

Multivariate Copula Modeling with Application in Software Project Management and Information Systems

Syed Muhammad Aqil Burney¹
Dept. of Actuarial Science Risk
Management
Institute of Business Management,
Karachi, Pakistan

Osama Ajaz²
Dept. of Sciences
FAST-National University,
Karachi, Pakistan

Shamaila Burney³
Dept. of Mgmt. Science
Hamdard University,
Karachi, Pakistan

Abstract—This paper discusses application of copulas in software project management and information systems. Successful software projects depend on accurate estimation of software development schedule. In this research, three major risk factors and their impact on software development schedule are considered. Software development schedule is calculated by COCOMO-II model. Two models are simulated 100000 times, model-I considered dependence among risk factors by T-copula and model-II considered risk factors independent. The comparison of the two risk models revealed that model-II always underestimate the software development schedule while model-I evaluated the software schedule risk accurately. Therefore it is necessary for software development experts to consider dependence among various risk factors. R-package copula is employed to implement the algorithm for multivariate T-copula. Multiplier goodness-of-fit test shows that T-copula is good choice for characterization of dependence among three risk factors.

KeyWords—T-Copula; COCOMO – II; software development schedule; risk analysis

I. INTRODUCTION

Software project management and information system consists of multiple activities under the umbrella of software engineering [1]. It involves planning, scheduling, budgeting and managing entire software development process. Each activity in a project management consumes time i.e. software project schedule. Further, entire software development life cycle depends on accurate estimation of software project schedule. Beside estimation of project schedule, it is the responsibility of a project manager to identify risk factors that results in project delays or failure. Therefore it is necessary to assess software project schedule accurately. Plethora of literatures is available on risk factors that results in schedule overrun or project failure. For further detail, see [2] [3] [4].

Generally software project manager just estimate project schedule and completely or partially ignore the impact of risks on estimated project schedule, even if project manager consider some risk factors; it is assumed to be independent. Positive and negative dependence may affect software risk severely. For example, it is possible that project manager inaccurately estimate the software schedule to 20 months and project ends in 26 months. This is also possible that one risk

may cause another risk to happen i.e. increase in customer requirements during software development duration strongly cause schedule overrun and in turn may lead to loss of key employee. Further, project delays have negative impact on customer satisfaction which results in bad reputation of an organization. This research paper analyzes software project schedule and consider dependence among risk factors by means of copula.

The reason to consider copulas in this research is that most of literatures available on theory of copulas discuss application in Econometrics, Finance and in Insurance [5]. Further, very few articles available that discusses copulas in software project management or in information systems. See [6] [7] [8] as examples and the references therein.

The remaining paper is organize as follows: Section 2 gives an overview to the theory of copulas. Section 3 discusses software schedule and associated risks. Section 4 discusses application of copula model for software houses based in Karachi and finally this research end up with Section 5 that discusses overall conclusion.

II. THEORY OF COPULAS

The discovery of copula is associated to the seminal research work of Sklar [9]. He derived the word “copula” for multivariate joint distribution that links to its marginal distribution. In this section, we present brief overview to the basic theory of copulas for higher dimensions. For further details and proofs about copulas and its historical development, please see [10] [11] [12] [13] and the references therein.

A. Multidimensional Copulas

A d-dimensional copula or d-copula is a function “C” from I^d to I where $I \in [0, 1]$, if and only if it satisfies the following conditions:

- (i) For every m in I^d , $C(m_1, m_2, \dots, m_d) = 0$ if at least one of m is equal to zero.
- (ii) If all m 's are set to one then for every $k \in \{1, 2, \dots, d\}$, $C(m) = m_k$.

(iii) C is d-increasing, for all $(a_1, a_2, \dots, a_d), (b_1, b_2, \dots, b_d)$ in I^d such that $a_i \leq b_i$,

$$V_c([\mathbf{a}, \mathbf{b}]) \geq 0$$

$$\sum_{i_1=1}^2 \sum_{i_2=1}^2 \dots \sum_{i_d=1}^2 (-1)^{i_1+i_2+\dots+i_d} C(m_{1_{i_1}}, m_{2_{i_2}}, \dots, m_{d_{i_d}}) \geq 0$$

All above properties of d-copula are multivariate version of bivariate copula. Property (i) defines grounded condition, property (ii) defines copula margin if $(d - 1)$ variables are known, and property (iii) defines C volume of rectangle $[a \times b]$ for any d-dimension. Further, it is easy to show that, any convex linear combination of copulas is a copula i.e. $\sum_{i=1}^d \alpha_i C_i$ for all $\alpha_i > 0$ and $\sum \alpha_i = 1$ [13]. Nelsen and Joe discusses some other important properties of copulas that we state as following theorems 1, 2, and 3 below [11] [13].

Theorem 1

For every d-copula, the Frechet-Hoeffding bound inequality is given by:

$$C_L^d \leq C(m_1, m_2, \dots, m_d) \leq C_U^d \tag{1}$$

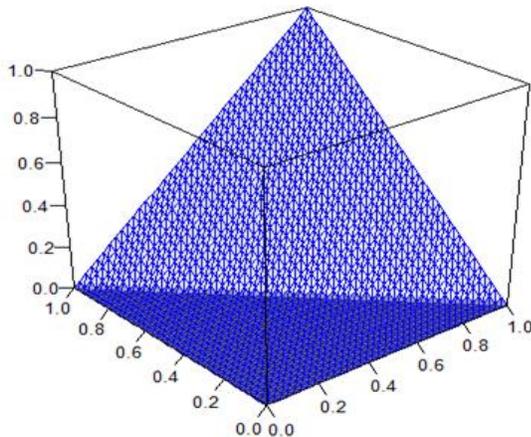


Fig. 1. Figure 1: Graphs of Frechet – Hoeffding Lower Bound Copula.

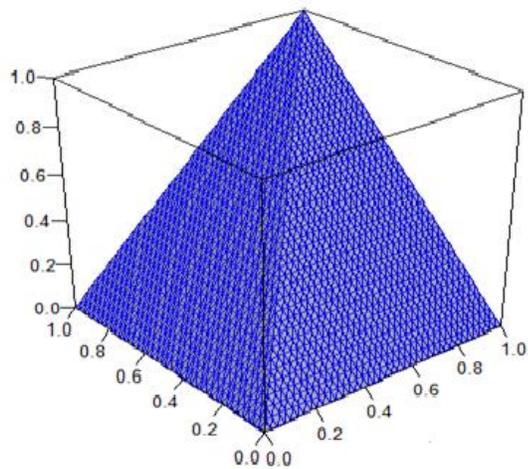


Fig. 2. Graph of Frechet-Hoeffding Upper Bound Copula.

Where C_L^d and C_U^d represents Frechet – Hoeffding lower and upper bounds, these bounds define as; $\max(\sum_{i=1}^d m_i - d + 1, 0)$ and $\min(m_1, m_2, \dots, m_d)$ respectively. Notice that, for $d > 2$, the lower bound C_L^d is not a copula. For further details see theorem 3.6 in [13]. For $d = 2$, the graphs of C_L^d and C_U^d are given in figure 1 and figure 2 below respectively.

Theorem 2

For all m in I , the copula is independent if and only if,

$$C^d(m) = m_1 \cdot m_2 \cdot \dots \cdot m_d \tag{2}$$

For all $d \geq 2$, the continuous random variables X_1, X_2, \dots, X_d are independent if and only if their d-copula is $C^d(m)$. For $d = 2$, the graph of independent copula is given as figure 3:

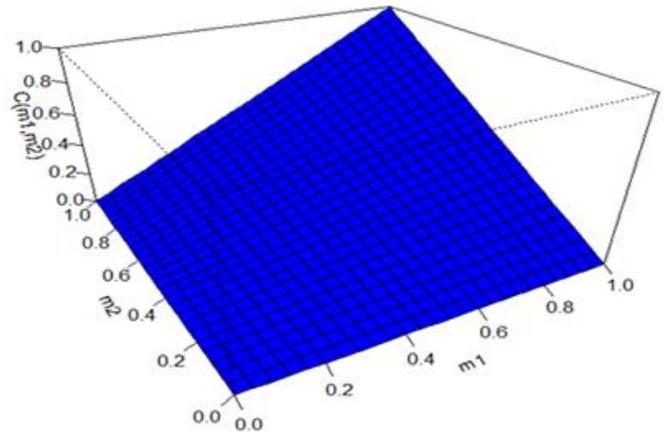


Fig. 3. Graph of Independent Copula.

Theorem 3

For all m, n in I^d , the d-copula satisfies Lipchitz condition,

$$|C(m) - C(n)| \leq \sum_{i=1}^d |m_i - n_i| \tag{3}$$

Thus d-copula is a continuous function from I^d to I .

Example 1

As an illustration, consider the following extended FGM copula for $d = 3$,

$$C^3(m_1, m_2, m_3) = m_1 m_2 m_3 [1 + \alpha_{12} (1 - m_1) (1 - m_2) + \alpha_{13} (1 - m_1) (1 - m_3) + \alpha_{23} (1 - m_2) (1 - m_3) + \alpha_{123} (1 - m_1) (1 - m_2) (1 - m_3)] \tag{4}$$

It is easy to show that, the above three dimensional FGM copula satisfies the basic requirements of d-copula.

Hence (4) is a copula, theorem 2 hold for $d = 3$ if dependence parameters α_{ij} equals to zero. For further details about d-dimensional extended FGM copula, see Drouet and Kotz [14].

B. Sklar's Theorem

Copulas are important because of Sklar's theorem [9]. According to this theorem, any multivariate joint distribution can be represented by its marginal distribution. Consider X_1, X_2, \dots, X_d be continuous random variables with their joint

distribution function J and univariate marginal distribution $F_i(X_i) = P(X_i \leq x_i)$, $i \in \{1, 2, \dots, d\}$. Then, there exists a d-copula C^d such that,

$$J(x_1, x_2, \dots, x_d) = C^d(F_1(x_1), F_2(x_2), \dots, F_d(x_d)) \quad (5)$$

For all $x \in R^d$, the function C^d is uniquely determined, if $F_i(X_i)$ are continuous. Otherwise C^d is uniquely determined on (range of $F_1(x_1)$) \times (range of $F_2(x_2)$) \times ... \times (range of $F_d(x_d)$).

Let J, C^d and $F_i(X_i)$ be as in above (5), let $F_1^{(-1)}, F_2^{(-1)}, \dots, F_d^{(-1)}$ are quasi-inverses of $F_i(X_i)$ where $i \in \{1, 2, \dots, d\}$, respectively. Then,

$$C(m_1, m_2, \dots, m_d) = J(F_1^{(-1)}(m_1), F_2^{(-1)}(m_2), \dots, F_d^{(-1)}(m_d)) \quad (6)$$

for all $m \in I^d$.

For the sake of simplicity, let F_1, F_2, \dots, F_d be continuous and differentiable distribution functions. Then the corresponding density function to (6) is

$$j(x) = c(F_1(x_1), F_2(x_2), \dots, F_d(x_d)) \prod_{i=1}^d j_d(x_d) \quad (7)$$

Where $j_d(x_d)$ is the marginal density of $F_i(X_i)$, $i \in \{1, 2, \dots, d\}$, and

$$c(F_1(x_1), F_2(x_2), \dots, F_d(x_d)) = \frac{\partial C(F_1(x_1), F_2(x_2), \dots, F_d(x_d))}{\partial F_1(x_1), \partial F_2(x_2), \dots, \partial F_d(x_d)} \quad (8)$$

According to (5), the copula function C^d is uniquely determined and can also be represented as,

$$C(m_1, m_2, \dots, m_d) = \int_0^{m_1} \int_0^{m_2} \dots \int_0^{m_d} c(n_1, n_2, \dots, n_d) \partial n_1 \partial n_2 \dots \partial n_d \quad (9)$$

Where $m_i = F_i(x_i)$ and corresponding density is c, then we have joint density function (7).

III. SOFTWARE DEVELOPMENT SCHEDULE & ASSOCIATED RISK FACTORS

A. Project Scheduling Risk

Risk is related to future happenings and it has two characteristics one is *Uncertainty* and other is *Losses* [1] [15]. If the risk is certain to occur then it has positive or negative impact on projects' objectives. In all, risk has two dimensions: probability of occurrence of event and its impact. If risk associated to software project schedule is certain to occur then estimated schedule exceeds deadline which results in financial losses and bad reputation of an organization. This research explores the relationship between risk and its impact on software project schedule. Scheduling risk is the probability of one or more events, if they occur has positive or negative impact on software development duration [16].

There are many uncertain risk factors that affect project schedule severely. However, for this research, three major risk factors are considered that every project manager must face [4]. These three major risks are defined below as:

a) Imprecise measurement of software effort:

Software effort is defined as number of working hours spend on the project. In this research, software effort is expressed as:

$$\text{Software Effort} = \left| \frac{\text{Actual Effort} - \text{Estimated Effort}}{\text{Estimated Effort}} \right| \times 10 \quad (10)$$

Incorrect measurement of software effort results in project delays. Incorrect estimation of effort is consequences of lack of experienced manager or inadequate knowledge about estimation tools.

b) Loss of Employees during project:

It is the usual turnover rate of employees during projects. It express simply in percentages. It includes resignations, death, medical leave, retirement or transfer of employees during project.

c) Change of Customer requirements:

Change of customer requirements includes Increment or decrement of customer requirements during software development duration and expressed as percentages.

B. Risk Assessment Model

Many risk assessment techniques exists. I will assess software development risk by risk assessment model. For this research, the following schedule risk model is considered:

$$\text{Schedule Risk} = (\text{Project duration}) \times R_1 \times R_2 \times R_3 \quad (11)$$

Where R_1 indicates impact of imprecise estimation of software effort, R_2 indicates impact of risk of Loss of Employees during project and R_3 indicates impact of Change of customer requirements during software development period on project schedule. As can be seen, the model is multiplicative in nature. For further details about this model, see [16] [17]. The probability distributions for these three risk factors are derived by using expert data from various software houses based in Karachi. Project cost and schedule is estimated by COCOMO-II model [18].

C. COCOMO – II

Boehm et al. [18] proposed a COCOMO-II model and it requires three stages to estimate software project cost, effort and schedule. For early stages of software project development, application composite model is used. When information packages, software architecture and infrastructure is finalized, early design model is used. Finally, post architecture model is used during software development duration.

We have calculated new software project schedule for 400000 SLOC using COCOMO – II. By setting variables to desire level, the COCOMO-II model estimated new software development schedule from 17 months to 40 months. Therefore, the new software development project can be completed at least in 17 months and at most 40 months while the actual schedule for this project was 30 months. This minimum and maximum duration represents best and worst case. The levels of COCOMO-II model for new software development schedule is set by consulting software development experts.

D. T – Copula Method to Model Dependence

Let n-dimensional random vector $Y = (Y_1, Y_2, \dots, Y_d)^T$ has d-variate t-distribution with ν degrees of freedom, μ mean

vector and correlation matrix R and its joint probability density function (PDF) is defined by:

$$f(y) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} \dots \int_{-\infty}^{y_d} \frac{\Gamma(\frac{v+d}{2})}{(\pi v)^{\frac{d}{2}} \Gamma(\frac{v}{2}) |R|^{\frac{1}{2}}} [1 + \frac{1}{v} (X - \mu)^T R^{-1} (X - \mu)]^{-\frac{v+d}{2}} \quad (12)$$

The multivariate t-copula is then defined as:

$$C(m_1, m_2, \dots, m_d) = T_{R,v} (T_v^{-1}(m_1), T_v^{-1}(m_2), \dots, T_v^{-1}(m_d)) \quad (13)$$

$$= \int_{-\infty}^{T_v^{-1}(m_1)} \int_{-\infty}^{T_v^{-1}(m_2)} \dots \int_{-\infty}^{T_v^{-1}(m_d)} \frac{\Gamma(\frac{v+d}{2})}{\Gamma(\frac{d}{2})(v\pi)^{\frac{d}{2}} |R|^{\frac{1}{2}}} (1 + \frac{1}{v} X^T R^{-1} X)^{-\frac{v+d}{2}} dx_1 \dots dx_d \quad (14)$$

Where $T_v^{-1}(u_i)$ is the inverse of the univariate distribution functions of t-distribution with v degrees of freedom. The corresponding canonical representation of multivariate T-copula density is given by:

$$c(m_1, m_2, \dots, m_d) = |R|^{-\frac{1}{2}} \frac{\Gamma(\frac{v+d}{2})}{\Gamma(\frac{v}{2})} \left(\frac{\Gamma(\frac{v}{2})}{\Gamma(\frac{v+1}{2})} \right)^d \frac{(1 + \frac{1}{v} \delta^T R^{-1} \delta)^{-\frac{v+d}{2}}}{\prod_{i=1}^d (1 + \frac{\delta_i^2}{v})^{-\frac{v+1}{2}}} \quad (15)$$

Where $\delta_i = T_v^{-1}(m_i)$.

Lot of studies has shown that T-copula is superior to any other copula such as Gaussian copula [19] [20]. It is symmetric but has tail dependence. We have modeled dependence among three risk factors by T-copula. Inference function for margins (IFM) method is used to estimate the parameters of T-copula. In the figure 4 below, contour plot for multivariate T-copula in two dimensions is given.

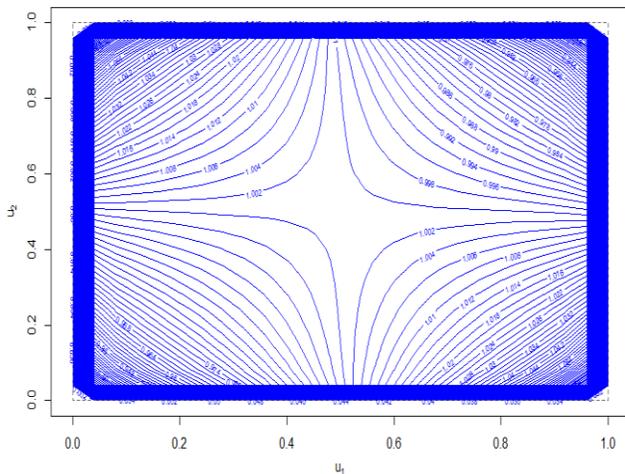


Fig. 4. Contour Plot for Multivariate T-Copula Density.

The contour plot shows moderate dependencies between variables. This is the beauty of copula that even if the correlation is zero nevertheless marginal distributions are related to joint distribution [9].

IV. APPLICATION OF COPULAS IN SOFTWARE PROJECT MANAGEMENT

In this research, two risk models are considered for new software project schedule. Model-I considered dependence among risk factors and Model-II assumes that risks are independent. For model -I, we have used multivariate T-Copula to model dependence among risk factors. Monte Carlo method is employed to simulate the two models. Simulations for both model executed 100000 times.

A. Distributions of Three Risk Factors

The results of copulas are useful if and only if we have fitted best distributions to marginal distributions. These marginal distributions are for Software Effort, Loss of employees and change of customer requirements during project. The marginal distributions for the three risks are derived using data of several software houses based in Karachi. Several statistical hypothesis testing tools and graphs are employed to assess goodness of fit for three marginal distributions. The results for goodness-of-fit tests for sample size 200 are provided in the table 1, 2 and 3.

TABLE I. PROBABILITY DISTRIBUTION FOR SOFTWARE EFFORT (R_1), H_0 : R_1 CONFORMS TO NORMAL DISTRIBUTION

Test	Test Statistic	p-value
Chi-Square test	0.52	0.9144
K-S test	0.053495	0.6162
Anderson & Darling Test	0.50791	0.7386
Cramer-Von Mises Test	0.06606	0.7765

All goodness-of-fit tests listed in above table 1 provide high p-values at 1% level of significance. Therefore, the risk distribution for imprecise measurement of software effort conforms to normal with parameters (40.49525%, 20.79036%)

TABLE II. PROBABILITY DISTRIBUTION FOR LOSS OF EMPLOYEES (R_2), H_0 : R_2 CONFORMS TO WEIBULL DISTRIBUTION

Test	Test Statistic	p-value
Chi-Square test	11.41	0.1795323
K-S test	0.08365265	0.1217015
Anderson & Darling Test	1.6798	0.1389
Cramer-Von Mises Test	0.2426781	0.1980663

All goodness-of-fit tests listed in the above table provide high p-values at 1% level of significance. Therefore, risk distribution for loss of employees during project conforms to Weibull with parameters (1.478794%, 27.747935%).

TABLE III. PROBABILITY DISTRIBUTION FOR CHANGE OF REQUIREMENTS (R₃), H₀: R₃ CONFORMS TO WEIBULL DISTRIBUTION

Test	Test Statistics	p-value
Chi-Square	2.8	0.5918327
K-S Test	0.06114199	0.4432991
Anderson & Darling Test	0.65045	0.6014
Cramer-Von Mises Test	0.08189777	0.6810586

All goodness-of-fit tests listed in the above table3 provide high p-values at 1% level of significance. Therefore, risk distribution for change of customer requirements conforms to Weibull distribution with parameters (1.56949%, 27.27296%).

B. Simulation Results of the Two Models

As described above, two risks models are considered. Model – I considered dependence among three risk factors by T-copula and model – II considered risk factors are independent. Both the models are simulated 100000 times and their histograms are presented in the fig. 5 and fig. 6 respectively.

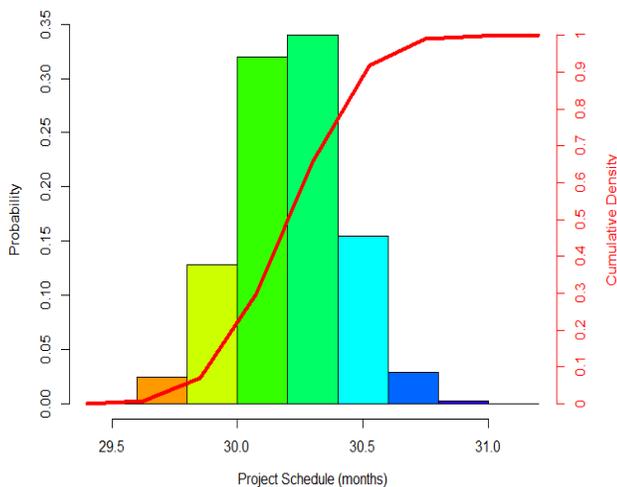


Fig. 5. Simulation Histogram for Model – I.

In the figure 5 above, the left y-axis represents probability density function and right y-axis represents cumulative density function for the simulated model-I. According to the above simulated histogram, the new software project schedule can vary from 29 months to 31 months. There is 100% chance that, the new software project can be completed in almost 30 months while there is approximately 38% chance that the new project can be completed in almost 29 months.

Multiplier goodness of fit test [21] is applied to the chosen T-copula. The goodness of fit provide test statistics $S_n^k = 0.036957$ with p-value 0.1004. Since p-value is high enough at 1% level of significance. Therefore the chosen T-copula model is appropriate for characterization dependence among three risk factors. Simulation histogram for model – II is shown below in the figure 6.

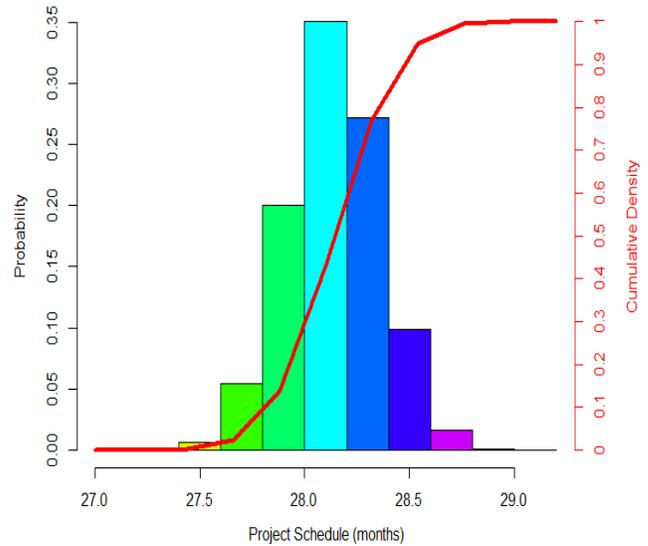


Fig. 6. Simulation Histogram for Model – II.

The simulated model-II shows that the new project schedule can vary from 27 months to 29 months. There is 100% chance that the new project can be completed in 28 months and less than 5% chance that the new project can be completed in 27 months.

The comparison of the two simulated risk model revealed that the new software schedule is underestimated without considering dependence among three risk factors. Further comparison of the simulated result revealed that, there is almost 100% chance for new project to be completed in 30 months in risk model-I and 100% chance to be completed in 28 months in risk model-II. There is almost 40% chance for the new software project to be completed in 29 months in risk model-I and almost 5% in risk model-II. The original duration for new project is 30 months. Its mean that the model-I which considered dependence among risk factors by T-copula evaluated software project duration accurately.

V. CONCLUSIONS

Project delays or failures are practicing routine in many software houses across Pakistan. In this research, we have considered three major risk factors that can negatively impact the estimated project schedule. The risk factors are evaluated by two models. Model-I assumed dependence among risk factors by multivariate T-copula and model-II assumed independence among risk factors. Both models implemented for some software houses based in Karachi and the analysis revealed that model – I which considered dependence among risk factors by T-copula, evaluated project schedule accurately. Multiplier goodness-of-fit test showed that, the chosen T-copula is appropriate for characterization of dependence among three risk factors. It is concluded that if a software manager do not consider dependence among risk factors then he may underestimate the software project schedule. Schedule overruns result in high budgeting cost, dissatisfaction of customers and sometimes failure of software project. Therefore copulas are important for characterization

of dependence among several variables and software houses in Pakistan can utilize theory of copulas for better outcomes of project management. Further, theory of copulas can be implemented to other risk factors defined by software project manager.

REFERENCES

- [1] Roger S. Pressman, "Software engineering a practitioner's approach (7th ed)." Doboque, IA McGraw-Hill, 2009.
- [2] Wen-Ming Han, Sun-Jen Huang, "An empirical analysis of risk components and performance on software projects," Journal of Systems and Software, Volume 80, Issue 1, 2007, Pages 42-50.
- [3] T. Christiansen, P. Wuttidittachotti, S. Prakancharoen and S. Arj-ong Vallipakorn "Prediction of risk factors of software development project by using multiple logistic regression" ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 3, Feb. 2015.
- [4] Tharwon Arnuphaptrairong "Top Ten Lists of Software Project Risks: Evidence from the Literature Survey" Proceedings of IMECS, Vol. 1, March 16 – 18, 2011, Hong Kong.
- [5] P. Jaworski et al. (eds.), Copula Theory and Its Applications, Lecture Notes in Statistics 198, DOI 10.1007/978-3-642-12465-5_1, © Springer-Verlag Berlin Heidelberg 2010.
- [6] D. Wu, H. Song, M. Li, C. Cai and J. Li, "Modeling risk factors dependence using Copula method for assessing software schedule risk," *The 2nd International Conference on Software Engineering and Data Mining*, Chengdu, 2010, pp. 571-574.
- [7] A. Burns, G. Bernat, I. Broster, "A probabilistic framework for schedulability analysis" *Proceedings of the Third International Conference on Embedded Software (EMSOFT 2003)*, pp. 1–15 (2003)
- [8] Guillem Bernat , Alan Burns , Martin Newby, Probabilistic timing analysis: An approach using copulas, Journal of Embedded Computing, v.1 n.2, p.179-194, April 2005.
- [9] Syed Muhammad Aqil Burney "Fuzzy Logic & Copulas applications in Financial and Actuarial Science" Talk delivered at Centre for Advanced studies in Mathematics (CASM) at LUMS Lahore, Pakistan, 2013.
- [10] A. Sklar "Fonctions de repartition a n dimensions et leurs marges" Publ. Inst. Statist. Univ. Paris vol. 8, 229-231 1959.
- [11] Roger B. Nelsen, An Introduction to Copulas (Springer Series in Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, 2006.
- [12] Durante, Fabrizio, Giovanni Puccetti, Matthias Scherer, et al. "My introduction to copulas" *Dependence Modeling*, 5.1 (2017): 88-98. Retrieved 20 Nov. 2018, from doi: 10.1515/demo-2017-0006.
- [13] Joe H., "Multivariate Models and Dependence Concepts," Chapman & Hall, London, 1997.
- [14] D. Drouet-Mari and S. Kotz, "Correlation and dependence," Imperial College Press, London, 2001.
- [15] Syed Muhammad Aqil Burney, "Risk Theory and Insurance: A stochastic approach," KU Press, Karachi, (2003).
- [16] B. W. Boehm "Software Risk Management Tutorial," IEEE Computer Society Press, Washington, 1989.
- [17] Paul L. Bannerman, "Risk and risk management in software projects: A reassessment," *The Journal of Systems and Software* 81 (2008) 2118–2133.
- [18] Barry W. Boehm, Clark, Horowitz, Brown, Reifer , Chulani , Ray Madachy , Bert Steece, "Software Cost Estimation with Cocomo II with Cdrom,," Prentice Hall PTR, Upper Saddle River, NJ, 2000.
- [19] S. Demarta and A. J. McNeil "The t copula and related copulas" Technical Report ETH Zurich, 2004.
- [20] Breymann, W., A. Dias and P. Embrechts, "Dependence structures for multivariate high frequency data in finance," *Quantitative Finance* 3, 1–14 (2003).
- [21] Kojadinovic, Ivan & Yan, Jun, "Modeling Multivariate Distributions with Continuous Margins Using the copula R Package," *Journal of Statistical Software*, Foundation for Open Access Statistics, vol. 34(i09), 2010.