\,|_{t_1=t_2=0}$$

$$\frac{1}{x_1!x_2!}f(x_1,x_2) = \sum_{k=0}^{x_1}\sum_{\ell=0}^{x_2}\binom{x_1}{k}\binom{x_2}{\ell}\frac{1}{(x_1-k)!(x_2-\ell)!}f_1(x_1-k,x_2-\ell)\frac{1}{k!\ell!}g_1(k,\ell)$$

The joint probability mass function of the bivariate distribution is easily obtained as

$$f(x_1,x_2) = \sum_{k=0}^{x_1}\sum_{\ell=0}^{x_2} f_1(x_1-k,x_2-\ell)\,g_1(k,\ell) \tag{2.5}$$

2.4.2 Trivariate Reduction

The method of trivariate reduction has been discussed in Mardia (1970) where three independent random variables from the same family of distributions have been combined in an appropriate way to form two correlated random variables, X_1 and X_2, expressed as

$$X_1 = (Y_1 + Y_2) \tag{2.6}$$

$$X_2 = (Y_1 + Y_3) \tag{2.7}$$

where Y_i, $i = 1,2,3$ are iid random variables. Let the pgf's of the random variable, Y_i be $\varphi_i(t)$, i=1, 2, 3. The random variable (X_1, X_2) defined by (2.6) and (2.7) has a joint pgf given by

$$\varphi(t_1,t_2) = \varphi_1(t_1)\varphi_2(t_2)\varphi_3(t_1t_2)\,. \tag{2.8}$$

Through successive differentiating of (2.8) we have

$$\varphi^{(x_1,x_2)}(t_1,t_2)=\sum_{i=0}^{x_1}\sum_{h=0}^{x_2}\sum_{k=0}^{\min(x_1-i,x_2-h)}\frac{x_1!x_2!}{i!h!k!(x_1-i-k)!(x_2-h-k)!}t_1^{x_2-h-k}t_2^{x_1-i-k}\varphi_1^{(i)}(t_1)\varphi_2^{(h)}(t_2)\varphi_3^{(x_1+x_2-i-h-k)}(t_1t_2) \tag{2.9}$$

and the joint pmf can be obtained by dividing (2.9) with $x_1!$ $x_2!$ and setting $t_1=t_2=0$. The joint pmf is

$$f(X_1=x_1,X_2=x_2)=\sum_{k=0}^{\min(x_1,x_2)}pr_1(x_1-k)pr_2(x_2-k)pr_3(k) \tag{2.10}$$

where $pr_i(k)$ is the pmf of the random variable X_i.

2.5 Properties of Bivariate Distributions

2.5.1 Probability Generating Function and Joint Probability Function

The use of transforms, in particular, pgf in the mathematical sciences is growing rapidly due to its simpler form and uniqueness. The joint pgf of the random variable (X_1, X_2) with joint pmf $f\left(x_1,x_2\right)$ is defined as

$$\varphi\left(t_1,t_2\right)=E(t_1^{x_1}t_2^{x_2})=\sum_{x_1,x_2}t_1^{x_1}t_2^{x_2}f(x_1,x_2) \tag{2.11}$$

Note that the series (2.11) converges absolutely if $|t_1|\le 1$, $|t_2|\le 1$. For a given joint pgf, we can apply series expansion or by taking derivatives to recover the joint pmf. These methods, as illustrated by Kocherlarkota & Kocherlarkoata (1992), are shown as follows.

(i) Expand the pgf, $\varphi\left(t_1,t_2\right)$ in powers of t_1 and t_2 as shown in (2.11). It is readily seen that the coefficient of the term $t_1^{x_1}t_2^{x_2}$ will provide us the joint pmf, $f(x_1,x_2)$.

(ii) Before evaluating at $t_1 = 0$ and $t_2 = 0$, the joint pgf is differentiated repeatedly with respect to t_1 and t_2 . The joint pmf is obtained as

$$f(x_1, x_2) = \frac{1}{x_1\,!}\frac{1}{x_2\,!}\frac{\partial^{x_1+x_2}}{\partial t_1^{x_1} \partial t_2^{x_2}} \varphi(t_1, t_2)\bigg|_{t_1=t_2=0}$$

2.5.2 Marginal and Conditional Distributions

Solving for marginal and conditional distributions are essential as the marginal distributions describe the individual behavior of the random variables, X_1 and X_2 while the conditional distributions give the underlying dependence structure between each pair of variables. In addition, they provide great assistance in computer simulation studies.

I. Marginal distributions

The marginal pgf of the bivariate distribution can be easily determined from its corresponding joint pgf. Let the joint pmf of $\left(X_1, X_2\right)$ be $f(x_1, x_2)$ and the marginal probability functions be $pr_1(X_1 = x_1) = \sum_{x_2} f(x_1, x_2)$ and $pr_2(X_2 = x_2) = \sum_{x_1} f(x_1, x_2)$.

Then, the marginal pgf's are

$$\varphi_{X_1}(t) = \sum_{x_1} pr_1(x_1)t^{x_1} = \sum_{x_1}\sum_{x_2} f(x_1, x_2)t^{x_1} = \varphi(t, 1) \tag{2.12}$$

$$\varphi_{X_2}(t) = \sum_{x_2} pr_2(x_2)t^{x_2} = \sum_{x_1}\sum_{x_2} f(x_1, x_2)t^{x_2} = \varphi(1, t) \tag{2.13}$$

II. Conditional distributions

Let $\varphi(t_1, t_2)$ be the joint pgf of (X_1 , X_2). Then, the conditional pmf of X_2 given $X_1 = x_1$ is $pr(X_2 \big| X_1 = x_1) = \dfrac{f(x_1, x_2)}{pr_1(x_1)}$.

Theorem 2.1: (Kocherlakota & Kocherlakota, 1992)

The pgf of the conditional distribution of X_2 given $X_1 = x_1$ is

$$\varphi_{X_2}(t|X_1 = x_1) = \frac{\varphi^{(x_1,0)}(0,t)}{\varphi^{(x_1,0)}(0,1)} \tag{2.14}$$

where $\varphi^{(x_1,x_2)}(t|X_1 = x_1) = \frac{\partial^{x_1+x_2}}{\partial t_1^{x_1} \partial t_2^{x_2}} \varphi(t_1,t_2)\Big|_{t_1=u,t_2=v}$.

A corollary of this theorem, gives the regression of X_2 on X_1.

Corollary 2.1

The regression of X_2 on X_1 is $E(X_2|X_1 = x_1) = \frac{\varphi^{(x_1,1)}(0,1)}{\varphi^{(x_1,0)}(0,1)}$

2.5.3 Covariance and correlation

The interdependence between a pair of random variables is always a concern in the statistical analysis and many possible measures of dependence have been proposed. In this thesis, the correlation coefficient is considered as a measure of dependence. The covariance of X_1 and X_2 is defined as

$$Cov(X_1, X_2) = E\left[\left(X_1 - E(X_1)\right)\left(X_2 - E(X_2)\right)\right] \tag{2.15}$$

The covariance (2.15) is standardized to be dimension free coefficient and is termed the correlation between X_1 and X_2

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sqrt{\sigma^2(X_1)\sigma^2(X_2)}} \tag{2.16}$$

where $\sigma^2(X_1)$ and $\sigma^2(X_2)$ are the variance for the random variables X_1 and X_2. Note that ρ_{X_1,X_2} must lies between -1 and 1.