Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26 - 28 February 2024; pp 61-96

# Modeling Multivariate Data Using Copula Theory: Analysis of an Environmental Dataset

**Rahul Chatterjee**[1] **and Nabendu Pal**[1,2]
[1]*Department of Mathematics,*
*University of Louisiana at Lafayette, Lafayette, Louisiana, USA*
[2]*Faculty of Mathematics and Statistics,*
*Ton Duc Thang University, Ho Chi Minh City, Vietnam*

Received: 13 May 2024; Revised: 02 June 2024; Accepted: 04 June 2024

**Abstract**

The normal distribution, though widely popular and heavily used in modelling datasets, has its own limitations, especially dealing with engineering and environmental data. In the univariate case, when the variable of interest is positively skewed, one can use a host of other distributions such as Gamma, Weibull, Lognormal *etc.*, just to name a few. However, in a multivariate set-up, the multivariate normal distribution appears to be the default choice, either by omission or by commission. The multivariate normal model has a host of advantages as its inferential problems are well studied, and the sampling distributions of its key statistics are relatively convenient to deal with. To be precise, the sample average follows a multivariate normal, and the sample cross-product matrix follows a Wishart distribution, and these two statistics are independent. Further, conditional expectation of any component given the remaining components is a linear function (of those remaining components) which is the foundation of the linear regression analysis.

But what happens if our multivariate data, which we commonly see in many applied problems, do not follow normal? The first casualty is the aforementioned mutual independence between the two commonly used statistics, let alone them being the minimal sufficient. Secondly, the linear regression model may not hold, thereby complicating the further conditional inferences. Also, multivariate normality forces one to assume marginally univariate normal distributions which may not seem reasonable as seen from the marginal empirical relative frequency histograms. One possible way out of this difficult situation is to transform the individual components to achieve multivariate normality, but this faces two big hurdles – (a) it would be an ad hoc approach to begin with; and (b) such ad hoc transformations may distort the natural association(s) among the components as well as the units being used, thus rendering the subsequent analyses questionable. On this backdrop, the copula theory comes handy in modelling multivariate data. Multiple individual components, apparently following skewed distributions, can be adequately combined by a suitable copula (also known as a link function) in order to model the given multivariate data. As opposed to the multivariate normal distribution's 'top-down' approach, the copula theory provides a 'ground-up'

Corresponding Author: Nabendu Pal
Email: nabendupal@gmail.com

approach where diversely distributed marginals can be combined into a suitable multivariate distribution for further inferences, including regressions.

Our study of the copula theory was motivated by an environmental dataset from the Mekong Delta Region (MDR) of Vietnam. In a bivariate set-up we have used a special copula, known as the Farlie-Gumbel-Morgenstern Copula (FGMC) to analyze the data. But this has also opened up a host of other research problems, such as estimation of the copula parameter, hypothesis testing, goodness of fit tests of FGMC, *etc.* Further, FGMC is just one of many, - possibly three dozen copulas, and thus this is a very rich emerging research field which has received relatively less attention, but has tremendous implications in 'Big Data' or 'Data Analytics'. We will also discuss some of the major challenges in copula theory which are related to heavy yet efficient computations in a reasonable amount of time. Thus this is a rich research area where experts in efficient algorithms and/or numerical analysis are very much welcome.

*Key words:* Copula; Jeffrey's prior; Parametric bootstrap method; Prediction mean absolute error; Prediction root mean squared error; Kolmogorov-Smirnov statistic.

**AMS Subject Classifications:** 62F10, 62F15, 62C05

## 1. Introduction

### 1.1. Why copula?

The normal distribution, though widely popular and heavily used in modelling datasets, has its own limitations. In the univariate case, when the variable of interest is positively skewed, one can use a host of non-normal distributions such as Gamma, Weibull, Lognormal *etc.*, just to name a few. However, in a multivariate set-up, the multivariate normal distribution appears to be the default choice, either by omission or by commission. The multivariate normal model has a host of advantages as its inferential problems are well studied, and the sampling distributions of its key statistics are relatively convenient to deal with. To be precise, let $\mathbf{X} = (X_1, X_2, ..., X_p)'$ be a p-variate random vector whose distribution is assumed to be $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} = ((\sigma_{ij})) > 0$ (p.d.). Based on a random sample $\mathbf{X}_i$, $1 \leq i \leq n$, (*i.e.*, $n$ copies of $\mathbf{X}$), assuming $n > p$, the maximum likelihood estimators (MLEs) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively $\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$, and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}/n$, where $\boldsymbol{S} = \sum_{i=1}^n (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'$. Further, under the above normality of $\mathbf{X} = (X_1, ...., X_p)'$, it is well known that $E(X_1|X_2, ..., X_p) = \beta_1 + \sum_{k=2}^p \beta_k X_k$, for suitable value of $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ which depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and this is the motivation behind the usual multiple linear regression where $X_1$ is intended to be explained as a linear function of $(X_2, ..., X_p)$ subject to some variation. But what happens if $\mathbf{X}$ does **not** follow $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$?
(a): Can we have the aforementioned $\overline{\mathbf{X}}$ and $\boldsymbol{S}/n$ as the MLE of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$?
- Possibly not.
(b): Can we have the independence of $\overline{\mathbf{X}}$ and $\boldsymbol{S}$ (which is the foundation of most of the normality based inferential results)?
- Most likely not.
(c): Does regressing $X_1$ on $(X_2, ..., X_p)$ through a linear function make sense?
- Doesn't seem so, since $E(X_1|X_2, ..., X_p)$ may not be linear at all if the distribution of $\mathbf{X}$ is non-spherically symmetric and/or does not follow homoscedasticity.

Also, if $\boldsymbol{X} = (X_1, ..., X_p) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then it yields $X_k \sim N(\mu_k, \sigma_{kk} = \sigma_k^2), 1 \leq k \leq p$. Thus, in dealing with a multivariate data set, if one assumes the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ model then inadvertently univariate normality is assumed for individual components, and this can be problematic. Yet, in multivariate modelling, ranging from psychology to anthropology, from agriculture to environmental science, especially in a 'Big Data' setting, multivariate normal distribution is being used hastily without paying closer attention to whether such model fitting is appropriate or not.

If the multivariate normal is found to be inappropriate for the data $\mathbf{X}_i, 1 \leq i \leq n$, then one may transform the variable(s) suitably hoping that the transformed data would follow normal. But there are two major issues with such transformations. There is no magic formula to tell us what transformation would be suitable for normality. Secondly, often such transformed variables are hard to interpret, and they lose significance to the original problem which gave rise to the dataset to begin with.

This study has been motivated by several datasets where component-wise histograms indicate that marginals are heavily skewed, and therefore the joint distribution of the marginals ought to be something other than a multivariate normal distribution (not even elliptically symmetric one). In such a situation, it makes sense to follow a 'ground-up' approach to build a multivariate model starting with marginals, rather than the 'top-down' approach of starting with a (questionable) multivariate model and then live with its consequences at the marginal level.

Copula theory is a convenient 'ground-up' approach where one theorizes a multivariate distribution for the random vector $\mathbf{X} = (X_1, X_2, ..., X_p)'$ based on the marginal of each $X_k, 1 \leq k \leq p$. This is based on the understanding that the desired joint distribution ought to obey a particular structure involving the marginals which we have much control over. The following subsection gives a brief introduction of the copula theory. The focus of this work is on the bivariate set-up; however, we may present some general multivariate results occasionally.

### 1.2. General copula framework

The path breaking theorem in Sklar (1959) plays the most important role in the Copula theory. In the simplest case of a bivariate distribution, it tells us that given a random vector $(X_1, X_2)$ with absolutely continuous marginal cumulative distribution functions (cdfs), $F_1$ and $F_2$, with corresponding probability density functions (pdfs) $f_1$ and $f_2$ respectively, and its joint cdf denoted by $F$, with joint pdf $f$, there exist unique copula $C$ (a functional), such that

$$\begin{aligned} F(x_1, x_2) &= C(F_1(x_1), F_2(x_2)), \\ i.e., f(x_1, x_2) &= \partial^2 F(x_1, x_2)/\partial x_1 \partial x_2 \\ &= C^{(x_1, x_2)}(F_1(x_1), F_2(x_2)) f_1(x_1) f_2(x_2), \end{aligned} \tag{1}$$

where $C^{(u,v)}(u, v) := \partial^2 C(u, v)/\partial u \partial v$.

In general, given a continuous random vector in $p$-dimension, *i.e.*, $\boldsymbol{X} = (X_1, X_2, ..., X_p)'$, with marginal cdfs $F_k, k = 1, 2, ..., p$, if we use the transformations such that $U_k := F_k(X_k), k = 1, 2, .., p$, then we have $U_k \sim \text{Uniform}(0, 1), k = 1, 2, ..., p$. The copula function $C : [0, 1]^p \to$

$[0,1]$ is a joint multivariate cdf of $\boldsymbol{U} := (U_1, U_2, ..., U_p)'$, *i.e.*,

$$C(u_1, u_2, ....., u_p) = P(U_1 \leq u_1, ...., U_p \leq u_p). \tag{2}$$

The joint cdf of $\mathbf{X}$, denoted by $F(x_1, ..., x_p)$, can be given in terms of $C(u_1, ..., u_p)$. By Sklar's theorem there exists a unique copula $C$ such that

$$F(x_1, x_2, ..., x_p) = C(F_1(x_1), F_2(x_2), ...., F_p(x_p). \tag{3}$$

Simply put, the copula $C$ is viewed as a dependence structure among the marginal cdfs.

Since the inception of the copula idea, one can find several copulas in the literature such as Gaussian copula, Exponential copula, Clayton copula, Frank Copula *etc.*, just to name a few. Out of the many available copula structures we focus on the Farlie-Gumbel-Morgenstern copula (FGMC) (see Morgenstern (1956)). The following section gives a brief introduction about the joint distribution based on FGMC, henceforth referred to as Farlie - Gumbel - Morgernstern Distribution (FGMD). The main reason behind our choice of FGMC (and subsequently that of FGMD) is its simplicity. Moreover, the nature of our investigation is completely new, and to the best of our knowledge the type of our investigation has not been carried out for FGMC. Therefore, this work of ours can be used as a template of future research for all other copulas as needed.

### 1.3. Farlie - Gumbel - Morgenstern Copula (FGMC) and the resultant model

As mentioned in the earlier section, a host of Copula structures have been discussed in the existing literature and one can find an overview of the available copula structures in Nelsen (2007). Several bivariate and multivariate non-normal probability distributions based on copula structures can be found in Kotz *et al.* (2004).

Morgenstern (1956) first introduced the following bivariate probability distribution on the square $[-1,1] \times [-1,1]$ of the form

$$f(x_1, x_2) = \frac{1}{4}(1 + \lambda x_1 x_2), \tag{4}$$

where $|\lambda| \leq 1$ and $-1 \leq x_1, x_2 \leq 1$. Farlie (1960) further studied various standard correlation coefficients between $X_1$ and $X_2$ for the bivariate distribution in (4). The limitations that a bivariate normal distribution brings to a dataset were first pointed out by Gumbel (1960) while he constructed a bivariate distribution with exponential marginals using Morgenstern's underlying copula in (4).

The pdf of the bivariate Farlie-Gumbel-Morgenstern distribution (FGMD) with general marginals based on the FGMC is given by

$$f(x_1, x_2) = f_1(x_1) f_2(x_2)[1 + \lambda(2F_1(x_1) - 1)(2F_2(x_2) - 1)], \tag{5}$$

where $|\lambda| \leq 1$ is the *association* parameter, $f_1, f_2$ are the marginal pdfs of the components $X_1$ and $X_2$, with corresponding marginal *cdfs* $F_1, F_2$ respectively. The range of $\lambda$ happens to be $[-1,1]$, similar to many common correlation coefficients.

As a special case of (5), D'este (1981) considered a special biavriate Gamma distribution with gamma marginals and studied the structures of the covariance, conditional

expectations as well as other distributional properties. Since the inception of the FGMC, it has undergone several modifications over the years leading to some wider family of FGMC by different researchers. All these modifications were done with the goal of capturing a wider range of dependence among the components through common dependence measures such as Pearson's Correlation Coefficient ($\rho$), Spearman's Correlation coefficient ($\rho_s$), Kendall's Tau ($\rho_K$), *etc.* In their modified FGMD Huang and Kotz (1999) showed that with a polynomial type single parameter extension of the FGMC with uniform marginals the maximal attainable range of $\rho$ is $[-0.39, 0.333...]$. Bairamov and Kotz (2002) proposed a new generalization of FGMD by introducing new association parameters and were able to attain a maximal positive (Pearson's) correlation of $\rho = 0.5021$ for some specific values of the model parameters. All these generalizations were made to accommodate a larger spectrum of the Pearson's correlation coefficient values. However the Pearson's correlation coefficient measures the strength of linear relationship between the components; therefore, paying attention only to this aspect of dependency, at the cost of adding more parameters to the model, is a rather narrow approach. Amblard and Girard (2009) gave a new family of copulas by generalizing the FGMC and highlighted the main feature of the proposed family as to permit modelling of data with high positive dependency, in particular over the range of $\rho_s \in [-0.75, 1]$. Another new generalization of the $FGMC$ was put forward by Bekrizadeh *et al.* (2012) and they were able to show the usefulness of the proposed generalized model in data with high negative dependence value by showing the (Spearman's rank correlation) values of $\rho_s \in [-0.5, 0.43]$. All these generalizations were made by introducing new parameters which only adds to the complexity of the statistical inferences of the FGMD model.

### 1.4. A motivational example with a real life dataset

This work has been motivated by an excellent investigation carried out by Merola *et al.* (2015) where the researchers have presented, among other things, a useful dataset on arsenic ($As$) concentration as well as a few other apparently benign elements from a survey carried out in Dong Thap province within the Mekong Delta Region (MDR) of Southern Vietnam. The complete dataset is given in Appendix A.1.

Vietnam is one of the worst affected countries where arsenic contamination in groundwater is particularly worrisome in two areas, - The Red River Delta (RDR) in the northern part, and the Mekong Delta Region (MDR) in the southern part. The MDR is the most economically vibrant region of the country which comprises twelve southern provinces and one major city (Can Tho) municipality. The provinces adjacent to Mekong river and its distributaries have been witnessing a very high concentration of arsenic in groundwater which is caused by both natural as well as man-made factors as discussed below.

As mentioned at the beginning of this section, Merola *et al.* (2015) collected data on arsenic concentration in groundwater in two subregions within Dong Thap province of MDR. Dong Thap, along with An Giang and Long An, is one of the provinces bordering Cambodia that has a high level of arsenic and poses a public health hazard. Thus, measuring arsenic in groundwater and issuing guidelines if and when needed is of paramount importance for the local administrations to mitigate arsenic poisoning. However, measuring arsenic level frequently and accurately is a time consuming and/or expensive exercise. Therefore it would be of great help to all the stakeholders if the level (or concentration) of arsenic could be predicted from the other benign elements when it is established, based on some existing

survey data, that in certain region there is an association between arsenic and one or more benign element(s) which can be measured easily (and cheaply), often through user friendly devices.

### 1.5. Scope of this research

The initial exploratory analysis points towards the fact that the MDR dataset consists of components which, firstly, have distinct distributions over the two sub-regions as mentioned above and secondly, have mostly skewed marginal distributions. We further delve into the exploration of the nature of pairwise association present among the variables in this dataset. We employ the FGMD model for the purpose.

The flexibility of the copula structure lies in allowing the freedom of choice of the desired marginal distributions. Hence, the association parameter $\lambda$ of the copula structure (5) becomes a pivotal parameter in conserving the dependency between the components. As a result, inferences on the association parameter $\lambda$ in (5) is of paramount interest. The basis of this current work has been the FGMD given in (5) with the goal of studying the inferential aspects of the association parameter $\lambda$ comprehensively, with known marginals.

The inferential aspect includes parameter estimation where we have discussed a host of estimators and recommend the most suitable ones. Secondly, we have studied the existence of association among the variates through hypothesis testing under the FGMD model. We proposed a family of parametric bootstrap (PB) tests which addresses the problem of $\lambda = 0$ vs $\lambda \neq 0$. Along with the regular asympototic tests, we have studied the proposed PB tests and have shown that they tend to attain the nominal level very accurately.

While various correlation measures reveal some interesting patterns in terms of association between Arsenic ($As$) and Chlorine ($Cl$), between $As$ and Hydrogen Potential ($pH$), and between $As$ and Redox Potential or level ($Eh$), they do not address the objective of this work, *i.e.*, predicting the value of $As$ when a suitable covariate, which is known to be significantly associated with $As$, is known. For example, in the southern region, where $As$ and $Cl$ are apparently strongly associated, can we predict the value (or, do we know the expected value) of $As$ when $Cl$ is equal to, say, 10 $ppm$? The prediction problem which has been posed above can be answered only by fitting an appropriate bivariate probability distribution to the given data on two relevant variates.

Let us denote the variate $As$ by $Y$ for the time being, and its suitable covariate by $X$ (where $X$ can be either $Cl$, or $pH$, or $Eh$). (For convenience in notation, these three covariates can be denoted by $X_1$, $X_2$ and $X_3$, respectively.) We addressed the suitable distribution of $(X, Y)$ through FGMD which fits the given data. Once that suitable distribution of $(X, Y)$ fits the data, then we use the conditional distribution of $(Y|X)$ to draw inferences on $Y$ when $X$ is given. As noted earlier, the joint probability distribution of $(X, Y)$ has to be a non-normal one because the univariate normality tests reject such a notion most of the time (six out of eight cases - four variables in two subregions).

Finally, one can raise the question of 'goodness of fit' (GoF) of FGMD. It is worthy of noting that there is no "one stop solution" for the goodness of fit problem for the host of available copula in the literature and it remains an open problem. Several goodness of fit tests are available across the literature but to the best of our knowledge there doesn't exist

one for FGMC which considers the parametric nature of the distribution. We have proposed and developed a novel data driven goodness of fit test for FGMD, which does not assume any known distribution of the test statistic under the null hypothesis. A detailed study of this goodness of fit test including the test procedures, as well as its performance is discussed which validates the application of FGMD model to the MDR dataset.

## 2. Point estimation of the association parameter

If one looks at the existing applications of the copula theory with real-life data sets then it becomes abundantly clear that the preferred estimator of the association parameter has always been the maximum likelihood estimator (MLE). But how good is the MLE? From an asymptotic point of view the MLE has nice tractable limiting distributional properties. But, for small to moderate sample sizes the performance of the MLE of the FGMD association parameter $\lambda$ is totally unknown. Worse, the existing literature is completely silent on other possible estimators, especially the Bayes ones under noninformative priors. In a parametric set up, one should study various estimators of all the model parameters simultaneously which include the association parameter $\lambda$ as well as other parameters of the marginal distributions. (For example, if one assumes a two parameter gamma model for each of the two marginals, then one ends up with a total five parameters.) It has been noted that estimating just the association parameter with known marginals itself is a research problem as it entails several point estimators with corresponding sampling distributions, followed by hypothesis testing which allows us to verify, under the FGMD assumption, whether the components are independent or not. The computational challenges that one faces with Bayes estimators in this simplistic scenario (*i.e.*, just for the association parameter) can be quite overwhelming. However, the simplistic model that we are using in this work can be applied in a totally non-parametric marginal set up where one can use the empirical marginal cdf of each component to replace the aforementioned known marginal, and then can proceed with the subsequent inferences. With that above objective in mind, the following subsections present a brief review of parametric estimation of the association parameter $\lambda$ as available in the existing literature. Also, the following lemma will be useful in deriving Bayes estimators under noninformative priors.

**Lemma 1:** Based on the iid observations $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ from (5) with marginals $f_1$ and $f_2$ completely known, the Fisher information $I(\lambda)$ is given as $I(\lambda) = nI^0(\lambda)$, where $I^0(\lambda)$ is the Fisher information per observation ($FIPO$), and

$$I^0(\lambda) = (1/4) \int_{-1}^{+1} \int_{-1}^{+1} u_1^2 u_2^2 (1 + \lambda u_1 u_2)^{-1} du_1 du_2. \tag{6}$$

Note that the $FIPO$ expression is free from $f_1$ and $f_2$. A further simplification yields

$$I^0(\lambda) = \sum_{m=0}^{\infty} \lambda^{2m}/(2m+3). \tag{7}$$

Note that the infinite sum in the above expression is convergent. Using that expression of the infinite sum, the final form of the FIPO, is given by

$$I^0(\lambda) = \{-\lambda + tanh^{-1}(\lambda)\}/\lambda^3, \tag{8}$$

where $tanh^{-1}(\lambda) = (0.5)log((1+\lambda)/(1-\lambda))$. See Chatterjee (2022) for the proof.

**Remark 1:** It is not at all surprising to see that the expression of $I^0$ in (7) or (8) is free from $f_k$'s $(k = 1, 2)$. Since the marginals are assumed to be completely known, without any loss of information one can look at $Y_{ik} = F_k(X_{ik})$, $k = 1, 2$, $1 \leq i \leq n$. Note that $Y_{ik}$'s are iid Uniform$(0, 1)$. Each $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})'$ then follows the $FGMD(\lambda)$ with joint *pdf*, say $g(\boldsymbol{y}) = [1+\lambda(2y_1-1)(2y_2-2)]$ on the unit square $[0, 1] \times [0, 1]$. The transformation $\boldsymbol{X}_i \to \boldsymbol{Y}_i$ does not change the problem as far as inference on $\lambda$ is concerned, and yields the $FIPO$ expression as stated above.

In the following subsections, we propose a wide variety of estimators of the association parameter $\lambda$ based on $n$ iid observations from (5) with known marginals $f_1$ and $f_2$.

### 2.1. Method of moment estimation

Method of moment estimator is attained essentially by equating the sample raw moment with the population moment. For the joint population moment, using the simple calculation of the expectation of the distribution in (5) and some further simplification lead us to the following form

$$E(X_1X_2) = E(X_1)E(X_2) + \lambda I_1 I_2, \tag{9}$$

where $I_k = \int_{-1}^{1}(u/2)F_k^{-1}((1+u)/2)\partial u$, $k = 1, 2$. For convenience define $\mu_k = E(X_k)$, $k = 1, 2$, *i.e.*, the means of the known marginals. Therefore from (9) it can be easily established that $Cov(X_1, X_2) = \lambda I_1 I_2$. For the method of moment estimator $\hat{\lambda}_{MM}$ we equate $\lambda I_1 I_2$ with the sample equivalent of $Cov(X_1, X_2)$ which is $(1/n)\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)$ where $\overline{X}_k = (1/n)\sum_{i=1}^{n} X_{ki}$, $k = 1, 2$. Therefore,

$$\hat{\lambda}_{MM} = (nI_1I_2)^{-1}\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2). \tag{10}$$

### 2.2. Maximum likelihood estimation

For the brevity in derivation, let us denote $2F_k(x_{ik}) - 1 = G_k(x_{ik}), k = 1, 2$. The log-likelihood function of the data denoted by $l(\lambda)$ is as follows

$$l(\lambda) = C + \sum_{i=1}^{n} ln(1 + \lambda G_1(x_{i1})G_2(x_{i2})), \tag{11}$$

where $C$ is a constant, free of $\lambda$. It is tempting to take derivative of $l(\lambda)$ and equating it with zero, *i.e.*,

$$\sum_{i=1}^{n} G_1(x_{i1})G_2(x_{i2})/(1 + \lambda G_1(x_{i1})G_2(x_{i2})) = 0, \tag{12}$$

to find the MLE of $\lambda$. But this can lead to a computational error as the solution may lie outside the parameter space which may go unnoticed in simulation studies. (We suspect that this issue may arise for other copula - based joint distributions as well, and may have gone unnoticed in applications.)

**Theorem 1:** The MLE of $\lambda$ *i.e.*, $\hat{\lambda}_{ML}$ as it is called here, which maximizes $l(\lambda)$ in (11), exists, and it is unique.

Proof. See Appendix of Chatterjee (2022).

**Remark 2:** Define $a_i = G_1(x_{i1})G_2(x_{i2})$, $1 \leq i \leq n$, and $G_k(x_{ik}) = 2F_k(x_{ik}) - 1$, $k = 1, 2$. Let $h(\lambda) = \sum_{i=1}^{n} a_i/(1 + \lambda a_i)$, $\lambda \in [-1, +1]$. As seen from the details of the proof of the above theorem, $\hat{\lambda}_{ML}$ takes the following form

$$\hat{\lambda}_{ML} = \begin{cases} -1 & \text{if } h(-1) < 0 \\ \text{solution of (12)} & \text{if } h(-1) > 0 \text{ and } h(+1) < 0 \\ +1 & \text{if } h(+1) > 0, \end{cases} \tag{13}$$

**Remark 3:** It is further seen that if all the $a_i$'s are $> 0$, which happens with probability $(0.5)^n$, then $l(\lambda)$ is monotonically increasing in $\lambda$. Hence $\hat{\lambda}_{ML}$ is $+1$. Thus, $\{(a_1, ..., a_n)|\, a_i > 0\, \forall\, i\} \subseteq \{(a_1, ..., a_n)|\, h(+1) = \sum_{i=1}^{n} a_i/(1 + a_i) > 0\}$. Similarly, if all the $a_i$'s are $< 0$, which again happens with probability $(0.5)^n$, then $l(\lambda)$ is monotonically decreasing in $\lambda$. Hence $\hat{\lambda}_{ML}$ is $-1$. Thus, $\{(a_1, ..., a_n)|\, a_i < 0\, \forall\, i\} \subseteq \{(a_1, ..., a_n)|\, h(-1) = \sum_{i=1}^{n} a_i/(1 - a_i) < 0\}$. We will see later in our simulation study that $\hat{\lambda}_{ML}$ can take $\pm 1$ with substantially high probabilities depending on the sample size as well as $\lambda$.

### 2.3. Bayes' estimators

For any suitable prior $\pi(\lambda)$ over the parameter space $[-1, +1]$, the posterior distribution of $(\lambda|data)$, denoted by $g(\lambda|data)$, is

$$g(\lambda|data) = \frac{\prod_{i=1}^{n}[1 + \lambda G_1(X_{i1})G_2(X_{i2})]\pi(\lambda)}{\int_{-1}^{1}\prod_{i=1}^{n}[1 + \lambda G_1(X_{i1})G_2(X_{i2})]\pi(\lambda)\partial\lambda}. \tag{14}$$

A natural choice of the prior for the association parameter is a modification of the beta distribution which is originally defined over the space $(0, 1)$. The beta-type prior density function defined over the parameter space $[-1, +1]$ is

$$\pi(\lambda) = 1/(2B(a, b))((1 + \lambda)/2)^{a-1}((1 - \lambda)/2)^{b-1}, \tag{15}$$

where $a, b$ are the hyper-parameters.

The most common loss function for estimating a parameter is the usual squared error loss. However, when a parameter is restricted to a finite range, as we have here for the association parameter $\lambda$, a weighted quadratic loss is more meaningful which can assign a heavy penalty near the boundary. Hence, we consider a general structure of the loss function of the form

$$L(\hat{\lambda}, \lambda) = w(\lambda)(\hat{\lambda} - \lambda)^2, \tag{16}$$

where $w(\lambda)$ is a suitable weight function. In this work we are going to consider weight function $w(\lambda)$ of the form

$$w_\delta(\lambda) = (1 - \lambda^2)^{-\delta},\ \delta \geq 0. \tag{17}$$

Note that $\delta = 0$ leads to the usual squared error loss. For any $\delta > 0$, the loss (16) goes to $\infty$ as $\lambda$ approaches $\pm 1$ and $|\hat{\lambda} - \lambda| > 0$. In other words, a small deviation of $\hat{\lambda}$ from $\lambda$ near the boundary can be very costly.

Under the general weighted quadratic loss (16), the general structure of the Bayes' rule is given as

$$\begin{aligned}\hat{\lambda}_B &= \frac{E(\lambda w(\lambda)|\lambda \sim g(\lambda|data))}{E(w(\lambda)|\lambda \sim g(\lambda|data))} \\ &= \frac{\int_{-1}^{1} \lambda w(\lambda) \prod_{i=1}^{n}[1 + \lambda G_1(X_{i1})G_2(X_{i2})]\pi(\lambda)d\lambda}{\int_{-1}^{1} w(\lambda) \prod_{i=1}^{n}[1 + \lambda G_1(X_{i1})G_2(X_{i2})]\pi(\lambda)d\lambda}.\end{aligned} \tag{18}$$

With the special structure of $w(\lambda) = w_\delta(\lambda) = (1 - \lambda^2)^{-\delta}$, we are now ready to derive the Bayes' rule, denoted by $\hat{\lambda}_{B\delta}$ as follows.

In order to attain a tractable structure of the Bayes' rule, we resort to a simple algebraic manipulation within Equation (18). Let us focus on the term $\prod_{i=1}^{n}(1 + \lambda G_1(X_{i1})G_2(X_{i2}))$ in the Equation (18). Recalling from Remark 2 that $a_i = G_1(X_{i1})G_2(X_{i2})$, the following product term can be rewritten as

$$\begin{aligned}\prod_{i=1}^{n}(1 + \lambda G_1(X_{i1})G_2(X_{i2})) &= (1 + \lambda a_1)(1 + \lambda a_2)....(1 + \lambda a_n) \\ &= 1 + \lambda \sum_{i_1=1}^{n} a_{i_1} + \lambda^2 \sum_{(1 \le i_1 < i_2 \le n)} a_{i_1}a_{i_2} + ... \\ &... + \lambda^k \sum_{(1 \le i_1 < i_2 < ... < i_k \le n)} a_{i_1}a_{i_2}...a_{i_k} + .... + \lambda^n a_{i_1}a_{i_2}...a_{i_n}.\end{aligned} \tag{19}$$

Call $\sum_{(1 \le i_1 < i_2 < ... < i_k \le n)} a_{i_1}a_{i_2}...a_{i_k} = D_k$, $1 \le k \le n$, and define $D_0 = 1$. Therefore $\prod_{i=1}^{n}(1 + \lambda G_1(X_{i1})G_2(X_{i2})) = \sum_{k=0}^{n} \lambda^k D_k$. Hence, the Bayes' rule in (18) can be simplified as -

$$\hat{\lambda}_B = \frac{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k+1}(1 - \lambda^2)^{-\delta}\pi(\lambda)d\lambda}{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^k(1 - \lambda^2)^{-\delta}\pi(\lambda)d\lambda}. \tag{20}$$

Further, we will consider the special case of $a = b = d$, which implies a symmetric prior about 0. We are going to introduce the notation $\beta$ as $\beta = d - \delta$ and the estimator (18) with the prior in (17) will be denoted as $\hat{\lambda}_{B\beta}$, *i.e.*,

$$\hat{\lambda}_{B\beta} = \frac{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k+1}(1 - \lambda^2)^{\beta-1}d\lambda}{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^k(1 - \lambda^2)^{\beta-1}d\lambda}. \tag{21}$$

### 2.3.1. Special case of $\beta = 1$ (Bayes estimator under flat prior or BFP)

A particular case of interest is $\beta = 1$ which can happen if $\delta = 0$ and $d = 1$ or $\delta = 1$ and $d = 2$ *etc.* Since $\beta = 1$ (due to $\delta = 0$ and $d = 1$) also implies the Bayes' estimator under

the flat prior $(FP)$ using the ordinary squared error loss function, we denote $\hat{\lambda}_{B1}$ as $\hat{\lambda}_{BFP}$ and is given by

$$\hat{\lambda}_{BFP} = \frac{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k+1} d\lambda}{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k} d\lambda} = \frac{\sum_{k=0}^{n} (D_k/(k+2))\{1-(-1)^k\}}{\sum_{k=0}^{n} (D_k/(k+1))\{1-(-1)^{k+1}\}}. \tag{22}$$

### 2.3.2. Bayes' estimator under Jeffrey's prior (BJP)

Let us step back to the initial form of the Bayes' estimator as mentioned in equation (20). A natural non-informative prior is the Jeffrey's prior, denoted by $\pi_{JP}(\lambda)$, which is

$$\pi_{JP}(\lambda) \propto (I(\lambda))^{1/2},$$

where $I(\lambda)$ = Fisher Information of $\lambda$ from a sample of size $n$. Hence, from (7), we have

$$\pi_{JP}(\lambda) \propto \sum_{m=0}^{\infty} \lambda^{2m}/(2m+3)$$

Therefore, the Bayes' estimator under Jeffrey's prior using $\delta = 0$ in the weight function in (18) and, denoting $\hat{\lambda}_{BJP}$, is given by

$$\hat{\lambda}_{BJP} = \frac{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k+1} (\sum_{m=0}^{\infty} \lambda^{2m}/(2m+3))^{1/2} d\lambda}{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k} (\sum_{m=0}^{\infty} \lambda^{2m}/(2m+3))^{1/2} d\lambda}. \tag{23}$$

### 2.3.3. Bayes' estimator under an approximate Jeffrey's prior (BAJP)

Note that in either of (20) or (18) the Bayes' estimator involves an infinite series. For the ease of simplification and being able to study the performance of a suitable Bayes' estimator analytically, we propose a simplistic approximation of the Jeffrey's prior which is given by $\sum_{m=0}^{\infty} |\lambda|^m/(2m+3)^{1/2}$. Also, note that this infinite series is convergent and has a finite value. In fact, the above series converges to $\sqrt{2}\Phi(|\lambda|, 1/2, 3/2)/2$, where $\Phi(x, y, z)$ is called the confluent hypergeometric function of the first kind (Abramowitz and Stegun (1964)), which is a function of $x$ when $y, z$ are held constants. Due to this fact, we can use this approximation as a new prior distribution. We call this as the approximate Jeffrey's prior and is given by

$$\pi_{AJP}(\lambda) \propto \sum_{m=0}^{\infty} |\lambda|^m/(2m+3)^{1/2}. \tag{24}$$

Hence, the Bayes' estimator with respect to (24), denoted by $\hat{\lambda}_{BAJP}$, is

$$\hat{\lambda}_{BAJP} = \frac{\sum_{k=0}^{n} \sum_{m=0}^{\infty} D_k (2m+3)^{-1/2} (1+(-1)^{k+1})(m+k+2)^{-1}}{\sum_{k=0}^{n} \sum_{m=0}^{\infty} D_k (2m+3)^{-1/2} (1+(-1)^{k})(m+k+1)^{-1}}. \tag{25}$$

The derivation of (25) is available in Chatterjee (2022).

### 2.4. Sampling distributions of various point estimators

One has to be extremely careful about obtaining the $MLE$ by maximizing the log-likelihood function by differentiation which yields (11). However, the solution of this equation exhibit a tendency to go outside of the parameter space $[-1, +1]$, especially when true $\lambda$ is near the boundary values, with a high probability. Therefore, $\hat{\lambda}_{ML}$ needs to be truncated at $\pm 1$ which shows a high probability concentration (*i.e.*, high relative frequency in the simulation study) at the boundaries. This feature hasn't been discussed by other researchers earlier. As the sample size increases, this behavior of $\hat{\lambda}_{ML}$ diminishes considerably, especially for $n \geq 50$ (see Chatterjee (2022)).

On the other hand, the Bayesian estimators are always strictly within the parameter space compared to the traditional estimator MLE showing a bimodal trend while estimating $\lambda$ close to the center of the parameter space, (see Figure 2.1 in Chatterjee (2022)). Figure 1 illustrates the simulated sampling distributions of the 4 estimators described earlier for a sample of size $n = 25$ based on $10^4$ replications.

As a demonstration, we apply the FGMD model to the MDR dataset. This gives us an opportunity to estimate the pairwise association among the variates in our dataset.

**Table 1: Estimates of the FGMD association parameter in two MDR subregions**

| Pair of Elements | North | | | | South | | | |
|---|---|---|---|---|---|---|---|---|
| | $\hat{\lambda}_{ML}$ | $\hat{\lambda}_{BFP}$ | $\hat{\lambda}_{BJP}$ | $\hat{\lambda}_{BAJP}$ | $\hat{\lambda}_{ML}$ | $\hat{\lambda}_{BFP}$ | $\hat{\lambda}_{BJP}$ | $\hat{\lambda}_{BAJP}$ |
| *As vs Cl* | 0.085 | 0.053 | 0.064 | 0.108 | −0.982 | −0.621 | −0.674 | −0.81 |
| *As vs Eh* | −1 | −0.587 | −0.646 | −0.803 | −1 | −0.872 | −0.892 | −0.941 |
| *As vs pH* | 0.746 | 0.423 | 0.475 | 0.666 | 0.611 | 0.431 | 0.469 | 0.648 |

**Remark 4:** (a) With the application of FGMD($\lambda$) we were able to estimate the underlying association among the pairwise variables using the four estimators. According to the estimates in Table 1 there exists a strong negative association between $Eh$ and $As$ in the northern region. $\hat{\lambda}_{ML}$ estimates the strongest negative association among the variables, followed by $\hat{\lambda}_{BAJP}$, $\hat{\lambda}_{BJP}$ and $\hat{\lambda}_{BFP}$. The highly negative association between $As$ and $Eh$ in the northern subregion, which was partially captured by the Spearman's and Kendall's, is ratified by the estimates of the association parameter of FGMD (see the details of the standard estimated correlation measures in Table 7 within Section 5).

(b) In the instance of $As$ vs $pH$ it is crucial to note that in the northern sub-region, Spearman's rho and Kendall's Tau contradicted pearson's correlation coefficient which showed a strong linear association. This is in agreement with our FGMD model.

(c) In the southern subregion, the standard correlation coefficients estimate a considerable negative linear association in ($As$, $Eh$). Although there is visible evidence of association present in ($As$, $Eh$) and to some extent in ($As$, $Cl$) but labeling it as a linear association will be an over simplification and inaccurate. The estimates in Table 1 of the association parameter $\lambda$ shows a strong negative association in ($As$, $Eh$) and in ($As$, $Cl$). The MLE registers the strongest association among the variables ($As$, $Eh$) followed by $\hat{\lambda}_{BAJP}$, $\hat{\lambda}_{BJP}$

and $\hat{\lambda}_{BFP}$. The same holds for $(As, Cl)$ as well. There is a positive association among the variables $(As, pH)$ as estimated by all the standard correlation measures, reiterating the same phenomenon by FGMD.

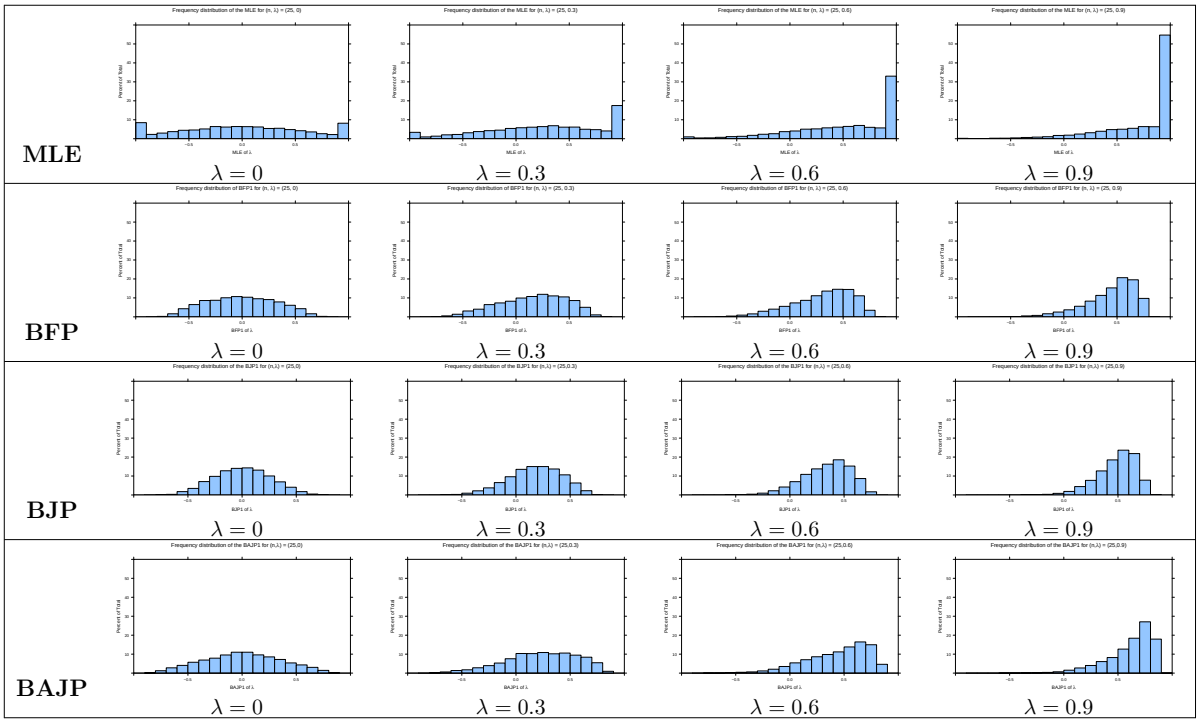



**Figure 1: Simulated relative frequency histograms of four estimators of $\lambda$, $n = 25$**

## 3. Hypothesis testing on the association parameter

### 3.1. The rationale behind hypothesis testing

In most of the applied cases one would be interested in knowing whether the components are associated or not. The copula based distributions such as FGMD which preserves the information of association through a single parameter $\lambda$ (in the bivariate case), if proven to be suitable, can provide an answer to this problem. In this Section, we study the performance of different types of hypothesis testing procedures to test the hypotheses $H_0 : \lambda = \lambda_0$ vs $H_A : \lambda \neq \lambda_0$. Hence to examine whether the association indeed exists or not, one particular value of $\lambda_0$ is of interest, that is $\lambda_0 = 0$. The following tests have been proposed and studied through size and power for the aforementioned hypotheses.

1. **Asymptotic tests**:

   (a) Asymptotic Normal Test. $(ANT)$

   (b) Asymptotic Likelihood Ratio Test $(ALRT)$

2. **Parametric bootstrap tests based on the LRT statistic** $(PBLRT)$

### 3.2. Asymptotic normal test (ANT)

While testing the values of $\lambda$, this is probably the simplest approach of developing a hypothesis test utilizing the asymptotic property of the $MLE$ of $\lambda$. Earlier we have already seen that for an iid sample of size $n$ from FGMD($\lambda$) the MLE exists and it is unique. It is a well known result that as $n \to \infty$, $\hat{\lambda}_{ML} \xrightarrow{d} N(\lambda,\ AV(\lambda))$, where $AV(\lambda)$ is the asymptotic variance of the $MLE$ and is given by the inverse of the fisher information of $\lambda$, *i.e.* $I^{-1}(\lambda)$, assuming that marginals are fully known. Therefore, if we assume that the null hypothesis is true, then $\hat{\lambda}_{ML} \xrightarrow{d} N(\lambda_0, I^{-1}(\lambda_0))$ as $n \to \infty$. Therefore we reject the null hypothesis if

$$|\sqrt{nI^0(\lambda_0)}(\hat{\lambda}_{ML} - \lambda_0)| > z_{(1-\alpha/2)},$$

where $z_{(1-\alpha/2)}$ is the right tail $(\alpha/2)$ - probability cutoff point of the standard normal distribution and $I^0(\lambda_0)$ is the $FIPO$ in the relation $I(\lambda_0) = nI^0(\lambda_0)$.

### 3.3. Asymptotic likelihood ratio test ($ALRT$)

Based on the iid observations derive the likelihood ratio statistic $\Lambda$ as

$$\Lambda = \frac{\underset{H_0}{Sup}\, L(\lambda|data)}{\underset{H_0 U H_A}{Sup}\ L(\lambda|data)} = \frac{L(\lambda_0|data)}{L(\hat{\lambda}_{ML}|data)}.$$

Define $\Lambda_* = -2ln(\Lambda)$. Asymtotically, as $n \to \infty$, $\Lambda_* \xrightarrow{d} \chi^2_1$ under $H_0$. So we reject the null hypothesis at level $\alpha$ if

$$\Lambda_* > \chi^2_{1;(1-\alpha)},$$

where $\chi^2_{1;(1-\alpha)}$ is the right tail $(\alpha)$ - probability cut off point of Chi squared distribution with 1 degree of freedom. The following Table 2 shows the simulated size values of the two asymptotic tests based on the $MLE$ $\hat{\lambda}_{ML}$.

**Table 2: Simulated size values of the two asymptotic tests for $\lambda_0 = 0, \alpha = 0.05$**

| Test | $n = 10$ | $n = 20$ | $n = 30$ | $n = 40$ | $n = 50$ | $n = 75$ | $n = 100$ |
|---|---|---|---|---|---|---|---|
| $ANT$ | 0.000 | 0.297 | 0.284 | 0.274 | 0.270 | 0.265 | 0.262 |
| $ALRT$ | 0.002 | 0.017 | 0.026 | 0.028 | 0.028 | 0.027 | 0.026 |

**Remark 5:** Both the asymptotic tests are far from satisfactory as far as size is concerned. For $n = 10$, both of them are hopelessly conservative. For $n \geq 20$, $ANT$ is overall a very liberal test and $ALRT$ on the other hand is a very conservative test. It is clearly visible that $ANT$ has a monotonically decreasing (albeit very slowly) size property with the increase in sample size, whereas $ALRT$'s size values indicate a conservative behavior. Even for sample of size 100, which are generally considered to be 'large', these tests are still unable to achieve the level condition satisfactorily.

### 3.4. Parametric bootstrap ($PB$) tests

As seen in the earlier section, the asymptotic tests do not perform well for small to moderately large sample sizes. Therefore, in this subsection we propose a class of four tests based on the idea of $LRT$ with the added parametric bootstrap ($PB$) concept.

The traditional $LRT$ calls for using $\Lambda_* = -2\ ln(\Lambda)$ which, under $H_0$, follows $\chi^2_1$ asymptotically. However, in this present $FGMD$ case, the null distribution of $\Lambda_*$ has been found to be way off from the asymptotic distribution $\chi^2_1$. Therefore, the cut-off point $\chi^2_{1;(1-\alpha)}$ is not applicable for the statistic $\Lambda_*$ in order to test $H_0$. A situation like this calls for coming up with different cut-off points for $\Lambda_*$ depending on sample size $n$ as well as the data $\mathbf{X}$ through a $PB$ method. Note that the expression $\Lambda$ has $\hat{\lambda}_{ML}$ in the denominator as an estimator of $\lambda$ while the numerator uses the null value $\lambda_0$ of $\lambda$. As a result, the value of $\Lambda$ is always between 0 and 1, and a value of $\Lambda$ closer to 1 implies a probable validity of $H_0$.

We extend the above traditional $LRT$ concept a bit further by incorporating the other three estimates of $\lambda$ which have shown considerable improvement over $\hat{\lambda}_{ML}$, especially in the mid region of the parameter space. In this regard we are going to consider $\hat{\lambda}_{BFP}$, $\hat{\lambda}_{BJP}$ and $\hat{\lambda}_{BAJP}$ (along with $\hat{\lambda}_{ML}$) in the $LRT$ structure. In its generic form, the structure of $\Lambda_*$ is going to be redefined as $\Lambda_*(\hat{\lambda}) = -2\ ln(\Lambda(\hat{\lambda}))$, where $\Lambda(\hat{\lambda}) = [\ L(\lambda_0|data)/L(\hat{\lambda}|data)\ ]$, where $\hat{\lambda}$ can be any one of the four aforementioned estimators of $\lambda$.

One difficulty with the above $\Lambda(\hat{\lambda})$ is that the denominator is not guaranteed to be greater or equal to the numerator unless $\hat{\lambda} = \hat{\lambda}_{ML}$. In other words, $\Lambda_*(\hat{\lambda}) = -2\ ln\Lambda(\hat{\lambda})$ is not guaranteed to be non-negative unless $\hat{\lambda} = \hat{\lambda}_{ML}$. However, a value of $\Lambda_*(\hat{\lambda})$ closer to 0 still conforms the validity of $H_0$. Therefore, to find suitable cut-off points for the statistic $\Lambda_*$, we consider

$$\Lambda_{**}(\hat{\lambda}) = |\Lambda_*(\hat{\lambda})|, \tag{26}$$

which is always nonnegative. The four versions of $\Lambda_{**}$ using four aforementioned estimators will be referred to as

$$\begin{aligned}
\Lambda_{**1}\ (or PBLRT\ 1) &= \Lambda_{**}(\hat{\lambda}_{ML}) \\
\Lambda_{**2}\ (or PBLRT\ 2) &= \Lambda_{**}(\hat{\lambda}_{BFP}) \\
\Lambda_{**3}\ (or PBLRT\ 3) &= \Lambda_{**}(\hat{\lambda}_{BJP}) \\
\Lambda_{**4}\ (or PBLRT\ 4) &= \Lambda_{**}(\hat{\lambda}_{BAJP})
\end{aligned} \tag{27}$$

**Algorithmic steps to implement $\Lambda_{**}(\hat{\lambda})$ as a test:**

**Step - 1**: For the given data $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$ from $FGMD$, compute $\hat{\lambda}$ (which is one of the above 4 estimators as mentioned earlier). Obtain the corresponding $\Lambda_{**}(\hat{\lambda})$.

**Step - 2**: Assume that $H_0 : \lambda = \lambda_0$ is true. Generate a bootstrap sample of size $n$ (say, $\mathbf{X}^*_1, \mathbf{X}^*_2, ..., \mathbf{X}^*_n$) from $FGMD(\lambda_0)$. Once this bootstrap data is generated, pretend that $\lambda$ is unknown, estimate $\lambda$ using the bootstrap data by $\hat{\lambda}$, and call it $\hat{\lambda}^*$, which in turn produces the value of $\Lambda_{**}(\hat{\lambda}^*)$. (See Chatterjee (2022) about generating data from FGMD($\lambda$).)

**Step - 3**: Repeat the above Step - 2 a large number of times (say, B times). This produces B copies of $\Lambda_{**}(\hat{\lambda}^*)$, and call them as $\Lambda^{(b)}_{**}(\hat{\lambda}) = \Lambda_{**}(\hat{\lambda}^{*(b)})$, $1 \le b \le B$, where $\hat{\lambda}^{*(b)}$ is the $b^{th}$

copy of $\hat{\lambda}^*$ as mentioned in Step - 2. These $\Lambda_{**}^{(b)}(\hat{\lambda})$ values are supposed to approximate the null distribution of $\Lambda_{**}(\hat{\lambda})$.

**Step - 4**: Order $\Lambda_{**}^{(b)}(\hat{\lambda})$, $1 \leq b \leq B$, and let $\Lambda_{**}(\hat{\lambda}|\alpha)$ be the $100(1-\alpha)^{th}$ percentile value of $\Lambda_{**}^{(b)}(\hat{\lambda})$, $1 \leq b \leq B$. This $\Lambda_{**}(\hat{\lambda}|\alpha)$ is the critical value for $\Lambda_{**}(\hat{\lambda})$ in Step - 1.

**Step - 5**: Reject $H_0$ if $\Lambda_{**}(\hat{\lambda})$ (from Step -1) $> \Lambda_{**}(\hat{\lambda}|\alpha)$, and retain $H_0$ if otherwise.

A complete comparison of the four $PB$ test in terms of size and power, for different sample size is detailed in Chatterjee (2022)

### 3.5. Application to MDR dataset

Using the estimates of $\lambda$ from the MDR dataset in Table 1, we proceed to perform $PBLRT1$, $PBLRT2$, $PBLRT3$ and $PBLRT4$ to test the hypothesis $H_0 : \lambda = 0$ vs $H_A : \lambda \neq 0$. The following Table 3 gives the $PBLRT$ test statistic values along with the simulated $P$-values.

**Table 3: $PBLRT$ test statistic and their $p$-values for the MDR data**

| Test | North | | | South | | |
|---|---|---|---|---|---|---|
| | *As vs Cl* | *As vs Eh* | *As vs pH* | *As vs Cl* | *As vs Eh* | *As vs pH* |
| $PBLRT1$ | 0.017 | 3.959* | 1.366 | 4.447* | 17.099* | 1.364 |
| | (0.899) | (0.047**) | (0.263) | (0.039**) | (0.000***) | (0.251) |
| $PBLRT2$ | 0.013 | 2.792 | 1.108 | 3.841 | 15.266* | 1.262 |
| | (0.208) | (0.101*) | (0.485) | (0.076*) | (0.002***) | (0.503) |
| $PBLRT3$ | 0.015 | 2.986 | 1.186 | 4.006 | 15.577* | 1.305 |
| | (0.216) | (0.101*) | (0.491) | (0.076*) | (0.000***) | (0.506) |
| $PBLRT4$ | 0.016 | 3.475 | 1.350 | 4.309 | 16.280* | 1.353 |
| | (0.239) | (0.098*) | (0.503) | (0.076*) | (0.000***) | (0.495) |

**Remark 6:** The results of Table 3 show that $As$ is not associated with $pH$ in both the regions based on the FGMD model. However, $As$ is significantly associated with both $Cl$ and $Eh$ in the southern subregion, and with $Eh$ in the northern subregion thereby opening up the possibility of further prediction (see more in Section 5).

## 4. Goodness of fit tests for FGMD

### 4.1. The rationale behind goodness of fit (GoF) tests

Since the pathbreaking work of Sklar (1959) about three dozen copulas have been proposed by various researchers for different applications. A particular copula presents a particular family of multivariate distributions of the random vector $\mathbf{X}$ which combines $p$ suitably hypothesized univariate marginal distributions of the components. Therefore, before adopting a particular copula for a specific dataset one must come up with a suitable GoF test for that copula, and this is where there appears to be an ample room for further research.

The problem of finding an optimal $GoF$ for a given copula is an open problem. It appears that there does not exist a robust test which can identify the most appropriate

copula for a given dataset. Therefore, one can take each copula, from a handful of copulas and see their applicability by running a GoF for a given dataset. Most of the available GoF tests in the literature are developed with either a specific copula or a specific family of copula based probability distributions in mind. A brief review of the available GoF tests and their inadequacy is discussed in the following section.

### 4.2. Inadequacy in the existing literature

There are several works on GoF tests involving copulas, such as Fermanian (2005), Genest *et al.* (2006), Genest and Favre (2007), Genest *et al.* (2009) (which is primarily a review of the existing methods with a limited power study), Genest *et al.* (2011) (which is a goodness of fit test for the bivariate extreme value copulas). However Genest *et al.* (2006) appears to encompass the overall GoF test methods for copulas.

Genest *et al.* (2006) provided two test statistics that have been developed to test the GoF of a given copula. These two test statistics, say $S_n$ and $T_n$, which are essentially Cramer-Von Mises and Kolmogorov-Smirnov statistics respectively, can be computed for FGMD through the following steps.

(i) Given the bivariate data $X_{ik}$, $i = 1, 2, ..., n$ and $k = 1, 2$, define the pseudo observations $V_1, V_2, ..., V_n$ as $V_i = (1/n)\sum_{l=1}^{n} I(X_{l1} \leq X_{i1}, X_{l2} \leq X_{i2})$, $1 \leq i \leq n$.

(ii) Define $K_n(t) = (1/n)\sum_{i=1}^{n} I(V_i \leq t) =$ (Number of $V_i$'s $\leq t)/n$.

(iii) Define $K(t|\lambda)$ as

$$K(t|\lambda) = \int_0^t \int_s^1 h(x, s|\lambda)dxds,$$

where

$$h(x, s|\lambda) = \frac{1}{(1-x)r(x, s|\lambda)} + \frac{1}{x} - \frac{1}{(1-x)},$$

with

$$r(x, s|\lambda) = [\{1 - \lambda(1-x)\}^2 + 4\lambda(1-x)(1-s/x)]^{1/2}.$$

Note that while implementation of the GoF tests, $\lambda$ in the above expression is to be replaced by a suitable estimate $\hat{\lambda}$.

(iv) Both $K_n(t)$ and $K(t|\hat{\lambda})$ are to be evaluated at $(j/n)$ as well as $((j+i)/n)$ with $j = 0, 1, 2, ..., (n-1)$ and $i = 0, 1$, such that the test statistics $S_n$ and $T_n$ have the desired expressions as follow (Genest *et al.* (2006))

$$S_n = \frac{n}{3} + n\sum_{j=1}^{n-1} K_n^2(\frac{j}{n})\{K(\frac{j+1}{n}|\hat{\lambda}) - K(\frac{j}{n}|\hat{\lambda})\}$$
$$-n\sum_{j=1}^{n-1} K_n(\frac{j}{n})\{K^2(\frac{j+1}{n}|\hat{\lambda}) - K^2(\frac{j}{n}|\hat{\lambda})\}$$

$$T_n = \sqrt{n}\max_{i=0,1;0\leq j\leq n-1}\{|K_n(\frac{j}{n}) - K(\frac{j+i}{n}|\hat{\lambda})|\}.$$

**Remark 7:** The above tests are shown to be applicable on samples coming from a few chosen copulas but the applicability of these tests has not been demonstrated when the data follow FGMD. In our work we have tried to implement the tests proposed by Genest *et al.* (2006) for a sample coming from FGMD. Following are the observations made while implementing these tests and why they do not work.

(a) In order to apply the above GoF tests, $K(t|\lambda = \hat{\lambda})$ needs to be evaluated at $t = 1$ which is encountered when $j = (n - 1)$. Note that we have used the extension of the Gaussian quadrature in two dimension to evaluate the definite integral. This numerical integration can be implemented by the 'quad2d' function in the 'pracma' package in R software.

(b) The function $K(t|\hat{\lambda})$ does not take a finite value, meaning the double integration is not convergent at the boundary for a given value of $\hat{\lambda}$. This issue is particularly evident at $t = 1$, but it can also occur at other values of $t$ depending on the value of $\hat{\lambda}$. This phenomenon should be taken into account when implementing the GoF tests.

(c) The double integration for $K(t = 1|\hat{\lambda})$ yields an "NaN" error in R since the integration fails to converge to a finite value. We provided a plot of this phenomenon in the following Figure 2 where $K(t|\lambda)$ has been plotted over $t$, $0 \leq t \leq 1$, for five different values of $\lambda$, $\lambda = -1, -0.5, 0, 0.5, 1$. It is evident from Figure 2 that the double integration fails to converge at the boundary value of $t = 1$ for all $\lambda$ values. Also, note that for $\lambda = 0.5$ and 1 the integration fails to converge not only at $t = 1$ but also at other values of $t$ between (0,1), thereby making the test statistic $S_n$ or $T_n$ questionable.

(d) The behavior of the function $K(t|\hat{\lambda})$ can be studied for any arbitrary $\hat{\lambda} \in [-1, 1]$. For example, at $\hat{\lambda} = 0$, this function fails to converge to a finite value. The function $K(t|\hat{\lambda})$ fails to attain a finite value when the upper limit of the definite integral is 1, and this is evident from Figure 3, where the function $K(t = 1|\hat{\lambda})$ has been redefined as $K(1 - 10^{-L}|\hat{\lambda})$ has been plotted against L such that the value of $t = 1$ is dependent on L through the relation $t = 1 - 10^{-L}$, *i.e.* as $L \to \infty$, $t \to 1$. This plot gives us an idea on how close to 1 we can achieve a finite value for the integration which defines the function $K(t|\hat{\lambda})$. The integration yields finite value approximately upto $L = 13$ *i.e.* the double integration would converge only upto $t = 1 - 10^{-13}$, not beyond that.

(e) Interestingly, $K(t|\hat{\lambda})$ is a distribution function of the iid pseudo observations $V_i$'s, and by definition it is supposed to demonstrate the non-decreasing property. However, going by the definition of Genest *et al.* (2006) as applied for the bivariate FGMD, $K(t|\hat{\lambda})$ as a function of $t$ for any given $\lambda$ fails to show the non-decreasing property as seen in Figure 2.

(f) The two test statistics $S_n$ and $T_n$ seem to work well for some non-FGMDs with large sample sizes of 100, 250 or 1000 in case of simulated data, or sample sizes of 1500 and 655 in case of application to real data Genest *et al.* (2006). It is crucial to note that FGMD, although briefly mentioned in Section 3.4 of Genest *et al.* (2006), it has neither been applied to any simulation exercise nor in the real data example. Hence, a GoF test for samples from FGMD became imperative, which has been developed and discussed in the following subsections.

(g) Finally, there is no evidence that the above mentioned tests based on $S_n$ and $T_n$ would work well in case of small and moderate sample sizes for any copula in terms of size as well as power. Table 5 and Table 6 of Genest *et al.* (2006) presents the size and power values only for large samples whereas in this study of ours, the datasets have sample sizes of 23 and 44.

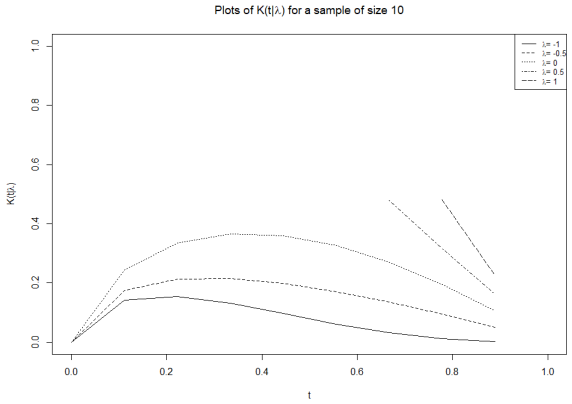



**Figure 2: Plot of $\mathbf{K}(t|\lambda)$ for different values of $\lambda$.**

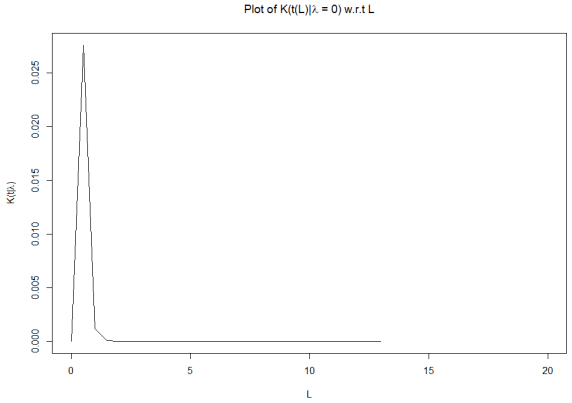



**Figure 3: Plot of $\mathbf{K}(1-10^{-L}|\lambda)$ as a function of L for $\lambda = 0$.**

### 4.3. A Bootstrap approach to GoF test for $FGMD$

#### 4.3.1. Developing the test statistic (bivariate case)

Suppose $\mathbf{X} = (X_1, X_2)'$ follows a bivariate distribution with *pdf* $f(\mathbf{x})$ and the corresponding *cdf* $F(\mathbf{x})$. How do we test that $f(\mathbf{x})$ is the $FGMD$ *pdf* given earlier?

For convenience let us denote the marginal *pdf* and *cdf* of $X_k$ by $f_k(.)$ and $F_k(.)$, respectively, $k = 1, 2$. If $\mathbf{X}$ follows $FGMD$, then the above joint *pdf* $f(\mathbf{x})$ and the corresponding bivariate *cdf* will be denoted by say $F_{FGMD}(\mathbf{x}|\lambda)$ as well. Our objective is to test

$$H_0 : F = F_{FGMD}(\mathbf{x}|\lambda), \text{for some } \lambda \quad vs \quad H_A : F \neq F_{FGMD}(\mathbf{x}|\lambda), \text{for any } \lambda, \tag{28}$$

where $\lambda \in [-1, +1]$.

Note that if we use the transformed data $\mathbf{Y}_i$, $1 \leq i \leq n$, where $\mathbf{Y}_i = (Y_{1i}, Y_{2i})'$, $i = 1, 2, ..., n$ and $Y_{ki} = F_k(x_{ki})$, $k = 1, 2$, $i = 1, 2, .., n$, then $\mathbf{Y}_i$'s are $n$ copies of the bivariate random vector $\mathbf{Y} = (Y_1, Y_2)'$ where marginally $Y_1$ and $Y_2$ are uniformly distributed over (0,1), and have the joint pdf, say $g(\mathbf{y})$ with the corresponding cdf, say $G(\mathbf{y})$ over the unite square $(0, 1) \otimes (0, 1)$.

If $\mathbf{X}$ has the specified distribution in (5), then equivalently $\mathbf{Y}$ has the distribution with *pdf*, say $g_{FGMD}(\mathbf{y}|\lambda)$, where

$$g_{FGMD}(\mathbf{y}|\lambda) = (1 + \lambda(2y_1 - 1)(2y_2 - 1)) \tag{29}$$

over the unit square $(0, 1) \otimes (0, 1)$. Testing (28) then boils down to testing

$$\overline{H}_0 : G = G_{FGMD}(\mathbf{y}|\lambda) \text{ for some } \lambda \;\; vs \;\; \overline{H}_A : G \neq G_{FGMD}(\mathbf{y}|\lambda) \text{ for any } \lambda, \tag{30}$$

based on the data $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n)$, where $G_{FGMD}$ is the *cdf* corresponding to the pdf $g_{FGMD}$ given in (29).

**Remark 8:** The broad idea of our testing mechanism will rely on finding a suitable distance between $G(\mathbf{Y})$ and $G_{FGMD}(\mathbf{Y}|\lambda)$. But since $F_i$'s $(i = 1, 2)$ are unknown, we are going to replace them by the corresponding marginal empirical cdfs, *i.e.*, we are going to work with

$$\begin{aligned} \hat{Y}_{ij} &= \hat{F}_{ij}(X_{ij}),\ i = 1, 2 \ and \ j = 1, 2, .., n. \\ &= (1/n)((Number\ of\ X_{ik}\ values \leq X_{ij})),\ 1 \leq k \leq n. \end{aligned} \tag{31}$$

Theoretically, $\mathbf{Y}$ is supposed to have a joint distribution with approximate *pdf* $g(\mathbf{y})$ and approximate *cdf* $G(\mathbf{y})$ whose marginals are uniform. The joint *cdf* $G(\mathbf{y})$ can be approximated by the observed empirical *cdf* $\hat{G}(\mathbf{y})$ as

$$\hat{G}(\mathbf{y}) = (1/n)\{Number\ of\ (\hat{Y}_{1s}, \hat{Y}_{2t})\ values\ \ni\ \hat{Y}_{1s} \leq y_1\ and\ \hat{Y}_{2t} \leq y_2\}, \tag{32}$$

$1 \leq s, t \leq n$. Notice that $\hat{G}_(\mathbf{y})$ is a bivariate step function which can be visualized on the grid points $(\hat{Y}_{1s}, \hat{Y}_{2t})$, $1 \leq s, t \leq n$.

At the same time, if we assume that $\mathbf{X} \sim f_{FGMD}(\mathbf{x}|\lambda)$ (*i.e.*, equivalently, $\mathbf{Y} \sim g_{FGMD}(\mathbf{y}|\lambda)$), then the *cdf* of $\mathbf{Y}$ under $\overline{H}_0$ can be approximated by

$$G_{FGMD}(\mathbf{y}|\hat{\lambda}) = \int_0^{y_1} \int_0^{y_2} g_{FGMD}(\mathbf{u}|\hat{\lambda})\ du_2\ du_1, \tag{33}$$

where $\hat{\lambda}$ is the estimated value of $\lambda$, and $g_{FGMD}(.|\hat{\lambda})$ expression is given in (29). It is easy to see that

$$G_{FGMD}(\mathbf{y}|\lambda) = y_1 y_2\{1 + \hat{\lambda}(y_1 - 1)(y_2 - 1)\}, \tag{34}$$

where $\mathbf{y} \in (0, 1) \otimes (0, 1)$. Similar to Kolmogorov - Smirnov test statistic, the distance between $\hat{G}(\mathbf{y})$ in (32) and $G_{FGMD}(\mathbf{y}|\hat{\lambda})$ can be measured by the statistic $\Delta(\mathbf{Y})$ for the given (transformed) data $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n)$; as

$$\Delta(\mathbf{Y}|\hat{\lambda}) = \underset{\mathbf{y} \in (0,1) \otimes (0,1)}{Sup} |\hat{G}(\mathbf{y}) - G_{FGMD}(\mathbf{y}|\hat{\lambda})|. \tag{35}$$

Numerical computation of $\Delta(\mathbf{Y}|\hat{\lambda})$ can be done easily by taking a very fine mesh over the unit square $(0,1) \otimes (0,1)$. Intuitively, one should reject $\overline{H}_0$ if $\Delta(\mathbf{Y}|\hat{\lambda})$ is too "large", and retain $\overline{H}_0$ otherwise. In the following we present a bootstrap method to find a data dependent cut-off value and the $p$-value.

### 4.3.2. Leveraging $\Delta(\mathbf{Y}|\hat{\lambda})$ to draw an inference via bootstrap

**Step - 1**: For the given data $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n)$, compute $\Delta(\mathbf{Y}|\hat{\lambda})$, where $\hat{\lambda}$ is the estimated value of $\lambda$ (from one of the four estimators as discussed in Section 2).
**Step - 2**: Assume that $\overline{H}_0$ holds. Using $\hat{\lambda}$ computed in Step - 1, generate a bootstrap sample $\mathbf{Y}_1^*, \mathbf{Y}_2^*, ..., \mathbf{Y}_n^*$ iid from $g_{FGMD}(\mathbf{y}|\hat{\lambda})$. [This is equivalent to generating $\mathbf{X}_1^*, \mathbf{X}_2^*, ..., \mathbf{X}_n^*$ iid from $f_{FGMD}(\mathbf{x}|\hat{\lambda})$, and then transforming them back to $\mathbf{Y}_j^*$'s.]
**Step - 3**: Using the bootstrap data $\mathbf{Y}^* = (\mathbf{Y}_1^*, \mathbf{Y}_2^*, ..., \mathbf{Y}_n^*)$, recalculate $\hat{\lambda}$ as in Step-1 (pretending that it were unknown). Call this estimate of $\lambda$ as $\hat{\lambda}^*$ (*i.e.*, $\hat{\lambda}$ based on $\mathbf{Y}^*$). Then compute $\Delta(\mathbf{Y}|\hat{\lambda})$ using this bootstrap data $\mathbf{Y}^*$, and $\hat{\lambda}^*$. Call this $\Delta^* = \Delta(\mathbf{Y}^*|\hat{\lambda}^*)$.
**Step - 4**: Repeat the above Step - 2 and Step - 3 a large number of times (say, B times). This yields B copies of $\Delta^*$, say $\Delta^{*b}$, $1 \leq b \leq B$, which are then ordered as $\Delta^{*(1)} \leq \Delta^{*(2)} \leq ... \leq \Delta^{*(B)}$.
**Step - 5**: Our $\alpha$-level bootstrap cut-off point is found as $\Delta_\alpha^B = \Delta^{*((1-\alpha)B)}$. If $\Delta$ (from Step - 1) $> \Delta_\alpha^B$, then reject $\overline{H}_0$, and retain it otherwise.

Alternatively, one can obtain the bootstrap $p$-value of the GoF test as

$$p_B = \{Number\ of\ \Delta^{*b}\ values\ >\ \Delta(\mathbf{Y}|\hat{\lambda})\}/B,$$

and compare it with $\alpha$. Essentially with the four proposed estimators of $\lambda$ we can have four different GoF tests which are named as follows - (i) $GoF_1$ when $\lambda = \hat{\lambda}_{ML}$, (ii) $GoF_2$ when $\lambda = \hat{\lambda}_{BFP}$, (iii) $GoF_3$ when $\lambda = \hat{\lambda}_{BJP}$ and (iv) $GoF_4$ when $\lambda = \hat{\lambda}_{BAJP}$.

### 4.4. Results of goodness of fit tests on MDR dataset

In the following table we present the bootstrap $p$-values of $FGMD$ goodness of fit test through the statistic $\Delta(\mathbf{Y}|\hat{\lambda})$ for the MDR data as discussed earlier. The results have been obtained for both north and south regions using all the four estimators of $\lambda$ (*i.e.*, $\hat{\lambda} = \hat{\lambda}_{ML}, \hat{\lambda}_{BFP}, \hat{\lambda}_{BJP}\ and\ \hat{\lambda}_{BAJP}$).

**Table 4: Goodness of fit $p$-values through bootstrap for testing $FGMD$**

| Variable Combination | North | | | | South | | | |
|---|---|---|---|---|---|---|---|---|
| | $MLE$ | $BFP$ | $BJP$ | $BAJP$ | $MLE$ | $BFP$ | $BJP$ | $BAJP$ |
| $As$ and $Cl$ | 0.499 | 0.498 | 0.496 | 0.496 | 0.647 | 0.661 | 0.665 | 0.660 |
| $As$ and $Eh$ | 0.571 | 0.558 | 0.557 | 0.562 | 0.838 | 0.826 | 0.829 | 0.837 |
| $As$ and $pH$ | 0.926 | 0.926 | 0.927 | 0.927 | 0.509 | 0.500 | 0.508 | 0.519 |

Table 4 clearly shows that bivariate FGMD is definitely an acceptable joint distribution to model $As$ along with each of the three other variables, *i.e.*, $Cl$, $Eh$ and $pH$.

Obviously a natural extension of this observation is that of studying the conditional distribution of $As$ given a benign element, and then making a suitable prediction of $As$. This aspect of prediction will be reported in a separate comprehensive work.

**Remark 9:** (a) In this section, we have made an attempt to address the elusive query of goodness of fit of a copula, specifically the FGMC. We have proposed a class of parametric bootstrap (PB) tests based on the Kolmogorov - Smirnov (KS) distance between the two $cdfs$, - the hypothesized FGMD and the empirical one (see (32)). To the best of our knowledge, this is first time that GoF of FGMD has been addressed in a comprehensive manner for small to moderate sample sizes.

(b) We have been able to show that our proposed parametric bootstrap tests not only adhere to the size criterion (see Chatterjee (2022)) but also there is no need to know the null distributions of the test statistics, either for a fixed sample or asymptotically. The performance of the tests in terms of power indicate that they are almost identical, and hence any one of the four can be used in applications.

(c) Though this section deals with GoF of a bivariate FGMD to model $As$ with another element, one should look at a possible extension in a multidimensional set up ($i.e.$, going beyond the dimension 2) so that $As$ can be modeled along with ($Cl$, $Eh$, $pH$) for a more meaningful analysis of the data. This is a future research problem which will be taken up later. Another potential avenue for further GoF study is to use a different distance measure (other than the KS one) and study the resultant implications.

## 5. Predictions under FGMD

### 5.1. The rationale behind prediction

Our in-depth analysis of the given data shows that each variate ($As$, $Cl$, $pH$ and $Eh$) individually has a vastly different probability distribution in each of the two subregions. The following Figure 4 shows the sample relative frequency histograms of the four variables in two subregions.

We have used two well-known and widely accepted formal test methods, namely - Anderson-Darling Test (ADT) and Shapiro - Wilk Test (SWT), to check if the sample histograms in Figure 4 conform to normality. Unfortunately six out of eight sample histograms rejected normality. Only the variate $pH$, and that too for the southern subregion, accepted normality (by both ADT and SWT) comfortably with very high $p$-values. In the north, $Eh$ appears to follow normality with moderately large $p$-values. Usually one should feel comfortable with the assumption of normality if both ADT and SWT show substantially large $p$-values. The following Table 5 shows the $p$-values for all the four variates in the two subregions when both the tests are applied.

Further rigorous investigation showed that not only the six out of eight subdatasets (four variables in two subregions) are non-normal, each variable's probability distributions in the two subregions are vastly different. In this regard we show the $p$-values of the well known Kolmogorov-Smirnov Test (KST) to test the equality of two distributions in the following Table 6.

**Table 5: *ADT* and *SWT* $p$-values to test the normality in two subregions**

| Test | Subregion | *As* | *Cl* | *Eh* | *pH* |
|---|---|---|---|---|---|
| North | ADT | $< 0.0001$ | 0.0414 | 0.1404 | 0.0029 |
| | SWT | $< 0.0001$ | 0.0286 | 0.0914 | 0.0012 |
| South | ADT | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ | 0.4363 |
| | SWT | 0.0002 | $< 0.0001$ | $< 0.0001$ | 0.6933 |

**Table 6: KST $p$-values to check equality of distributions in two subregions**

| | *As* | *Cl* | *Eh* | *pH* |
|---|---|---|---|---|
| $p$-value | $< 0.0001$ | 0.0004 | $< 0.0001$ | $< 0.0001$ |

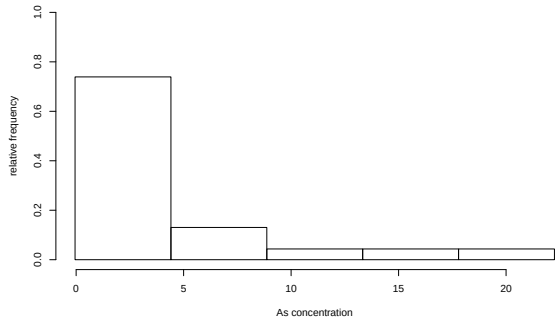
The above observations about the distributional properties of four variables now set the ground for bivariate scatter plots between *As* and each of the other three variables. Figure 5 comprehensively shows the six scatterplots in the two subregions.

Notice that out of six bivariate scatterplots, four do not show any linear trend (and these are (a), (b), (d) and (f)). Plots (c) and (e) are somewhat linear, but the variations (or dispersion) of *As* against *Eh* (in (c)), and *pH* (in (e)) do not look uniform (*i.e.*, the conditional probability distribution of *As* given another variable appears to be heteroscedastic). As a result, the standard Pearson's correlation coefficient $\rho_P$ is not going to be an adequate measure to assess the association between *As* and other variables in a bivariate framework.

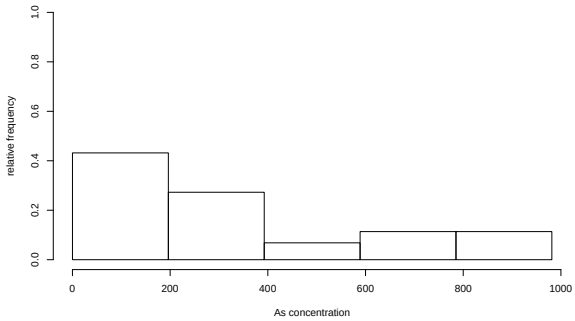
Yet, for the sake of argument, one can compute the three standard correlation estimates, including $\rho_P$, while the other two being the Spearman's rank correlation (denoted by $\rho_S$) and Kendall's 'Tau' (or, Kendall's rank correlation), denoted by $\rho_K$, to get an overall sense of these associations. While $\rho_P$ measures the strength of linear association, $\rho_S$ and $\rho_K$ are much more robust, and indicate the strength of monotonic association between the two variables of interest. The following Table 7 provides the three sample correlation measures for three pairs of variables in the two subregions. The value in parentheses under each entry is the $p$-value for testing the null hypothesis ($H_0$) which states that the true (or population) correlation measure is zero, against the alternative hypothesis ($H_A$), which negates the null. Note, in Table 7: '***' implies $p$ - value $\leq 0.01$; '**' for $\leq 0.05$; and '*' for $\leq 0.10$.

**Table 7: Estimated standard correlation coefficients with corresponding $p$-values**

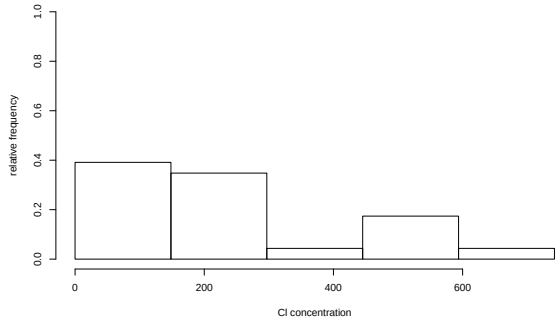
| | North ($n_N = 23$) | | | South ($n_S = 43$) | | |
|---|---|---|---|---|---|---|
| | *As vs Cl* | *As vs Eh* | *As vs pH* | *As vs Cl* | *As vs Eh* | *As vs pH* |
| $\hat{\rho}_P$ | 0.182 | 0.525 | 0.754 | $-0.325$ | $-0.668$ | 0.119 |
| | (0.405) | (0.010*) | (0.000***) | (0.031**) | (0.000***) | (0.442) |
| $\hat{\rho}_S$ | 0.018 | $-0.414$ | 0.260 | $-0.320$ | $-0.753$ | 0.156 |
| | (0.936) | (0.050**) | (0.231) | (0.035**) | (0.000***) | (0.314) |
| $\hat{\rho}_K$ | 0.012 | $-0.323$ | 0.188 | $-0.230$ | $-0.577$ | 0.101 |
| | (0.937) | (0.032**) | (0.213) | (0.028**) | (0.000***) | (0.336) |



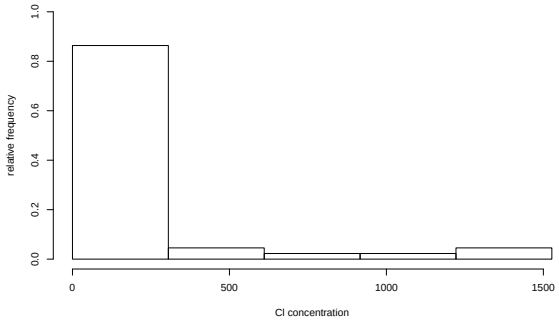

(a) *As* (North)



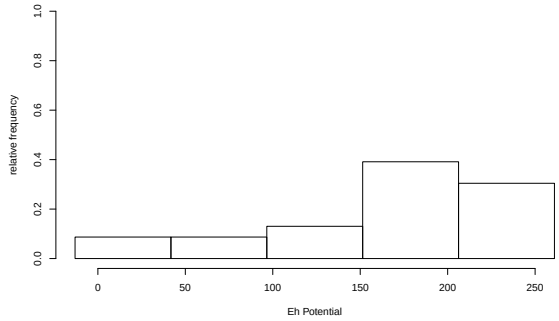

(b) *As* (South)



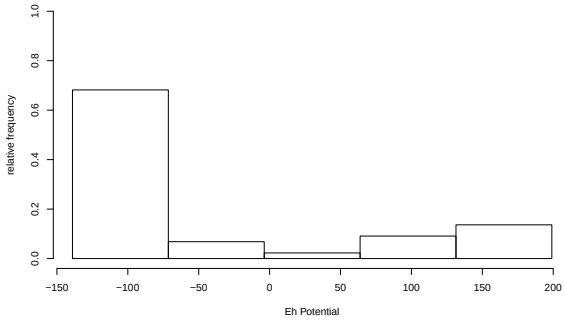

(c) *Cl* (North)



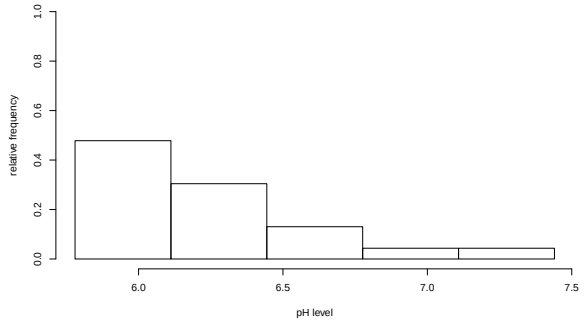

(d) *Cl* (South)



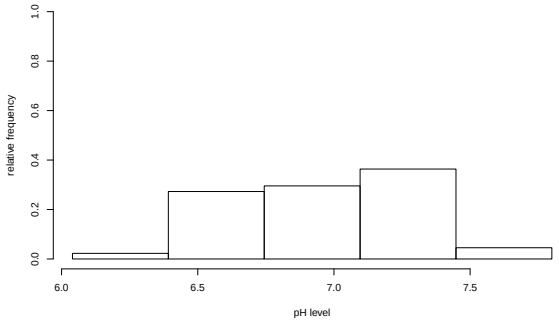

(e) *Eh* (North)




(f) *Eh* (South)

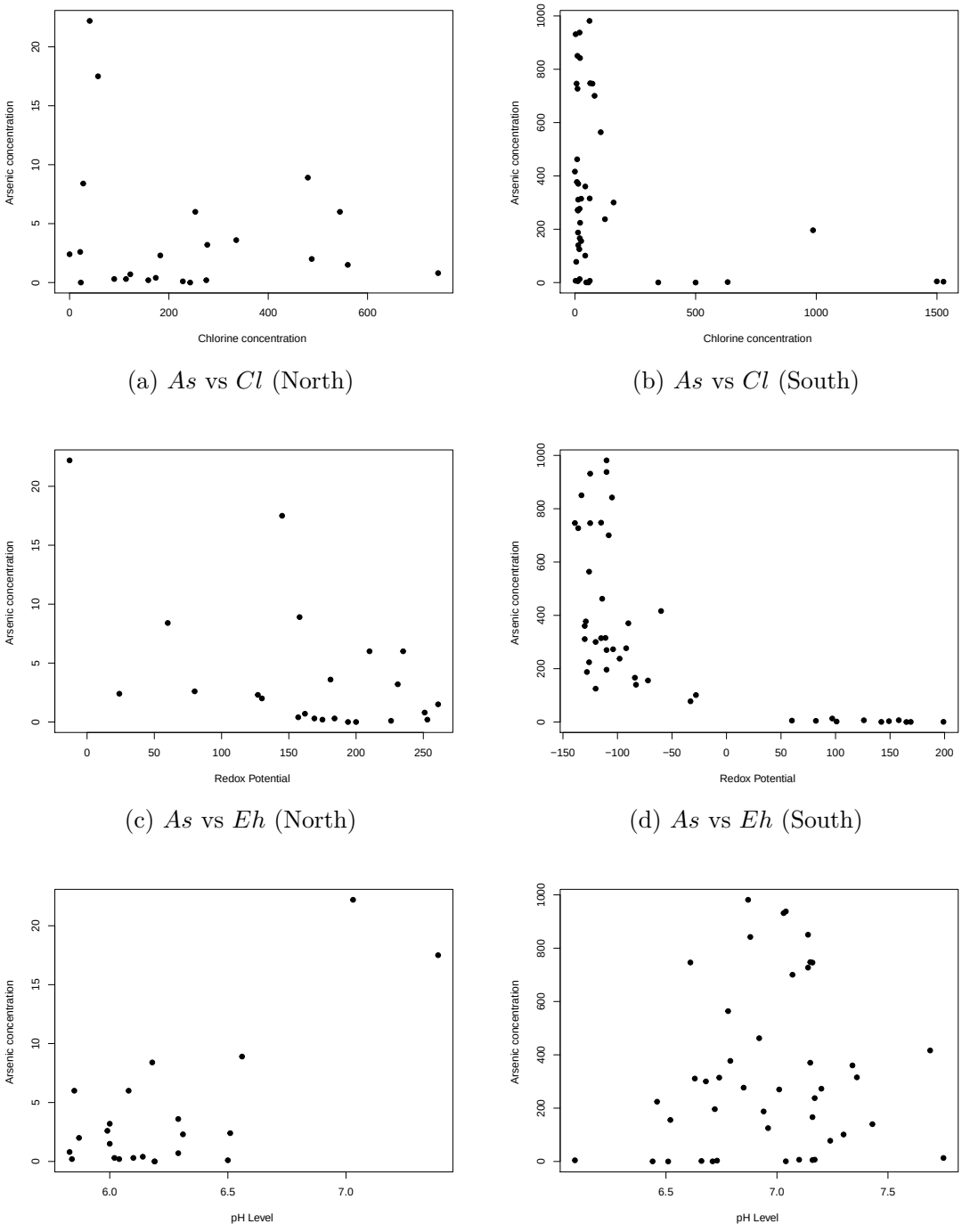



(g) *pH* (North)




(h) *pH* (South)

**Figure 4: Relative frequency histograms of the four variables in two subregions**




(a) $As$ vs $Cl$ (North)

(b) $As$ vs $Cl$ (South)

(c) $As$ vs $Eh$ (North)

(d) $As$ vs $Eh$ (South)

(e) $As$ vs $pH$ (North)

(f) $As$ vs $pH$ (South)

**Figure 5: Scatter plots of $As$ against each of the other three benign variables**

**Remark 10:** The three estimated standard correlation measures portray some interesting scenarios as summarized below.

(a) Between $As$ and $Cl$, in the North, all three correlations indicate that there is no association. However, in the South, they all indicate a significant negative association.

(b) Between $As$ and $Eh$, there appears to be a significant negative association in both the subregions mostly. Only a conflicting picture is provided by $\rho_P$ in the north which shows a significant positive linear association.

(c) Between $As$ and $pH$, there appears to have no significant association mostly as all but one $p$-values are quite high (more than 20%). However, only $\rho_P$ shows a strong linear positive association in the north.

The usual correlation measures show interesting associations between $As$ and $Cl$, $As$ and $pH$, and $As$ and $Eh$, but they do not help predict $As$ values based on these variables. For instance, in regions where $As$ and $Cl$ are strongly associated, predicting $As$ when $Cl$ is $10ppm$ or $Eh$ is $100mv$ requires more than just linear regression, which assumes normality and homoscedasticity. The problem is better addressed by fitting a suitable bivariate probability distribution to the data. Specifically, we need a non-normal bivariate distribution of $(X, Y)$, where $Y$ is $As$ and $X$ can be $Cl$, $pH$, or $Eh$, using copula theory. This approach allows us to use the conditional distribution of $Y$ given $X$ to make predictions. The goal is to explore the association between $As$ and the other variables beyond linear correlation and to exploit this association for prediction, recognizing that simplistic linear models might lead to incorrect conclusions.

### 5.2. Prediction of Y using a covariate under FGMD

The parameter $\lambda$, by its appearance, has some similarities with the three standard correlation measures discussed earlier. If $\lambda = 0$, then $X$ and $Y$ are independent; if $\lambda < 0$ $(> 0)$, then they are negatively (positively) associated.

Note that, marginally the pdfs $f_X$ and $f_Y$ of $X$ and $Y$ are unknown, and so are the cdfs $F_X$ and $F_Y$. It is possible to adopt a suitable parametric model for $f_X$ and $f_Y$, but that is not the main focus of our study. We want to bypass this aspect of unknown marginals by replacing $F_X$ (and $F_Y$) by $\hat{F}_X$ (and $\hat{F}_Y$) where $\hat{F}_X$ (and $\hat{F}_Y$) represents the empirical distribution function defined as (where the subscript has been dropped for generalization)

$$\hat{F}(t) = \{\text{number of observed data points} \leq t\}/n, \tag{36}$$

$n$ being the sample size. Therefore, we pretend that the marginals ($F_X$ and $F_Y$) are known, and they can be replaced by their estimates whenever needed.

The conditional distribution of $Y$ given $X$ is

$$f_{Y|X}(y|x) = f_Y(y)[1 + \lambda(2F_X(x) - 1)(2F_Y(y) - 1)], \tag{37}$$

and this will be used in predicting the value of $Y$ when $X = x$ is given. The main challenge here is the estimation of $\lambda$. Some noteworthy works along this line, *i.e.*, regressing $Y$ based

on $X$ from copula based predictive models is discussed as follows. We now proceed with the conditional distribution given in (37) to predict the value of $Y$ when $X = x$ is given. We have discussed the structures of the predictors of $Y$ based on the three standard measures of center of a conditional probability distribution as follows.

### 5.3. Conditional mean as a predictor

The first predictor is the conditional mean, denoted by $\hat{Y}_{mean}(x)$, given as

$$\hat{Y}_{mean}(x) = E(Y|X = x) = \int y f_{Y|X}(y|x)dy, \tag{38}$$

where the integration is over the appropriate range of $Y$ given $X = x$. In the current context where $Y$ represents the $As$ level, this range of $Y$ is $(0, \infty)$. It can be shown that (see Chatterjee (2022)) the expression (38) simplifies to

$$\hat{Y}_{mean}(x) = \mu_Y + \lambda G_X(x) I_Y(F_Y), \tag{39}$$

where $\mu_Y$ is the unconditional mean of $Y$, $G_X(x) = (2F_X(x) - 1)$ and

$$I_Y(F_Y) = \int_{-1}^{1} (t/2) F_Y^{-1}((1 + t)/2)dt. \tag{40}$$

In applications, $\mu_Y$ will be replaced by $\hat{\mu}_Y = \overline{Y}$ = sample mean of Y observations, and $G_X(x)$ will be replaced by $\hat{G}_X(x) = 2\hat{F}_X(x) - 1$ with $I_Y(F_Y)$ being replaced by $I_Y(\hat{F}_Y)$. Also, $\lambda$ will be replaced by a suitable estimator as mentioned earlier.

### 5.4. Conditional median as a predictor

Another simple predictor is the median calculated from the conditional distribution of $Y|X$. When the conditional distribution is skewed, which is expected in real life applications, the conditional median tends to be a robust predictor than the mean. We get the conditional median $\hat{Y}_{median}(x)$ of $Y|X$ by solving the following Equation (41) in terms of $M$, the desired median of the conditional distribution.

$$0.5 = \int_{-\infty}^{M} f_Y(y)(1 + \lambda(2F_X(x) - 1)(2F_Y(y) - 1)\partial y. \tag{41}$$

For brevity, we use the following notation $\lambda G_X(x) = A$ and $F_Y(M) = w$. It can be shown (see Chatterjee (2022)) that solving Equation (41) boils down to solving

$$2Aw^2 + 2(1 - A)w - 1 = 0. \tag{42}$$

The feasible solution from the above quadratic equation and inverting the cdf for $Y$ gives us our estimate of the $Y$ for a given $x$ based on the median of the conditional distribution. Hence the predictor is as follows

$$\hat{Y}_{median}(x) = F_Y^{-1}[\frac{\{\lambda(2F_X(x) - 1) - 1\} + \sqrt{1 + \lambda^2(2F_X(x) - 1)^2}}{2\lambda(2F_X(x) - 1)}]. \tag{43}$$

In applications, $\lambda$ will be replaced by $\hat{\lambda}$, the marginal $F_X(x)$ is replaced by $\hat{F}_X(x)$, and $F_Y^{-1}$ should be replaced by $\hat{F}_Y^{-1}$.

### 5.5. Conditional mode as a predictor

One can use a predictor of the third kind, *i.e.*, the conditional mode, which can be found simply by deriving the mode of the conditional distribution of $Y|X$, which is found by differentiating the conditional pdf and equating it with zero, *i.e.*,

$$(\partial/\partial y)(f_{Y|X}(y|x)) = 0, \tag{44}$$

assuming that the $f_{Y|X}$ is absolutely continuous. Using the usual notation of $f_Y(y) = f_Y$ and $\lambda(2F_X(x) - 1) = \lambda G_X(x) = A$, the equation (44) above then can be written as $f'_Y + A(2(f_Y)^2 + (2F_Y(y) - 1)f'_Y) = 0$. Substituting for $2F_Y(y) - 1 = u(y) = u$(say), we have $2f_Y = u'$ and $2f'_Y = u''$. The above equation (44) thus yields a second-order ordinary differential equation as $(u''/2) + 2A(u'/2) + Auu''/2 = 0$, *i.e.*, $(u')^2 + (B + u)u'' = 0$ where $B = 1/A$. Let $(B + u) = v$, *i.e.*, $u' = v'$, *i.e.*, $u'' = v''$. Then the above equation boils down to the differential equation $(v')^2 + vv'' = 0$, *i.e.*, $\partial(vv') = 0$, *i.e.*, $vv' = c$, for some constant $c$, *i.e.*, $v\partial v/ = c\partial y$, *i.e.*, $v^2/2 = cy + d$, *i.e.*,

$$(B + u)^2 = c_1 y + d_1, \tag{45}$$

where $c,\ d,\ c_1,\ d_1$ are suitable constants. The above final expression (45) gives the general solution of the differential equation (44). In order to find the values of the constants in the solution, specific boundary values were chosen, say $y = y^*$ and $y = y^{**}$, where $y^*$ and $y^{**}$ are two suitable small and large extreme values of the variable $y$ over its support. For the purpose of computational convenience we have taken $y^* = y_{(1)}$ and $y^{**} = y_{(n)}$, the smallest and the largest observed values of the variable $Y$ respectively. Then, $u(y^*) \approx -1$ and $u(y^{**}) \approx 1$ respectively. Plugging-in these choices of $y = y^*$ and $y = y^{**}$ as boundary values in our general solution (45), we get the following solution with $c_1 = c_1^* = 4B/y^{**}$ and $d_1 = d_1^* = (B - 1)^2$ as

$$4(\hat{F}_Y(y))^2 + 2(B - 1)(\hat{F}_Y(y)) = c_1^* y. \tag{46}$$

**Remark 11:** The solution of (46) in terms of $Y$ gives an approximate mode of the conditional distribution of $Y|X$. Further, note that this conditional mode depends on $X = x$, through the term $B = 1/A$, which involves $x$. Thus the solution of the Equation (46) will be the intersection of the plots of the left hand side (LHS) and the right hand side (RHS) of the said equation within the range of $Y$. But the plot of the LHS depends on the sign of $B$. The sign of $B$ in turn is dictated by the sign of $\lambda$ and the sign of the term $(2F_X(x) - 1)$. For example, in the southern subregion, the data on *As* and *Eh*, all the estimates of $\lambda$ are negative. This phenomenon is true for the estimate calculated for the entire data and all the estimates calculated by the "Leave-One-Out Bootstrap" (LOOB) computation (elaborated in Section 5.6). This means, the sign of $B$ is determined on the basis of three distinct scenarios eventually giving rise to three distinct cases: (*i*) $x >$ median $(X)$; (*ii*) $x <$ median $(X)$; (*iii*) $x \equiv$ median $(X)$.

Eventually, the mode predictor of the conditional distribution, denoted by $\hat{Y}_{mode}(x)$,

is defined in the following way

$$\hat{Y}_{mode}(x) = \begin{cases} \text{(possible) solution(s) of (46)} & \text{if } x \neq \text{median}(X) \\ \text{unconditional mode of } (Y) & \text{if } x \equiv \text{median}(X). \end{cases} \tag{47}$$

**Remark 12:** Regarding the above predictor expression in (47) note the following -

(a) The sign of $B$ (negative or positive) is determined by whether $x > \text{median}(X)$ or $x < \text{median}(X)$ from the definition of $B$ as noted earlier.

(b) For example, in As vs Eh in the southern subregion, we observe that $\text{median}(X) = -106.5\ mV$. When $x < \text{median}(X)$, $B$ is positive and we have noted the plots of both RHS and LHS are monotonically increasing. The opposite happens when $x > \text{median}(X)$. Representative plots for both the cases have been considered in Figure 6 and Figure 7 *i.e.*, *As* vs *Eh* data from the southern subregion in MDR for $x = -126\ mV$ and $x = 126\ mV$ which are less than and greater than the $median(X)$ respectively.

(c) Analytically, there is a possibility of having multiple solutions of (46) due to multiple intersections between LHS and RHS. The intuition behind finding the mode of the conditional density function is as follows. One can collect all the multiple solutions and check which one yields the maximum value of the conditional density given in (37). Under the assumed parameter free model of the marginals, the next step is to estimate the marginal density in the expression (37), *i.e.* how to get $f_Y(y)$. This can be achieved in a multiple ways but we have presented a simple and straightforward way in our *LOOB* calculations. Assume one obtains multiple solutions as $y_i^*$, $i = 1, 2, ..., k$ in (47) and $k < n$ within the range of $Y$, then the unconditional density function at $y_i^*$, by definition, is the rate of change of the cumulative distribution function at $y_i^*$. In light of this definition, one can approximate $f_Y(y_i^*)$ as $f_Y(y_i^*) \approx (1/nh) \sum_{i=1}^{n} K((y_i^* - y_i)/h)$, where $K(.)$ is a suitable kernel - a non-negative function, $y_i$, $i = 1, 2, ..., n$ are the sample observations and $h > 0$ is a smoothing parameter called the bandwidth.

(d) An in-built R-package has been used for the above density estimation which uses the Gaussian kernel function. While choosing the bandwidth in Kernel, which is still an open topic of research, the default choice of bandwidth rule selection in the R - package is by Silverman's 'rule of thumb' (Silverman (1986, page 48, eq.(3.31)). This choice is more appropriate if the original distribution (*i.e.* the true marginals) is bell-shaped and symmetric in nature. In contrast, none of our marginals are symmetric and bell-shaped. Therefore, in our case Sheather and Jones method (Sheather and Jones (1991)) is more applicable which is a more robust and data dependant approach. Moreover, in theory, a finer kernel bandwidth reveals more intricacies in the true distribution. But there is a risk of under-smoothing by choosing a too small '$h$'. On the other hand, a risk of over-smoothing exists if '$h$' is too large. We have examined several bandwidths under the Sheather and Jones rule and have chosen $h = 10^{10}$.

(e) Finally, when $x \equiv \text{median}(X)$, it is straightforward to note from the conditional distribution in (37) that the predictor would be the unconditional mode of Y, say $Y_{mode}$.

However, the exact sampling distributions of these three predictors are intractable theoretically. Therefore, the performance of these three predictors have been evaluated

through the 'Leave-One-Out-Bootstrap' ($LOOB$) method, which has been discussed and applied on the groundwater data from MDR in the following subsection. Extending on the existing idea and incorporating ten estimators of $\lambda$ the association parameter of FGMD (one traditional estimator - $MLE$, and nine Bayes' estimators - under each of the three types of priors - Flat, Jeffrey's and approximate Jeffrey's prior, three types of central tendency measure (expectation, median and mode) of the three resulting posterior distribution) and three predictors, we have achieved a collection of thirty predicted values of $Y$ for a given value of $X$.

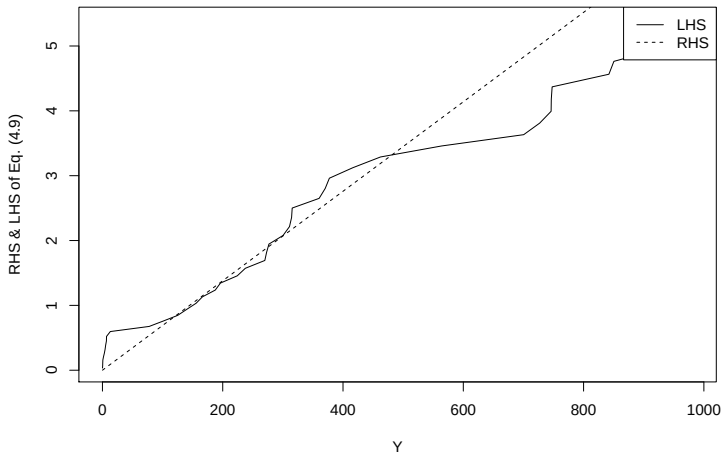



**Figure 6: Plots of LHS and RHS of equation (46) when $x = -126\ mV$ ($<$ median$(X)$), where $X = Eh$ in the southern subregion.**

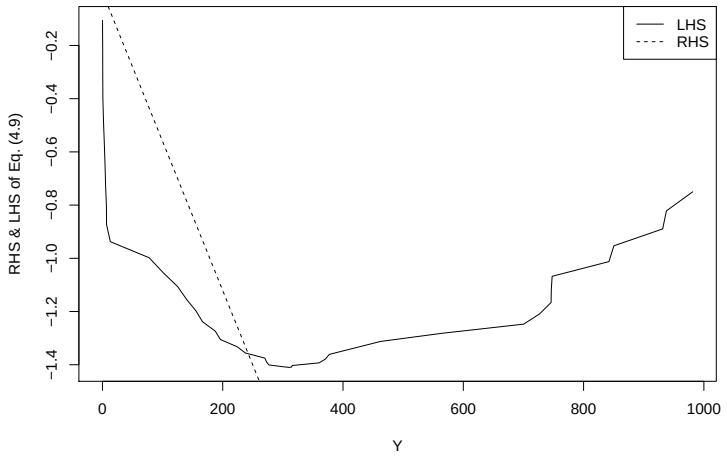



**Figure 7: Plots of LHS and RHS of equation (46) when $x = 126\ mV$ ($>$ median$(X)$), where $X = Eh$ in the southern subregion.**

### 5.6. Applications: Predicting arsenic from $Cl$

In this section we are going to demonstrate how to predict $As$ from the three covariates but focusing mostly on $Cl$ in the southern subregion just as an example. Before delving into the performances of the different predictors in the $As$ prediction study, let us look at the $FGMD$ association parameter estimates from the two subregions in Table (8). Note that, for a pair of variables that includes $As$, we can employ a total of ten estimates of $\lambda$. Selection of an estimator of $\lambda$ will depend on the overall $LOOB$ performance to be explained below.

**Table 8: Estimates of the FGMD association parameter in the MDR subregions**

(a) Northern subregions

| Pair of Elements | $MLE$ | Posterior | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $BFP$ | | | $BJP$ | | | $BAJP$ | | |
| | | Mean | Median | Mode | Mean | Median | Mode | Mean | Median | Mode |
| *As vs Cl* | 0.089 | 0.051 | 0.065 | 0.061 | 0.064 | 0.075 | 0.134 | 0.108 | 0.195 | 0.560 |
| *As vs Eh* | −1 | −0.585 | −0.665 | −0.254 | −0.646 | −0.745 | −0.341 | −0.803 | −0.905 | −0.114 |
| *As vs pH* | 0.749 | 0.418 | 0.475 | 0.371 | 0.475 | 0.555 | 0.722 | 0.666 | 0.815 | 0.311 |

(b) Southern subregion

| Pair of Elements | $MLE$ | Posterior | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $BFP$ | | | $BJP$ | | | $BAJP$ | | |
| | | Mean | Median | Mode | Mean | Median | Mode | Mean | Median | Mode |
| *As vs Cl* | −0.979 | −0.624 | −0.675 | −0.176 | −0.674 | −0.745 | −0.242 | −0.810 | −0.895 | −0.083 |
| *As vs Eh* | −1 | −0.869 | −0.905 | −0.001 | −0.892 | −0.925 | −0.001 | −0.941 | −0.975 | −0.001 |
| *As vs pH* | 0.593 | 0.425 | 0.465 | 0.302 | 0.469 | 0.525 | 0.754 | 0.648 | 0.765 | 0.344 |

We present the findings on the performance of the predictors using $As$ as $Y$ and $Cl$ as $X$. In a generic sense, when we apply FGMD to a data set, we should first estimate the association parameter $\lambda$ in the model, and call it $\hat{\lambda}$. We now apply the three predictors as discussed in Section 3, to predict $As$ from $Cl$. We implement a Leave-One-Out-Bootstrap ($LOOB$) method to evaluate the performance of the three predictors. The scheme of $LOOB$ is simple where we drop one observation (a pair of $As$ and the corresponding $Cl$ observation) from the dataset and then we fit the FGMD model onto that reduced dataset with $(n-1)$ observations. We can estimate the association parameter $\lambda$ by making use of any suitable estimator as mentioned in Section 2. Finally, with the estimated model, we use $Cl$ (the independent variable or $X$ in our study) of the dropped off observation to estimate the corresponding $As$ (the dependent variable or $Y$ in our study). This $LOOB$ mechanism is applied to all the $n$ observations which in turn helps us to see how a predictor fared against all the true observations.

The performance of a predictor in conjunction with an estimator of $\lambda$ is then evaluated by the Prediction Mean Absolute Error ($PMAE$) and Prediction Root Mean Squared Error ($PRMSE$). The following Table 9 presents $LOOB-PMAE$ and $LOOB-PRMSE$ of the three predictors each with one of the ten estimators of $\lambda$. Let $Y_i$ be the $i^{th}$ observation of Y (= $As$), and $\hat{Y}_i^{(-i)}$ is the predicted value of $Y_i$ based on the remaining $(n-1)$observations (after fitting the $FGMD$) and using $X_i$ (= the $i^{th}$ value of $Cl$), then $PMAE = \sum_{i=1}^{n} |Y_i - \hat{Y}_i^{(-i)}|/n$, and $PRMSE = [\sum_{i=1}^{n}(Y_i - \hat{Y}_i^{(-i)})^2/n]^{1/2}$.

**Table 9: Performance of the predictors of $As$ from $Cl$ in the southern subregion**

| | | $\lambda$ Estimate Used | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Predictor | $MLE$ | Posterior Mean | | | Posterior Median | | | Posterior Mode | | |
| | | | $BFP$ | $BJP$ | $BAJP$ | $BFP$ | $BJP$ | $BAJP$ | $BFP$ | $BJP$ | $BAJP$ |
| PMAE | $Mean$ | 248.12 | 247.15 | 246.90 | 245.86 | 247.37 | 247.15 | 245.73 | 248.98 | 245.12 | 251.70 |
| | $Median$ | 242.06 | 243.81 | 244.16 | 242.64 | 244.54 | 244.14 | 241.19 | 244.54 | 244.14 | 241.19 |
| | $Mode$ | **230.76**[2] | 241.98 | 249.63 | 253.55 | 248.69 | 247.41 | **237.87**[3] | 241.42 | 251.36 | **204.33**[1] |
| PRMSE | $Mean$ | 306.47 | 305.55 | 305.26 | 304.46 | 305.60 | 305.41 | 304.34 | 305.52 | **302.03**[3] | 307.69 |
| | $Median$ | 314.65 | 313.65 | 314.22 | 314.63 | 314.55 | 315.18 | 313.50 | 314.55 | 315.18 | 313.5 |
| | $Mode$ | **298.86**[2] | 319.35 | 325.18 | 334.94 | 324.57 | 328.74 | 309.14 | 319.5 | 330.34 | **286.28**[1] |

**Remark 13:** (a) Overall, the mean predictor and the mode predictor based on the conditional distribution show better performance than the mode predictor in terms of both $PRMSE$ as well as $PMAE$. Out of the ten estimators of the association parameter, $MLE$ of $\lambda$ is consistently the top performer followed by $BAJP2$ in the second place and $BJP3$ in the third place. This $LOOB$ based work is highly data dependent, *i.e.*, for another dataset the performance evaluation measures can vary drastically and hence one must apply all the three predictors and all the ten estimators of the association parameter to see which predictor (along with which $\hat{\lambda}$) has the best overall performance.

(b) If the parabolic shape of the scatterplot between $As$ and one of the covariates is ignored and the usual simple linear regression model is force-fitted, then it can cause several theoretical as well as practical complexities, such as: (i) the normality assumption of the errors which is implicit in linear regression, is violated; (ii) the homoscedasticity of the error variance is not tenable; and (iii) the predicted value of $As$ may result in negative values as seen for several of our data points.

(c) Forcing a linear regression upon ignoring the previously stated concerns, and truncating the $As$ value at 0 for negative predicted values (unrealistic in nature though) may still result in poor performance.

(d) One has to keep in mind that these aforementioned computational results are based on using a single predictor ($Eh$, $Cl$ or $pH$). Things will definitely improve if we use two predictors (say, $Eh$ and $Cl$) or all the three predictors ($Eh$, $Cl$ and $pH$). This requires upgrading our bivariate $FGMD$ to a trivariate or a quadravariate $FGMD$ and this is currently under investigation. While a bivariate $FGMD$ has a single association parameter $\lambda$ ($= \lambda_{12}$) between the components 1 and 2, a trivariate $FGMD$ has four association parameters $\lambda_{12}$, $\lambda_{13}$, $\lambda_{23}$ and $\lambda_{123}$. Expanding it further, a quadravariate $FGMD$ has a total eleven association parameters. Ota and Kimura (2021) considered the three variate $FGMC$ and the resultant $FGMD$ mainly from an asymptotic point of view. More specifically, they considered the special case of $\lambda_{12} = \lambda_{23} = \lambda_{13} = \lambda_{123} = \lambda$ (say), and considered estimation of the common association parameter $\lambda$. However, more work needs to be done to investigate the exact sampling distribution of the $MLE$ either for all the four parameters or the single common parameter in dimension three. How the behavior of high probability concentration of MLE near the boundary, as we have seen in the bivariate case and discussed in Section 2, permeates to 3 or higher dimensions, needs to studied extensively especially for small to moderate sample sizes.

Further, the Bayesian estimation of the association parameter vector in dimension greater than 2 may lead to interesting results.

(e) The main challenge in dealing with a $p$-dimensional ($p > 2$) $FGMD$ is to carry out a very complex set of computations within a reasonable amount of time which requires sophisticated computational codes. We are currently studying the trivariate $FGMD$ and how it can be used for the arsenic prediction study. This will be reported in near future as we sort out the computational complexities. The case of $p = 2$ is the springboard for the higher dimensional generalizations. Even for $p = 3$, in order to find the maximum likelihood estimates of $\lambda_{12}$, $\lambda_{13}$, $\lambda_{23}$ and $\lambda_{123}$ is a computational nightmare as the optimization is to be done in a 4-dimensional space over a feasible region subject to 8 linear inequalities (*i.e.*, the feasible region has a 'diamond cut' shape).

## 6. Conclusion

With the onset of copula theory which brought about an influx of several copula based joint distributions and its growing application across several disciplines, it is of paramount interest to investigate the copula models more closely. The flexibility of the copula model lies in producing a unique link function (in the continuous random variate case) which essentially joins the marginals. This copula function preserves the entire information about the mutual dependence between two marginals through a single association parameter.

In our work, we have provided a template of a comprehensive inferential investigation of the association parameter of $FGMD$. In our application, we have taken up the bivariate case *i.e.*, we have studied the pairwise components of the groundwater data of MDR. There is, in fact, an array of future directions that are in the works for this research stream -

(a) The generalization of the copula model to a $p$ - dimensional ($p > 2$) set up. Investigating the sampling distribution of the different estimators in the general case and construction of confidence bands.

(b) We have seen the superiority in performance of the Bayes' estimators but the computational challenge was stifling at times. It is intuitive that this challenge will only grow as $p$ increases. Tackling this computational challenge in itself will be an interesting data science research problem.

(c) Development of higher dimensional predictors and subsequent GoF test will be another research problem. Our GoF tests which show adherence to the size criteria in the bivariate case, need to be studied in higher dimensional cases.

(d) The nature of our study for bivariate $FGMD$ has been comprehensive and covers several inferential aspects. This template of investigation can be extended to other commonly used Archimedean and non-Archimedean copulas.

(e) Even though our comprehensive study on $FGMD$ was motivated by an environmental dataset, one might be interested to study how the copula based models can reveal some hidden information for other datasets especially sparse gene expression datasets.

**Acknowledgements**

A version of this paper was presented by the second author as an invited talk at the 26th Annual Conference of The Society of Statistics, Computer & Applications (SSCA), hosted and co-organized by the Department of Mathematics and Statistics & Center for Artificial Intelligence, Banasthali Vidyapith, Rajasthan, India (February 26 - 28, 2024). The authors are deeply indebted to the organizers for their hospitality and generous support.

**Conflict of interest**

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

**References**

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. US Government printing office.

Amblard, C. and Girard, S. (2009). A new extension of bivariate fgm copulas. *Metrika*, **70**, 1–17.

Bairamov, I. and Kotz, S. (2002). Dependence structure and symmetry of huang-kotz fgm distributions and their extensions. *Metrika*, **56**, 55–72.

Bekrizadeh, H., Parham, G. A., and Zadkarmi, M. R. (2012). The new generalization of farlie–gumbel–morgenstern copulas. *Applied Mathematical Sciences*, **6**, 3527–3533.

Chatterjee, R. (2022). *Inferences for the Bivariate Probability Distribution Using Farlie - Gumbel -Morgenstern Copula*. PhD thesis. Available at `https://www.proquest.com/dissertations-theses/inferences-bivariate-probability-distribution/docview/2882152890/se-2`; Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-10-3.

D'este, G. (1981). A morgenstern-type bivariate gamma distribution. *Biometrika*, **68**, 339–340.

Farlie, D. J. (1960). The performance of some correlation coefficients for a general bivariate distribution. *Biometrika*, **47**, 307–323.

Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, **95**, 119–152.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, **12**, 347–368.

Genest, C., Kojadinovic, I., Nešlehová, J., and Yan, J. (2011). A goodness-of-fit test for bivariate extreme-value copulas. *Bernoulli*, **17**, 253–275.

Genest, C., Quessy, J.-F., and Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scandinavian Journal of Statistics*, **33**, 337–366.

Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and economics*, **44**, 199–213.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, **55**, 698–707.

Huang, J. S. and Kotz, S. (1999). Modifications of the farlie-gumbel-morgenstern distributions. a tough hill to climb. *Metrika*, **49**, 135–145.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2004). *Continuous Multivariate Distributions, Volume 1: Models and Applications*. John Wiley & Sons.

Merola, R., Hien, T., Quyen, D., and Vengosh, A. (2015). Arsenic exposure to drinking water in the mekong delta. *Science of the Total Environment*, **511**, 544–552.

Morgenstern, D. (1956). Einfache beispiele zweidimensionaler verteilungen. *Mitteilingsblatt fur Mathematische Statistik*, **8**, 234–235.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer Science & Business Media.

Ota, S. and Kimura, M. (2021). Effective estimation algorithm for parameters of multivariate farlie–gumbel–morgenstern copula. *Japanese Journal of Statistics and Data Science*, **4**, 1049–1078.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, **53**, 683–690.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, **8**, 229–231.

## A. Appendix

### A.1. Application data

**NOTE**: Well ID starting with TH are located in the northern subregion and Well ID starting with DT or TB are from the southern subregion.

**Table 10: MDR groundwater data**

| Well ID | As (ppb) | Cl (ppm) | Eh (mv) | pH | Well ID | As (ppb) | Cl (ppm) | Eh (mv) | pH |
|---|---|---|---|---|---|---|---|---|---|
| DT7 | 563.9 | 107 | -126 | 6.78 | TB19 | 300.3 | 160.3 | -120 | 6.68 |
| DT6 | 0.5 | 56.1 | 142 | 6.71 | TBE10 | 700.4 | 81.4 | -108 | 7.07 |
| DT5 | 0.7 | 46.8 | 199 | 7.04 | TBE9 | 196.2 | 986.6 | -110 | 6.72 |
| DT3 | 0.4 | 345.2 | 169 | 6.44 | TBE7 | 166.3 | 20 | -84 | 7.16 |
| DT4 | 0.1 | 500.1 | 165 | 6.51 | TBE4 | 4.4 | 1499.6 | 82 | 6.09 |
| DT2 | 1.8 | 632.8 | 101 | 6.66 | TBE5 | 981.4 | 60.4 | -110 | 6.87 |
| DT1 | 13.1 | 19.7 | 97 | 7.75 | TBE3 | 6.8 | 2.7 | 158 | 7.17 |
| TB11 | 462.3 | 9.2 | -114 | 6.92 | TBE1 | 6.6 | 61.7 | 126 | 7.1 |
| TB18 | 155.7 | 25.9 | -72 | 6.52 | TBE11 | 5.3 | 12.2 | 60 | 7.16 |
| TB9 | 187.6 | 12.8 | -128 | 6.94 | TBE6 | 3.2 | 1527 | 149 | 6.73 |
| TB2 | 850.4 | 10.5 | -133 | 7.14 | TH16 | 0.4 | 173.6 | 157 | 6.14 |
| TB24 | 370.4 | 13.9 | -90 | 7.15 | TH9 | 0.2 | 275.3 | 253 | 5.84 |
| TB26 | 139.9 | 13.8 | -83 | 7.43 | TH13 | 0 | 22.7 | 194 | 6.19 |
| TB27 | 77.7 | 5.4 | -33 | 7.24 | TH14 | 0.3 | 113.6 | 184 | 6.02 |
| TB21 | 842.1 | 21.1 | -105 | 6.88 | TH22 | 0.1 | 228.1 | 226 | 6.5 |
| TB1 | 276.8 | 19.6 | -92 | 6.85 | TH21 | 0.3 | 89.9 | 169 | 6.1 |
| TB10 | 377.3 | 8.2 | -129 | 6.79 | TH5 | 0.8 | 742.1 | 251 | 5.83 |
| TB25 | 272.9 | 11.9 | -104 | 7.2 | TH12 | 2.3 | 182.7 | 127 | 6.31 |
| TB13 | 746 | 72.7 | -125 | 7.16 | TH15 | 8.4 | 27.5 | 60 | 6.18 |
| TB22 | 311 | 13.5 | -130 | 6.63 | TH1 | 6 | 544.4 | 210 | 6.08 |
| TB15 | 937.7 | 19.3 | -110 | 7.04 | TH10 | 3.2 | 277.3 | 231 | 6 |
| TB16 | 314.5 | 25.8 | -115 | 6.74 | TH2 | 2 | 487.6 | 130 | 5.87 |
| TB20 | 746.3 | 6.9 | -139 | 6.61 | TH23 | 0.2 | 158.5 | 175 | 6.04 |
| TB23 | 270 | 12.7 | -110 | 7.01 | TH3 | 1.5 | 560.2 | 261 | 6 |
| TB17 | 224.2 | 21.5 | -126 | 6.46 | TH4 | 2.6 | 21.4 | 80 | 5.99 |
| TB3 | 727 | 10.8 | -136 | 7.14 | TH11 | 8.9 | 479.8 | 158 | 6.56 |
| TB12 | 931.5 | 2.9 | -125 | 7.03 | TH18 | 3.6 | 335.6 | 181 | 6.29 |
| TB14 | 747.7 | 63.4 | -115 | 7.15 | TH8 | 6 | 253.2 | 235 | 5.85 |
| TB5 | 416.3 | 0 | -60 | 7.69 | TH7 | 0.7 | 122.3 | 162 | 6.29 |
| TB4 | 360.3 | 42.9 | -130 | 7.34 | TH6 | 0 | 242.8 | 200 | 6.19 |
| TB6 | 315.5 | 61.2 | -111 | 7.36 | TH17 | 22.2 | 40.5 | -13 | 7.03 |
| TB7 | 101.1 | 42.7 | -28 | 7.3 | TH19 | 17.5 | 57.3 | 145 | 7.39 |
| TB8 | 237.6 | 124.4 | -98 | 7.17 | TH20 | 2.4 | 0 | 24 | 6.51 |