

# EXPLORATORY DATA ANALYSIS WITH BIVARIATE DEPENDENCE FUNCTIONS

J. Y. Shin, N. J. Park, Nohpill Park, and K.M. George  
Oklahoma State University  
Computer Science  
Stillwater, OK 74078, USA  
Phone: 405-744-5668, Fax: 405-744-9097  
E-mail: {jyious, noh, npark, kmg}@a.cs.okstate.edu

## Abstract

*Dependence functions are used to construct joint distributions with fixed marginals. They can shed light on relationships among associated random variables. Many dependence functions have been proposed and standardized. However, there has not been an attempt to understand why certain dependence functions are used and what makes certain dependence functions better than others in solving practical problems. In this paper, we compare an approach to dependence function which identified and characterized the density weighting function and a class of bivariate densities constructed from polygonal covariance characteristic that would be flexible enough to capture various dependence structures. And, we also use the copulas for dependent random variables to compare the effects behind the dependence function.*

## Keywords

Dependence function, Bivariate extreme value, Density weighting function, Covariance characteristic, Dependence structure, Copulas.

## 1. Introduction

Dependence functions are used to construct joint distributions with fixed marginals. They can shed light on relationships among associated random variables. Many methods of constructing bivariate distributions with specified marginals have been proposed and standardized by several authors. One of the simplest forms to construct bivariate distribution function is Copulas which link joint distributions to their marginals and reveals their dependence structure. A number of well known classical distributions such as the generalized Farlie-Gumbel-Morgenstern (FGM) distribution[5] are developed in early 1990. Kotz and Seeger[3] published a treatise on a class of bivariate densities in 1991. They identified and characterized the

density weighting function for known classes of bivariate distributions such as FGM, Gumbel type I, Gaussian and Pareto etc.

In 1995, Dou Long and Roman Krzysztofowicz[1] offered a model for bivariate density which is constructed from specified marginals and a dependence structure modelled in terms of a functional. Various forms of this functional lead to different families of models. In their paper, one family is investigated in depth; its bivariate densities assume a polygonal dependence structure and prescribe positive (or negative) mutual regression dependence between variates. In this paper, we use simulation results of a class of bivariate densities which are identified and characterized by three different ways such as i) copulas, ii) density weighting function, and iii) dependence structure function, and determine the effects behind the dependence function.

Our work is motivated by the need to develop methodologies for identifying suitable dependence models for practical applications. In this paper, we present results of three models applied to two random variables. Results are presented with varying parameter values. The resulting graphs illustrate the significance of the parameter values. This paper is organized as follows. In section 2, literature related to dependence function is reviewed. The basic principles and details of the applied model are described in section 3. The experimental simulation of the applied model is evaluated in section 4. In the final section, remarks and conclusion are presented.

## 2. Literature Review

Various methods of constructing bivariate distributions with specified marginals have been recently investigated [1, 3, 4, 8, 11, 12, 13]. The study of bivariate distributions was long confined to the normal case. Gumbel[2] says it seems appropriate to study bivariate distributions where the marginal distributions, namely the logistic are similar to

the normal one and to compare their properties to those of the classical bivariate normal one. The logistic distribution closely resembles the normal one, and both are symmetrical. Two logistic bivariate distributions are studied by Gumbel[2]. Kotz and Seeger[3] published a treatise on a class of bivariate densities. They also re-examined the concept of density weighting function and a new constructive approach to the generation of dependence between random variables based on this concept is proposed. According to Frank, a two parameter family of binary operation which consists of measurable function and copula (or dependence function) arises naturally as the distributional counterpart of operations on random variables[11]. Recently, Johnson and Whitt[4,7] investigated the construction of families of bivariate distributions with specified marginals, and, especially, Johnson presents the weighted linear combination method (WLC). The class of bivariate distribution resulted from standard normal marginals is compared with the standard normal bivariate distribution, and it is found that the two joint distributions and the two conditional distributions each agree tolerably well by Placktt[5]. Furthermore, Shaked, Marshall and Olkin[6, 8] introduced a family of concepts of stochastic dependence for bivariate distribution functions and a family of bivariate distributions with marginals as parameters. Previous empirical studies of the weighting function have suggested an inverse S-shaped function, first concave and then convex. According to Wu and Gonzalez[12], the probability weighting function let probabilities to be weighted nonlinearly. The construction of a probabilistic model is a key step in most decision and risk analyses. Clemen and Reilly[13] constructed the probabilistic model with an approach using a copula to construct joint distributions and correlations to incorporate dependence among the variables. In this paper, we study three dependence models. They are described in detail in the next section.

### 3. Dependence functions

In this section, we present the theoretical models on dependence functions considered in this paper. To evaluate the models, we used publicly available stock market data.

#### 3.1. Dependence by copulas

In this section, we review the definition and state the properties of copula for bivariate case[14].

**Definition of copula:** A copula is a function

$C : [0,1]^2 \rightarrow [0,1]$  which satisfies:

- i) for every  $u, v$  in  $[0,1]$ ,  $C(u,0) = 0 = C(0,v)$ , and  $C(u,1) = u$  and  $C(1,v) = v$
- ii) for every  $u_1, u_2, v_1, v_2$  in  $[0,1]$  such that  $u_1 \leq u_2$  and  $v_1 \leq v_2$ ,  $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$ .

The importance of coplas is described in Sklar's Theorem [14].

**Sklar's theorem:** Let  $X$  and  $Y$  be random variables with joint distribution function  $H$  and marginal distribution function  $F$  and  $G$ , respectively. Then there exists a copula  $C$  such that

$$H(x, y) = C(F(x), G(y)), \text{ for all } x, y \text{ in } \mathfrak{R} \quad \text{--- (1)}$$

In the other direction, given  $C$  in  $\Omega$  which is the set of copulas, and distributions  $F$  and  $G$ , the function  $K$  defined by (1) is a bivariate distribution with margins  $F$  and  $G$ . In particular, if  $X$  and  $Y$  are extended real valued random variables, defined on a common probability space, with individual distribution  $F_X$  and  $F_Y$  and joint distribution  $F_{X,Y}$ , then there is a  $C_{X,Y}$  in  $\Omega$  such as  $F_{X,Y}(u, v) = C_{X,Y}(F_X(u), F_Y(v))$ . If  $F_X$  and  $F_Y$  are continuous,  $C_{X,Y}$  is unique. The copula in (2) is said to connect  $X$  and  $Y$ . We shall refer to  $C_{X,Y}$  as a copula of  $X$  and  $Y$ . The copula of two random variables thus reveals their dependence structure. For instance,  $X$  and  $Y$  are independent precisely when they are connected by the copula as follows:

$$C(s, t) = st \quad \text{--- (2)}$$

#### 3.2. Dependence by density weighting function

Kotz and Seeger[3] published a treatise on a class of bivariate densities of the form  $h(x, y) = f(x)g(y)\varphi(x, y)$ . They identified and characterized the density weighting function  $\varphi$  for known classes of bivariate distributions such as Farlie-gumbel-Morgenstern (F.G.M.). Then, we can approach by using F.G.M. class with absolutely continuous, strictly increasing marginals  $F_1$  and  $F_2$  with respective densities  $f_1$  and  $f_2$ . We shall denote the F.G.M. c.d.f.  $H_\theta(x, y)$  as follows:

$$H_\theta(x, y) = F_1(x) \times F_2(y) [1 + \theta(1 - F_1(x))(1 - F_2(y))] \quad \text{--- (3)}$$

and by  $h_\theta(x, y)$  the corresponding density which is known to be

$$h_\theta(x, y) = f_1(x) \times f_2(y) [1 + \theta(1 - 2F_1(x))(1 - 2F_2(y))] \quad \text{--- (4)}$$

where  $\theta \in [-1,1]$ . In the expression for  $h_\theta(x, y)$ , the independent density,  $f_1(x) \times f_2(y)$ , is multiplied by a

function of  $x$  and  $y$  dependent on  $\theta$  which we shall denote by  $\varphi_\theta(x, y)$ .

Expanding, we get

$$\varphi_\theta(x, y) = 1 + \theta - 2\theta\{F_1(x) + F_2(y)\} + 4\theta F_1(x) \times F_2(y) \quad \text{--- (5)}$$

### 3.3. Dependence structured by functionality

Long and Krzysztofowicz[1] introduced covariance characteristic as a polygonal dependence structure. The bivariate density  $h$  we are applying takes the form as follows:

$$h(x, y) = f(x)g(y)[1 + \theta c(F(x), G(y))] \quad \text{--- (6)}$$

where  $c$ , termed a covariance characteristic, is a continuous function on the unit square  $[0,1]^2$  satisfying the following property: for all  $0 \leq u, v \leq 1$ , as follows:

$$\int_0^1 c(u, v) du = \int_0^1 c(u, v) dv = 0 \quad \text{--- (7)}$$

For a specified  $c$ , constant  $\theta$ , termed the covariance scalar, is constrained, either below or above, in order that for all  $0 \leq u, v \leq 1$ , as follows:

$$1 + \theta c(u, v) \geq 0 \quad \text{--- (8)}$$

It can be verified immediately that under conditions (7) and (8), function  $h$  is a proper bivariate density having  $f$  and  $g$  as its marginals. The challenge en route to operationalizing model (6) is the development of parametric forms for the covariance characteristic  $c$  that would be flexible enough to capture various dependence structures and that would lend themselves to estimation from data with a reasonable expenditure of computational effort. In the sequel we detail the development of  $c$  having a polygonal structure in covariance characteristic case. Let  $(U, V)$  denote a pair of uniform random variables obtained through transformations of the original random variables  $(X, Y)$  such that  $U = F(X)$ ,  $V = G(Y)$ , and  $0 \leq U, V \leq 1$ . The sample space of uniform random variables is partitioned into four polygons by drawing two lines. Then, let  $k$  be a continuous and monotonic function on the unit interval;  $k$  is termed the regression characteristic. For every  $0 \leq \omega \leq 1$ , define as follows:

$$K(\omega) = \int_0^\omega k(t) dt \quad \text{--- (9)}$$

Next, define two bivariate functions,  $c_1$  and  $c_2$ , such that for any  $0 \leq u, v \leq 1$ ,

$$c_1(u, v) = k(u - v) \quad \text{if } v \leq u$$

$$= k(v - u) \quad \text{if } v \geq u \quad \text{--- (10)}$$

and

$$c_2(u, v) = k(u + v) \quad \text{if } u \leq 1 - v$$

$$= k(2 - u - v) \quad \text{if } u \geq 1 - v \quad \text{--- (11)}$$

and let

$$c(u, v) = c_1(u, v) + c_2(u, v) - 2K(1) \quad \text{--- (12)}$$

It can be verified that  $c$  specified by (10) – (12) satisfies the necessary property (7).

The covariance characteristic so constructed is a four-piece continuous function on the polygonal partition of the unit square.

## 4. Performance evaluation through simulation

The analytic models presented in the previous section provide a basic understanding of the performance impact of dependence functions constructed by three different ways such as i) copulas in (2), ii) density weighting function in (4), and iii) dependence structure function in (6), respectively shown as below. Also, the model does not consider  $\beta$  value, sufficiently, which tends to get more convex surface as  $\beta$  value gets larger.

We apply the presented models under the hypothesis that the type of distribution has no effect for given  $\beta$  value.

The models investigated in this study are listed as follows:

i) copula:

$$h(x, y) = f(x)g(y)$$

$$\text{where } f(x) = \frac{1}{\alpha} \exp\left(\frac{x - \mu}{\alpha}\right) \exp\left(-\exp\left(\frac{x - \mu}{\alpha}\right)\right),$$

$$g(y) = \frac{1}{\alpha} \exp\left(\frac{y - \mu}{\alpha}\right) \exp\left(-\exp\left(\frac{y - \mu}{\alpha}\right)\right)$$

$$\mu : \text{location parameter } (= \bar{x} - \alpha\gamma)$$

$$\alpha : \text{scale parameter } (= \sqrt{\frac{\sigma_x \times 6}{\pi^2}})$$

$$\gamma : \text{Euler-Mascheroni constant } = 0.57721\dots$$

ii) density weighting function:

$$h(x, y) = f(x)g(y)[1 + \theta(1 - F(x))(1 - G(y))]$$

$$\text{where } \theta = \rho \frac{(\beta + 1)(\beta + 3)(\beta + 4)}{2\beta(\beta + 7)}$$

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{E(xy) - E(x)E(y)}{\sqrt{E(x^2) - E^2(x)} \sqrt{E(y^2) - E^2(y)}}$$

$$\beta : \text{constant}$$

$$f(x) = \frac{1}{\alpha} \exp\left(\frac{x-\mu}{\alpha}\right) \exp\left(-\exp\left(\frac{x-\mu}{\alpha}\right)\right),$$

$$g(y) = \frac{1}{\alpha} \exp\left(\frac{y-\mu}{\alpha}\right) \exp\left(-\exp\left(\frac{y-\mu}{\alpha}\right)\right)$$

$$F(x) = \exp\left(-\exp\left(\frac{x-\mu}{\alpha}\right)\right),$$

$$G(y) = \exp\left(-\exp\left(\frac{y-\mu}{\alpha}\right)\right),$$

$\mu$  : location parameter ( $=\bar{x} - \alpha\gamma$ )

$\alpha$  : scale parameter ( $=\sqrt{\frac{\sigma_x \times 6}{\pi^2}}$ )

$\gamma$  : Euler-Mascheroni constant = 0.57721...

iii) regression characteristic:

$$h(x, y) = f(x)g(y)[1 + \theta \times c(F(x), G(y))]$$

where  $\theta = \rho \frac{(\beta+1)(\beta+3)(\beta+4)}{2\beta(\beta+7)}$

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{E(xy) - E(x)E(y)}{\sqrt{E(x^2) - E^2(x)}\sqrt{E(y^2) - E^2(y)}}$$

$\beta$  : constant

$$c(u, v) = c_1(u, v) + c_2(u, v) - 2k(1)$$

$$c_1(u, v) = k(u - v) \quad \text{if } v \leq u$$

$$= k(v - u) \quad \text{if } v \geq u$$

$$c_2(u, v) = k(u + v) \quad \text{if } u \leq 1 - v$$

$$= k(2 - u - v) \quad \text{if } u \geq 1 - v$$

$$k(w) = 91 - w)^\beta, 0 \leq w \leq 1, \beta > 0$$

$$f(x) = \frac{1}{\alpha} \exp\left(\frac{x-\mu}{\alpha}\right) \exp\left(-\exp\left(\frac{x-\mu}{\alpha}\right)\right),$$

$$g(y) = \frac{1}{\alpha} \exp\left(\frac{y-\mu}{\alpha}\right) \exp\left(-\exp\left(\frac{y-\mu}{\alpha}\right)\right)$$

$$F(x) = \exp\left(-\exp\left(\frac{x-\mu}{\alpha}\right)\right),$$

$$G(y) = \exp\left(-\exp\left(\frac{y-\mu}{\alpha}\right)\right),$$

$\mu$  : location parameter ( $=\bar{x} - \alpha\gamma$ )

$\alpha$  : scale parameter ( $=\sqrt{\frac{\sigma_x \times 6}{\pi^2}}$ )

$\gamma$  : Euler-Mascheroni constant = 0.57721...

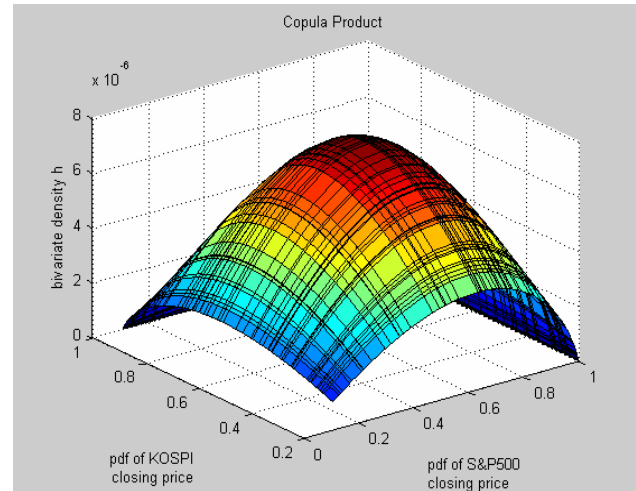
In our study, we used industrial data sets S&P500 (from Jan. 1999 to Dec. 2005, basis: monthly)[9] and KOSPI (Korea Composite Stock Price Index, from Jan. 1999 to Dec. 2005, basis: monthly) with Extreme value distribution as marginals. We use monthly closing price of each data set as the random variables, respectively. In addition, some

of the common measurements of dependence are shown in Table 1. The models are evaluated for several values of  $\beta$ .

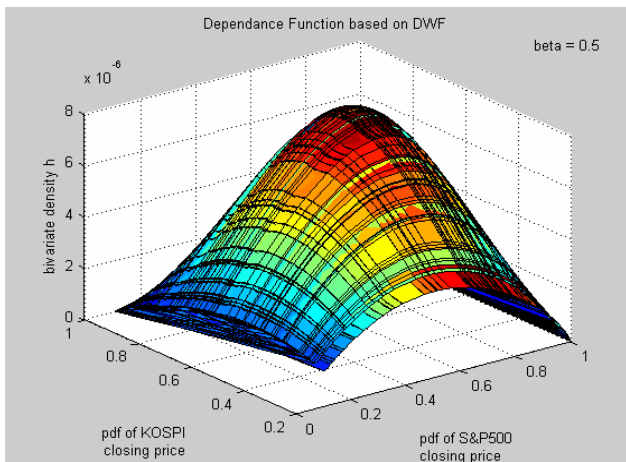
**Table 1** Measurement of dependence between two data sets

Peason's corr	Spearman's rho	Kendall's tau
-0.4523	-0.5371	-0.2070

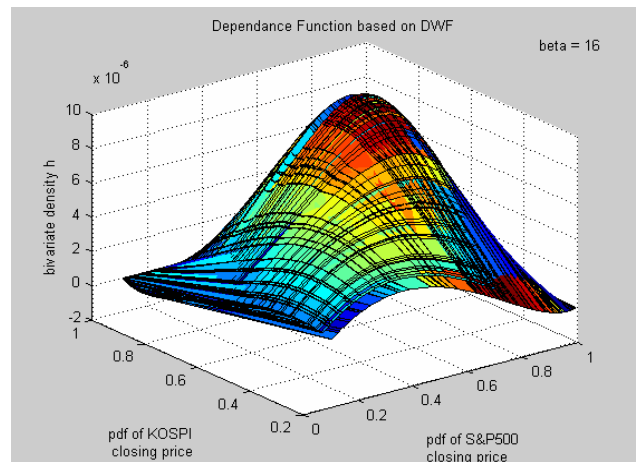
All the figures shown below are joint probability density functions for the same random variables with different dependence functions. All graphs are generated using MATLAB. Figure 1 shows the dependence function with copula, and Figure 2-1 through 2-6 show the dependence function with density weighting function (d.w.f.), and Figure 3.1 through 3-6 show dependence function with covariance characteristic as functionals, and it is helpful to run simulation with the  $\beta$  value 0.5, 1, 2, 16, 128, 256, respectively, in extreme value distribution. It tends to get more convex surface as larger  $\beta$  value it gets. But, in its shape, they differ among three different approaches to construct dependence function. This simulation result reveals the hypothesis that the type of dependence construction has no effect for given  $\beta$  value, is rejected.



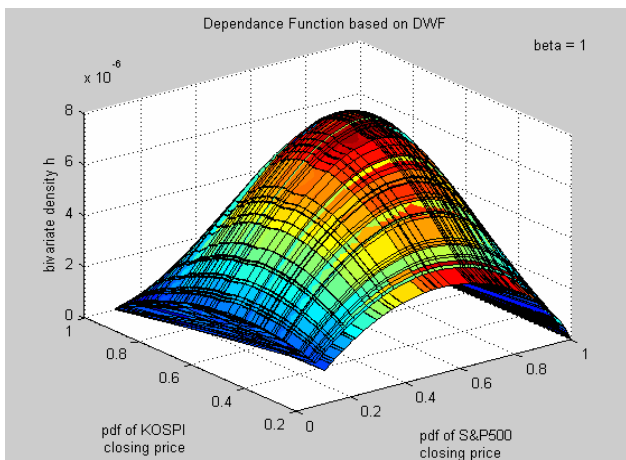
**Figure 1** JPDF - copula with S&P 500 & KOSPI data.



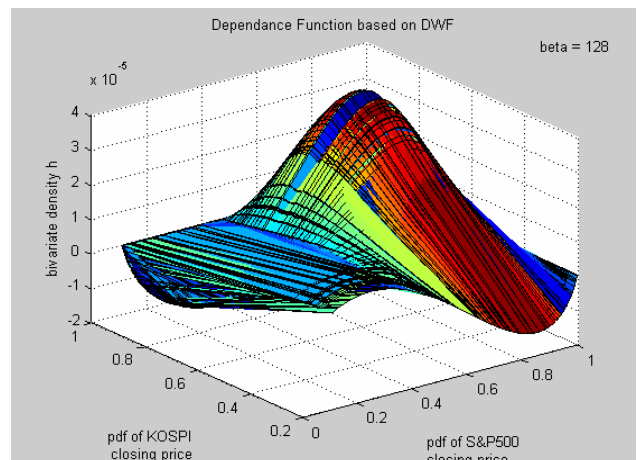
**Figure 2-1** JPDF - density weighting function ( $\beta=0.5$ ) with S&P 500 & KOSPI data.



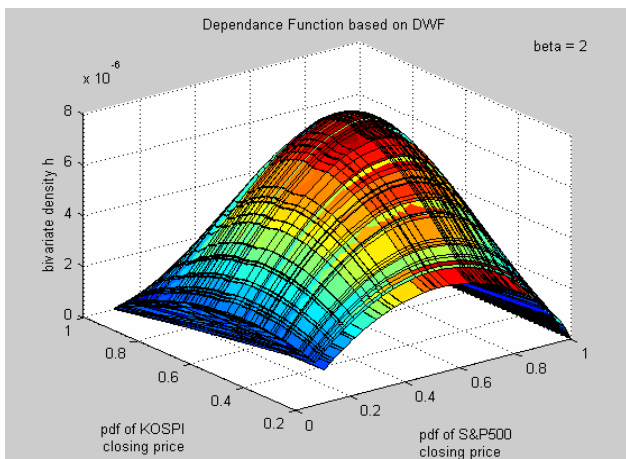
**Figure 2-4** JPDF - density weighting function ( $\beta=16$ ) with S&P 500 & KOSPI data.



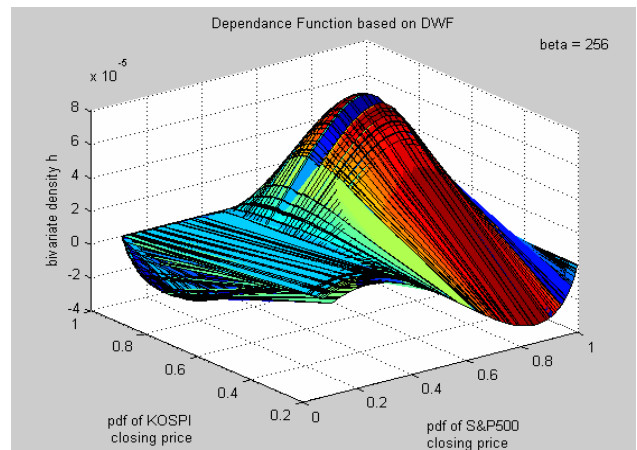
**Figure 2-2** JPDF - density weighting function ( $\beta=1$ ) with S&P 500 & KOSPI data.



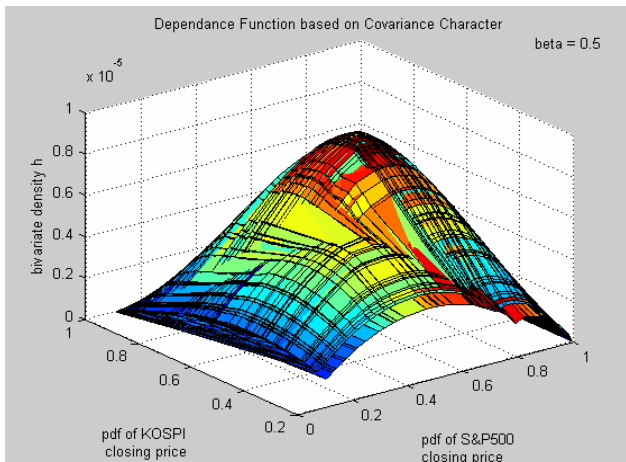
**Figure 2-5** JPDF - density weighting function ( $\beta=128$ ) with S&P 500 & KOSPI data.



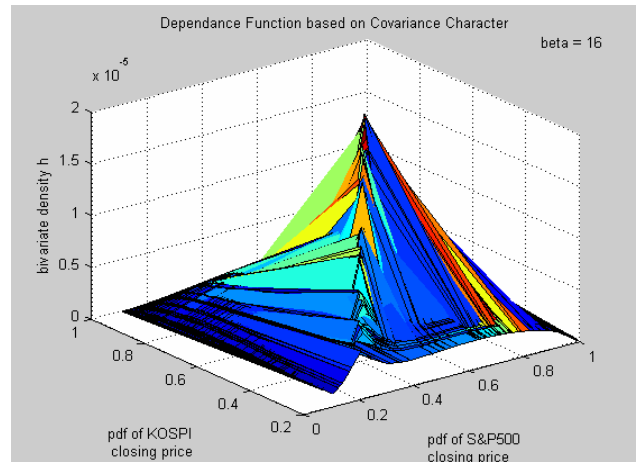
**Figure 2-3** JPDF - density weighting function ( $\beta=2$ ) with S&P 500 & KOSPI data.



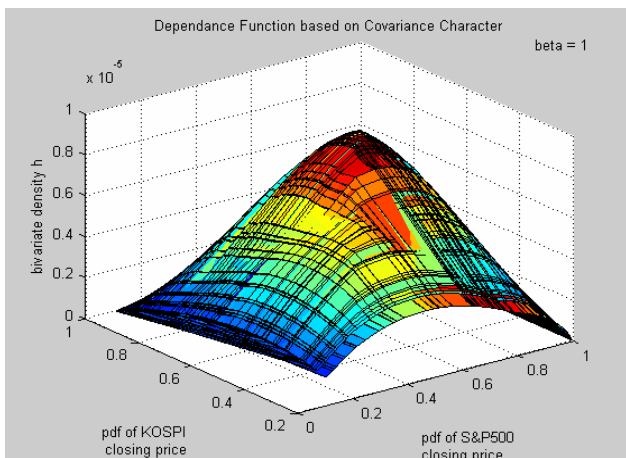
**Figure 2-6** JPDF - density weighting function ( $\beta=256$ ) with S&P 500 & KOSPI data.



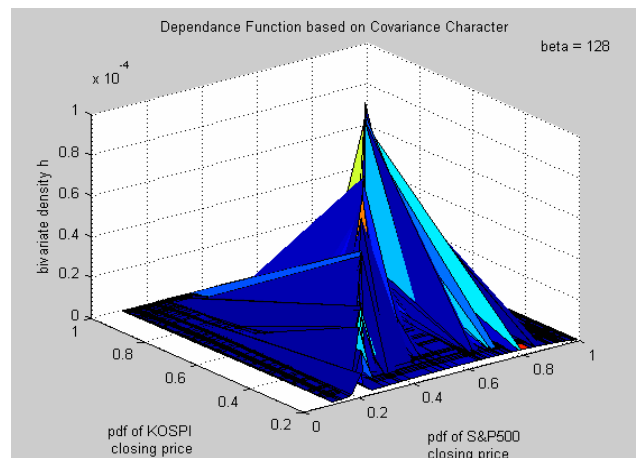
**Figure 3-1** JPDP - covariance characteristic ( $\beta=0.5$ ) with S&P 500 & KOSPI data.



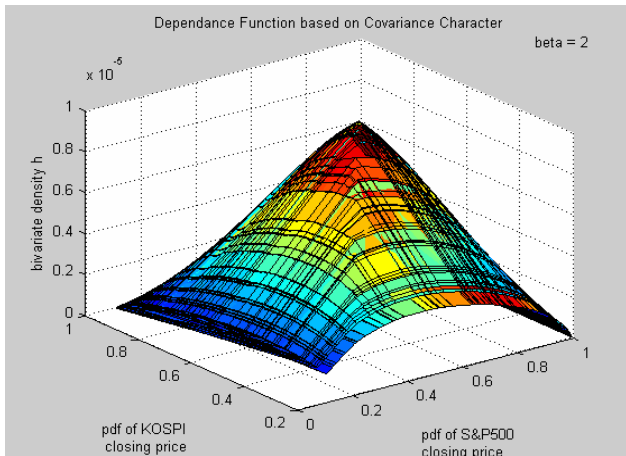
**Figure 3-4** JPDP - covariance characteristic ( $\beta=16$ ) with S&P 500 & KOSPI data.



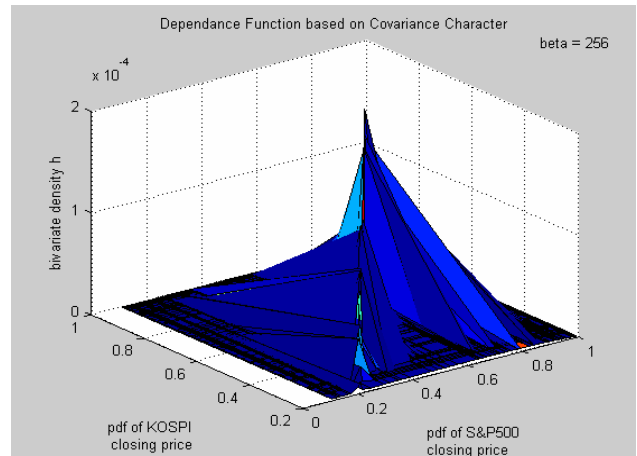
**Figure 3-2** JPDP - covariance characteristic ( $\beta=1$ ) with S&P 500 & KOSPI data.



**Figure 3-5** JPDP - covariance characteristic ( $\beta=128$ ) with S&P 500 & KOSPI data.



**Figure 3-3** JPDP - covariance characteristic ( $\beta=2$ ) with S&P 500 & KOSPI data.



**Figure 3-6** JPDP - covariance characteristic ( $\beta=256$ ) with S&P 500 & KOSPI data.

## 5. Summary and Concluding Remarks

In this paper we studied three classes of dependence functions constructed by three different ways such as i) copulas, ii) density weighting function and iii)

dependence structure function. Copulas link joint distributions to their margins: Let  $K$  be a bivariate distribution function with margins  $F$  and  $G$  in (1). Based on F.G.M. distribution, the density weighting function (d.w.f.) is identified and characterized. And, the limits of dependence was explained by d.w.f. In case of dependence structure, the joint density  $h$  of continuous variable  $(X, Y)$  with specified marginals is defined by (6), is constructed from specified marginals and a dependence structure modeled in terms of functional. In addition to the marginal densities and distributions of  $X$  and  $Y$ , the model contains a covariance scalar  $\theta$ . This scalar is expressed in terms of two parameters such as the shape parameter  $\beta$  and the association parameter  $\rho$ , which are directly estimable.

This study is conducted under the hypothesis that the type of distribution has no effect for given  $\beta$  value on this model. The results seems to reveal the hypothesis that the type of dependence function has no effect for given  $\beta$  value, is rejected. Furthermore, higher values of  $\beta$  seem to distinguish the models dramatically. Thus, estimation of  $\beta$  is a significant problem to be considered in future work. Also, the structure of the dependence function influences the shape of jpdf.

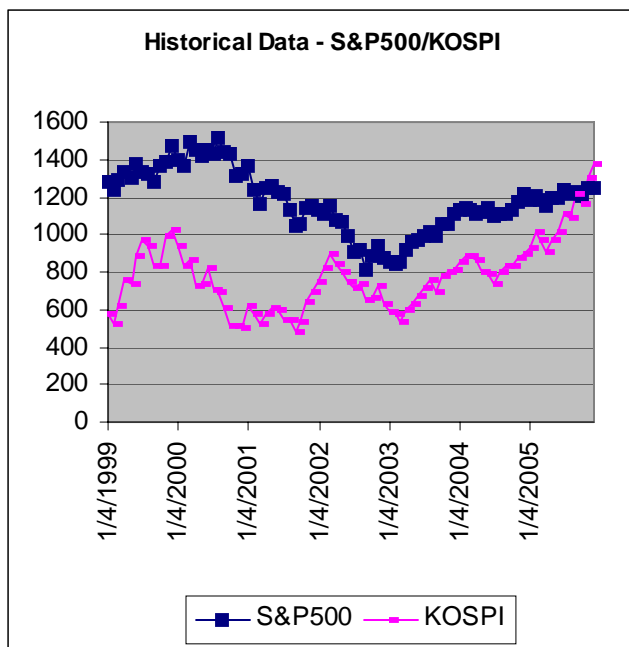


Figure 4. Historical Data of S&P500 and KOSPI

## References:

[1] D. Long, R. Krzysztofowicz, A family of bivariate densities constructed from marginals, 739-746, Journal of the American Statistical Association, Vol. 90, No. 430, 1995.

[2] E. J. Gumbel, Bivariate Logistic Distributions, *Journal of the American Statistical Association*, vol. 56, no. 294, pp. 335-349, 1961.

[3] S. Kotz, J. P. Seeger, A new approach to dependence in multivariate distributions. In: *Advances in probability distributions with given marginals*, 113-127, Mathematics and its Applications 67, Kluwer Acad. Publ., Dordrecht, 1991.

[4] M. E. Johnson, A. Tenenbein, A Bivariate Distribution Family with Specified Marginals, *Journal of the American Statistical Association*, vol. 76, no. 373, pp. 198-201, 1981.

[5] R.L. Plackett, A Class of Bivariate Distributions, *Journal of the American Statistical Association*, vol. 60, no. 310, pp. 516-522, 1965.

[6] M. Shaked, A Family of Concepts of Dependence for Bivariate Distributions, *Journal of the American Statistical Association*, vol. 72, no. 359, pp. 642-650, 1977.

[7] W. Whitt, Bivariate Distributions with Given Marginals, *The Annals of Statistics*, vol. 4, no. 6, pp. 1280-1289, 1976.

[8] A. W. Marshall, I. Olkin, Families of Multivariate Distributions, *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 834-841, 1988.

[9] <http://finance.yahoo.com/q/hp?s=%5EGSPC&a=00&b=1&c=1999&d=11&e=31&f=2005&g=m>, creation date: Not Available, date accessed: February 1, 2006.

[10] <http://finance.yahoo.com/q/hp?s=%5EKS11&a=00&b=1&c=1999&d=11&e=31&f=2005&g=m>, creation date: Not Available, date accessed: February 2, 2006.

[11] M.J. Frank, Convolutions for dependent random variables, *Advances in Probability Distributions with Given Marginals*, Kluwer, 1991, 75-93.

[12] G. Wu, R. Gonzalez, Curvature of the Probability Weighting Function, *Management Science*, vol. 42, no. 12, pp. 1676-1690, 1996.

[13] R. T. Clemen, T. Reilly, Correlations and Copulas for Decision and Risk Analysis, *Management Science*, vol. 45, no. 2, pp. 208-224, 1999.

[14] R. B. Nelsen, An Introduction to Copulas, Springer, New York, 1998.