Fast and Efficient Feature Selection Method Using Bivariate Copulas

K. Femmam
Applied Mathematics, Mohamed Khider Institution, Biskra, Algeria
Email: karima.femmam@univ-biskra.dz

S. Femmam
UHA University & Polytechnic Engineers School, Sceaux, France
Email: smain.femmam@uha.fr

***Abstract*—Handling datasets nowadays has become a crucial task, since today's world is heavily dependent on data information. However, many data tend to be big and contain redundancy which makes them difficult to deal with. Due to that, data pre-processing became almost necessary before using any data, and one of the main tasks in data pre-processing is dimensionality reduction. In this paper we propose a new approach for dimensionality reduction using feature selection method based on bivariate copulas. This approach is a direct application of copulas to describe and model the inter-correlation between any two dimensions - bivariate analysis. The study will first show how we use the bivariate method to detect redundant dimensions and eliminate them, and then compare the quality of the results against most-known selection methods in term of accuracy, using statistical precision and classification models.**

***Index Terms*—bivariate copulas, data pre-processing, dimensionality reduction, feature selection**

I. Introduction

Every field has its hands wet with big data, aiming for optimization and efficiency, but most of the time useless and redundant observations are added to datasets, which increase the time complexity and decrease models accuracy. But fortunately, we always apply pre-processing to any data, and dimensionality reduction is part of it, hence various methods were introduced to perform this step, including many feature selection methods. Feature selection is a dimensionality reduction technique that filters the data in order to choose which one to select and which one to eliminate without losing important information. Many well-known methods were introduced in this field, however most of them suffer from a high time complexity as a result of the sequential search method. Feature selection techniques were applied in different fields, recently, Authors proposed an effective feature selection method for clinical treatments that improved the classification and reduced the running time [1], while others published a comparative analysis in the Network Intrusion Detection System (NIDS) using the conventional Genetic Algorithm (GA), Genetic Algorithm with Improved Feature Selection (GA-IFS) technique using the Support Vector Machine (SVM) classifier and GA-IFS with Naïve Bayes classifier (NBC) [2], they concluded that GA-IFS with SVM classifier outperformed both methods in term of accuracy. In 2016, a dimensionality reduction technique based on copulas and LU-decomposition was proposed [3], this approach gives good results against well-known methods, however it includes a complex optimization problem and passes by a lot of operations, which leads to a long processing time $(O(n^2))$, to improve that we propose a new filter method with less complex model and time complexity, it is built using an algorithm programmed in R, and uses bivariate copula as a tool to detect redundancy between each two attributes in order to eliminate one of them. Similar work has been submitted for publication but using multivariate copulas instead "unpublished" [4].

Our method will have to outperform other methods of dimensionality reduction by having more accuracy, and better reduction of data.

II. Basic Concept

In this part, we introduce the mathematical background and the necessary tools that constitute our method. We choose to use bivariate copula because it separates the marginal distributions from the dependency structure of a given bivariate distribution, which makes detecting inter-correlation easier.

Let X and Y be random variables with the continuous cumulative functions $F_1 = P(X \le x)$ and $F_2 = P(Y \le y)$ respectively. By using the integral transform of probability for each individual variable, we get the uniformly distributed variables presented in (1).

$$(U,V) = (F_1(X), F_2(Y)) \tag{1}$$

From this transformation, we are able to generate pseudo-random samples from continues random variables as in (2).

$$(X,Y) = (F_1^{-1}(U), F_2^{-1}(V)) \tag{2}$$

Manuscript received November 9, 2021; revised April 20, 2022.

doi: 10.12720/jait.13.3.301-305

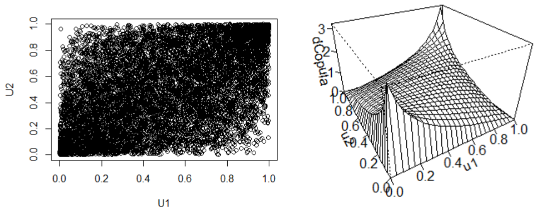
The bivariate copula is a cumulative distribution function with Uniform [0,1] margin. It is used to describe the inter-correlation (dependency structure) between two random variables by combining the bivariate distribution function with their one-dimension marginal distribution function. Following Sklar's theorem [5], any bivariate joint distribution can also be written as univariate marginal distribution functions (a unique Copula C in $[0,1]^2$), and standard uniform marginal distributions (U_1,U_2) which display the dependencies between the variables. This relationship is presented in (3) (the formula of the bivariate theoretical copula C), while its corresponding bivariate empirical copula C_n is defined in (4). In these equations, F represents the joint cumulative distribution of the couple (X,Y) , and n represents the number of observations in the random variables.

$$F(X,Y)=C(F_1(X),F_2(Y)) \tag{3}$$

$$C_n(u,v)=\frac{1}{n}\sum_{i=1}^{n}1(U_i\le u,V_i\le v) \tag{4}$$

There are several families of bivariate copulas, among them, the Elliptical copulas are bivariate distributions and come in different types, the one we use is the Gaussian copula, a symmetrical type of elliptical copulas. Equation (5) represents the theoretical bivariate Gaussian copula, where α defines the correlation parameter (also called copula's parameter) and Φ is the cumulative standard Gaussian distribution function. This copula is visualized in Fig. 1, where Fig. 1.a shows the scatter plot of (5) with the parameter $\alpha=0.5$, while Fig. 1.b presents the corresponding density copula, these plots describe the dependency (inter-correlation) between the two random variables.

$$C(u,v,\alpha)=\frac{1}{\sqrt{1-\alpha^2}}\cdot\exp\left[\frac{\alpha^2(\Phi^{-1}(u)^2+\Phi^{-1}(v)^2)-2\alpha\Phi^{-1}(u)\Phi^{-1}(v)}{2(1-\alpha)^2}\right] \tag{5}$$



(a) Theoretical copula. (b) Density copula.

Figure 1. The bivariate Gaussian copula ($\alpha=0.5$).

To describe the dependency between the variables, we introduce the relationship between Kendall's tau τ and the copula's parameter as a tool to detect the inter correlation. For this study, we use the relationship between Kendall's tau τ_{ij} and the Gaussian bivariate copula's parameter α_{ij} defined in (6), where $i,j\in\{1,...,m\}$ are the indices of the variables.

$$\tau_{ij}=\frac{2}{\pi}\arcsin\alpha_{ij} \tag{6}$$

III. Proposed Approach

This section focuses on introducing the Proposed Approach (PA) and the algorithm behind it.

Let X be the input matrix of $n\times m$ dimensions containing redundant variables. In order to transform the matrix X 's attributes into random variables between $[0,1]$, we use the pseudo observation transformation defined in (7) by forcing the variates to fall inside the open unit hypercube.

$$u_{ij}=\frac{r_{ij}}{n+1} \tag{7}$$

where $i\in\{1,...,n\}$, $j\in\{1,...,m\}$ and r_{ij} denotes the rank of X_{ij} among all X_{kj} where $k\in\{1,...,n\}$. Next, in order to visualize the dependency between the pairs attributes, we use (4) to calculate and plot the bivariate empirical copulas for each pair of attributes, after that we follow these steps:

1) Determine the bivariate theoretical copulas for each pairs using the data based on the scatter plot of the bivariate empirical copula and the marginal distributions of the datasets.
2) Pick the first pair of attributes.
3) Calculate Kendall's tau τ .
4) Deduce the bivariate theoretical copula's parameter α using (6).
5) Eliminate one of the correlated attributes if $|\alpha|>=0.5$, otherwise skip to the next pair of attributes and go back to step 3.
6) After all the attributes are tested, we get a new reduced data as output with uncorrelated attributes holding the same information as the input matrix X .

ALGORITHM I. Dimensionality Reduction Using the PA

```
Input: Data matrix $X$ .
Output: Matrix of reduced data $X$ .
Begin
$\alpha$ = NULL.
for i: =1 to m do
  for j: =1 to m do
    $\alpha_{ij}=\sin(\pi/2\times\tau_{ij})$ .
    if $|\alpha_{ij}|>0.5$ then
      Delete one of the attributes.
    end
  end
end
end
```

Algorithm I and Fig. 2 represent the Proposed Approach (PA). Taking as input the matrix X , this algorithm checks for inter-correlation between two attributes and eliminate one attribute each time correlation is detected. The choice of which to eliminate between the two is random, as the first one detected will be directly flagged for elimination. We then apply the same procedure to all the possible pairs, leaving us with a new relevant and uncorrelated dataset representing the same information as the input matrix X.

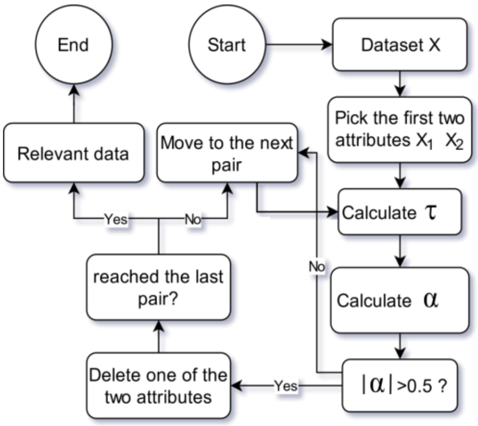


Figure 2. Flow chart of the PA.

To improve our algorithm, and to reduce its time complexity, we use a method proposed in [6] and has been described with more details in [7] and [8] named Fast Kendall's tau instead of the commonly used method for calculating the Kendall's tau τ (time complexity of O(n^2)). It uses a process called sorting by exchanging that decreases the time complexity to O($n \log n$). The equation of Fast Kendall's tau is presented in (8).

$$\tau = \frac{4c}{n(n-1)} - 1 \tag{8}$$

where c defines the concordant pairs. This leads to an initial time complexity of O($m^2 n \log n$) for the entire Algorithm I, but due to the nature of this algorithm, the time complexity is variable and decreases each time an attribute is eliminated. The memory complexity on the other hand is O($m \times n$).

In Fig. 3, we can see an illustration of how our approach treats the data, using the matrix X as input for algorithm I, we eliminate k redundant attributes where $1 \le k \le m-1$, as an output we get a reduced and relevant data where $1 \le l \le m-k$.

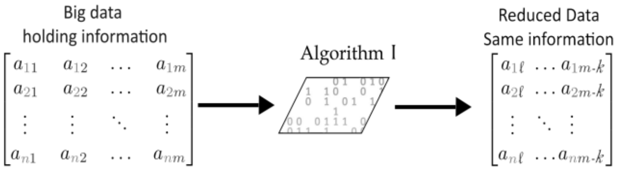

Big data
holding information
$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$
Algorithm I
Reduced Data
Same information
$\begin{bmatrix} a_{1\ell} & \cdots a_{1m\text{-}k} \\ a_{2\ell} & \cdots a_{2m\text{-}k} \\ \vdots & \ddots \vdots \\ a_{n\ell} & \cdots a_{nm\text{-}k} \end{bmatrix}$

Figure 3. Illustration of the PA.

IV. Experimental Results

The results shown in this study were obtained from simulations on RStudio using R version 4.0.3 [9] (64bit), and a PC with the following specs: PU: Intel Core i5-9300H (4 Cores, 8 Threads, up to 4.10 GHz), RAM: 8 GB DDR4 (2666 MHz), GPU: GTX 1050, Disk: SSD and OS: Win 10 (64bit).

To demonstrate the performance of our method, we apply it on real data, then compare the results with baseline methods. For that, we choose datasets taken from UCI machine learning repository [10], which are: "Crop mapping using fused optical-radar" dataset [11] with 174 attributes and 325834 rows, "First-Order Theorem Proving" datasets [12] with 52 attributes and 4589 rows, where the last column defines the class, and lastly "Vehicle" datasets [13] with 18 attributes and 846 rows. Fig. 4, Fig. 5 and Fig. 6 are the plots of the empirical copulas [14], Gaussian theoretical copulas, and copula densities respectively for these datasets. The graphs are obtained using the transformation in (7). By using the goodness of fit test [15] between the bivariate empirical copula and the bivariate theoretical Gaussian copula, we can assume that they belong to the same distribution.

A. Dimensionality Reduction

Table I represents the number of attributes left after performing reduction with the Proposed Approach (PA), using the package [16] for fast Kendall's tau, and the two other selection methods: LASSO technique and Stepwise Selection Method (SW) using the 3 datasets. Stepwise selection is a combination between the Forward Selection and the Backward Elimination [17], the reduction is performed following the best model criteria [18]. While LASSO technique is applied using 10 folds cross-validation [19]. These two methods are performed using the packages "MASS" [20] and "glmnet" [21] respectively.

TABLE I. Dimensionality Reduction Results

Datasets	Original data	PA	LASSO	SW
Crop mapping	174	24	67	141
First order theorem proving	51	14	32	27
Vehicle	18	5	16	16

B. Fitting to the Classification Models

In order to show the performance of the proposed approach against other methods, we fit the dataset "First-order theorem proving" to several classification models. Before that, we normalize and shuffle the reduced dataset in order to reduce the risk of overfitting. We also perform 10 folds cross-validation for each model to make sure we pick the best parameters for the models. The classification models that we chose are: "Artificial Neural Network (ANN)" [22], "Random Forest" [23] and "Adaboost". To run these models, we use the packages "neuralnet" [24], "caret" [25], "dplyr" [26] and "fastAdaboost" [27] respectively. The obtained results are shared in Table II.

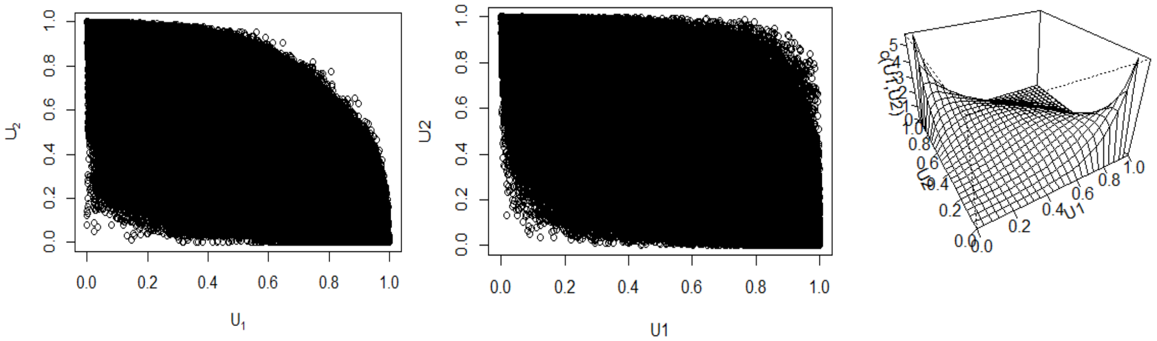
TABLE II. Accuracy of "The First Order Theorem Proving" Datasets

		Original data	PA	LASSO	SW
Dimensions		51	14	32	27
Accuracy	Neural Network	0.761	0.724	0.719	0.715
	Random Forest	0.843	0.837	0.829	0.830
	Adaboost	0.818	0.817	0.815	0.812

C. Discussion

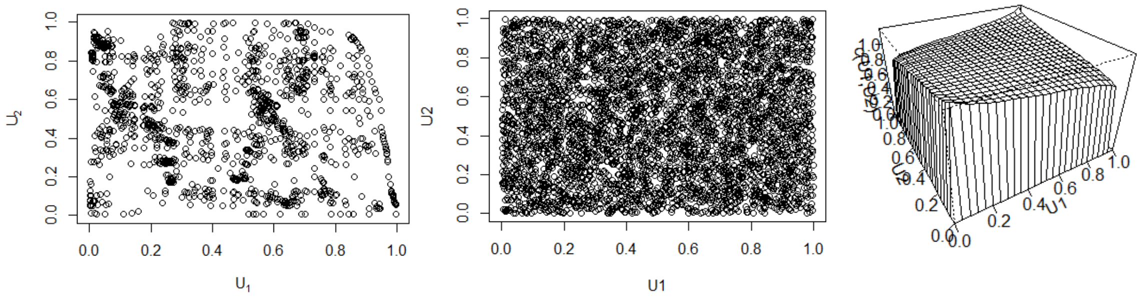
After performing the feature selection methods (Lasso, Stepwise, and PA) for all the datasets, and also the accuracy checks through several models, we obtained the results that will allow us to determine which method is best for each desired type of performance:

- Crop mapping dataset: Starting with the Stepwise selection, it selected 141 dimensions out of 174. Next comes the Lasso technique which left 67 dimensions, and finally our PA, leaving out only 24 features.
- First-Order Theorem Proving datasets: the dataset is reduced to 27 variables by the Stepwise method, giving an accuracy of 0.715 on Neural Network model, 0.83 on Random Forest and 0.812 on Adaboost. Followed by Lasso technique which managed to lower it to 32 variables, which retain 0.719 accuracy on Neural Network, 0.829 on Random Forest, and 0.815 on Adaboost. Finally, our PA cleared out most of redundancy, leaving only 14 variables while also maintaining an accuracy of 0.761 on Neural Network, 0.843 on Random Forest, and 0.818 on an Adaboost model.
- Vehicle dataset: from 18 features in the original data, both Stepwise selection and Lasso technique eliminated enough to leave 16 features, while PA reduced their number down to 5.
- Accuracy test for the “Crop mapping” and “Vehicle” datasets was not included, because these datasets didn’t have a proper class column that could be used to calculate accuracy after prediction.



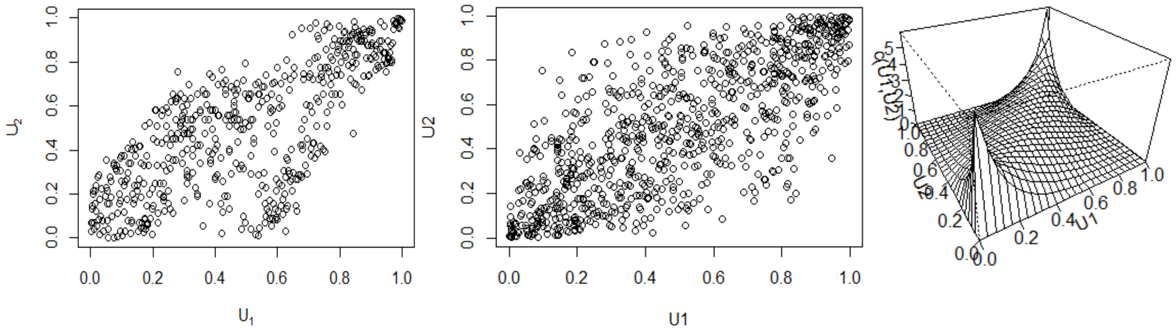
(a) The empirical copula. (b) The theoretical copula. (c) The copula density.

Figure 4. The attributes pair (X_{80}, X_{81}), $\alpha = -0.75$ from “Crop Mapping” dataset.



a) The empirical copula. (b) The theoretical copula. (c) The copula density.

Figure 5. The attributes pair (X_1, X_2), $\alpha = 0.06$ from “First-Order Theorem Proving” dataset.



(a) The Empirical Copula. (b) The Theoretical Copula. (c) The Copula Density.

Figure 6. The attributes pair (X_1, X_2), $\alpha = 0.77$ from “Vehicle” dataset.

Random Forest model showed high accuracy values for all the methods, and that's because it is a good fit to the "First Order Theorem" dataset.

V. Conclusion

Even though reduction methods react in a different way to different datasets, our Proposed Approach (PA) was able to maintain a slightly higher accuracy within different models, with results being close to the other tested feature selection methods. But in term of dimensionality reduction, it is way ahead of them as it was able to reduce much more features, cleaning a lot of redundancy and noise in the way. Other data and studies can be found in [28].

Future work will focus on which attribute of the detected pair to eliminate for the best result. It will use a special algorithm to predict the final outcome before eliminating.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

K. Femmam carried out the research under the supervision of S. Femmam, and all authors approved the final version.

References

[1] S. Jang and S. Lee, "Feature selection based on euclid distance and neuro-fuzzy system," *Journal of Advances in Information Technology*, vol. 11, no. 3, pp. 155-160, August 2020.

[2] M. Kadhum, S. Manaseer, and A. L. A. Dalhoum, "Evaluation feature selection technique on classification by using evolutionary ELM wrapper method with features priorities," *Journal of Advances in Information Technology*, vol. 12, no. 1, pp. 21-28, February 2021.

[3] R. Houari, A. Bounceur, M. T, Kechadi, and R. Euler, "Dimensionality reduction in data mining: A copula approach," *Expert Systems with Applications*, vol. 64, pp. 247-260, July 2016.

[4] K. Femmam, M. T, Kechadi, and S. Femmam, "Dimensionality reduction using multivariate copulas," *Expert System with Application Journal*, 2021.

[5] R. B. Nelsen, *An Introduction to Copulas*, 2nd ed., Springer Science & Business Media, Jun. 2007.

[6] W. R. Knight, "A computer method for calculating Kendall's tau with ungrouped data," *Journal of the American Statistical Association*, vol. 61, no. 314, pp. 436-439, Jun. 1966.

[7] J. Abrevaya, "Computation of the maximum rank correlation estimator," *Economics Letters*, vol. 62, no. 3, pp. 279-285, Mar. 1999.

[8] D. Christensen, "Fast algorithms for the calculation of Kendall's T," *Computational Statistics*, vol. 20, no. 1, pp. 51-62, Mar. 2005.

[9] D. Dua and C. Graff, "UCI machine learning repository," University of California, School of Information and Computer Science, Irvine, CA, 2019.

[10] R. C. Team, "R: A language and environment for statistical computing," 2020.

[11] I. Khosravi and S. K. Alavipanah, "A random forest-based framework for crop mapping using temporal, spectral, textural and polarimetric observations," *International Journal of Remote Sensing*, vol. 40, no. 18, pp. 7221-7251, Sep. 2019.

[12] J. P. Bridge, S. B. Holden, and L. C. Paulson, "Machine learning for first-order theorem proving," *Journal of Automated Reasoning*, vol. 53, no. 2, pp. 141-172, Aug. 2014.

[13] J. P. Siebert, "Vehicle recognition using rule based methods," *Turing Institute*, 1987.

[14] I. Kojadinovic and J. Yan, "Modeling multivariate distributions with continuous margins using the copula R package," *Journal of Statistical Software*, vol. 34, no. 1, pp. 1-20, May 2010.

[15] C. Genest, B. Rémillard, and D. Beaudoin. "Goodness-of-fit tests for copulas: A review and a power study," *Insurance: Mathematics and Economics*, vol. 44, no. 2, pp. 199-213, Apr. 2009.

[16] P. Filzmoser, H. Fritz, K. Kalcher, and M. V. Todorov, "Package 'pcaPP'," *Journal of the American Statistical Association*, vol. 314, no. 61, part 1, pp. 436-439, Apr. 2021.

[17] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd ed., New York: Springer, 2015.

[18] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*, CRC Press, 2019.

[19] V. Fonti and E. Belitser, "Feature selection using lasso," *VU Amsterdam Research Paper in Business Analytics*, vol. 30, pp. 1-25, Mar. 2017.

[20] W. R Venables and B. D. Ripley, *Modern Applied Statistics with S-PLUS*, 4th ed., New York: Springer Science & Business Media, 2013, ch. 7, pp. 183-206.

[21] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, Aug. 2010.

[22] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of Artificial Neural Network (ANN) modeling and its application in pharmaceutical research," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 22, no. 5, pp. 717-727, Jun. 2000.

[23] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18-22, Dec. 2002.

[24] S. Fritsch, F. Guenther, and M. N. Wright, "Neuralnet: Training of neural networks," R package version 1.44.6, 2019.

[25] M. Kuhn, *et al.*, "Caret: Classification and regression training," R package version 6.0-86, Mar. 2020.

[26] H. Wickham, R. François, L. Henry, and K. Müller, "Dplyr: A grammar of data manipulation," R package version 1.0.4, 2021.

[27] S. Chatterjee, "fastAdaboost: A fast implementation of Adaboost," R package version 1.0.0, 2016.

[28] R. Aschheim, S. Femmam, and M. F. Zerarka, "New "Graphiton" model: A computational discrete space, self-encoded as a trivalent graph," *Computer and Information Science J.*, vol. 5, no. 1, pp. 1-12, 2012.

K. Femmam is currently a PhD student in statistics at the Mohamed Khider university of Biskra Algeria. She received her Master degree in applied mathematics, her research interests include applications of the estimation of copula's parameter in different fields such as maching learning and modeling, data mining and financial time series.

S. Femmam is Director of research at the University of Haute-Alsace France and responsible of the Research team on Signals & Safety Systems of Polytechnic Engineers School Sceaux France. He joined the CMU Carnegie Mellon University & West Virginia University as Postdoc Fellow and Distinguished Visiting Professor. His main research area is signal processing, safety systems, communication and embedded systems. He has a strong interest in perception and characterization of signals, optimal filtering, spectral analysis, wavelets and perception haptics. Dr. Femmam is a senior member of IEEE, Board of Director of the Institute for Engineering and Technology Innovations in the World. He is the (Academic, Chief) Editor-in-Chief, Editor, Editorial Board, Guest Editor & Advisory Board members of more than 20 International Journal.