STATISTICAL PROPERTIES OF LIQUID PROTEIN-WATER MOLECULAR SYSTEM DYNAMICS

JUTHARATH VORAPRATEEP

Doctor of Philosophy

January, 2017

Abstract

It is considered an established fact that water plays the major role in protein motion, there is a close connection between the water dynamics and the protein conformational dynamics.

We report on statistical analysis of such conformational dynamics obtained using classical molecular dynamics simulations with explicit water. We investigate specific moments in time when one of the dihedral angles of a simulated protein (a peptide dialanine) makes a large amplitude change causing a conformational transition in the peptide. We are interested in finding statistical correlations between the values of the angle at the moment of transition and several moments in advance of the transition (between 0.0 and 50.1ps). We also investigate how these correlations change when conditioned on the presence of water at different locations in space around the peptide. The challenge is in a large number of parameters that influence the conformational dynamics, which leads to multivariate probabilities. As statistical tools, we use pair-copulas and the Kendall's tau correlation.

Copulas are a special way of representing multivariate probabilities. Pair-copulas construction (PCC) decomposes a multivariate probability density into bivariate copulas, so-called pair-copulas. D-vine is one of graphical models that give a specific way of decomposing the probability density. The dependency structure is determined by the bivariate copulas and a nested set of trees using pair-copula. For this research, we apply the D-vine to study the statistical correlations between variables describing molecular conformation of a peptide and the properties of water molecules surrounding the peptide.

We have found that the dynamics of peptides conformation possesses temporal correlations well in advance of the moments of conformational transitions. Moreover, when conditioned on the presence of water molecules at a few very specific locations in the first hydration shell of the peptide, these correlations become stronger and longer in time. This quantifies the influence of water on the conformational transitions and specifies water molecules that appear critical for the peptide to make successful conformational transition.

Acknowledgements

I would like to deeply thank my supervisor, Dr. Dmitry Nerukh, for providing me with the topic and for his patience and help during the work undertaken herein. I would like to thank my family and all my friends for their encouragement. I am also grateful to Aston University for the opportunity to take the course of research skills and professional development that formed the basis of the thesis. Moreover, I am indebted to the Royal Thai Government, Ministry of Science and Technology and Ramkhamhaeng University for financial support during my study at Aston University.

Contents

List of Figures

List of Tables

List of Symbols

$F(x)$	marginal distribution function of X
$G(y)$	marginal distribution function of Y
$H(x, y)$	joint distribution function of X and Y
$C(u, v)$	two-dimensional copula of U and V
$C(u_1, \ldots, u_d)$	d-dimensional copula of $U_1, U_2, \ldots, U_d$
$C_{\Omega}^{Gauss}(u)$	Gaussian copulas with correlation matrix Ω
$C_{\nu,\Omega}^{t}(u)$	Student-t copulas with correlation matrix Ω and degrees of freedom ν
ϕ	copulas generator (Chapter 2)
θ	copulas parameter
$\rho_{X,Y}$	Pearson Product-Moment Correlation Coefficient
r_s	Spearman's Rank Correlation Coefficient
τ	Kendall's Tau Correlation Coefficient
$\rho_{XY\|Z}$	conditional correlation of X and Y given Z
α	shape parameter of Beta distribution
β	shape parameter of Beta distribution
ψ	psi: one of dihedral angles of dialanine molecule (Chapter 4)
ϕ	phi: one of dihedral angles of dialanine molecule (Chapter 4)
X_0	a grid point of the maximum probability of hydrogen
X_1	an opposite grid point of X_0
X_2	a neighbor grid point of X_0
X_3	a neighbor grid point of X_0
ps	picoseconds: unit of delay time before the transition
$\sin(\psi_i)$	sine of psi at delay time i
$\sin(\psi_j)$	sine of psi at delay time j
δ_i	hydrogen density at delay time i
$\rho_{\psi_i\psi_j\|\delta_i}$	conditional correlation of $\sin(\psi_i)$ and $\sin(\psi_j)$ given δ_i
$\sin(\psi_{(i+j)/2})$	sine of psi at delay time $(i+j)/2$
$\rho_{\psi_i\psi_j\|\psi_{(i+j)/2}}$	conditional correlation of $\sin(\psi_i)$ and $\sin(\psi_j)$ given $\sin(\psi_{(i+j)/2})$
$\delta_{0(k)}$	the hydrogen density at delay time 0.0 on a chosen grid point k
$\rho_{\psi_i\psi_j}$	unconditional correlation of $\sin(\psi_i)$ and $\sin(\psi_j)$
$\rho_{\psi_i\psi_j\|\delta_{0(k)}}$	conditional correlation of $\sin(\psi_i)$ and $\sin(\psi_j)$ given $\delta_{0(k)}$

1

Introduction

1.1 Introduction

In real life, proteins are a particular type of biological molecule from which most living things are made of. They are interesting to study for many reasons (Echenique [31]):

- proteins can be found in all living beings on Earth,
- proteins show an apparently limitless capacity for assuming different shapes and for creating particular catalytic regions on their surfaces,
- proteins constitute the working power of chemistry of living beings,
- as hormones, proteins transmit information and signals among cells and organs,
- as antibodies, proteins defend the organism against intruders,
- proteins are an essential components of muscles,
- proteins are fascinating molecular devices,
- proteins are proving to be a powerful centre of interdisciplinary research,
- most proteins perform their work under particular native shapes which involve many twists, loops and bends of the linear chain of amino acids.

Clearly, proteins are important and useful molecules in humans and animals. In the next section, we give a short literature review, mainly pertaining to the understanding of proteins structure and we describe the previous research on proteins relevant to this work in section 1.3.

1.2 Protein Structure

Echenique [31] stated that proteins are a rather homogeneous class of molecules from the chemical point of view, they are **linear heteropolymers**. The important components or building units for proteins are called **amino acids** which can exist as stand-alone stable molecules. As Figure 1.1 illustrates, amino acids are built up of a central α-carbon with four groups attached to it; an amino group ($-NH_2$), a carboxyl group ($-COOH$), a hydrogen atom and a fourth arbitrary group ($-R$). When the group ($-R$) is not equal to one of the other three groups attached to the α-carbon, the amino acids are called **chiral**. Our hands, for example, exist in two different forms, which are mirror images of one another and cannot be overlaid by switching one of them in space. We cannot wear the left-hand glove on our right hand. Moreover, the α-carbon constitutes an **asymmetric centre** and the amino acids may exist as two different **enantiomers** called L- and D- forms. L- and D- stand for **Levorotatory** and **Dextrorotatory**, respectively as illustrated in Figure 1.2. In general, we will write the prefixes, L and D letters, in small capitals, as in L- and D- (Echenique [31]).

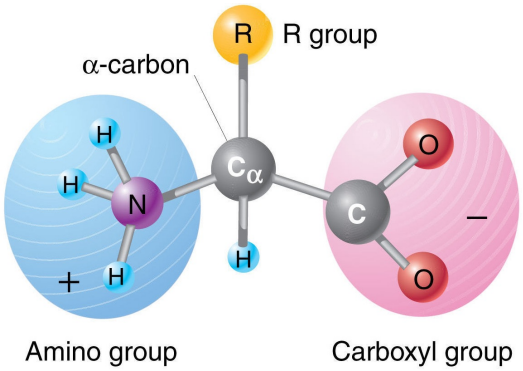

Figure 1.1: Amino acids - The figure shows amino acids structures, retrieved from Stereoisometry of Amino Acids.

In principle, Echenique [31] mentioned that amino acids may be L- and D-, and the group (-R) may be anything provided that the resultant molecule is stable. The amino acid sequence of the resultant protein, read from the **amino terminus** to the **carboxyl terminus**, is called **primary structure**; and the amino acids included in such a polypeptide chain are normally termed **amino acid residues**, or simply **residues**, in order to distinguish them for their isolated form. The main chain formed by the repetition of α-carbon and the C and N atoms at the peptide bond is called **backbone** and the (-R) groups branching out from it are called **side chains**. The specificity of each protein is provided by the different properties of the twenty side chains and their particular positions in the sequence. The vast majority of these changes (the chemical (covalent) structure of the protein chain) either depend on the existence of some chemical agent external to the protein or are catalysed by an enzyme (Echenique [31]).

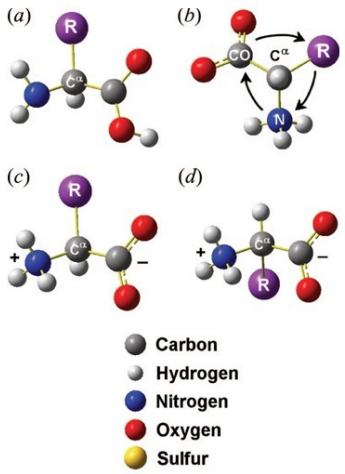

Figure 1.2: Variety of amino acids - The figure shows a variety of amino acids (a) Uncharged L-enantiomer (b) CORN mnemotechnic rule to remember which one is the L-form (c) Charged L-enantiomer (the predominant form found in living beings) (d) Charged D-enantiomer (Source: Echenique [31], p.86).

1.2.1 Definition of Protein Terms

For the protein study, there are many technical terms that are essential in the research. However, three protein terms that are useful in this research are as follows: (Echenique [31])

- **Peptide** is a small protein that comes from a combination of several amino acids. Figure 1.3 illustrates a peptide bond formation reaction.
- **Internal coordinates** are the coordinates between two atoms that consist of three coordinates: ***bond length, bond angle and dihedral angle***. All of these describe rotations around triple, double and partial double bonds that may be considered to be determined by the covalent structure. Figure 1.4 shows the internal coordinates as follows:
  - r_{21} is the **bond length** between atoms 2 and 1.
  - θ_{321} is the **bond angle** formed by the bonds (2,1) and (3,2) which ranges from 0° to 180°.
  - ϕ_{4321} is the **dihedral angle** that describes the rotation around the bond (3,2); it is defined as the angle formed by the plane containing atoms 1, 2 and 3 and the plane that contains atoms 2, 3 and 4; it ranges from -180° to 180° or from 0° to 360°.

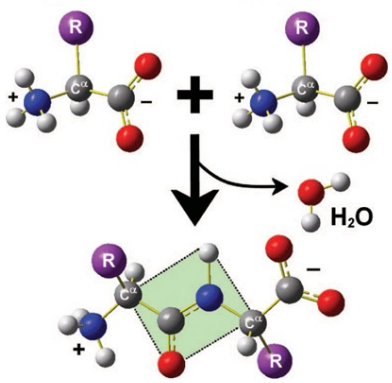

Figure 1.3: Peptide - The figure shows the peptide bond formation reaction. The enclosed dotted area is a peptide plane (Source: Echenique [31], p.87).

The internal coordinates can refer to the Euler angles of a rigid body rotation. There are three independent quantities are needed to characterize the rotation, they are called Euler angles: α (or ϕ), β (or θ), γ (or ψ). Figure 1.5 shows Euler angles which the axes of the original frame are denoted as x,y,z and the axes of the rotated frame are denoted as X,Y,Z. The line of nodes: N is the intersection of the planes xy and the XY, it can also be defined as the common perpendicular to the axes z and Z and then written as the vector product $N = Z \times z$. The three Euler angles are defined as follows:

1. α (or ϕ) is the angle between the x axis and the N axis, it is called x-convention. It could also be the angle between the y axis and the N axis, it is called y-convention.
2. β (or θ) is the angle between the z axis and the Z axis.
3. γ (or ψ) is the angle between the N axis and the X axis, it is called X-convention.

- **Non-superimposable** is the feature for an object cannot be placed over another object, for example, the molecules are non-superimposable which means that the molecules cannot be placed on top of one another and give the same molecule as illustrated in Figure 1.6.
- **Conformations** are the non-superimposable three-dimensional arrangements of the molecule.

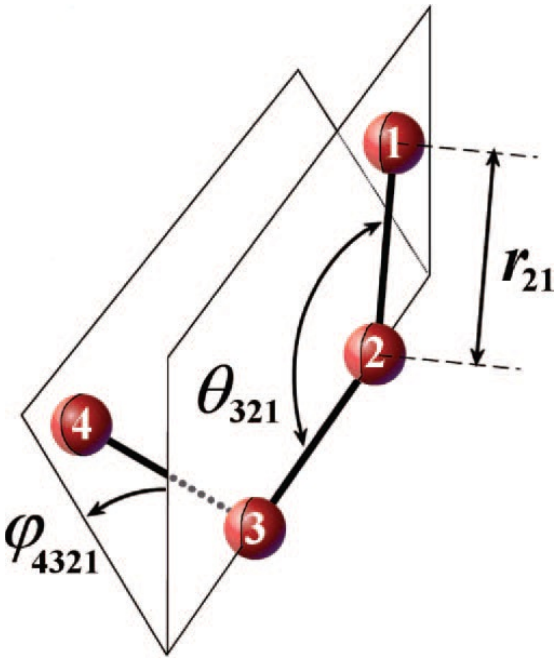

Figure 1.4: Internal coordinates - The figure shows the bond length, bond angle and dihedral angle (Source: Echenique [31], p.90).

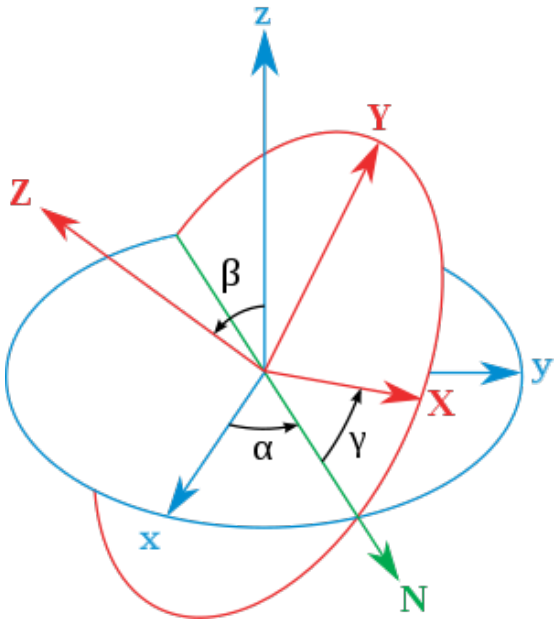

Figure 1.5: Euler angles - The figure shows three Euler angles: α (or ϕ), β (or θ), γ (or ψ). The xyz (fixed) frame is shown in blue, the XYZ (rotated) frame is shown in red. The line of nodes (N) is shown in green, retrieved from Euler angles of a rigid body rotation.

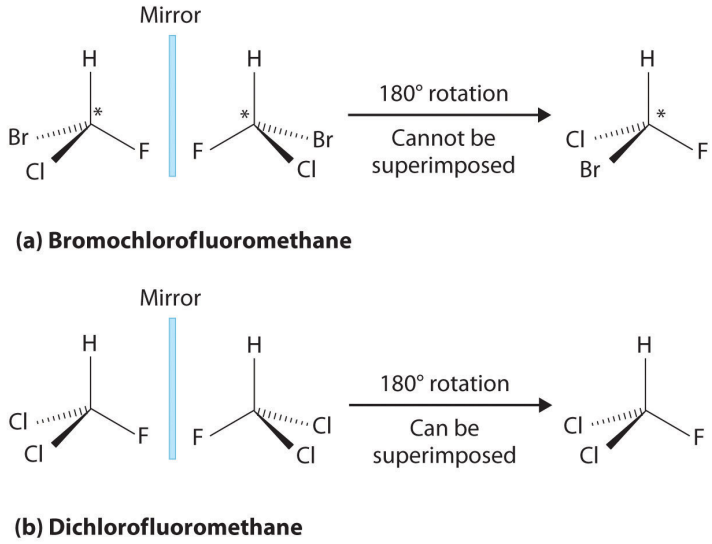

Figure 1.6: Non-superimposable - The figure shows (a) Bromochlorofluoromethane are non-superimposable molecules and rotation of its mirror image does not give the same structure. (b) Dichlorofluoromethane are superimposable molecules and its mirror image can be rotated and give the same structure, retrieved from Enantiomers.

1.3 Protein Research

There are many fields in chemistry and physics that are related to the movement of proteins, for example, protein folding problem is one of the most important yet unsolved issues of modern science. The protein folding problem is the prediction of the three dimensional native structure of proteins from the knowledge of their amino acids sequence. The previous studies about proteins and related fields are following.

Atamas et al. [6] studied the effect of water dynamics on conformation changes of albumin in pre-denaturation state: photon correlation spectroscopy and simulation. Water is essential for protein three-dimensional structure, conformational dynamics, and activity. Human serum albumin (HSA) is one of major blood plasma proteins, and its functioning is fundamentally determined by the dynamics of surrounding water. The goal of this study was to link the conformational dynamics of albumin to the thermal motions in water taking place in the physiological temperature range. They reported the results of photon correlation spectroscopy and molecular dynamics simulations of HSA in aqueous solution. The experimental data processing produced the temperature dependence of the HSA hydrodynamic radius and its zeta potential. Molecular dynamics reproduced the results of experiments and revealed changes in the secondary structure caused by the destruction of hydrogen bonds in the macromolecule's globule. The conclusions for this study are following.

- The conformation changes of albumin are ultimately correlated with water dynamics in the vicinity of the temperature point 42°C.
- On the one hand the experimental (hydrodynamic radius) and simulation (gyration radius, stability of disulfide bonds and salt bridges) data support the fact of the macromolecule is stabile in the temperature range 25-50°C with approximately the same size.
- The study suggested that the hydrodynamic radius includes both solvent and shape effects.
- On the other hand the zeta potential behaviour from temperature 35°C up to denaturation indicated that the hydration layer beyond the Stern layer (an inner region, where surrounding molecules and the ions are strongly boundwith the protein macromolecule of HSA) is drastically changed in the vicinity of 42°C.
- The temperature dependence of the zeta potential for HSA may imply conformational changes in the macromolecule structure, surface modifications, and, as a result the initial stage of denaturation process.
- The simulation data just revealed the conformational changes of the secondary structures, which are stabilized namely by hydrogen bonds. Thus it can mean that the temperature range 40 - 42°C is the initial stage of denaturation for HSA.

Nerukh [81] applied a hidden Markov state model to classical molecular dynamics simulated small peptide in explicit water. The hidden Markov processes framework was developed from Markov processes theory. In this research, he developed a variant of the hidden Markov description of conformational transitions and used the methodology called "Computational Mechanics" (CM). The hidden markov state model was called non-Markov state model (nMSM) and applied to protein dynamics. The detection of short lived "transition" states was possible by reducing the time step by an order of magnitude compared to the usual Markov state model (MSM). For Computational Mechanics, it was applied to molecular systems in several contexts. A zwitterion L-alanyl-L-alanine was studied with three reasons: 1) the conformation of the molecule is completely defined by the two dihedral angles ψ and ϕ, 2) in water the conformation $\psi \approx 2.5$, $\phi \approx$ -2.2 radians is prevalent, and 3) the transitions only happen in water because the molecule's charged ends lock it in a "loop" like conformation in vacuum. Therefore, this molecule is a good model of protein conformational changes at the same time being technically advantageous: the transitions are clearly defined and well separated in time. The conclusions for this study are following.

- The methodology provides a possibility of reducing the time step, thus, increasing the time resolution of the model.
- The resolution is only restricted by the amount of available data for the analysis, whereas in the standard MSM the time resolution restriction is conceptual.

- The method provides a detailed description of the mechanisms of the transitions. It identifies the importance of multiple recrossing and quantifies their probabilities.
- The mechanisms of the transitions can now be elucidated with high time resolution, a detailed molecular picture of conformational motions can be obtained.
- The mechanisms of the transitions for this particular peptide show that there is, probably, other important degrees of freedom that need to be included in the picture. He hypothesised that such degrees of freedom should include water molecules.

Nerukh et al. [83] studied molecular phase space transport in water using molecular dynamics (MD). Molecular transport in liquids can be considered from two different viewpoints. The first one, the familiar diffusion, the mean squared displacement of atoms in Euclidean three dimensional space, describes how, on average, the atoms move in the liquid. The second one, the high-dimensional phase space transport of the dynamical system trajectories comprising all the coordinates and velocities of all the particles in the volume of interest can be analysed. Molecular transport in phase space is crucial for chemical reactions because it defines how pre-reactive molecular configurations are found during the time evolution of the system. Using Molecular Dynamics (MD) simulated atomistic trajectories, they analysed the statistical properties of the phase space transport for bulk water at room temperature. They tested the assumption of the normal diffusion in the phase space for bulk water at ambient conditions by checking the equivalence of the transport to the random walk model. The conclusions for this study are following.

- Molecular trajectories of bulk water fill the phase space in a very non-uniform manner and hence not randomly.
- Contrary to the three-dimensional space, some statistical features of the transport in the phase space differ from those of the normal diffusion models. This implies a non-random character of the path search process by the reacting complexes in water solutions.
- A significant long period of non-stationarity in the transition probabilities of the segments of molecular trajectories can account for the observed non-uniform filling of the phase space.
- The characteristic periods in the model non-stationarity constitute hundreds of nanoseconds, that is much longer time scales compared to typical lifetime of known liquid water molecular structures (several picoseconds).
- It is worth noting that several characteristic motions of molecules are known to exist in water.

Nerukh et al. [84] studied complex temporal patterns in molecular dynamics: a direct measure of the phase-space exploration by the trajectory at macroscopic time scales. The computational mechanics approach utilizing information theoretic concepts of the ϵ machine and statistical complexity were used for describing the high-dimensional molecular dynamics of an ensemble of 392 water molecules. The problem of finding hidden regular patterns in a time series that appears to be a random process was addressed by the analysis of the phase-space filling property by an individual trajectory. The presence of patterns was judged by significant deviations from the uniform coverage of the phase space expected for the case of a random process. Long-range memories present in the molecular dynamics simulations were detected and investigated by the means of statistical complexity analysis. A direct measure based on the notion of statistical complexity that describes how the trajectory explores the phase space and independent from the particular molecular signal used as the observed time series was introduced. The conclusions for this study are following.

- The method estimated the information contained in the molecular trajectory by detecting and quantifying temporal patterns present in the simulated data (velocity time series).
- Two types of temporal patterns were found. The first, defined by the short-time correlations corresponding to the velocity autocorrelation decay times ($le0.1$ ps), remained asymptotically stable for time intervals longer than several tens of nanoseconds. The second was caused by previously unknown longer-time correlations (found at longer than the nanoseconds time scales) leading to a value of statistical complexity that slowly increases with time.
- It was shown that arbitrary long memories (much longer than one can expect from a spectral or correlation analysis) are present in the recorded time series, manifesting themselves as groups of causal states in the velocity-defined phase space.
- It should be stressed that these long-range correlations cannot be detected using the usual linear two-point statistics: the correlation function is essentially zero at all times for the data points spaced with intervals longer than a few picoseconds.
- Statistical complexity turns out to be a universal measure of dynamical structures present in the observed data.
- A comparison to surrogate data sets with broken dynamica correlations supports the hypothesis that the patterns are not caused by the details of the computational procedure, intrinsic statistical errors, or insufficient data, but by the complex dynamics of the system.

Ryabov and Nerukh [89] studied behaviour of symbolic subsequences (words) of nine symbols long at a very long time of 1 μs. In this research, a trajectory of liquid water simulated using classical molecular dynamics was analysed in the framework of symbolic

dynamics. In this framework, the molecular trajectory was converted into a sequence of symbols from an alphabet consisting of only a few symbols. The resulting symbolic sequence was analysed using various statistical methods. The trajectory was analysed not as isolated symbols but as a sequence of symbolic strings (words). This ensues the extraction of detailed information from the initially continuous trajectory despite seemingly very coarse grained representation of it with only a few symbols. Also, they analysed the dynamics of liquid water using such symbolic representation and found surprisingly slow convergence of calculated statistical indicators. The statistics of water time series behaved fundamentally different compared to those of a pseudorandom sequences (for example, the digits of the number) or a random surrogate signal that has the correlation function (and hence the power spectrum) identical to that of water. Moreover, water dynamics resembles the behaviour found in a simple chaotic system, the ChirikovTaylor or Standard map. They hypothesised that the origin of such non-random properties of molecular trajectory can be in its deterministically chaotic character. The conclusions for this study are following.

- Despite the homogeneous nature of the system composed of identical molecules, there is significant deviations from "simple" random behaviour at the times of the order of 1 μs.
- They observed certain similarities between the trajectories calculated from the simulation of water dynamics and the dynamics of a classical two dimensional map modelling the kicked rotator (the Standard map).
- Their statistical approach was focused on finding the signatures of chaotic dynamics in such a high dimensional dynamical system as theensemble of interacting molecules.
- The statistical characteristics of the trajectories in the molecular system occupy intermediate position between the random surrogate and the chaotic map.
- When the dimensionality of the system becomes large, the transport properties are no longer defined by impenetrable barriers formed by tori, but some essentially new features such as Arnold diffusion emerge as a result of torus break-up.
- The destroyed tori has complex structure, and some of them are unstable (analogous to saddle points in the case of Standard map).
- The chaotic trajectories can be trapped by such structures, therefore the observed statistical properties of an arbitrary chaotic trajectory can strongly depend on their presence in the phase space.

Ryabov and Nerukh [90] studied computational mechanics of molecular systems for quantifying high-dimensional dynamics by distribution of Poincaré recurrence times. They stated that computational mechanics (CM) is a promising new concept which aims at building a statistical and at the same time dynamical description. It combines

the well-developed theoretical framework of generalised Markov chains which is called the ϵ-machine with the concept of short time predictability characteristic of dynamical systems. Their research goal was to find persistent structures in the phase space formed by the trajectories and interpret typical behaviour of structures in terms of the statistical theory and the dynamical systems approach. They analysed the application of computational mechanics to Hamiltonian dynamics of molecular systems. A conceptually important connection of the causal states of the ϵ-machine built on an initially symbolised trajectory to the areas of phase space that are optimal in the sense of predicting the trajectorys behaviour was analysed. It was shown that the areas in the phase space defined by the causal states possess special properties in the dynamical sense, that is their recurrence time distributions follow Poincaré law with two distinct exponents. This allows classifying the causal states into quasi-periodic and chaotic types. The most difficult problems in the analysis of the high-dimensional molecular trajectories which is the definition of the notion of structure or cluster in the phase space. The conclusions for this study are following.

- Their approach provided a new quantitative characteristic that allows to separate the motion in the phase space into two distinct classes.
- The result on the distribution of recurrence rates over the ensemble of causal states suggested that the phase space of the dynamical system corresponding to water has more complex structure than can be concluded from average statistical analysis of return times.
- The parameter D introduced as an indicator of deviation from Poincaré law thus provided a more subtle distinction between periodic and chaotic phases of motion.
- Several causal states demonstrated much slower decay rate than can be expected from Poincaré law. This fact evidences the presence of the areas in the phase space where the trajectory spends longer time compared to the rest of the accessible volume.
- Such areas can not be detected easily by other methods, most probably due to abundance of resonant areas in the high-dimensional phase space that makes difficult a clear distinction between chaotic and quasi-periodic motions.
- From a different perspective, their method also had a special importance for the problem of quantifying transport properties in high-dimensional molecular systems, since it reveals a (small) number of areas in the phase space playing crucial importance for particle motion through the phase space.
- Finding such areas from the analysis of a single scalar time series can be very useful in numerical experiments with large number of interacting particles that typically generate huge volumes of data.

- Extracting the essential information from the trajectory of a single test particle thus looks a promising approach, for example, in modeling the process of protein folding or dynamics of complex biomolecules.

Nerukh and Karabasov [82] studied transitions between metastable conformations of a dipeptide using classical molecular dynamics (MD) simulation with explicit water molecules. In this work, they showed that water indeed drives the changes and elucidate the specific mechanisms of this phenomenon. The main goal of their work was to investigate the probability distributions of water atoms corresponding to the conformational states of the peptide. The summary of this study is following.

1. A dipeptide: zwitterion L-alanyl-L-alanine was selected to study that is a very convenient model because (1) the conformation of molecule is completely defined by the two dihedral angles ψ and ϕ, (2) in water the conformation $\psi \approx 2.5$, $\phi \approx$ -2.2 radians is prevalent, however, very rare transitions to two other metastable conformations ($\psi \approx$ -1, $\phi \approx$ -2.2 and $\psi \approx 2.5$, $\phi \approx 1$) take place, and (3) the transitions only happen in water because in vacuum the negative charged COO^- group strongly interacts with the positively charged NH_3^+ group, excluding all conformations except the one with the groups at the minimal distance from each other.

2. Two different molecular dynamics (MD) models of the system were studied. One is the united atom forcefield GROMOS, the other is OPLS (Optimized Potentials for Liquid Simulations). These are among the most popular MD models for peptides and proteins. They both showed the same results in investigations despite significantly different representation of atoms and their interactions for the peptide and water molecules.

3. The definition of dynamically metastable states is the conformations in which the molecule spends significantly more time compared with the time it spends for transitions between the conformations. This is reflected in the probabilities of the molecules conformations calculated as the probability of finding an MD trajectory point with specific values of ψ and ϕ for the whole MD simulation time. There are three well-separated metastable states, clearly visible in the space of conformation probabilities. These allow one to introduce a simple natural discretization of the conformations, which is designated as "A", "B", or "C".

4. The time evolution of the water distributions during the conformational transitions between the states reflects the role of water in conformational rearrangements.

5. The distribution of the surrounding water at different moments before the transitions and the dynamical correlations of water with the peptide's configurational motions indicated that the water molecules represent an integral part of the molecular system during the conformational changes, in contrast with the metastable periods when water and peptide dynamics are essentially decoupled.

6. It is possible to identify the moments of transitions between the conformational states A, B and C which defined boundaries of states. So, one could define the moments of transitions between the states when the trajectory crosses the boundaries (the boundaries describe the probabilities of conformations averaged over the whole trajectory).

7. The individual pieces of the trajectory do not go directly from one state to another but they wind in a complicated manner, often crossing the boundaries many times before setting in a new conformation. This results in flickering states when they form sequences of very short-lived alternating states, for example, "ABABAB" which clearly do not satisfy the desired property of metastability. Sometimes after several such crossing, it returns to the original state without settling in the new state for long enough time.

8. The problem can be solved by increasing the time step between state observations such that the step becomes larger than the time required for the transition to complete. Thus, by discretizing time with a step Δt the continuous MD trajectory was converted into a string of symbols $\{s_i\}$, $i = 0, \ldots, N$, where s_i equals "A", "B", or "C" depending on where the trajectory point falls at the time moment t_i and N is the number of such steps in the simulation. The **Markov State Model (MSM)** was used to analyse the sequence of symbols, in addition to the static probabilities of the conformations, takes into account their dynamics. The model specifies the probabilities of each of the discrete states as well as the probabilities of the transitions between them. The MSM transition matrix can be calculated from the MD trajectory by counting the state changes.

9. Since the MSM model is valid only for relatively large time steps, which follows from the requirement for the transitions to be history-independent (Markovian, that is, statistically uncorrelated) and this requirement ensures that "flickering" is hidden from the analysis but at the same time it excludes the information about the actual process of transition, therefore, the MSM has to be augmented to be able to describe the dynamics at significantly shorter time steps, that is the **hidden Markov model** (For the studied peptide, the minimal valid time step is $\sim$6 ps which is the same order as the period of fluctuations within each conformational state and, most importantly, this is approximately the duration of the process of the trajectory passing from state to state).

10. This work calculated the distribution of oxygen (hydrogen) atoms in space by averaging over the selected time frames. The probability equals the number of atoms in the small volume divided by the total number of atoms in the system. Moreover, the structure of water was defined by the hydrogen bonds network, which implied approximately the same distance between water molecules everywhere.

11. The overall changes of the water structure proceed concurrently with the change of the dihedral angle, reflecting the conformational transition. Calculating the

dynamical correlations between the high-probability water areas and the dihedral angles of the peptide was quantified the degree of the dependence between them, this is done using the linear stochastic estimation (LSE) technique.

12. From 10 to 1 ps before the transition, when the dihedral angles change the most, the water molecules tend to be located at more specific positions around the peptide compared with more uniform distribution at other times.

13. During the transitions, the dynamics of water distribution becomes highly correlated with the dynamics of the dihedral angles and these correlations are completely absent during the stable conformation periods.

14. The water and the peptide behave as an integral dynamical system. During the conformational transition the peptide and the surrounding water undergo transitions together. This is in contrast with the metastable periods when their dynamics is essentially decoupled. The transition is characterized by a more specifically defined hydrogen bonds network of water reflected in more definite positions of water atoms around the peptide.

Recently, Cuendet et al. [24] studied the allostery landscape: quantifying thermodynamic couplings in biomolecular systems. Allostery plays a fundamental role in most biological processes and has been suggested to be present in nearly all proteins. They used a statistical mechanical approach to show that the allosteric coupling between two collective variables is not a single number, but instead a twodimensional thermodynamic coupling function that is directly related to the mutual information from information theory and the copula density function from probability theory. They showed how to quantify the contribution of specific energy terms to this thermodynamic coupling function, enabling an approximate decomposition that reveals the mechanism of allostery. Also, they illustrated the thermodynamic coupling function and its use by showing how allosteric coupling in the alanine dipeptide molecule contributes to the overall shape of the Φ/Ψ free energy surface, and by identifying the interactions that are necessary for this coupling. This research is one of very few examples of using copulas in molecular simulations. The conclusions for this study are following.

- A thermodynamic coupling function based on the allosteric efficacy that quantifies the allosteric coupling between two continuous or discrete collective variables (CVs) was derived and found that the thermodynamic coupling function is related to both the pointwise mutual information and the copula, and is best represented in the form of an allostery landscape, in units of free energy.

- The allostery landscape of the Φ and Ψ dihedral angles of the alanine dipeptide contains positive allosteric couplings that appear to stabilize the α_L and C_{7ax} conformations, and negative allosteric couplings that coincide with the high regions of the Φ/Ψ free energy landscape.

- On the basis of the formalism they developed, they were able to attribute features of this thermodynamic coupling function to specific interaction energy terms, thus allowing interpretation of the allosteric landscape.
- The concepts that they developed are very general and are applicable to larger molecular systems, provided enough sampling is available and the functionally relevant CVs are known. This second condition is especially noteworthy for cases in which a complete functional description involves multimolecular considerations.
- Their new theoretical formalism and its computational implementation remains applicable despite such complications, and can serve as a powerful tool in understanding the molecular mechanisms of the many proteins in which allostery is essential to biological function.

From the proteins research reviews, especially the research of Nerukh and Karabasov in water-peptide dynamics during conformational transitions, the result indicated that water plays the main role in protein motion and there is an association between water dynamics and proteins conformations. Therefore, further research about water and proteins dynamics during conformational transitions would be interesting and challenging.

In this thesis, the atomistic dynamics of liquid aqueous solutions of small proteins are studied and investigated using state of the art statistics. Particular attention is given to statistical dependencies: copulas and correlation analysis in

1. the dynamics of the protein: the dihedral angles of a peptide dialanine and surrounding water molecules: oxygen and hydrogen atoms,
2. collective long space- and time-range correlations in the molecular trajectories.

In the next section, we give a short literature review, mainly pertaining to the understanding of general information on copulas and copulas research.

1.4 General Information on Copulas

In this section, we describe an introduction to copulas. The copulas are statistical method that is used to analyse the data in this research. Moreover, we also review some of previous research in copulas.

1.4.1 Introduction

The word "copula" is a Latin noun that means a link, tie, bond (Cassells Latin Dictionary) and is used in grammar and logic to describe that part of a proposition which connects the subject and predicate (Oxford English Dictionary). In 1959, the word copulas was first employed in a mathematical or statistical sense by Abe Sklar in the theorem (which now bears his name) describing the functions that join together from

multivariate distribution functions to one-dimensional distribution functions (Nelsen [80]).

Nelsen [80] stated the meaning of copulas in two ways as follows:

- Copulas are functions that join or 'couple' multivariate distribution functions to their one-dimensional marginal distribution functions.
- Copulas are multivariate distribution functions whose one-dimensional marginals are uniform on the interval (0,1).

Nelsen [80] referred to Fisher answers in his article in the Encyclopedia of Statistical Sciences, copulas are interesting to study in probability and statistics for two main reasons: firstly, as a way of studying scale-free measures of dependence (copulas are flexible method that can apply to parametric, semi-parametric or non-parametric statistics); and secondly, as a starting point for constructing families of bivariate distributions. Moreover, copulas are popular in statistical applications as they allow one to easily model and estimate the distribution of random vectors by estimating marginals and copulas separately. There are many parametric copulas families available, which usually have parameters that control the strength of dependence (Joe [59]).

1.4.2 Copulas Research

The copulas literature is still interesting and expanding. In recent years, copulas have been applied in many fields of research and studies, for example, actuarial science, finance (Cherubini et al. [21]), hydrology, etc. Some of previous copulas researches can be categorized as follows:

- **actuarial science**

  In analyzing the impact of future contingent events, actuaries are faced with problems involving multivariate outcomes. Frees and Valdez [40] introduced actuaries to the concept of copulas, a tool for understanding relationships among multivariate outcomes. A copula is a function that links univariate marginals to their full multivariate distribution. Copulas were introduced in 1959 in the context of probabilistic metric spaces. The literature on the statistical properties and applications of copulas has been developing rapidly in recent years. Their article explored some of these practical applications, including estimation of joint life mortality and multidecrement models. In addition, they described basic properties of copulas, their relationships to measures of dependence, and several families of copulas that have appeared in the literature. An annotated bibliography provided a resource for researchers and practitioners who wish to continue their study of copulas. For those who wish to use copulas for statistical inference, they illustrated statistical inference procedures by using insurance company data on losses and expenses. For the data analysis, they (1) showed how

to fit copulas and (2) described their usefulness by pricing a reinsurance contract and estimating expenses for pre-specified losses. The conclusions for their article are following.

- They reviewed the problems of (1) estimating distributions of joint lifetimes of paired individuals, useful in the analysis of survivorship insurance protection, and (2) investigating mortality experience, for the actuary who needs to distinguish among causes of death.
- They introduced and provided a solution for, the problem of dependence between an insurers losses and expenses. Failures of ignoring dependencies can lead to mispricing. Thus, it is important for actuaries to be able to adequately model multivariate outcomes.
- The tool used to study multivariate outcomes is the copula function; it couples univariate marginals to the full multivariate distribution.
- The biological frailty models and the mathematical Archimedean models can motivate copulas. A statistical mixture of powers model serves as a bridge between these two sets of families. Because copulas are parametric families, standard techniques such as maximum likelihood can be used for estimation.
- Other statistical tools have been recently developed to help fit copulas. They described a graphical tool to identify the form of the copula.
- They discussed how copulas could be used to simulate multivariate outcomes, an important tool for actuaries.
- They also developed the connection between copulas and the regression function, a widely used tool for summarizing what we expect based on conditional distributions.
- Their article has focused on the connection between copulas and statistics, the theory of data. Much of the development of copulas has historically arisen from probability theory.
- To recognize this connection, they briefly reviewed topics in applied probability theory pertaining to copulas that are of the greatest interest to actuaries: stochastic ordering, fuzzy set theory, and insurance pricing.
- Copulas offer a flexible structure that can be applied in many situations. They hoped that their article encourages actuaries to seek new applications for this promising tool.

For Frees and Valdez's article, Genest et al. [45] also gave the discussion on it.

- **finance**

  Understanding and quantifying dependence are at the core of all modelling efforts in financial econometrics. The linear correlation coefficient, which is the far most used measure to test dependence in the financial community and also elsewhere, is

only a measure of linear dependence. This means that it is a meaningful measure of dependence if asset returns are well represented by an elliptical distribution. Outside the world of elliptical distributions, using the linear correlation coefficient as a measure of dependence may lead to misleading conclusions. Therefore, alternative methods for capturing co-dependency should be considered. One class of alternatives are copula-based dependence measures.

Aas [1] studied the dependence structure of financial assets by four copulas: Gaussian, Student-t, Clayton and Gumbel copulas. Two assets, the Norwegian and Nordic geometric returns were fitted by Gaussian, Clayton and a Student-t copula (since the data obviously did not exhibit greater dependence in the positive tail than in the negative, the Gumbel copula was not fitted to this data set). The parameter of Gaussian and Clayton copula were estimated to be 0.64 and 1.14. For Student-t copula, the parameters and degree of freedom were estimated to be 0.64 and 7.

Berg and Aas [12] studied models for construction of higher-dimensional dependence in finance: precipitation values and equity returns by comparison study. They reviewed two classes of structures for construction of higher-dimensional dependence; the nested Archimedean constructions (NACs) and the pair-copula constructions (PCCs). For both structures, a multivariate data set was modelled using a cascade of lower-dimensional copulae. The constructions were compared, and estimation- and simulation techniques were examined. The fit of the two constructions was tested on two different four-dimensional data sets; precipitation values and equity returns, using state of the art copula goodness-of-fit procedures. They showed that the PCCs in general are more computationally efficient than the NACs. The NACs was strongly rejected for both data sets, while the pair-copula construction provided a much better fit.

Brechmann and Czado [16] studied risk management with high-dimensional vine copulas by analysis of the Euro Stoxx 50. The aim of this study was to present the use and usefulness of vine copulas in financial risk management. They developed a flexible R-vine based factor model for stock market dependencies, the regular vine market sector (RVMS) model, and discussed passive and active portfolio management using vine copula models. The developed methodology was used to analyze the dependence structure among important European stocks as represented in the Euro Stoxx 50 index. In these analyses, their models were critically compared to relevant benchmark models such as the dynamic conditional correlation (DCC) model and the state-of-the art dependency model, the Student-t copula. It turned out that vine copula models provide good fits of the data and accurately and efficiently forecast the Value-at-Risk at the high levels, as they are frequently used in practice. Similarly, active portfolio management can benefit from the more accurate assessment of tail risk using vine copulas.

Brechmann et al. [17] studied truncated regular vines in high dimensions with applications to financial data. They considered the problem of determining whether

R-vine copulas can be pairwisely truncated or alternatively, simplified with Gaussian pair-copulas, after a certain tree. In extensive simulations different procedures for truncation and simplification were proposed and evaluated. The results showed that Vuong test based procedures performed particularly well. They also considered truncating or simplifying the special case of a C-vine copula. In this case, the remaining dependencies may be captured by a multivariate copula; the independence copula for the truncation alternative and the Gaussian copula for the simplification one. Therefore, simplification/truncation levels may be determined using a multivariate copula goodness-of-fit-test. However, simulations showed that their procedures developed for the general R-vine copula overall seemed to detect the simplification/truncation levels more accurately than the multivariate goodness-of-fit-tests. They investigated whether it is possible to simplify or truncate the R-vine copula specification corresponding to a 19-dimensional data set consisting of Norwegian and international market variables. This study showed that the most important dependencies in the Norwegian data set are captured in the first 4-6 trees, meaning that the corresponding R-vine copula may be truncated at level 6, or even at level 4. Moreover, simplification at level 2 seemed to be appropriate, indicating that all important (asymmetric) tail dependencies are captured in the first two trees.

Chollete et al. [23] studied modeling international financial returns with a multivariate regime switching copula. They provided further evidence on asymmetric dependence in international financial returns by estimating a multivariate regime-switching model of copulas for the dependence of the stock indices in the G5 (Germany, France, the UK, the US and Japan) and four Latin American countries (Brazil, Mexico, Argentina and Chile). They used regime switching copulas, which allowed them to model dependence in a much more flexible and realistic way than previously-suggested switching models based on the Gaussian copula. They modelled dependence with one Gaussian and one canonical vine copula regime. Canonical vines are constructed from bivariate conditional copulas and provide a very flexible way of characterizing dependence in multivariate settings. For this study, there are two main findings: (1) they discovered that models with canonical vines generally dominate alternative dependence structures (2) the choice of copula is important for risk management, because it modifies the Value at Risk (VaR) of international portfolio returns.

Czado et al. [26] studied maximum likelihood estimation of mixed C-vines with application to exchange rates. For this study, they introduced the class of mixed C-vine copulas and provided sequential and maximum likelihood (ML) estimation procedures for the unknown parameters. Mixed C-vines allow the variables to be ordered according to their influence. Vines are build from bivariate copulas only and the term mixed refers to allowing the pair-copula family to be chosen individually for each term. Two extensive simulation studies showed very satisfactory behavior of the ML estimation for many different mixed and non-mixed C-vine copulas. In addition there are many C-vine structure specifications possible and

they proposed a novel data driven sequential selection procedure, which selects both the C-vine structure and its attached pair-copula families with parameters. For the selection of the appropriate pair-copula families they followed standard test approaches involving goodness-of-fit tests for bivariate copulas, Vuong and Clarke tests suitable for non-nested models and finally explorative tools based on scatter and contour plots. An extensive simulation study showed a satisfactory performance of ML estimates in small samples. Finally an application involving US-exchange rates demonstrated the need for mixed C-vine models.

De Melo Mendes et al. [28] studied pair-copulas modeling in finance. They explored the potentials of pair-copulas modeling using dependent financial data. A fully flexible multivariate distribution was obtained by combining univariate fits and D-vines. They gave a broad view of the problem of modeling multivariate financial log-returns using pair-copulas, gathering theoretical and computational results scattered among many papers on canonical vines. They showed to the practitioner the advantages of modeling through pair-copulas and sent the message that this is a possible methodology to be implemented in a daily basis. Moreover, all steps (model selection, estimation, validation, simulations and applications) were given in a level reached by all data analysts.

Dißmann et al. [30] studied selecting and estimating regular vine copula and application to financial returns. This study provided a significant contribution towards making R-vine copulae a standard building block for copula based models. They provided a general selection approach to sequentially choose the tree representation together with choosing the copula type for each copula term from a large class of bivariate copula families and estimated the corresponding parameters. The selection procedure was completely operational, it was implemented in the statistical software R and was capable to handle medium sized dimensions of up to 20 dimensions. This comprehensive search strategy was evaluated in a large simulation study and applied to a 16-dimensional financial data set of international equity, fixed income and commodity indices which were observed over the last decade, in particular during the recent financial crisis. The analysis provided economically well interpretable results and interesting insights into the dependence structure among these indices.

Embrechts et al. [33] studied modelling dependence with copulas and applications to risk management. They introduced and reviewed the theory of copulas, dependence concepts and three classes of copulas: Marshall-Olkin, elliptical and archimedean copulas. For the applications, they gave some examples for modelling external events in practice for insurance and market risks. For example, the problem of measuring the risk of holding an equity portfolio over a short time horizon (one day, say) without the possibility of rebalancing. Moreover, there are numerous alternative applications of copula techniques to integrated risk management.

Fischer et al. [37] studied an empirical analysis of multivariate copula models

in finance. Whereas copulas are well-studied in the bivariate case, the higher-dimensional case still offers several open issues and it is far from clear how to construct copulas which sufficiently capture the characteristics of financial returns. For this reason, elliptical copulas (i.e. Gaussian and Student-t copula) still dominate both empirical and practical applications. The aim of this work was to empirically investigate whether these models are really capable of outperforming its benchmark, i.e. the Student-t copula and, in addition, to compare the fit of these different copula classes among themselves.

- **hydrology**

  Favre et al. [34] studied multivariate hydrological frequency analysis using copulas. They proposed an approach based on copulas applied to bivariate frequency analysis that has not been used in hydrology. The approach allowed them to model the dependence structure independently of the marginal distributions, which is not possible with standard classical methods. The model was applied to two different problems in hydrology: flow combination and joint modeling of flow and volume. The copulas were applied on two different problems in hydrology. The first application was concerned with the combined risk in the framework of frequency analysis. Four copulas: independence, Farlie-Gumbel-Morgenstern, Clayton and Frank copulas were tested on peak flows from the watershed of Peribonka in Québec, Canada. The second application related to the joint modeling of peak flows and volumes. Three copulas: independence, Clayton and Frank copulas were applied to the watershed of the Rimouski River in Québec, Canada. The conclusions for this study are following.

  - In the first case the measures of correlation were low and even in this case the difference between the independence case and the Frank copula is of 5.5%. If the correlation would be higher they believed that the difference would be significantly increased. Differences between the copulas should increase as well.
  - The second application emphasized on the conditional return probability of the flow given the volume. The obtained bivariate probabilities were more precise, which is a serious gain for water resources managers.
  - The approach using copulas was promising since it allowed them to take into account a wide range of correlation which can happen in hydrology. In fact the classical multivariate models can not reproduce all type of correlations. Moreover, the standard models are limited, especially because the choice of the marginal distributions is restricted.

From the copulas research reviews, copulas were applied to many fields of research. Copulas are a method of multivariate statistics for measure of dependence, in particular, the recent developments of copulas, the approach expresses n-dimensional distribution function as functions of its univariate marginals. Clearly, copulas are a useful tool to

study the dependence problem. Therefore, copulas will be mainly studied in this thesis and applied to study the relationship between the dynamics of the small protein and water molecules and also collective long space- and time-range correlations for high dimensions in molecular trajectories.

1.5 Summary

In this chapter, we have introduced and given a short literature review of proteins and copulas. We noted that protein-water molecular system dynamics have been studied for many years. Most of the previous protein-water studies were to investigate the behaviour between protein motion and water. Moreover, there is a close connection between the water dynamics and the protein conformations. In this thesis, we will study the statistical dependencies of protein-water molecular system dynamics using copulas and correlation analysis. Thesis objectives, thesis novelty and thesis outline are described in the next three subsections.

1.5.1 Thesis Objectives

Our interest in this thesis is to study statistical properties of liquid protein-water molecular system dynamics. The thesis objectives are:

1. to study time correlations of liquid protein-water molecular system dynamics,
2. to study conditional correlations of liquid protein-water molecular system dynamics on selected points,
3. to study conditional correlations of liquid protein-water molecular system dynamics on grid points,
4. to compare un- and conditional correlations of liquid protein-water molecular system dynamics.

1.5.2 Thesis Novelty

From the previous protein research, there is no research which has been done to study statistical properties of liquid protein-water molecular system dynamics using copulas and correlation analysis. In this thesis, we will obtain new research results of the statistical dependencies: time correlations, conditional correlations of liquid protein-water molecular system dynamics on selected points and grid points in space. Moreover, comparing between un- and conditional correlations of liquid protein-water molecular system dynamics will be obtained. This will be advantageous to future research.

1.5.3 Thesis Outline

In this thesis, we focus on and apply copulas and correlation analysis to study statistical properties of liquid protein-water molecular system dynamics. In the second chapter, we describe and review theory of copulas, vines, correlation, conditional correlation and comparing correlations.

In the third chapter, we apply copulas to test systems which are sample data from molecular dynamics (MD) simulation. We do the analysis both N-variate copulas and D-vine for angle and amplitude dataset in two and five dimensions. Moreover, the test systems results and summary are also given.

In the fourth chapter, the main results of this thesis are presented. We calculate Pearson product-moment, Spearman's rank and Kendall's tau correlation for liquid protein-water molecular system dynamics. Also, conditional correlations on selected points, for example, middle points in time and closed point to peptide. Furthermore, comparing correlations between un- and conditional correlations of all grid points in space are calculated. The last chapter presents thesis conclusions and suggestions for future research.

2

Theory of Copulas and Correlation

In Chapter 1, we introduced copulas and reviewed some of previous research in copulas. In this chapter, we will describe the theory of copulas: definition and theorem and pair-copulas constructions of multiple dependence in section 1. In section 2, we will also describe a graphical tool for copulas, statistical software and graphical analysis. Furthermore, the correlations, conditional correlations and comparing correlations will be proposed in section 3, 4 and 5 respectively.

2.1 Theory of Copulas

In this section, we describe definition, theorem of copulas both two and d-dimensional copulas and pair-copulas constructions following (Marshall [79], Sklar [100], Trivedi and Zimmer [107], Valle [108]).

2.1.1 Two-dimensional Copula

Let $F(x) = P[X \leq x]$ and $G(y) = P[Y \leq y]$ be marginal distribution functions and $H(x, y) = P[X \leq x, Y \leq y]$ be joint distribution function of random variables X and Y. For each pair (x, y), we can associate three numbers: $F(x)$, $G(y)$ and $H(x, y)$.

Definition 2.1. If (X, Y) is a pair of continuous random variables with distribution function $H(x, y)$ and marginal distributions $F(x)$ and $G(y)$ respectively, then $U = F(x)$ $\sim$ U(0,1) and $V = G(y) \sim$ U(0,1). A two-dimensional copula is a distribution function of (U,V) or $C(u, v)$ on $[0, 1]\times[0, 1]$ with standard uniform marginal distributions with the following properties: C: $[0, 1]^2 \rightarrow$ [0,1]

1. $C(u, 0) = C(0, v) = 0$.
2. $C(u, 1) =$ u and $C(1, v) =$ v.

3. If v_1, v_2, u_1, $u_2 \in [0,1]$; $u_2 \geq u_1$, $v_2 \geq v_1$, then $C(u_2, v_2)$ - $C(u_1, v_2)$ - $C(u_2, v_1)$ + $C(u_1, v_1) \geq 0$.

Remark. A two-dimensional independent copula is a function C with the property $C(u_1, u_2) = u_1 \times u_2, \mathbf{u} \in [0,1]^2$.

Theorem 2.1 (Sklar's Theorem)**.** Let H be a joint distribution function with marginal distribution functions F and G, then there exists a copula C such that

$$H(u, v) = C(F(u), G(v)).$$

Remark. If F and G are continuous, then the copula is unique (Nelsen [80]).

2.1.2 d-dimensional Copulas

Definition 2.2. A d-dimensional copulas $C : [0,1]^d \to [0,1]$ is a function which is a distribution function with uniform marginals.

In general, the notation $C(u) = C(u_1, \ldots, u_d)$ will always be used for a copulas. The condition that C is a distribution function immediately leads to the following properties: (Gijbels and Mielniczuk [53])

- As distribution functions are always increasing, $C(u_1, \ldots, u_d)$ is increasing in each component u_i.
- The marginal in component i is obtained by setting $u_j = 1$ for all $j \neq i$ and as it must be uniformly distributed,

$$C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i.$$

Theorem 2.2 (Sklar's Theorem)**.** Consider a d-dimensional distribution function F with marginals $F_1, \ldots, F_d$. There exists a copula C, such that

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \ ,$$

for all x_i in (-∞, ∞), $i = 1, \ldots, d$. If F_i is continuous for all $i = 1, \ldots, d$ then C is unique.

Remark. The **d-dimensional independent copulas** are the copulas of d independent uniform (0,1) random variables. It equals

$$C^{ind}(u_1, \ldots, u_d) = u_1 \cdots u_d$$

and has a density that is uniform on $[0,1]^d$, that is, its density is $C^{ind}(u_1, \ldots, u_d) = 1$ on $[0,1]^d$.

2.1.3 Copulas Families

The well-known families of copulas are described in this section following.

2.1.3.1 Gaussian Copulas

Multivariate normal distributions offer a convenient way to generate families of copulas. Let $X = (x_1,\ldots,x_d)$ have a multivariate normal distribution. Since C_X depends only on the dependencies within X, not the univariate marginal distributions, C_X depends only on the correlation matrix of X, which will be denoted by Ω. Therefore, there is a one-to-one correspondence between correlation matrices and Gaussian copulas. The Gaussian copulas with correlation matrix Ω will be denoted by $C_{\Omega}^{Gauss}(u)$ and can be written as:

$$C_{\Omega}^{Gauss}(u) = P(\Phi(X_1) \leq u_1,\ldots,\Phi(X_d) \leq u_d) = \Phi_{\Omega}^{d}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_d)),$$

where Φ is the standard univariate normal distribution, Φ_{Ω}^{d} is the joint distribution function of the d-variate standard normal distribution function with the correlation matrix Ω, $\Omega \in$ [-1, 1]$^{d\times d}$ and Φ^{-1} denotes the inverse of the distribution function of the standard univariate normal distribution.

For the bivariate Gaussian copulas can be written as:

$$C_{\rho}^{Gauss}(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} exp\Big\{ -\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \Big\} dxdy,$$

where ρ is the linear correlation coefficient of the corresponding bivariate normal distribution (the parameter of the copulas) and Φ^{-1} is the inverse of the standard univariate Gaussian (normal) distribution function.

Remarks.

1. A d-dimensional Gaussian copulas whose correlation matrix is the identity matrix, so that all correlations are zero, is called the **d-dimensional independence copulas.**
2. The copulas that have perfect positive dependence will be called the **co-monotonicity copulas**, as the copulas that have perfect negative dependence will be called the **counter-monotonicity copulas.**
3. A Gaussian copulas will converge to the co-monotonicity copulas if all correlations converge to 1, as the correlation converges to -1, the copulas converge to the counter-monotonicity copulas.

2.1.3.2 Student-t Copulas

Similarly, let $C^t(u \mid \nu,\Omega)$ be the copulas of a multivariate Student-t distribution with correlation matrix Ω and degrees of freedom ν. The degrees of freedom ν, is a shape parameter that affects both the univariate marginal distributions and the copulas, so ν

is a parameter of the Student-t copulas. The Student-t copulas with correlation matrix Ω will be denoted by $C^t_{\nu,\Omega}(u)$ and can be written as: (Demarta and McNeil [27])

$$C^t_{\nu,\Omega}(u) = t^d_{\nu,\Omega}(t^{-1}_\nu(u_1), \ldots, t^{-1}_\nu(u_d)),$$

where $t^d_{\nu\Omega}$ is the joint distribution function of the d-variate Student-t distribution function with the degrees of freedom ν and correlation matrix Ω, $\Omega \in [-1,1]^{d\times d}$ and t^{-1}_ν denotes the inverse of the distribution function of the univariate Student-t distribution with the degrees of freedom ν.

For the bivariate Student-t copulas can be written as:

$$C^t_{\nu\rho}(u,v) = \int_{-\infty}^{t^{-1}_\nu(u)} \int_{-\infty}^{t^{-1}_\nu(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \left\{1 + \frac{x^2 - 2\rho xy + y^2}{\nu(1-\rho^2)}\right\}^{-(\nu+2)/2} dxdy,$$

where ρ is the linear correlation coefficient of the corresponding bivariate Student-t distribution with $\nu > 2$, that means that ρ and ν are the parameter of the Student-t copulas and t^{-1}_ν is the inverse of the univariate Student-t distribution with degrees of freedom ν.

2.1.3.3 Archimedean Copulas

Archimedean copulas are an alternative families of copulas. Archimedean copulas are popular because they allow modeling dependence in arbitrarily high dimensions with only one parameter that is the strength of dependence. Generally, Archimedean copulas with a strict generator has the form (Nelsen [80])

$$C(u_1, \ldots, u_d) = \phi^{-1}\{\phi(u_1) + \ldots + \phi(u_d)\}, \tag{2.1}$$

where the function ϕ is the generator of the copulas and satisfies

1. ϕ is a continuous, strictly decreasing, and convex function mapping [0,1] onto [0,∞],
2. $\phi(0) = \infty$ and
3. $\phi(1) = 0$.

There are many families of Archimedean copulas, but we will only focus at three, the Frank, Clayton and Gumbel copulas. Moreover, Archimedean copulas are most useful in the bivariate case or in applications where we expect all pairs to have similar dependencies (Genest and MacKay [44]).

- **Frank copulas**

  The Frank copulas have generator

$$\phi^{Fr}(u) = -log\left\{\frac{e^{-\theta u}-1}{e^{-\theta}-1}\right\}, -\infty < \theta < \infty,$$

and the inverse generator is

$$(\phi^{Fr})^{-1}(y) = -\frac{log[e^{-y}(e^{-\theta}-1)+1]}{\theta}.$$

Therefore, by (2.1), the bivariate Frank copulas are (Genest [41])

$$C^{Fr}(u_1,u_2) = -\frac{1}{\theta}log\left\{1+\frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right\}, \tag{2.2}$$

where θ is the Frank copulas parameter. The case $\theta = 0$ requires some care, since plugging this value into (2.2) gives 0/0. Instead, one must evaluate the limit of (2.2) as $\theta \to 0$. Using the approximations $e^x - 1 \approx x$ and $\log(1 + x) \approx x$ as $x \to 0$, one can show that as $\theta \to 0$, $C^{Fr}(u_1, u_2) \to u_1u_2$, this is the bivariate independence copulas. Therefore, for $\theta = 0$ we define the Frank copulas to be the independence copulas.

- **Clayton copulas**

  The Clayton copulas, with generator $(t^{-\theta} - 1)/\theta, \theta > 0$ are

  $$C^{Cl}(u_1,\ldots,u_d) = (u_1^{-\theta} + \ldots + u_d^{-\theta} - d + 1)^{-1/\theta}.$$

  We define the Clayton copulas for $\theta = 0$ as the limit

  $$\lim_{\theta\to 0} C^{Cl}(u_1,\ldots,u_d) = u_1 \cdots u_d,$$

  which is the independence copulas.

  As $\theta \to$ -1, the bivariate Clayton copulas converge to the counter-monotonicity copulas (perfect negative dependence), and as $\theta \to \infty$, the Clayton copulas converge to the co-monotonicity copulas (perfect positive dependence).

- **Gumbel copulas**

  The Gumbel copulas have generator $[-log(t)]^\theta$, $\theta \geq 1$, and consequently is equal to

  $$C^{Gu}(u_1,\ldots,u_d) = exp[-\{(logu_1)^\theta + \ldots + (logu_d)^\theta\}^{1/\theta}].$$

  The Gumbel copulas are the independence copulas when $\theta = 1$ and converge to the co-monotonicity copulas (perfect positive dependence) as $\theta \to \infty$, but the Gumbel copulas cannot have negative dependence.

Clearly, copulas are one of interesting and convenient statistical method to express joint distributions in two parts: the marginal distributions of the individual variables and the interdependency of the probabilities (Venter [109]). Moreover, copulas are flexible method that can apply to parametric, semi-parametric or non-parametric statistics, for example, Capéraá et al. [20] applied copulas to a nonparametric estimation procedure for bivariate extreme value copulas.

2.1.4 Pair-Copulas Construction

Aas et al. [2] pointed out that pair-copulas is a method to model multivariate data using a cascade of simple building blocks. The principle is to model dependency using simple local building blocks based on conditional dependence. However, building higher-dimensional copulas is generally recognized as a difficult problem because there are a large number of parametric bivariate copulas to estimate, but the set of higher-dimensional copulas is rather limited. There have been some attempts to construct multivariate extensions of Archimedean bivariate copulas. Fortunately, the pair-copulas decomposition is a very flexible way to construct higher-dimensional copulas (Schirmacher and Schirmacher [95], Czado [25]).

We consider a pair-copulas decomposition of a general multivariate distribution. Let $X = (X_1,\ldots,X_n)$ be a vector of random variables with a joint density function $f(x_1,\ldots,x_n)$. This density can be factorized as (Hobæk Haff et al. [57], Hobæk Haff [56])

$$f(x_1,\ldots,x_n) = f_n(x_n)\cdot f(x_{n-1} \mid x_n)\cdot f(x_{n-2} \mid x_{n-1},x_n)\ldots f(x_1 \mid x_2,\ldots,x_n) \quad (2.3)$$

and this decomposition is unique up to a re-labelling of the variables.

From Theorem 2.2, we found that every multivariate distribution F with marginals $F_1,\ldots,F_n$ can be written as

$$F(x_1,\ldots,x_n) = C(F_1(x_1),\ldots,F_n(x_n)), \quad (2.4)$$

for some appropriate n-dimensional copulas C. In fact, the copulas from (2.4) have the expression

$$C(u_1,\ldots,u_n) = F\{F_1^{-1}(u_1),\ldots,F_n^{-1}(u_n)\},$$

where the $F_i^{-1}(u_i)$'s are the inverse distribution function of the marginals.

Passing to the joint density function f, for an absolutely continuous F with strictly increasing, continuous marginal densities $F_1,\ldots,F_n$ using the chain rule we have

$$f(x_1,\ldots,x_n) = c_{1\ldots n}\{F_1(x_1),\ldots,F_n(x_n)\}\cdot f_1(x_1)\ldots f_n(x_n), \quad (2.5)$$

for some (uniquely identified) n-variate copulas density $c_{1\ldots n}(\cdot)$.

For the bivariate case, from (2.5), we simplify to

$$f(x_1, x_2) = c_{12}\{F_1(x_1), F_2(x_2)\} \cdot f_1(x_1) \cdot f_n(x_n),$$

where $c_{12}(\cdot, \cdot)$ is the appropriate pair-copulas density for the pair of transformed variables $F_1(x_1)$ and $F_2(x_2)$.

2.2 Vines

Copulas are functions that join or 'couple' multivariate distribution functions to their one-dimensional marginal distribution functions. However, standard multivariate copulas can become inflexible in high dimensions and do not allow for different dependency structures between pairs of variables. For high-dimensional distributions, there are a significant number of possible pair-copulas constructions or we can say that the pair-copulas decomposition is a very flexible way to construct higher-dimensional copulas. Aas et al. [2] suggested a graphical model denoted as the regular vine. The class of regular vines is still very general and embraces a large number of possible pair-copulas decompositions. There are two special cases of regular vines; the **canonical vine** and the **D-vine**. Each model gives a specific way of decomposing the density (Czado et al. [26]). The dependency structure is determined by the bivariate copulas and a nested set of trees. Vines approach is more flexible, as we can select bivariate copulas from a wide range of (parametric) families (Bedford and Cooke [11]). For this research, we are interested in the D-vine and will apply the D-vine to liquid protein-water molecular system dynamics.

Given a d-dimensional density, we can decompose it into products of marginal densities and bivariate copulas densities. Vines represent the decomposition with nested set of trees that fulfill a proximity condition. A d-dimensional regular vine is a sequence of d - 1 trees which consists of (Kurowicka and Joe [71])

1. **tree 1**
   - d nodes: $X_1,\ldots,X_d$
   - (d - 1) edges: pair-copulas densities between $X_1,\ldots,X_d$
2. **tree** j
   - (d + 1 - j) nodes: edges of tree (j - 1)
   - ($d - j$) edges: conditional pair-copulas densities

Proximity condition: If two nodes in tree (j + 1) are joined by an edge, the corresponding edges in tree j share a node.

The general expression for the three-dimensional D-vine structures is

$$\begin{aligned} f(x_1, x_2, x_3) = {} & f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3) \\ & \cdot c_{12}\{F_1(x_1), F_2(x_2)\} \cdot c_{23}\{F_2(x_2), F_3(x_3)\} \\ & \cdot c_{13|2}\{F(x_1 \mid x_2), F(x_3 \mid x_2)\}. \end{aligned} \tag{2.6}$$

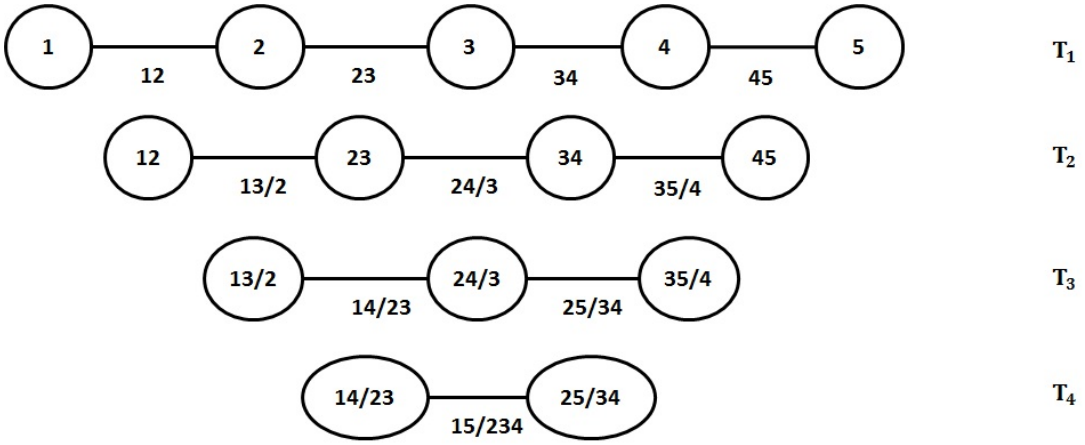

Figure 2.1: D-vine - The figure shows the five-dimensional D-vine.

The general expression for the four-dimensional D-vine structures is (Bedford and Cooke [10])

$$\begin{aligned} f(x_1,x_2,x_3,x_4) &= f_1(x_1)\cdot f_2(x_2)\cdot f_3(x_3)\cdot f_4(x_4) \\ &\cdot c_{12}\{F_1(x_1),F_2(x_2)\}\cdot c_{23}\{F_2(x_2),F_3(x_3)\}\cdot c_{34}\{F_3(x_3),F_4(x_4)\} \\ &\cdot c_{13|2}\{F(x_1\mid x_2),F(x_3\mid x_2)\}\cdot c_{24|3}\{F(x_2\mid x_3),F(x_4\mid x_3)\} \\ &\cdot c_{14|23}\{F(x_1\mid x_2,x_3),F(x_4\mid x_2,x_3)\}. \end{aligned} \tag{2.7}$$

The general expression for the five-dimensinal D-vine structure is

$$\begin{aligned} f(x_1,x_2,x_3,x_4,x_5) &= f_1(x_1)\cdot f_2(x_2)\cdot f_3(x_3)\cdot f_4(x_4)\cdot f_5(x_5) \\ &\cdot c_{12}\{F_1(x_1),F_2(x_2)\}\cdot c_{23}\{F_2(x_2),F_3(x_3)\} \\ &\cdot c_{34}\{F_3(x_3),F_4(x_4)\}\cdot c_{45}\{F_4(x_4),F_5(x_5)\} \\ &\cdot c_{13|2}\{F(x_1\mid x_2),F(x_3\mid x_2)\}\cdot c_{24|3}\{F(x_2\mid x_3),F(x_4\mid x_3)\} \\ &\cdot c_{35|4}\{F(x_3\mid x_4),F(x_5\mid x_4)\} \\ &\cdot c_{14|23}\{F(x_1\mid x_2,x_3),F(x_4\mid x_2,x_3)\} \\ &\cdot c_{25|34}\{F(x_2\mid x_3,x_4),F(x_5\mid x_3,x_4)\} \\ &\cdot c_{15|234}\{F(x_1\mid x_2,x_3,x_4),F(x_5\mid x_2,x_3,x_4)\}. \end{aligned} \tag{2.8}$$

Figure 2.1 shows the specification corresponding to the five-dimensional D-vine. It consists of four trees T_j, $j = 1, \ldots, 4$. Tree T_j has 6 - j nodes and 5 - j edges. Each edge corresponds to a pair-copulas density and the edge label corresponds to the subscript of the pair-copulas density, for example, edge 14|23 corresponds to the copulas density $c_{14|23}(\cdot)$. The whole decomposition is defined by the n(n - 1)/2 edges and the marginal densities of each variable.

2.2.1 Statistical Software for Copulas

In this research, copulas and CDVine R package are applied to analyze datasets. Package copulas provides a carefully designed and easily extensible platform for multivariate modeling with copulas in R. Also, package CDVine provides functions and tools for statistical inference of canonical vine (C-vine) and D-vine copulas (Brechmann and Schepsmeier [18]). For example, Figure 2.2 shows the workflow and provided functionality in the package CDVine. Both packages can be easily extended by user-defined copulas and margins to solve problem (Yan [111]).

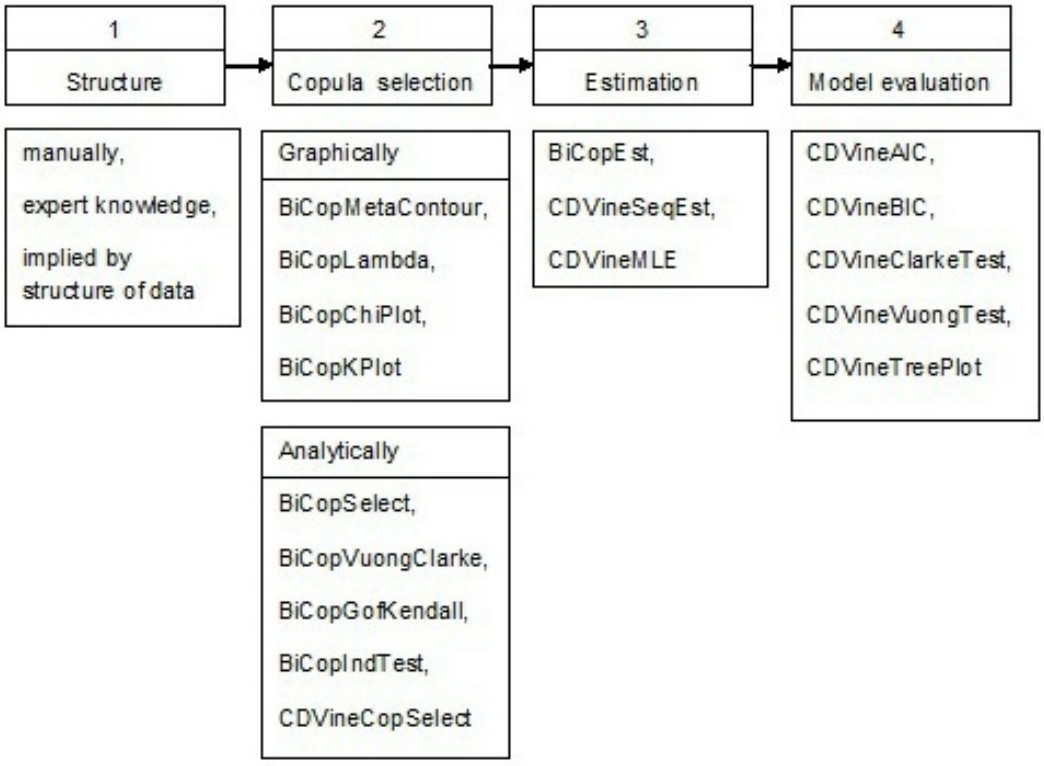

Figure 2.2: Workflow - The figure shows the proposed data analysis and model building workflow and provided functionality in the package CDVine (Source: Brechmann and Schepsmeier [18], p.6).

Brechmann and Schepsmeier [18] stated that C- and D-vine copulas are based on bivariate copulas as same as building blocks. CDVine package has varieties of tools for bivariate data analysis and inference of bivariate copulas families. The main idea for copulas analysis is the data which we work with has approximately standard uniform distribution or uniform on [0,1], that we call **copulas data**.

We can classify the data analysis for CD-vine package in 2 groups as follows: (Schepsmeier [93], Schepsmeier and Brechmann [94])

1. **Bivariate data analysis methods** include many useful tools:
   - bivariate copulas families,
   - tools for bivariate exploratory data analysis - graphical tools and analytical tools, and
   - estimation of bivariate copulas families.
2. **Statistical inference of C- and D-vine copulas** include many useful tools:
   - specification of C- and D-vine copulas models and data simulation,
   - estimation,
   - selection among vines copulas models and
   - implementation and numerical issues.

Table 2.1: Some copulas families, generator and parameter range included in CDVine.

Family	Generator	Parameter range
Gaussian	-	(-1,1)
Student-t	-	(-1,1), $\nu > 2$
Clayton	$\frac{1}{\theta}(t^{-\theta} - 1)$	(0,∞)
Gumbel	$(-log\, t)^{\theta}$	[1,∞)
Frank	$-log[\frac{e^{-\theta t}-1}{e^{-\theta}-1}]$	$\mathbf{R} \setminus \{0\}$
Joe	$-log[1-(1-t)^{\theta}]$	(1,∞)

CDVine package is a useful package and has many commands to analyze data and check statistical properties for pair-copulas construction. For bivariate data analysis in CDVine package, there are important graphical tools for pair-copulas which are used for test of dependence. Three graphical tools for test of dependence in CDVine package are very useful and will be described in next subsection.

2.2.2 Graphical Analysis for Test of Dependence

For D-vine, graphical analysis is one of determination methods for bivariate copulas data in pair-copulas construction. There are three graphical tools for detecting dependence: Chi-plot, K-plot and Lambda function plot following (Anscombe [5]).

1. **Chi-plot**

   Chi-plot was originally proposed by Fisher and Switzer [38] and more fully illustrated by Fisher and Switzer [39]. Chi-plot is based on the Chi-square statistic for independence in a two-way table (Genest and Favre [43]).

   For observations $u_{i,j}, i = 1, \ldots, n, j = 1, 2$, where $u_{i,j} \sim$ (0,1), the Chi-plot is based on the following two quantities: (Schepsmeier and Brechmann [94])

the chi-statistics

$$\chi_i = \frac{\hat{F}_{U_1U_2}(u_{i,1}, u_{i,2}) - \hat{F}_{U_1}(u_{i,1})\hat{F}_{U_2}(u_{i,2})}{\sqrt{\hat{F}_{U_1}(u_{i,1})(1 - \hat{F}_{U_1}(u_{i,1}))\hat{F}_{U_2}(u_{i,2})(1 - \hat{F}_{U_2}(u_{i,2}))}}$$

and the lambda-statistics

$$\lambda_i = 4sgn(\tilde{F}_{U_1}(u_{i,1}), \tilde{F}_{U_2}(u_{i,2})) \cdot \max(\tilde{F}_{U_1}(u_{i,1})^2, \tilde{F}_{U_2}(u_{i,2})^2),$$

where $\hat{F}_{U_1}$, $\hat{F}_{U_2}$ and $\hat{F}_{U_1U_2}$ are the empirical distribution functions of the uniform random variables U_1 and U_2 and of (U_1, U_2), respectively. Further, $\tilde{F}_{U_1} = \hat{F}_{U_1} - 0.5$ and $\tilde{F}_{U_2} = \hat{F}_{U_2} - 0.5$.

These quantities only depend on the ranks of the data and are scaled to the interval (0,1). λ_i measures a distance of a data point $(u_{i,1}, u_{i,2})$ to the center of the bivariate data set, while χ_i corresponds to a correlation coefficient between dichotomized values of U_1 and U_2. Under independence it holds that $\chi_i \sim N(0, 1/n)$ and $\lambda_i \sim U(-1, 1)$ asymptotically, i.e., values of χ_i close to zero indicate independence corresponding to $F_{U_1U_2} = F_{U_1}F_{U_2}$.

When plotting these quantities, the pairs of (λ_i, χ_i) will tend to be located above zero for positively dependent margins and vice versa for negatively dependent margins. Control bounds around zero indicate whether there is significant dependence present. The examples of Chi-plot for positive dependence, independence and negative dependence are illustrated in Figure 2.3.

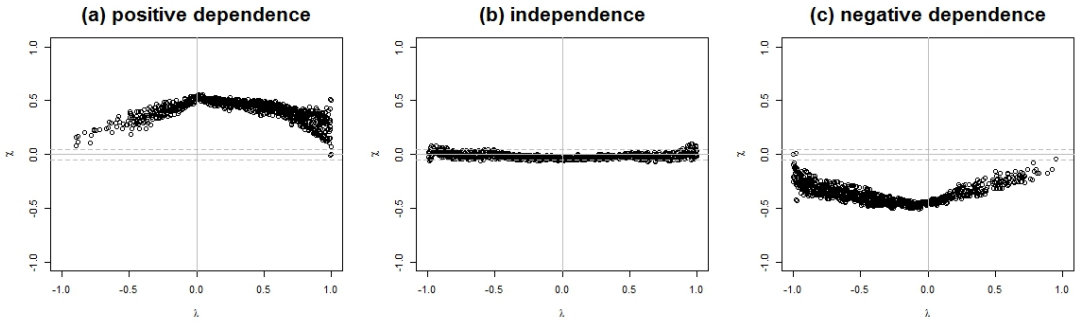

Figure 2.3: Chi-plot for positive dependence, independence and negative dependence - The figure shows three types of Chi-plot (a) positive dependence (b) independence (c) negative dependence.

2. **Kendall's Tau Plot (K-plot)**

   Kendall's Tau Plot (K-plot) is a rank-based graphical tool for visualizing dependence that was proposed by Genest and Boies [42] and inspired by the notion of QQ-plot (Genest and Favre [43]).

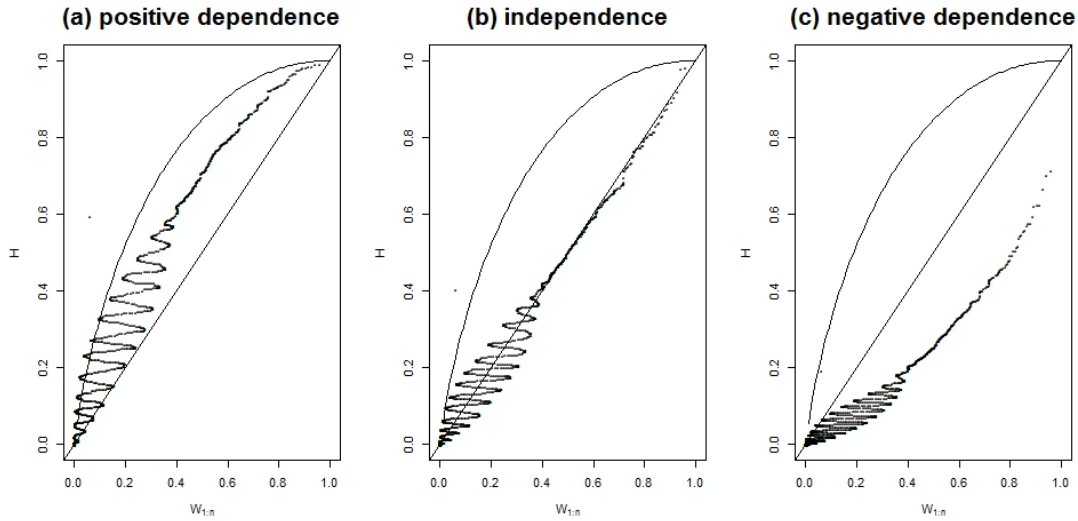

Figure 2.4: Kendall's tau plot (K-plot) for positive dependence, independence and negative dependence - The figure shows three types of Kendall's tau-plot (K-plot) (a) positive dependence (b) independence (c) negative dependence.

For observations $u_{i,j}, i = 1, \ldots, n, j = 1, 2$, where $u_{i,j} \sim (0,1)$, the K-plot considers two quantities: (Schepsmeier and Brechmann [94])

First, the ordered values of the empirical bivariate distribution function $H_i := \hat{F}_{U_1U_2}(u_{i,1}, u_{i,2})$ and, second, $W_{i:n}$, which are the expected values of the order statistics from a random sample of size n of the random variable $W = C(U_1, U_2)$ under the null hypothesis of independence between U_1 and U_2. $W_{i:n}$ can be calculated as follows

$$W_{i:n} = n\binom{n-1}{i-1}\int_0^1 \omega k_0(\omega)(K_0(\omega))^{i-1}(1-K_0(\omega))^{n-i}d\omega,$$

where

$$K_0(\omega) = \omega - \omega log(\omega),$$

and $k_0(\cdot)$ is the corresponding density.

K-plot can be seen as the bivariate copula equivalent to QQ-plots. If the points of a K-plot lie approximately on the diagonal $y = x$, then U_1 and U_2 are approximately independent. Any deviation from the diagonal line points towards dependence. In case of positive dependence, the points of the K-plot should be located above the diagonal line, and vice versa for negative dependence. The larger the deviation from the diagonal, the stronger is the degree of dependency. There is a perfect positive dependence if points $(W_{i:n}, H_i)$ lie on the curve $K_0(\omega)$

located above the main diagonal. If points $(W_{i:n}, H_i)$ however lie on the x-axis, this indicates a perfect negative dependence between U_1 and U_2. The examples of K-plot for positive dependence, independence and negative dependence are illustrated in Figure 2.4.

3. **lambda function plot**

   The λ-function is another graphical tool for visualizing dependence and is characteristic for each bivariate copulas family and defined by Kendall's distribution function K: (Schepsmeier and Brechmann [94])

$$\lambda(\upsilon, \theta) := \upsilon - K(\upsilon, \theta)$$

   with

$$K(\upsilon, \theta) := P(C_\theta(U_1, U_2) \leq \upsilon),\ \upsilon \in (0, 1).$$

   For Archimedean copulas one has the following closed form expression in terms of the generator function φ of the copulas C_θ:

$$\lambda(\upsilon, \theta) = \frac{\varphi(\upsilon)}{\varphi'(\upsilon)},$$

   where φ' is the derivative of φ. For more details see Genest and Rivest [48] or Schepsmeier [93].

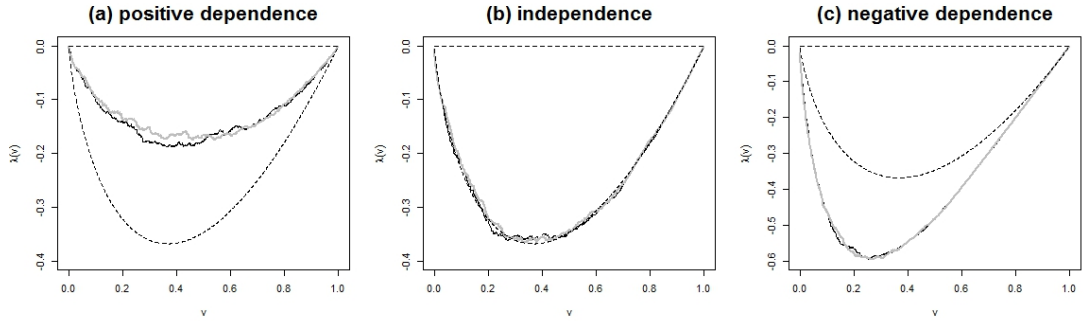

Figure 2.5: Lambda function plot for positive dependence, independence and negative dependence - The figure shows three types of lambda function plot (a) positive dependence (b) independence (c) negative dependence.

   For the bivariate Gaussian and t-copulas no closed form expression for the theoretical λ-function exists. Therefore it is simulated based on samples of size 1000. For all other implemented copulas families there are closed form expressions available. The plot of the theoretical λ-function also shows the limits of the λ-function corresponding to Kendall's Tau = 0 and Kendall's Tau = 1 ($\lambda = 0$). The examples of lambda function plot for positive dependence, independence and negative dependence are illustrated in Figure 2.5.

2.3 Correlation

In this section, we review three correlation coefficients that are used and calculated in this research following Siegel [98], Siegel [99], Kullback [70], Liebetrau [76], Schweizer and Wolff [96], Blum et al. [13], Chen and Popovich [22], Mari and Kotz [78], Sheskin [97].

2.3.1 Pearson Product-Moment Correlation Coefficient

Pearson product-moment correlation coefficient is a measure of the linear correlation between two variables X and Y , the range of coefficient is between -1 and +1, where -1 is perfect negative correlation, +1 is perfect positive correlation and 0 is no correlation (Hays [54], Lewis-Beck [74]).

Definition 2.3. The Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, \tag{2.9}$$

where $cov(X,Y)$ is the covariance between X and Y, σ_X and σ_Y is the standard deviation of X and Y, respectively.

2.3.2 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient is a nonparametric measure of statistical dependence between two variables X and Y, the range of coefficient is between -1 and +1, where -1 is perfect negative correlation, +1 is perfect positive correlation and 0 is no correlation (Kruskal [69], Lehmann [72], Lehmann [73], Gibbons [51], Gibbons [52], Roberts and Kunst [86], Borkowf [15], Sprent and Smeeton [101]).

Definition 2.4. The Spearman correlation coefficient is defined as same as the Pearson correlation coefficient between the ranked variables.

$$r_s = \rho_{rk(X),rk(Y)} = \frac{cov(rk(X), rk(Y))}{\sigma_{rk(X)} \sigma_{rk(Y)}}, \tag{2.10}$$

where ρ denotes the Pearson correlation coefficient, but applied to the rank variables, $cov(rk(X), rk(Y))$ is the covariance of the rank variables X and Y, and $\sigma_{rk(X)}$ and $\sigma_{rk(Y)}$ are the standard deviations of the rank variables X and Y.

2.3.3 Kendall's Tau Correlation Coefficient

Kendall's tau correlation coefficient is a nonparametric measure of statistical dependence between two ranked variables X and Y same as Spearman correlation coefficient but the calculation is different, the range of coefficient is between -1 and +1, where -1 is perfect negative correlation, +1 is perfect positive correlation and 0 is no correlation (Kendall [62], Kendall [63], Kendall and Stuart [64], Abdi [3]).

Definition 2.5. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be paired observations of the random variables X and Y, such that all the values of X_i and Y_i are unique. Any pair of observations (x_i, y_i) and (x_j, y_j), where $i \neq j$, are said to be concordant if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant (Scarsini [92], Barbe et al. [9]).

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}, \tag{2.11}$$

where n is the number of paired observations.

Remark. When $n \geq 10$, the sampling distribution of τ converges towards a normal distribution (the convergence is satisfactory for values of $n \geq 10$) with mean and standard deviation as follows: (Samara and Randles [91], Walker [110], Abdi [3])

$$\begin{aligned} E(\tau) &= 0 \\ \sigma_\tau &= \sqrt{\frac{2(2n+5)}{9n(n-1)}}. \end{aligned}$$

For the Kandall's Tau significance test: (Fieller et al. [36], Taylor and Karlin [106])

$$\text{Ho: } \tau = 0 \quad \text{vs.} \quad \text{Ha: } \tau \neq 0,$$

the test statistic for τ when $n \geq 10$ is

$$Z = \frac{\tau}{\sigma_\tau} = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}},$$

where Z is normally distributed with mean 0 and standard deviation 1 or $Z \sim$ N(0,1) (Hoeffding [58]).

Clearly, correlation analysis is a statistical method to study the association or relationship between two variables. Correlation and copulas are related. Venter [109] stated that some measures of association, for example, Kendalls tau and Spearmans rank correlation depend only on the copula and not on the marginal distributions. As, Pearson product-moment correlation depends on the marginal distributions. Correlation coefficients measure the overall strength of the association, but give no information about how that varies across the distribution. Therefore, it is a good idea to do parallel study both correlation and copulas, for example, Embrechts et al. [32] studied correlation and dependence in risk management and in 2003, Embrechts et al. [33] also studied modelling dependence with copulas and applications to risk management.

Furthermore, we are also interested in conditional correlation and comparing correlation between un- and conditional correlations of protein-water molecular system dynamics in this research. Therefore, conditional correlation and comparing correlations will be reviewed in next two sections.

2.4 Conditional Correlation

Baba et al. [7] studied and proposed partial correlation and conditional correlation as measure of conditional independence. The definition of the conditional correlation is

Definition 2.6. The conditional correlation of X and Y given Z: $\rho_{XY|Z}$, is the product moment correlation calculated with the conditional distributions X and Y given Z.

$$\rho_{XY|Z} = \frac{cov[(X,Y) \mid Z]}{\sigma_{X|Z}\sigma_{Y|Z}}, \tag{2.12}$$

where $cov[(X,Y) \mid Z]$ is the covariance between X and Y given Z, $\sigma_{X|Z}$ and $\sigma_{Y|Z}$ is the standard deviation of X given Z and Y given Z, respectively.

2.5 Comparing Correlations

When conducting correlation analysis by two independent groups of different sample sizes, typically, a comparison between the two correlations is examined. This is recommended when the correlations are conducted on the same variables by two different groups, and if both correlations are found to be statistically significant. The way to do this is by transforming the correlation coefficient values, or r values, into z scores. This transformation, also known as Fisher r to z transformation, is done so that the z scores can be compared and analyzed for statistical significance by determining the observed z test statistic. The process of comparing two correlations is following (Steiger, [102]).

1. To tranform the two correlation coefficients: r_1 and r_2 to $r_1^{'}$ and $r_2^{'}$ by

$$r_i^{'} = (0.5)log_e\left[\frac{1+r_i^{'}}{1-r_i^{'}}\right].$$

2. To compute the observed z test statistic by

$$z_0 = \frac{r_1^{'} - r_2^{'}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}},$$

where n_1 and n_2 are sample sizes of group 1 and 2, respectively.

3. To test hypothesis about the equality of two population correlations:

$$\text{Ho: } \rho_1 = \rho_2 \quad \text{vs.} \quad \text{Ha: } \rho_1 \neq \rho_2$$

where ρ_1 and ρ_2 are the population correlations from the population 1 and 2, respectively. The criteria to reject the null hypothesis Ho: $\rho_1 = \rho_2$ at significance level 0.05 or $\alpha = 0.05$ is $|z_0| > Z_{\alpha/2} = Z_{0.025} = 1.96 \approx 2.0$.

3

Test Systems

3.1 Introduction

In Chapter 1 and 2, we have introduced and given a short literature review of protein and copulas. We have learned about the importance and necessity of protein for humans and animals. Also, protein structure and some of protein term definitions are reviewed to clarify basic knowledge for this research. Furthermore, the interesting previous research is given to form the concept of protein research. The focus of the literature review in Chapter 2 is copulas that is one of statistical methods in this research. Copulas is a popular method for modeling multivariate distributions or the dependence between many variables in a multivariate distributions, and we can say that copulas is a way to construct multivariate probabilities from the marginals. Furthermore, the measures of association, correlations and related tools are also reviewed.

Our interest in this chapter now is to study copulas of test systems for liquid protein-water molecular system dynamics. Test systems are test datasets which we use for the preliminary study of copulas in this research. In section 3.2, we describe the plan of the test systems. It is clear that the copulas are multivariate distribution functions whose one-dimensional margins are uniform on the interval (0,1). This is illustrated through test systems that are carried out in the section 3.3. The test systems in D-vine is given in the section 3.4. A summary of this chapter is given in the last section.

3.2 The Algorithms of Test Systems

One of the aims of the test systems is to investigate the statistical dependencies of the dynamics of the protein and surrounding water molecules and collective long space- and time-range correlations in the molecular trajectories. The large scale molecular dynamics simulations are performed using the hardware and expertise of the Nonlinearity and Complexity Research Group (NCRG) and the collaboration of the group with RIKEN, Japan, Prof. Makoto Taiji team at K-Computer. The research fits with the research directions of the Systems Analytics Research Institute (SARI) created by

NCRG and CSRG (Computer Science Research Group). Big systems are exemplified by large scale molecular dynamics simulations that produce big data, the molecular trajectories amounting petabytes of data.

For test systems studies, we simulated simple water system. The obtained data includes two datasets: the angle and the amplitude. The datasets come from the process of converting the continuous atomic velocity (coordinate) of the oxygen and hydrogen atoms of one of the water molecules that was used as a 3-dimensional signal. At the locations where the velocity pierces the xy plane, the points of a 2-dimensional map were generated and used as the original continuous signal for analysis. The symmetry of the 2-dimensional set of points can be further illustrated by tranforming the data to the polar coordinates (x,y) $\longrightarrow$ (ρ,ϕ) as illustrated in Figure 3.1. From each dataset, we are interested in two and five variables (or dimensions) for copulas, five and ten variables for D-vine. The example of the layout of the variables following.

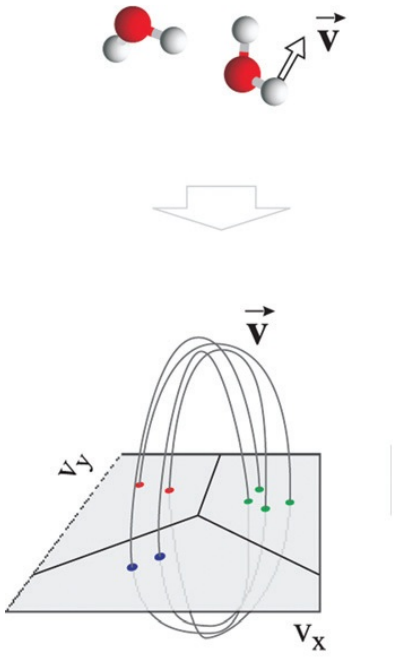

Figure 3.1: The process of transformation the continuous atomic velocity signal $\vec{V}$ - The figure shows the process of transformation the continuous atomic velocity signal $\vec{V}$ (x,y) to (ρ,ϕ) (Source: Ryabov and Nerukh [90], p.037113-5).

Let $a_1, a_2, \ldots$ be a molecular trajectory. Let X_1, X_2 and X_1, X_2, X_3, X_4, X_5 be the two and five variables with sample sizes n in two and five dimensions respectively, where

$$
\begin{aligned}
X_1 &= a_1, a_2, \ldots, a_{n-1}, a_n \\
X_2 &= a_2, a_3, \ldots, a_n, a_{n+1} \\
X_3 &= a_3, a_4, \ldots, a_{n+1}, a_{n+2} \\
X_4 &= a_4, a_5, \ldots, a_{n+2}, a_{n+3} \\
X_5 &= a_5, a_6, \ldots, a_{n+3}, a_{n+4}.
\end{aligned}
$$

The algorithms for copulas analysis are following.

1. To examine the marginal distribution of the variables separately by plotting their histograms and densities.
2. To visualize the relationship between variables by scatter plot.
3. To calculate the Pearson product-moment correlation matrix between variables.
4. To fit the marginal distribution of the variables (optional).
5. To visualize the fitted distribution of the variables by drawing the corresponding distribution function and density function (optional).
6. To take statistical integral transformation to transform any continuous variable to a uniform (0,1) variable via its distribution function (optional).
7. To check the goodness of fit, plotting the fitted distribution function against empirical distribution function (optional).
8. To draw the scatter plot of transformed responses and calculate the Spearman's rho/Kendall's tau correlations between these uniform variables (optional).
9. To fit any copulas families/CD-vine to uniform variables, for example, Gaussian, Student-t and Frank etc.

In Section 3.3 we exhibit the copulas analysis for two and five dimensions through test systems. The R codes (R Development Core Team [85], Yan [111]) for the test systems analysis are given in the Appendix.

3.3 N-variate Copulas

In this section, the test systems study is performed by taking the same sample sizes both angle and amplitude datasets, that are equal to 10000. The results of the test systems study are classified by the datasets and dimensions.

3.3.1 Angle Dataset

We summarized the results of data analysis according to the algorithms following.

3.3.1.1 dimension = 2

1. We examine the marginal distribution of X_1 and X_2 separately by plotting their histograms and densities. From Figure 3.2, the histograms and densities show that both X_1 and X_2 are right skewed distributions (Bobée and Ashkar [14]).

2. The relationship between X_1 and X_2 by scatter plots are given in Figure 3.3. From the scatter plots, they show that X_1 and X_2 do not have a linear association (Balakrishnan and Lai [8]).

3. The Pearson product-moment correlation matrix between X_1 and X_2 is given in Table 3.1 that shows that correlation is quite small.

Table 3.1: The Pearson product-moment correlation matrix between X_1 and X_2 for angle dataset.

	X_1	X_2
X_1	1.00000000	0.08939178
X_2	0.08939178	1.00000000

4. From Figure 3.2, we consider a Beta distribution to fit the marginal distribution of X_1 and X_2 and then extract the fitted model coefficients.

5. The fitted marginal distribution of X_1 and X_2 are shown in Figure 3.4 and 3.5, we plug in the estimated parameters and draw the corresponding distribution function and density function.

6. We take the inverse transformation method to transform any continuous variable to a uniform (0,1) variable via its distribution function. Thus, we transform the variable X_1 and X_2 that have Beta distribution to the variable U_1 and U_2 which follow a uniform distribution on [0,1] (Devroye [29]).

7. We check the goodness of fit by plotting the fitted distribution function against empirical distribution function. The plots are given in Figure 3.6 and 3.7, the right panel of both plots show that the Beta distribution is a satisfactory distribution assumption to X_1 and X_2 (Fermanian [35]).

8. We first draw the scatter plot of transformed responses (U_1 and U_2) (Kimeldorf and Sampson [66]) that is given in Figure 3.8 and calculate the Spearman's rho correlation between these two uniform variables that is equal to 0.1044906.

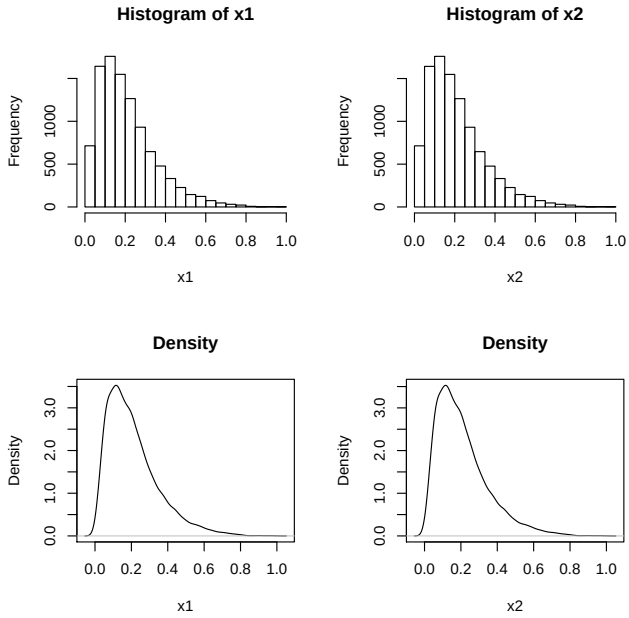

Figure 3.2: Histograms and densities of X_1 and X_2: angle dataset - The figure shows histograms and densities of X_1 and X_2 for angle dataset.

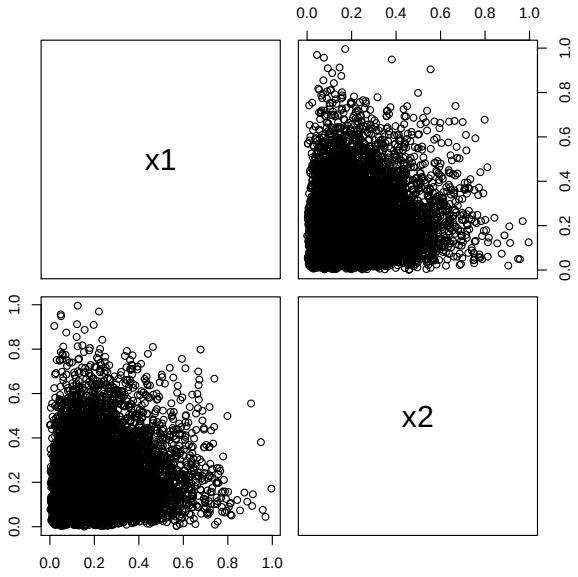

Figure 3.3: Scatter plots of X_1 and X_2: angle dataset - The figure shows scatter plots of X_1 and X_2 for angle dataset.

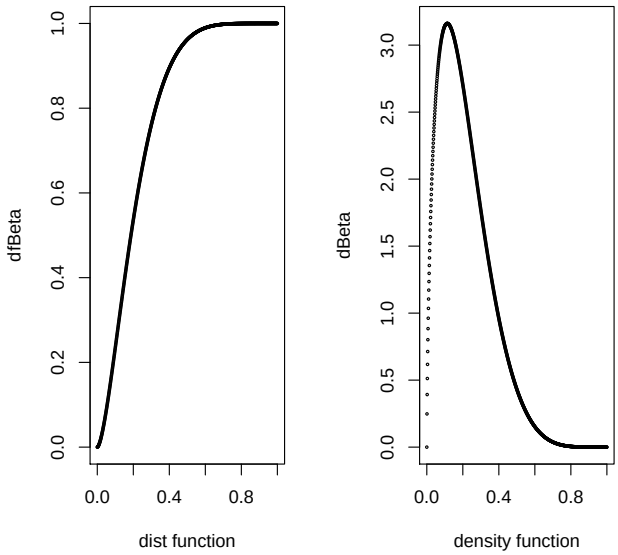

Figure 3.4: The fitted distribution function of X_1: angle dataset - The figure shows the fitted distribution function and density function of X_1 for angle dataset.

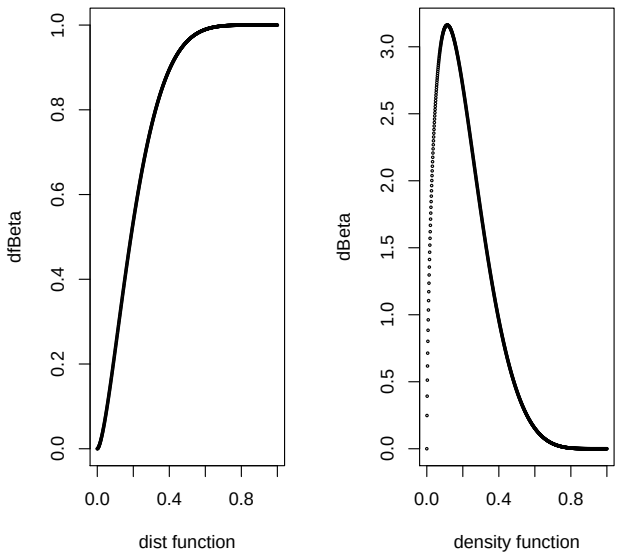

Figure 3.5: The fitted distribution function of X_2: angle dataset - The figure shows the fitted distribution function and density function of X_2 for angle dataset.

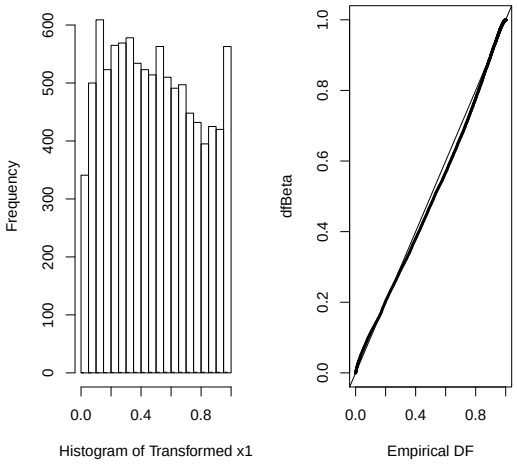

Figure 3.6: The histogram and plot of the fitted distribution function of tranformed X_1: angle dataset - The figure shows the histogram and plot of the fitted distribution function against empirical distribution function of transformed X_1 for angle dataset.

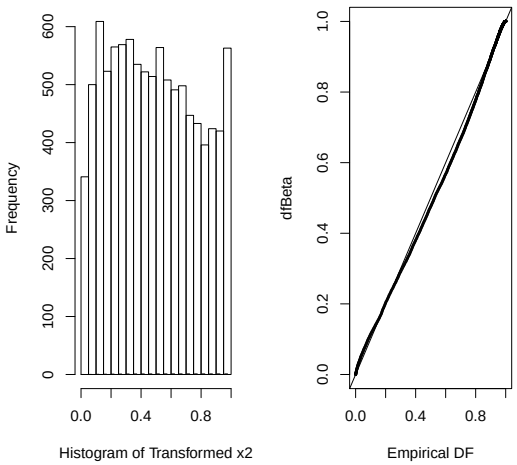

Figure 3.7: The histogram and plot of the fitted distribution function of tranformed X_2: angle dataset - The figure shows the histogram and plot of the fitted distribution function against empirical distribution function of transformed X_2 for angle dataset.

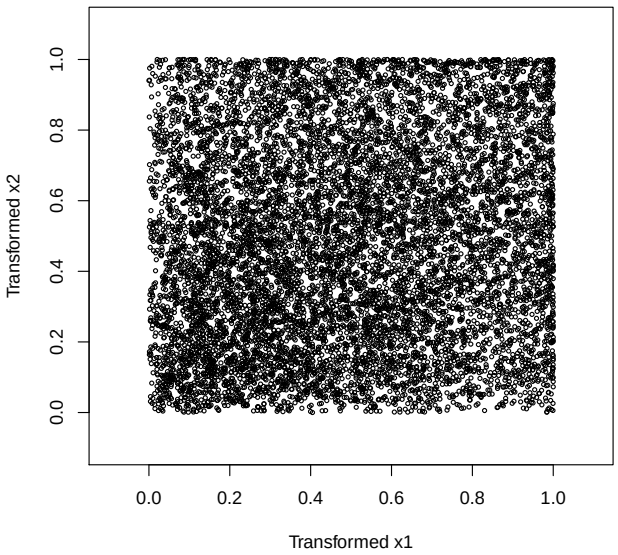

Figure 3.8: The scatter plot of transformed X_1 and X_2: angle dataset - The figure shows the scatter plot of transformed X_1 and X_2 (U_1 and U_2) for angle dataset.

9. We fit Gaussian, Student-t, Frank and Clayton family to U_1 and U_2 using maximum likelihood method and consider the maximum likelihood estimator (MLE) of the copulas parameter, standard error (standard deviation of estimator) and p-value. If p-value < 0.05 then the test is significantly, that means that the copulas parameter is significant different from 0. The statistical hypothesis to test the copulas parameter (θ) is

$$\text{Ho: } \theta = 0 \quad \text{vs.} \quad \text{Ha: } \theta \neq 0.$$

The MLE of copulas parameter (Akaike [4]), standard error, Z-value and p-value of Gaussian, Student-t, Frank and Clayton families are given in Table 3.2. We found that in all copulas families, the fitted model implies that X_1 and X_2 are positive dependent (p-value < 0.05, the copulas parameter is significantly different from 0).

Conclusion: For angle dataset in two dimensions, we conclude that all copulas parameters from fitted copulas family are significant. That is X_1 and X_2 are positive correlated with small value of copulas parameters: 0.094746, 0.098643, 0.671645 and 0.138741, respectively as illustrated in Table 3.2.

Table 3.2: The MLE of copulas parameter, standard error, Z-value and p-value of Gaussian, Student-t, Frank and Clayton families for angle dataset, dimension = 2.

Family	Parameter	Standard Error	Z-value	p-value
Gaussian	0.094746	0.009904	9.566726	$<$ 2e-16
Student-t	0.098643	0.010241	9.631914	$<$ 2e-16
Frank	0.671645	0.063037	10.65471	$<$ 2e-16
Clayton	0.138741	0.015591	8.898719	$<$ 2e-16

Remarks.

1. **The Beta distribution**

   If $X \sim B(\alpha, \beta)$ then the probability density function of X is (Johnson and Kotz [60], Johnson et al. [61], Kotz et al. [67])

   $$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, 0 \leq x \leq 1, \alpha, \beta > 0,$$

   where

   $$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

2. **The Inverse transformation method**

   If X is a random variable with a continuous cumulative distribution function $F(x)$ $= P(X \leq x)$, then the random variable

   $$U = F(X),$$

   has a uniform (0,1) distribution. This fact provides a very simple relationship between a uniform random variable U and a random variable X with cumulative distribution function F:

   $$X = F^{-1}(U).$$

3. Genest and Rivest [49] studied the multivariate probability integral transformation and Genest et al. [46] studied the goodness-of-fit procedures for copula models based on the probability integral transformation.

4. For copulas analysis, Genest and Rivest [48] studied statistical inference procedures for bivariate Archimedean copulas. In 2004, Genest and Rémillard [47] studied tests of independence and randomness based on the empirical copula process and in 2005, Genest and Verret [50] also studied locally most powerful rank tests of independence for copula models.

3.3.1.2 dimension = 5

We still carry out the test systems study for five dimensions that is same as two dimensions. The results are following.

1. We examine the marginal distribution of X_1, X_2, X_3, X_4 and X_5 separately by plotting their histograms and densities. The histograms and densities from Figure 3.9 and 3.10 show that all of variables are right skewed distribution.

2. The scatter plots of X_1, X_2, X_3, X_4 and X_5 is given in Figure 3.11. From the scatter plots, they show that all of variables are not linear correlation.

3. The Pearson product-moment correlation matrix of X_1, X_2, X_3, X_4 and X_5 is given in Table 3.3 that shows that paired-correlations are quite small.

Table 3.3: The Pearson product-moment correlation matrix of X_1, X_2, X_3, X_4 and X_5 for angle dataset.

	X_1	X_2	X_3	X_4	X_5
X_1	1.000000e+00	0.089391781	0.22730582	7.297037e-05	0.149340996
X_2	8.939178e-02	1.000000000	0.08938141	2.275232e-01	0.000204394
X_3	2.273058e-01	0.089381410	1.00000000	8.925085e-02	0.227450278
X_4	7.297037e-05	0.227523194	0.08925085	1.000000e+00	0.089259234
X_5	1.493410e-01	0.000204394	0.22745028	8.925923e-02	1.000000000

4. From Figure 3.9 and 3.10, we consider a Beta distribution to fit the marginal distribution of X_1, X_2, X_3, X_4 and X_5 and then extract the fitted model coefficients.

5. All fitted distribution of X_1, X_2, X_3, X_4 and X_5 are same as X_1 and X_2 in two dimensions (see Figure 3.4 and 3.5). For example, the fitted distribution function and density function of X_3 are given in Figure 3.12.

6. We take statistical integral transformation to transform any continuous variable to a uniform (0,1) variable via its distribution function. Thus, we transform the variable X_1, X_2, X_3, X_4 and X_5 that have Beta distribution to the variable U_1, U_2, U_3, U_4 and U_5 which follow a uniform distribution on [0,1].

7. We check the goodness of fit by plotting the fitted distribution function against empirical distribution function of X_1, X_2, X_3, X_4 and X_5 which the plots are same as X_1 and X_2 in two dimensions (see Figure 3.6 and 3.7). For example, the fitted distribution function against empirical distribution function of X_4 is given in Figure 3.13, the right panel of plots show that the Beta distribution is a satisfactory distribution assumption to X_4.

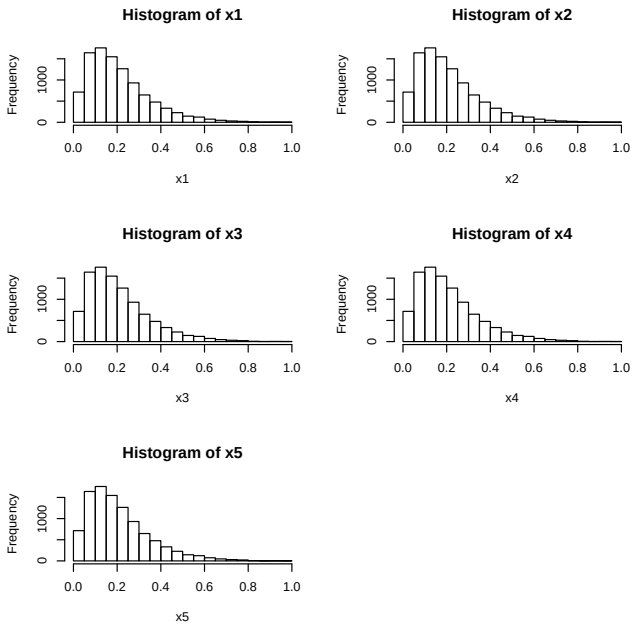

Figure 3.9: Histograms of X_1, X_2, X_3, X_4 and X_5: angle dataset - The figure shows histograms of X_1, X_2, X_3, X_4 and X_5 for angle dataset.

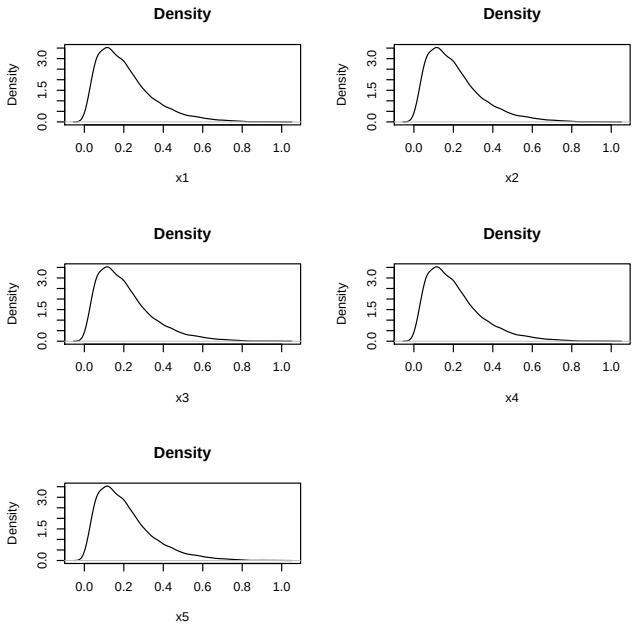

Figure 3.10: Densities of X_1, X_2, X_3, X_4 and X_5: angle dataset - The figure shows densities of X_1, X_2, X_3, X_4 and X_5 for angle dataset.

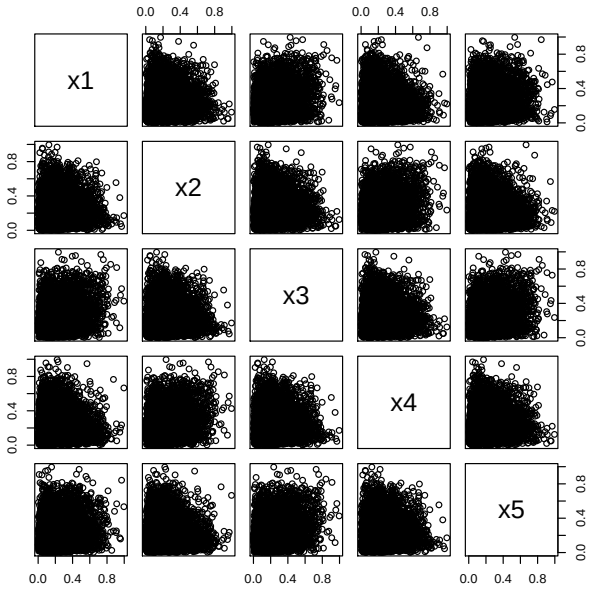

Figure 3.11: Scatter plots of X_1, X_2, X_3, X_4 and X_5: angle dataset - The figure shows scatter plots of X_1, X_2, X_3, X_4 and X_5 for angle dataset.

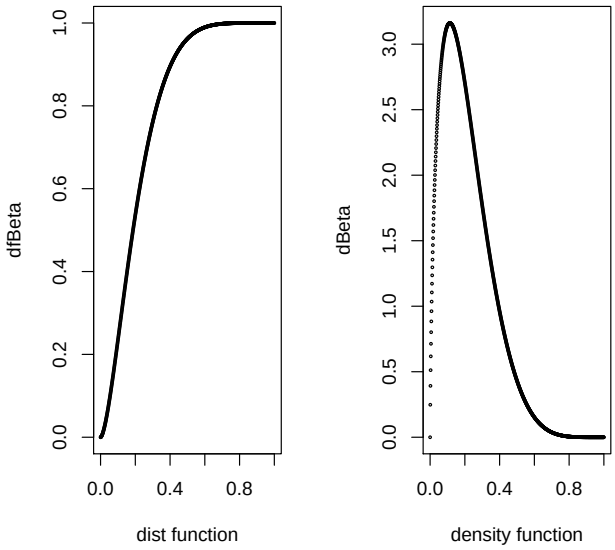

Figure 3.12: The fitted distribution function of X_3: angle dataset - The figure shows the fitted distribution function and density function of X_3 for angle dataset.

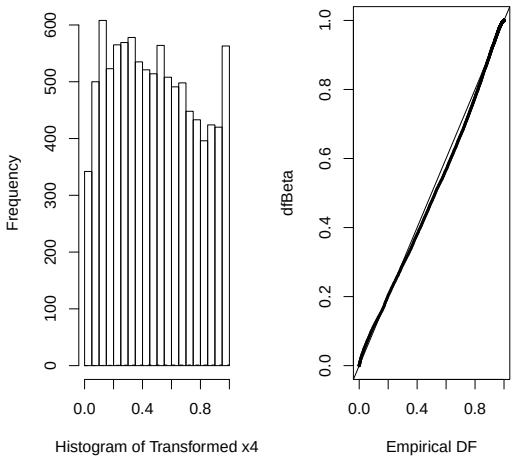

Figure 3.13: The histogram and plot of the fitted distribution function of tranformed X_4: angle dataset - The figure shows the histogram and plot of the fitted distribution function against empirical distribution function of transformed X_4 for angle dataset.

8. We fit Frank and Clayton family to U_1, U_2, U_3, U_4 and U_5 using maximum likelihood method and consider the maximum likelihood estimator (MLE) of the copulas parameter, standard error (standard deviation of estimator) and p-value.

   The MLE of copulas parameter, standard error, Z-value and p-value of Frank and Clayton families are given in Table 3.4. We found that in all copulas families, the fitted model implies that X_1, X_2, X_3, X_4 and X_5 are positive dependent.

Table 3.4: The MLE of copulas parameter, standard error, Z-value and p-value of Frank and Clayton family for angle dataset, dimension = 5.

Family	Parameter	Standard Error	Z-value	p-value
Frank	0.466360	0.023240	20.06786	$<$ 2e-16
Clayton	0.101211	0.005613	18.03212	$<$ 2e-16

Conclusion: For angle dataset in five dimensions, we conclude that all copulas parameters from fitted copulas family are significant. That is X_1, X_2, X_3, X_4 and X_5 are positive correlated with small value of copulas parameters: 0.466360 and 0.101211, respectively as illustrated in Table 3.4.

3.3.2 Amplitude Dataset

We summarized the results of data analysis for amplitude dataset following.

3.3.2.1 dimension = 2

1. We examine the marginal distribution of X_1 and X_2 separately by plotting their histograms and densities. From Figure 3.14, the histograms and densities show that both X_1 and X_2 are uniform (0,1) distribution.

2. The relationship between X_1 and X_2 by scatter plots are given in Figure 3.15. From the scatter plots, they show that X_1 and X_2 are not linear correlation.

3. The Pearson product-moment correlation matrix between X_1 and X_2 is given in Table 3.5 that shows that correlation is quite small.

Table 3.5: The correlation matrix between X_1 and X_2 for amplitude dataset.

	X_1	X_2
X_1	1.00000000	0.01952086
X_2	0.01952086	1.00000000

4. Since the marginal distribution of X_1 and X_2 are uniform (0,1) distribution, thus we fit Gaussian, Student-t, Frank and Clayton families to X_1 and X_2 using maximum likelihood method and consider the maximum likelihood estimator (MLE) of the copulas parameter, standard error (standard deviation of estimator) and p-value.

   The MLE of copulas parameter, standard error, Z-value and p-value of Gaussian, Student-t, Frank and Clayton families are given in Table 3.6. We found that in all copulas families, the fitted model implies that X_1 and X_2 are positive dependent.

Table 3.6: The MLE of copulas parameter, standard error, Z-value and p-value of Gaussian, Student-t, Frank and Clayton family for amplitude dataset, dimension = 2.

Family	Parameter	Standard Error	Z-value	p-value
Gaussian	0.020567	0.00444	4.63085	3.64e-06
Student-t	0.020758	0.00452	4.59271	4.38e-06
Frank	0.112870	0.02673	4.22316	2.41e-05
Clayton	0.016350	0.004604	3.55098	3.84e-04

Conclusion: For amplitude dataset in two dimensions, we conclude that all copulas parameters from fitted copulas family are significant. That is X_1 and X_2 are positive correlated with a little value of copulas parameters: 0.020567, 0.020758, 0.112870 and 0.016350, respectively as illustrated in Table 3.6.

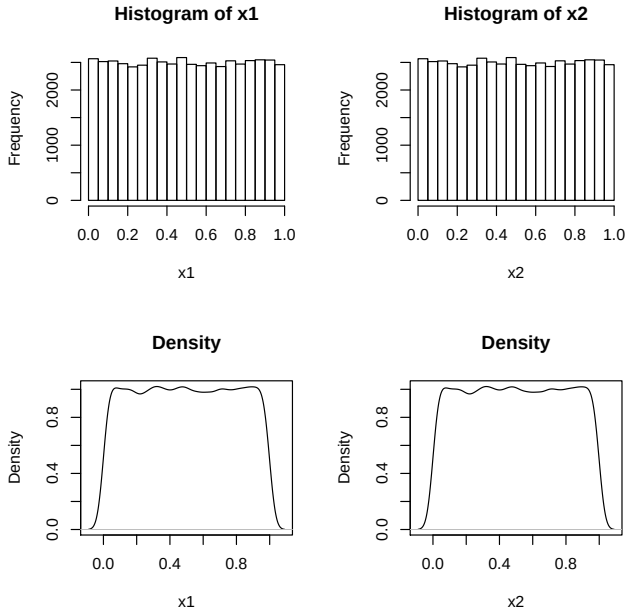

Figure 3.14: Histograms and densities of X_1 and X_2: amplitude dataset - The figure shows histograms and densities of X_1 and X_2 for amplitude dataset.

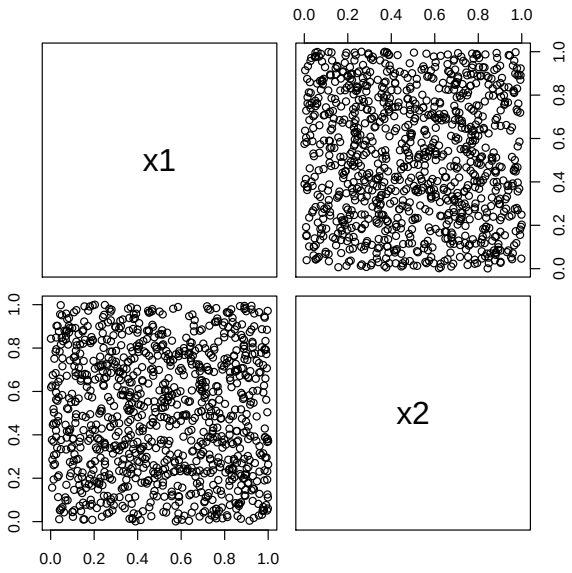

Figure 3.15: Scatter plots of X_1 and X_2: amplitude dataset - The figure shows scatter plots of X_1 and X_2 for amplitude dataset.

3.3.2.2 dimension = 5

We still carry out the test systems study for five dimensions that is same as two dimensions. The results are following.

1. We examine the marginal distribution of X_1, X_2, X_3, X_4 and X_5 separately by plotting their histograms and densities. From Figure 3.16 and 3.17, the histograms and densities show that X_1, X_2, X_3, X_4 and X_5 are uniform (0,1) distribution.

2. The relationship of X_1, X_2, X_3, X_4 and X_5 by scatter plots is given in Figure 3.18. From the scatter plots, they show that X_1, X_2, X_3, X_4 and X_5 are not linear correlation.

3. The Pearson product-moment correlation matrix of X_1, X_2, X_3, X_4 and X_5 is given in Table 3.7 that shows that paired-correlations are quite small.

Table 3.7: The correlation matrix of X_1, X_2, X_3, X_4 and X_5 for amplitude dataset.

	X_1	X_2	X_3	X_4	X_5
X_1	1.00000000	0.01952086	0.03817455	0.05481715	0.03483997
X_2	0.01952086	1.00000000	0.01924201	0.03842021	0.05501616
X_3	0.03817455	0.01924201	1.00000000	0.01913014	0.03818375
X_4	0.05481715	0.03842021	0.01913014	1.00000000	0.01950141
X_5	0.03483997	0.05501616	0.03818375	0.01950141	1.00000000

4. Since the marginal distribution of X_1, X_2, X_3, X_4 and X_5 are uniform (0,1) distribution, thus we fit Frank and Clayton copulas family to X_1, X_2, X_3, X_4 and X_5 using maximum likelihood method and consider the maximum likelihood estimator (MLE) of the copulas parameter, standard error (standard deviation of estimator) and p-value.

   The MLE of copulas parameter, standard error, Z-value and p-value of Frank and Clayton family are given in Table 3.8. We found that in all copulas families, the fitted model implies that X_1, X_2, X_3, X_4 and X_5 are positive dependent.

Table 3.8: The MLE of copulas parameter, standard error, Z-value and p-value of Frank and Clayton family for amplitude dataset, dimension = 5.

Family	copulas Parameter	Standard Error	Z-value	p-value
Frank	0.064029	0.007713	8.301803	1.025e-16
Clayton	0.013627	0.001502	9.071046	1.179e-19

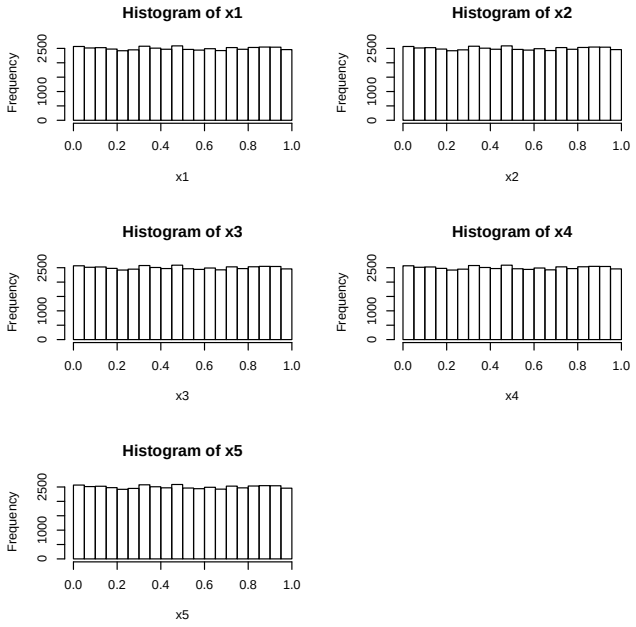

Figure 3.16: Histograms of X_1, X_2, X_3, X_4 and X_5: amplitude dataset - The figure shows histograms of X_1, X_2, X_3, X_4 and X_5 for amplitude dataset.

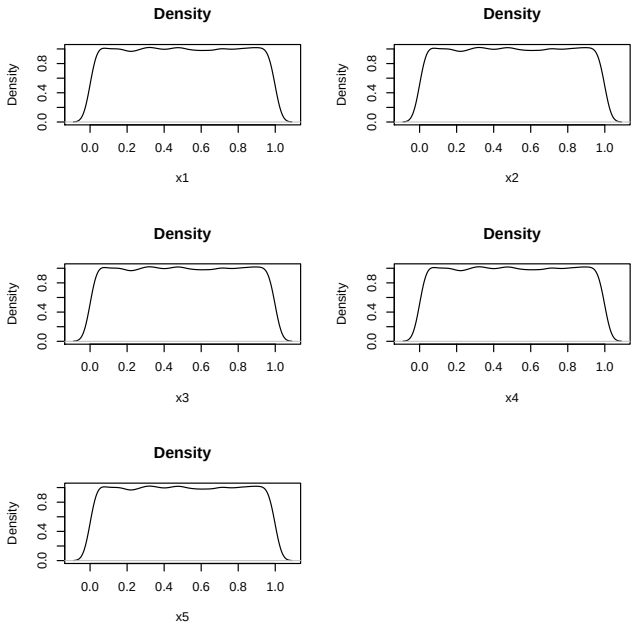

Figure 3.17: Densities of X_1, X_2, X_3, X_4 and X_5: amplitude dataset - The figure shows densities of X_1, X_2, X_3, X_4 and X_5 for amplitude dataset.

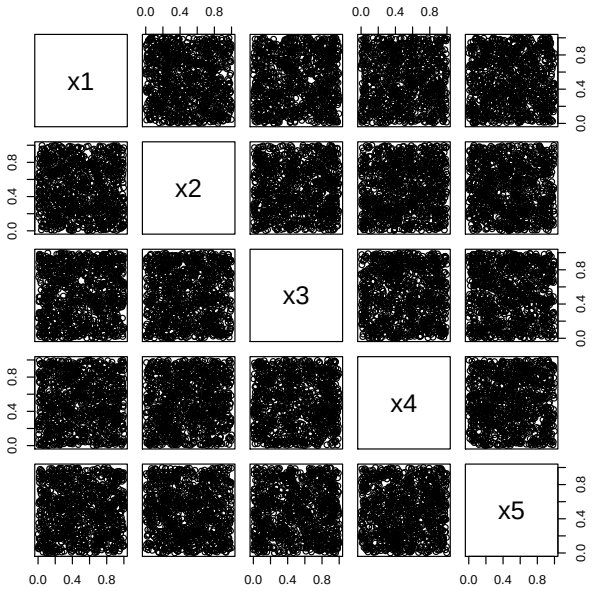

Figure 3.18: Scatter plots of X_1, X_2, X_3, X_4 and X_5: amplitude dataset - The figure shows scatter plots of X_1, X_2, X_3, X_4 and X_5 for amplitude dataset.

Conclusion: For amplitude dataset in five dimensions, we conclude that all copulas parameters from fitted copulas family are significant. That is X_1, X_2, X_3, X_4 and X_5 are positive correlated with a little value of copulas parameters: 0.064029 and 0.013627, respectively as illustrated in Table 3.8.

3.4 D-vine

D-vine analysis for five and ten dimensions in angle and amplitude datasets are carry out in this section. The test systems study is performed by taking variety of sample sizes (n) both angle and amplitude datasets, for example, n = 10000, 50000 etc. Our first interest is to investigate whether the D-vine result is different when sample sizes are not equal.

According to the effect of sample sizes, Henson and Smith [55] studied the art in statistical significance and effect size and found that the significant test is different when sample size varies. As, Rosenthal [87] also stated that significance tests can be usefully viewed as simple products of effect size estimates and sample sizes, any nonzero effect size will reach statistical significance given a sufficiently large sample size. Moreover, Knapp [68] gave an interesting comment on the statistical significance testing and sample size following.

"Researchers have actual sample sizes and actual values for their statistics; speculating as to what might have happened if they had bigger or smaller sample sizes, or the population correlations had been bigger or smaller, or whatever, is the sort of thinking that should be gone through before a study is carried out, not after."

From the examples of previous research which studied about statistical significance test, clearly, they showed that sample size affects the statistical hypothesis testing. For this research, we expect that the D-vine results are different when sample sizes are also different.

Besides sample sizes, we are also interested in differences of D-vine results when observations from each dataset are selected according to a fixed interval. This interval we call the **interval**, for example interval = 1, 2 etc. Therefore, our second interest is to investigate whether the D-vine result is different when the interval is different. As same as sample sizes, we expect that the D-vine results are different when the interval is also different. For the analysis, CD-Vine package in R is a statistical package that we use for modeling analysis pair-copulas construction. The results of the test systems study are classified by the dataset and dimension in subsection 3.4.1 and 3.4.2.

3.4.1 Angle Dataset

3.4.1.1 Dimension = 5

First of all, we carry out the data analysis according to the algorithms from the previous section following.

1. We examine the marginal distribution of X_1, X_2, X_3, X_4 and X_5 separately by plotting their histograms and densities. The histograms and densities from Figure 3.9 and 3.10 show that all of variables are right skewed distribution.

2. From Figure 3.9 and 3.10, we consider a Beta distribution to fit the marginal distribution of X_1, X_2, X_3, X_4 and X_5 and then extract the fitted model coefficients.

3. We take statistical integral transformation to transform any continuous variable to a uniform (0,1) variable via its distribution function. Thus, we transform the variable X_1, X_2, X_3, X_4 and X_5 that have Beta distribution to the variable U_1, U_2, U_3, U_4 and U_5 which follow a uniform distribution on [0,1].

4. We check the goodness of fit by plotting the fitted distribution function against empirical distribution function. We found that the Beta distribution is a satisfactory distribution assumption to all of variables.

5. We fit D-vine to U_1, U_2, U_3, U_4 and U_5 using maximum likelihood method. The fitted family, MLE of pair-copulas parameter and independent test of each edge are given in Table 3.9.

From Table 3.9, we fit D-vine for angle dataset when sample size = 10000, 50000, 100000 and no interval. The results can summarize following.

- For n = 10000 and 50000, the best fit family of each edge for all trees are similar: Frank is the best fit family for the first and third tree of each edge, as Gumbel is the best fit family for the second tree of each edge and Gaussian is the best fit family for the fourth tree.

- For n = 100000, the best fit family of each edge for the first two trees are similar to n= 10000 and 50000 cases: Frank and Gumbel. As, Gaussain is the best fit family for the third and fourth trees of each edge, as Gumbel is the best fit family for the second tree of each edge and Gaussian is the best fit family for the fourth tree.

- When we consider the estimator of pair-copulas parameter of each edge, tree and sample size, the estimator is quite similar, for example, n = 10000, the estimator for the first tree is between 0.67000287 and 0.67164488, the difference is approximately 0.0016 etc. See Table 3.10 for the differences of estimator.

- From Table 3.10, when the sample size increases, the difference decreses at each tree, for example, when n = 10000, 50000, 100000, the differences at tree 2 are equal to 0.00013573, 0.00002074, 0.00001942 respectively, etc.

- For the independent test, that means that the pair-copulas parameter (θ) of each edge is tested whether it is independent. The statistical hypothesis is

$$\text{Ho: } \theta = 0 \quad \text{vs.} \quad \text{Ha: } \theta \neq 0,$$

Table 3.9: The best fit family, MLE of pair-copulas parameter and independent test of each edge for D-vine of angle dataset when sample size = 10000, 50000, 100000 and interval = 0.

n	Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
10000	1	12	Frank	0.67164488	Dependent
		23	Frank	0.67131425	Dependent
		34	Frank	0.67000287	Dependent
		45	Frank	0.67001662	Dependent
	2	13/2	Gumbel	1.09872956	Dependent
		24/3	Gumbel	1.09886529	Dependent
		35/4	Gumbel	1.09883150	Dependent
	3	14/23	Frank	-0.15675825	Dependent
		25/34	Frank	-0.15588928	Dependent
	4	15/234	Gaussian	0.09127093	Dependent
50000	1	12	Frank	0.76436410	Dependent
		23	Frank	0.76464839	Dependent
		34	Frank	0.76465051	Dependent
		45	Frank	0.76455606	Dependent
	2	13/2	Gumbel	1.10181720	Dependent
		24/3	Gumbel	1.10181887	Dependent
		35/4	Gumbel	1.10179813	Dependent
	3	14/23	Frank	-0.16473762	Dependent
		25/34	Frank	-0.16493373	Dependent
	4	15/234	Gaussian	0.09672307	Dependent
100000	1	12	Frank	0.79401441	Dependent
		23	Frank	0.79396905	Dependent
		34	Frank	0.79395414	Dependent
		45	Frank	0.79389445	Dependent
	2	13/2	Gumbel	1.10652738	Dependent
		24/3	Gumbel	1.10653314	Dependent
		35/4	Gumbel	1.10654680	Dependent
	3	14/23	Gaussian	-0.02978778	Dependent
		25/34	Gaussian	-0.02978373	Dependent
	4	15/234	Gaussian	0.09614729	Dependent

Table 3.10: The estimator differences of each tree of angle dataset when sample size = 10000, 50000, 100000 and interval = 0.

n	Tree	Best fit family	max.	min.	difference
10000	1	Frank	0.67164488	0.67000287	0.00164201
	2	Gumbel	1.09886529	1.09872956	0.00013573
	3	Frank	-0.15588928	-0.15675825	0.00086897
50000	1	Frank	0.76465051	0.76436410	0.00028641
	2	Gumbel	1.10181887	1.10179813	0.00002074
	3	Frank	-0.16473762	-0.16493373	0.00019611
100000	1	Frank	0.79401441	0.79389445	0.00011996
	2	Gumbel	1.10654680	1.10652738	0.00001942
	3	Gaussian	-0.02978373	-0.02978778	0.00000405

where θ is the pair-copulas parameter of each edge.

From Table 3.9, all pair-copulas parameters of each edge, tree, sample size are dependent when the test systems observations do not interval.

From Table 3.9, we fit D-vine to the angle dataset without skipping the observations. In addition to our interest, we have tried to interval the observations, say, interval = 2 and 5, after that, fitting and investigation the D-vine results, see Table 3.11, 3.13.

For interval = 2, the results are showed in Table 3.11 and we summarize as follows.

- For n = 50000 and 100000, the best fit family of each edge for all trees are similar: Clayton is the best fit family for the first tree and Gaussian is the best fit family for the second and third tree.
- For n = 10000, the best fit family of each edge for the first and fourth tree are similar, that is Gaussian, as Gumbel is the best fit family for all edges in the second tree. For the third tree, Frank is the best fit family of each edge.
- When we consider the estimator of pair-copulas parameter of each edge, tree and sample size, the estimator is quite similar, for example, n = 50000, the estimator for the second tree is between 0.09826101 and 0.09829687, the difference is 0.00003586 etc. See Table 3.12 for the differences of estimator.
- From Table 3.12, for n = 50000 and 100000, when the sample size increases, the difference decreses at each tree, for example, when n = 50000, 100000, the differences at tree 1 are equal to 0.00005445, 0.00004969 respectively, etc.
- For n = 50000 and 100000, most of all independent test for pair-copulas parameter of each edge are similar, that is, all edges in tree 1, 2 and 4 are independent,

Table 3.11: The best fit family, MLE of pair-copulas parameter and independent test of each edge for D-vine of angle dataset when sample size = 10000, 50000, 100000 and interval = 2.

n	Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
10000	1	12	Gaussian	0.02168114	Dependent
		23	Gaussian	0.02185148	Dependent
		34	Gaussian	0.02166794	Independent
		45	Gaussian	0.02159354	Independent
	2	13/2	Gumbel	1.04226104	Dependent
		24/3	Gumbel	1.04223025	Dependent
		35/4	Gumbel	1.04231607	Dependent
	3	14/23	Frank	-0.22375583	Dependent
		25/34	Frank	-0.22521606	Dependent
	4	15/234	Gaussian	0.07387519	Dependent
50000	1	12	Clayton	0.02176477	Independent
		23	Clayton	0.02179607	Independent
		34	Clayton	0.02181922	Independent
		45	Clayton	0.02180741	Independent
	2	13/2	Gaussian	0.09827332	Dependent
		24/3	Gaussian	0.09826101	Dependent
		35/4	Gaussian	0.09829687	Dependent
	3	14/23	Gaussian	-0.02093765	Dependent
		25/34	Gaussian	-0.02098103	Independent
	4	15/234	Gaussian	0.06984853	Dependent
100000	1	12	Clayton	0.01960320	Independent
		23	Clayton	0.01963013	Independent
		34	Clayton	0.01965289	Independent
		45	Clayton	0.01962888	Independent
	2	13/2	Gaussian	0.09729499	Dependent
		24/3	Gaussian	0.09729200	Dependent
		35/4	Gaussian	0.09731779	Dependent
	3	14/23	Gaussian	-0.02256822	Dependent
		25/34	Gaussian	-0.02259098	Dependent
	4	15/234	Gaussian	0.07131709	Dependent

Table 3.12: The estimator differences of each tree of angle dataset when sample size = 10000, 50000, 100000 and interval = 2.

n	Tree	Best fit family	max.	min.	difference
10000	1	Gaussian	0.02185148	0.02159354	0.00025794
	2	Gumbel	1.04231607	1.04223025	0.00008582
	3	Frank	-0.22375583	-0.22521606	0.00146023
50000	1	Clayton	0.02181922	0.02176477	0.00005445
	2	Gaussian	0.09829687	0.09826101	0.00003586
	3	Gaussian	-0.02093765	-0.02098103	0.00004338
100000	1	Clayton	0.01965289	0.01960320	0.00004969
	2	Gaussian	0.09731779	0.09729200	0.00002579
	3	Gaussian	-0.02256822	-0.02259098	0.00002276

dependent and dependent, respectively. As the independent test for edge 25/34 of each sample size is different: for n= 50000, the test is independent, but for n = 100000, the test is dependent.

- For n = 10000, the independent test for edge 34 and 45 from tree 1 are independent. As, the independent test of rest of edges are dependent.

For interval = 5, the results are showed in Table 3.13 and we summarize following.

- For n = 50000 and 100000, the best fit family of each edge for tree 1, 2 and 3 are Gaussian and the best fit family for the fourth tree is Gaussian when n = 50000 and Frank when n = 100000.
- For n = 10000, the best fit family of each edge for the first tree is Student-t, as Gumbel is the best fit family for edge 15/234 in the fourth tree. For tree 2 and 3, Gaussian is the best fit family of each edge.
- When we consider the estimator of pair-copulas parameter of each edge, tree and sample size, the estimator is not too much different, for example, n = 100000, the estimator for the first tree is between 0.10709450 and 0.10711847, the difference is 0.00002397 etc. See Table 3.14 for the differences of estimator.
- From Table 3.14, the difference of each tree from all sample size is very small.
- For the independent test for pair-copulas parameter, all edges are dependent, excepting, when n = 10000, edge 15/234 is independent.

Table 3.13: The best fit family, MLE of pair-copulas parameter and independent test of each edge for D-vine of angle dataset when sample size = 10000, 50000, 100000 and interval = 5.

n	Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
10000	1	12	Student-t	0.10286037	Dependent
		23	Student-t	0.10270977	Dependent
		34	Student-t	0.10295820	Dependent
		45	Student-t	0.10279326	Dependent
	2	13/2	Gaussian	0.07632154	Dependent
		24/3	Gaussian	0.07635409	Dependent
		35/4	Gaussian	0.07632808	Dependent
	3	14/23	Gaussian	0.05304979	Dependent
		25/34	Gaussian	0.05324412	Dependent
	4	15/234	Gumbel	1.00699653	Independent
50000	1	12	Gaussian	0.10393600	Dependent
		23	Gaussian	0.10394122	Dependent
		34	Gaussian	0.10395245	Dependent
		45	Gaussian	0.10394202	Dependent
	2	13/2	Gaussian	0.06889225	Dependent
		24/3	Gaussian	0.06891066	Dependent
		35/4	Gaussian	0.06891015	Dependent
	3	14/23	Gaussian	0.06080905	Dependent
		25/34	Gaussian	0.06082768	Dependent
	4	15/234	Gaussian	0.03945292	Dependent
100000	1	12	Gaussian	0.10709450	Dependent
		23	Gaussian	0.10709528	Dependent
		34	Gaussian	0.10711726	Dependent
		45	Gaussian	0.10711847	Dependent
	2	13/2	Gaussian	0.07134126	Dependent
		24/3	Gaussian	0.07136261	Dependent
		35/4	Gaussian	0.07137140	Dependent
	3	14/23	Gaussian	0.05983978	Dependent
		25/34	Gaussian	0.05984671	Dependent
	4	15/234	Frank	0.24974652	Dependent

Table 3.14: The estimator differences of each tree of angle dataset when sample size = 10000, 50000, 100000 and interval = 5.

n	Tree	Best fit family	max.	min.	difference
10000	1	Student-t	0.10295820	0.10270977	0.00024843
	2	Gaussian	0.07635409	0.07632154	0.00003255
	3	Gaussian	0.05324412	0.05304979	0.00019433
50000	1	Gaussian	0.10395245	0.10393600	0.00001645
	2	Gaussian	0.06891066	0.06889225	0.00001841
	3	Gaussian	0.06082768	0.06080905	0.00001863
100000	1	Gaussian	0.10711847	0.10709450	0.00002397
	2	Gaussian	0.07137140	0.07134126	0.00003014
	3	Gaussian	0.05984671	0.05983978	0.00000693

Conclusion: For angle dataset in five dimensions, we conclude as follows.

1. When we select the observations without interval (interval = 0), the best fit family and independent test for all trees from all sample sizes are similar (except for tree 3 when n = 100000, the best fit family is different). We may say that D-vine results for test systems are not different when we take the observations without interval and any sample size (Table 3.9).

2. When we select the observations with interval = 2 and n = 50000 and 100000, the best fit family and independent test for all trees are similar (except for tree 3, edge 25/34 when n = 50000, the independent test is different). We may say that D-vine results are not different when we take the observations with interval = 2 and large sample size (Table 3.11).

3. When we select the observations with interval = 5 and n = 50000 and 100000, the best fit family and independent test for all trees are similar (except for tree 4 when n = 100000, the best fit family is different). We may say that D-vine results for test systems are not different when we take the observations with interval = 5 and large sample size (Table 3.13).

4. We found that D-vine results for angle dataset in five dimensions are different when we take the observations without or with interval. When we take the observations with any interval and large sample size, the D-vine results are similar and at the high level of tree, the best fit family is Gaussian.

3.4.1.2 Dimension = 10

In ten dimensions, we carry out the data analysis according to the algorithms as same as in five dimensions as follows.

1. We examine the marginal distribution of X_1, X_2, X_3, ..., X_{10} separately by plotting their histograms and densities. The histograms and densities show that all of variables are right skewed distribution.

2. We consider a Beta distribution to fit the marginal distribution of X_1, X_2, X_3, ..., X_{10} and then extract the fitted model coefficients.

3. We take statistical integral transformation to transform any continuous variable to a uniform (0,1) variable via its distribution function. Thus, we transform the variable X_1, X_2, X_3, ..., X_{10} that have Beta distribution to the variable U_1, U_2, U_3, ..., U_{10} which follow a uniform distribution on [0,1].

4. We check the goodness of fit by plotting the fitted distribution function against empirical distribution function. We found that the Beta distribution is a satisfactory distribution assumption to all of variables.

5. We fit D-vine to U_1, U_2, U_3, ..., U_{10} using maximum likelihood method. The fitted family, MLE of pair-copulas parameter and independent test of each edge with skipping = 1, 2, 3, 5 are showed in Figure 3.19, 3.20, 3.21 and 3.22.

From Figure 3.19, 3.20, 3.21 and 3.22, we fit D-vine for 10 uniform (0,1) variables of angle dataset with interval = 1, 2, 3 and 5. The 10-dimensional D-vine has 9 trees and 45 edges and the results are following.

1. **interval = 1**

   - The best fit family for most trees is similar both n = 10000 and 50000, excepting, when n = 10000, the best fit family for tree 5 is Gumbel, as n = 50000, Gaussian is the best fit family.
   - The estimator of pair-copulas parameter of each edge for tree 1 - 4, 7 - 9 increases when sample size increases but for tree 6, the estimator decreases when sample size increases.
   - For the independent test for pair-copulas parameter, most edges for all trees both n = 10000 and 50000 are dependent. As n = 10000, all edges from tree 7 are independent but when n = 50000, all edges are dependent.

2. **interval = 2**

   - The best fit family for most trees is different both n = 10000 and 50000, excepting, when tree = 4, 5, 6 and 9, the best fit family is similar, that is Gaussian.
   - The estimator of pair-copulas parameter of each edge for tree 4, 5 and 9 decreases when sample size increases but for tree 6, the estimator increases when sample size increases.

D-Vine Angle Dataset, Dimension = 10, Skip = 1

	n = 10000		n = 50000	
Pair-copula(Edge)	**Best fit family**	**Estimated**	**Best fit family**	**estimated**
12	Gumbel	1.10620051*	Gumbel	1.11122339*
23	Gumbel	1.10632430*	Gumbel	1.11124068*
34	Gumbel	1.10629388*	Gumbel	1.11123642*
45	Gumbel	1.010634391*	Gumbel	1.11124376*
56	Gumbel	1.10638950*	Gumbel	1.11126031*
67	Gumbel	1.10640111*	Gumbel	1.11126865*
78	Gumbel	1.10632263*	Gumbel	1.11126595*
89	Gumbel	1.10634497*	Gumbel	1.11126194*
910	Gumbel	1.10617025*	Gumbel	1.11121872*
13/2	Gaussian	.07998308*	Gaussian	.09066024*
24/3	Gaussian	.07979491*	Gaussian	.09063484*
35/4	Gaussian	.07978876*	Gaussian	.09064434*
46/5	Gaussian	.07986084*	Gaussian	.09066578*
57/6	Gaussian	.07993519*	Gaussian	.09066986*
68/7	Gaussian	.07986309*	Gaussian	.09067473*
79/8	Gaussian	.08004171*	Gaussian	.09069564*
810/9	Gaussian	.08037421*	Gaussian	.09074134*
14/23	Gaussian	.04525652*	Gaussian	.06746236*
25/34	Gaussian	.04513136*	Gaussian	.06744377*
36/45	Gaussian	.04505036*	Gaussian	.06743529*
47/56	Gaussian	.04510724*	Gaussian	.06742197*
58/67	Gaussian	.04506394*	Gaussian	.06743167*
69/78	Gaussian	.04513326*	Gaussian	.06743734*
710/89	Gaussian	.04527113*	Gaussian	.06746619*
15/234	Frank	.27416759*	Frank	.35334819*
26/345	Frank	.27429153*	Frank	.35340471*
37/456	Frank	.27312874*	Frank	.35336763*
48/567	Frank	.27198212*	Frank	.35334860*
59/678	Frank	.27173976*	Frank	.35319355*
610/789	Frank	.27159698*	Frank	.35353019*
16/2345	Gumbel	1.01457067	Gaussian	.04224718*
27/3456	Gumbel	1.01460183	Gaussian	.04228181*
38/4567	Gumbel	1.01461597*	Gaussian	.04227267*
49/5678	Gumbel	1.01463313*	Gaussian	.04226405*
510/6789	Gumbel	1.01460252*	Gaussian	.04227595*
17/23456	Gaussian	.04862847*	Gaussian	.04754477*
28/34567	Gaussian	.04883379*	Gaussian	.04756773*
39/45678	Gaussian	.04895098*	Gaussian	.04759549*
410/56789	Gaussian	.04870248*	Gaussian	.04767215*
18/234567	Gaussian	.02542544	Gaussian	.03299304*
29/345678	Gaussian	.02516106	Gaussian	.03294285*
310/456789	Gaussian	.02550702	Gaussian	.03295014*
19/2345678	Frank	.11288487	Frank	.17641889
210/3456789	Frank	.11242082*	Frank	.17597262*
110/23456789	Frank	.16939812*	Frank	.18528238*

*Independent test is significant at 0.05

Figure 3.19: D-vine of angle dataset when sample size = 10000, 50000 and interval = 1 - The figure shows the best fit family, MLE of pair-copulas parameter and independent test in each edge for D-vine of angle dataset when sample size = 10000, 50000 and interval = 1.

D-Vine Angle Dataset, Dimension = 10, Skip = 2

n = 10000			n = 50000	
Pair-copula(Edge)	**Best fit family**	**Estimated**	**Best fit family**	**Estimated**
12	Gaussian	.021681143*	Clayton	.02176477
23	Gaussian	.021851479*	Clayton	.02179607
34	Gaussian	.021667939	Clayton	.02181922
45	Gaussian	.021593543	Clayton	.02180741
56	Gaussian	.021763274	Clayton	.02182469
67	Gaussian	.021629221*	Clayton	.02181479*
78	Gaussian	.021569500	Clayton	.02182962
89	Gaussian	.021672171	Clayton	.02182062
910	Gaussian	.021566341*	Clayton	.02176954
13/2	Gumbel	1.042261044*	Gaussian	.09827332*
24/3	Gumbel	1.042230246*	Gaussian	.09826101*
35/4	Gumbel	1.042316072*	Gaussian	.09826987*
46/5	Gumbel	1.042294651*	Gaussian	.09830405*
57/6	Gumbel	1.042325458*	Gaussian	.09830470*
68/7	Gumbel	1.042297846*	Gaussian	.09828228*
79/8	Gumbel	1.042306799*	Gaussian	.09826220*
810/9	Gumbel	1.042355589*	Gaussian	.09830649*
14/23	Frank	-.223755829*	Gaussian	-.02093765
25/34	Frank	-.225216065*	Gaussian	-.02098103*
36/45	Frank	-.224786901*	Gaussian	-.02099686*
47/56	Frank	-.223468696*	Gaussian	-.02098866*
58/67	Frank	-.222209931*	Gaussian	-.02094590*
69/78	Frank	-.221524732*	Gaussian	-.02091511
710/89	Frank	-.219759997*	Gaussian	-.02088303*
15/234	Gaussian	.073875191*	Gaussian	.06984853*
26/345	Gaussian	.074045059*	Gaussian	.06987492*
37/456	Gaussian	.073639686*	Gaussian	.06987849*
48/567	Gaussian	.073667672*	Gaussian	.06989034*
59/678	Gaussian	.073517738*	Gaussian	.06984575*
610/789	Gaussian	.073527744*	Gaussian	.06985718*
16/2345	Gaussian	-.005836210	Gaussian	-.02012665*
27/3456	Gaussian	-.005842996	Gaussian	-.02012839*
38/4567	Gaussian	-.005876284	Gaussian	-.02015636
49/5678	Gaussian	-.005913228	Gaussian	-.02017179
510/6789	Gaussian	-.006217392	Gaussian	-.02021975
17/23456	Gaussian	.047051734*	Gaussian	.06531737*
28/34567	Gaussian	.047214743*	Gaussian	.06532834*
39/45678	Gaussian	.047264234*	Gaussian	.06536702*
410/56789	Gaussian	.047331265*	Gaussian	.06535744*
18/234567	Frank	.023514631	Gaussian	-.02130559*
29/345678	Frank	.022478671	Gaussian	-.02131038*
310/456789	Gaussian	-.003836455	Gaussian	-.02129719*
19/2345678	Gumbel	1.010164619	Gaussian	.04416252*
210/3456789	Gumbel	1.010040637*	Gaussian	.04410723*
110/23456789	Gaussian	-.003102843	Gaussian	-.01045968

*Independent test is significant at 0.05

Figure 3.20: D-vine of angle dataset when sample size = 10000, 50000 and interval = 2 - The figure shows the best fit family, MLE of pair-copulas parameter and independent test in each edge for D-vine of angle dataset when sample size = 10000, 50000 and interval = 2.

D-Vine Angle Dataset, Dimension = 10, Skip = 3

n = 10000			n = 50000	
Pair-copula(Edge)	**Best fit family**	**Estimated**	**Best fit family**	**Estimated**
12	Gaussian	.11272121*	Gumbel	1.06469865*
23	Gaussian	.11260939*	Gumbel	1.06470124*
34	Gaussian	.11256604*	Gumbel	1.06471927*
45	Gaussian	.11253488*	Gumbel	1.06472028*
56	Gaussian	.11272101*	Gumbel	1.06475277*
67	Gaussian	.11289378*	Gumbel	1.06479118*
78	Gaussian	.11294201*	Gumbel	1.06480685*
89	Gaussian	.11355430*	Gumbel	1.06481102*
910	Gaussian	.11379957*	Gumbel	1.06481846*
13/2	Frank	.50203355*	Gaussian	.07512475*
24/3	Frank	.50341437*	Gaussian	.07513785*
35/4	Frank	.50380339*	Gaussian	.07513715*
46/5	Frank	.50518151*	Gaussian	.07516961*
57/6	Frank	.50435991*	Gaussian	.07514905*
68/7	Frank	.50409404*	Gaussian	.07511387*
79/8	Frank	.50182715*	Gaussian	.07507515*
810/9	Frank	.50198766*	Gaussian	.07509203*
14/23	Gaussian	.06456690*	Gaussian	.05608811*
25/34	Gaussian	.06466590*	Gaussian	.05609484*
36/45	Gaussian	.06455751*	Gaussian	.05610178*
47/56	Gaussian	.06447320*	Gaussian	.05609338*
58/67	Gaussian	.06458345*	Gaussian	.05610368*
69/78	Gaussian	.06502937*	Gaussian	.05615087*
710/89	Gaussian	.06504945*	Gaussian	.05616798*
15/234	Frank	.25665330*	Frank	.28748734*
26/345	Frank	.25619326*	Frank	.28722519*
37/456	Frank	.25696288*	Frank	.28725355*
48/567	Frank	.25842538*	Frank	.28740422*
59/678	Frank	.25682378*	Frank	.28711625*
610/789	Frank	.25674503*	Frank	.28714560*
16/2345	Frank	.14194878	Frank	.22540174*
27/3456	Frank	.14243748	Frank	.22550434*
38/4567	Frank	.14106322*	Frank	.22564492*
49/5678	Frank	.14136687*	Frank	.22548642*
510/6789	Frank	.14171082*	Frank	.22557487*
17/23456	Frank	.18607377	Gaussian	.03289064
28/34567	Frank	.18605915*	Gaussian	.03287274*
39/45678	Frank	.18506288*	Gaussian	.03287774*
410/56789	Frank	.18573679*	Gaussian	.03289413*
18/234567	Frank	.14350168	Frank	.18176378*
29/345678	Frank	.14609264	Frank	.18193374
310/456789	Frank	.14556407	Frank	.18210831*
19/2345678	Gaussian	.02716039	Gaussian	.02261333
210/3456789	Gaussian	.02715994*	Gaussian	.02261057*
110/23456789	Frank	.15720177*	Gaussian	.02268634*

*Independent test is significant at 0.05

Figure 3.21: D-vine of angle dataset when sample size = 10000, 50000 and interval = 3 - The figure shows the best fit family, MLE of pair-copulas parameter and independent test in each edge for D-vine of angle dataset when sample size = 10000, 50000 and interval = 3.

D-Vine Angle Dataset, Dimension = 10, Skip = 5

	n = 10000		n = 50000	
Pair-copula(Edge)	Best fit family	Estimated	Best fit family	Estimated
12	Student-t	.10286037, 27.07826*	Gaussian	.10393600*
23	Student-t	.10270977, 27.08157*	Gaussian	.10394122*
34	Student-t	.10295820, 27.04640*	Gaussian	.10395245*
45	Student-t	.10279326, 27.06626*	Gaussian	.10394202*
56	Student-t	.10300582, 27.08519*	Gaussian	.10398432*
67	Student-t	.10309739, 27.07450*	Gaussian	.10394777*
78	Student-t	.10320442, 27.10399*	Gaussian	.10397772*
89	Student-t	.10320432, 27.08848*	Gaussian	.10389618*
910	Student-t	.10337966, 27.07792*	Gaussian	.10390628*
13/2	Gaussian	.07632154*	Gaussian	.06889225*
24/3	Gaussian	.07635409*	Gaussian	.06891066*
35/4	Gaussian	.07632808*	Gaussian	.06891015*
46/5	Gaussian	.07637075*	Gaussian	.06894336*
57/6	Gaussian	.07637119*	Gaussian	.06892379*
68/7	Gaussian	.07649505*	Gaussian	.06891066*
79/8	Gaussian	.07656622*	Gaussian	.06883280*
810/9	Gaussian	.07664987*	Gaussian	.06884634*
14/23	Gaussian	.05304979*	Gaussian	.06080905*
25/34	Gaussian	.05324412*	Gaussian	.06082768*
36/45	Gaussian	.05307066*	Gaussian	.06078236*
47/56	Gaussian	.05287210*	Gaussian	.06077623*
58/67	Gaussian	.05280260*	Gaussian	.06074868*
69/78	Gaussian	.05304571*	Gaussian	.06087660*
710/89	Gaussian	.05311488*	Gaussian	.06088391*
15/234	Gumbel	1.00699653	Gaussian	.03945292
26/345	Gumbel	1.00694070	Gaussian	.03939329*
37/456	Gumbel	1.00690724	Gaussian	.03943167*
48/567	Gumbel	1.00687873	Gaussian	.03941708*
59/678	Gumbel	1.00685039	Gaussian	.03941832*
610/789	Gumbel	1.00683379	Gaussian	.03939255*
16/2345	Frank	.19017376*	Gumbel	1.01004300
27/3456	Frank	.19116193*	Gumbel	1.01004984
38/4567	Frank	.19051654*	Gumbel	1.01005804*
49/5678	Frank	.18872851*	Gumbel	1.01005885
510/6789	Frank	.18962480*	Gumbel	1.01006404*
17/23456	Gaussian	.03314633	Gaussian	.02874120
28/34567	Gaussian	.03327982*	Gaussian	.02876373*
39/45678	Gaussian	.03323116*	Gaussian	.02871264*
410/56789	Gaussian	.03319030*	Gaussian	.02872456*
18/234567	Gaussian	.02403336	Gaussian	.03023786*
29/345678	Gaussian	.02428541	Gaussian	.03024403
310/456789	Gaussian	.02413502	Gaussian	.03023998
19/2345678	Gaussian	.01746271	Gaussian	.01982124
210/3456789	Gaussian	.01741214*	Gaussian	.01980184*
110/23456789	Gaussian	-.00376644	Gumbel	1.00608288

*Independent test is significant at 0.05

Figure 3.22: D-vine of angle dataset when sample size = 10000, 50000 and interval = 5 - The figure shows the best fit family, MLE of pair-copulas parameter and independent test in each edge for D-vine of angle dataset when sample size = 10000, 50000 and interval = 5.

- For the independent test for pair-copulas parameter, some edges both n = 10000 and 50000 can be dependent or independent. As all edges from tree 4 and 6 are dependent both n = 10000 and 50000.

3. **interval = 3**

   - The best fit family for most trees is quite similar both n = 10000 and 50000, for example, when tree = 3 and 8, the best fit family is Gaussian and the best fit family of each edge for three 4, 5 and 7 is Frank.
   - The estimator of pair-copulas parameter of each edge for tree 4, 5 and 7 increases when sample size increases but for tree 3 and 8, the estimator decreases when sample size increases.
   - For the independent test for pair-copulas parameter, most edges for all trees both n = 10000 and 50000 are dependent.

4. **interval = 5**

   - The best fit family for most trees is quite similar both n = 10000 and 50000, for example, when tree = 2, 3, 6, 7 and 8, the best fit family is Gaussian.
   - The estimator of pair-copulas parameter of each edge for tree 3, 7 - 8 increases when sample size increases but for tree 2 and 6, the estimator decreases when sample size increases.
   - For the independent test for pair-copulas parameter, most edges for all trees both n = 10000 and 50000 are dependent.

Conclusion: For angle dataset in ten dimensions, we conclude results following.

1. When we select the observations with interval = 1, the best fit family and independent test for all trees from both sample sizes (n =10000 and 50000) are similar (except for tree 5 when n = 100000, the best fit family is different). We may say that D-vine results for test systems are not different when we take the observations with interval = 1 and any sample size (Figure 3.19).

2. When we select the observations with interval = 2 and n = 10000 and 50000, the best fit family and independent test for all trees are different (except for tree 4 and 5 from both sample sizes, the best fit family is similar, as the independent test is quite similar). We may say that D-vine results are different when we take the observations with interval = 2 (Figure 3.20).

3. When we select the observations with interval = 3 and n = 10000 and 50000, the best fit family and independent test for all trees are different (except for tree 3, 4, 5, 7 and 8 from both sample sizes, the best fit family is similar, as the independent test is quite similar). We may say that D-vine results are different when we take the observations with interval = 3 (Figure 3.21).

4. When we select the observations with interval = 5 and n = 10000 and 50000, the best fit family and independent test for all trees are different (except for tree 2, 3, 6, 7 and 8 from both sample sizes, the best fit family is similar, as the independent test is quite similar). We may say that D-vine results are different when we take the observations with interval = 5 (Figure 3.22).

5. We found that D-vine results for angle dataset in ten dimensions are different when we take the observations with interval. When we take the observations with high values interval (interval = 3 and 5) and n = 10000 and 50000, the D-vine results are quite similar and at the high level of tree, the best fit family is Gaussian.

Clearly, all D-vine results for angle dataset both 5 and 10 dimensions give useful information, for example, the best fit family, estimator and independent test of pair-copulas parameter of each edge. For the following part, we are interested in the graphical analysis from three plots: Chi-plot, K-plot and Lambda-function plot that are the graphical tools for independent test. We fit the D-vine for angle dataset from small sample size: n = 500 and make the independent test, the results are given in Table 3.15.

Table 3.15: The best fit family, MLE of pair-copulas parameter and independent test of each edge for D-vine of angle dataset when sample size = 500 and interval = 5.

Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
1	12	Frank	0.17058261	Independent
	23	Frank	0.15235078	Independent
	34	Frank	0.15469816	Independent
	45	Frank	0.14455129	Independent
2	13/2	Gaussian	-0.02244981	Independent
	24/3	Gaussian	-0.02338721	Independent
	35/4	Gaussain	-0.02419226	Independent
3	14/23	Joe	1.03677501	Independent
	25/34	Student-t	0.1822808, 10.9061(df.)	Independent
4	15/234	Frank	0.06419995	Independent

From Table 3.15, the best fit family of tree 2 and 3 are different, as Frank is the best fit family for tree 1 and 4. For the independent test of pair-copulas parameter of each edge and tree, all edges are independent. The Chi-plot, K-plot and Lambda-function plot for the independent test of each edge from Table 3.15 are given in Figure 3.23 - 3.32. Clearly, from Figure 3.23 - 3.32, all Chi-plot, K-plot and Lambda-function plot are similar and show that all egdes from all trees are independent.

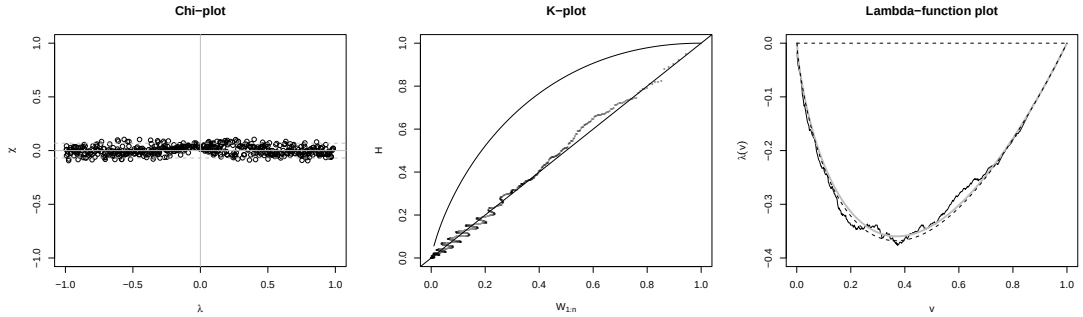

Figure 3.23: Chi-plot, K-plot and Lambda-function plot of angle dataset: edge 12 - The figure shows Chi-plot, K-plot and Lambda-function plot of angle dataset for edge: 12.

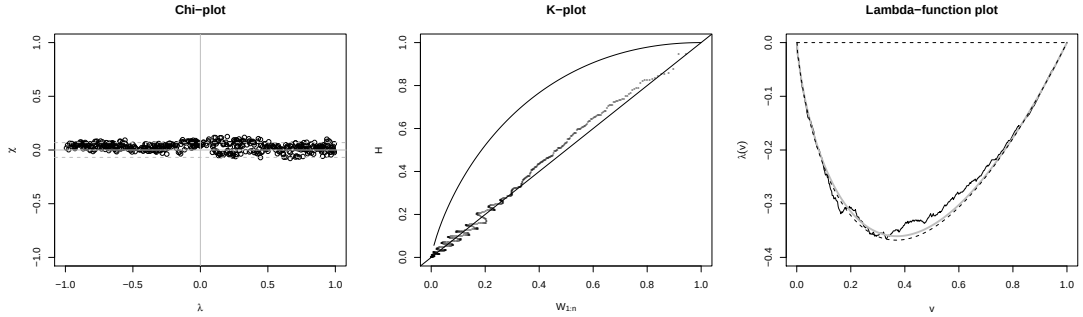

Figure 3.24: Chi-plot, K-plot and Lambda-function plot of angle dataset: edge 23 - The figure shows Chi-plot, K-plot and Lambda-function plot of angle dataset for edge: 23.

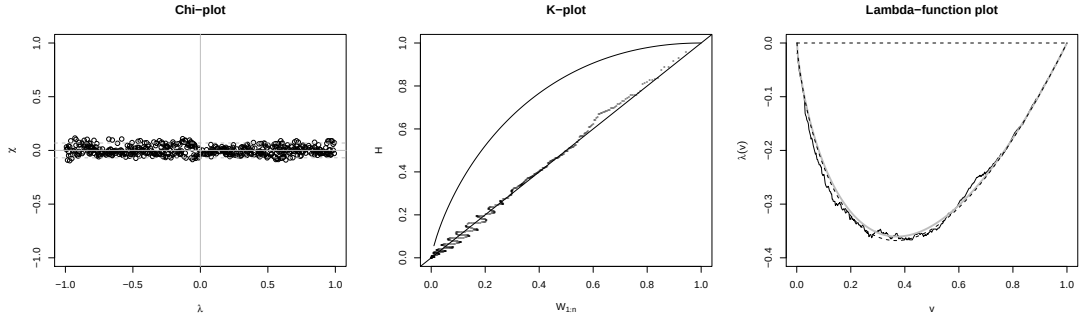

Figure 3.25: Chi-plot, K-plot and Lambda-function plot of angle dataset: edge 34 - The figure shows Chi-plot, K-plot and Lambda-function plot of angle dataset for edge: 34.

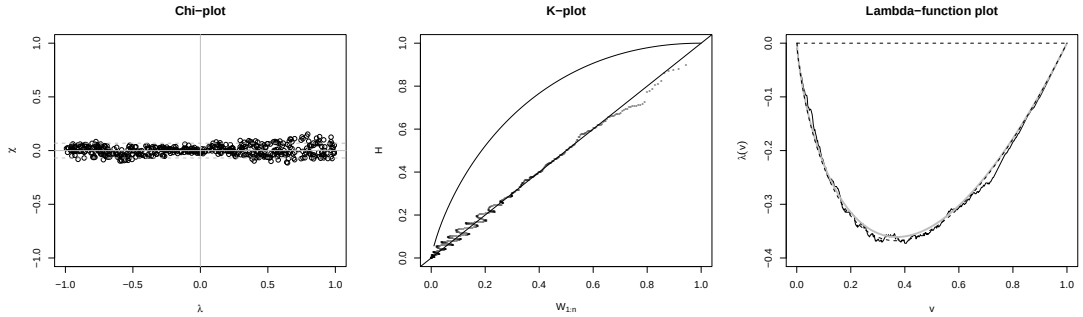

Figure 3.26: Chi-plot, K-plot and Lambda-function plot of angle dataset: edge 45 - The figure shows Chi-plot, K-plot and Lambda-function plot of angle dataset for edge: 45.

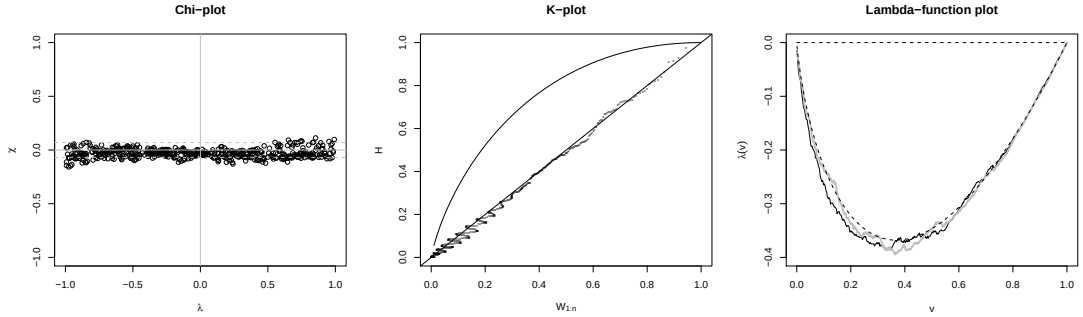

Figure 3.27: Chi-plot, K-plot and Lambda-function plot of angle dataset: edge 13/2 - The figure shows Chi-plot, K-plot and Lambda-function plot of angle dataset for edge: 13/2.

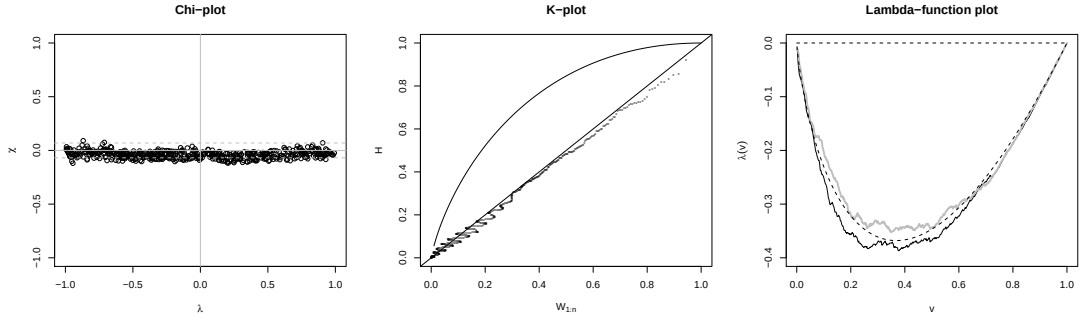

Figure 3.28: Chi-plot, K-plot and Lambda-function plot of angle dataset: edge 24/3 - The figure shows Chi-plot, K-plot and Lambda-function plot of angle dataset for edge: 24/3.

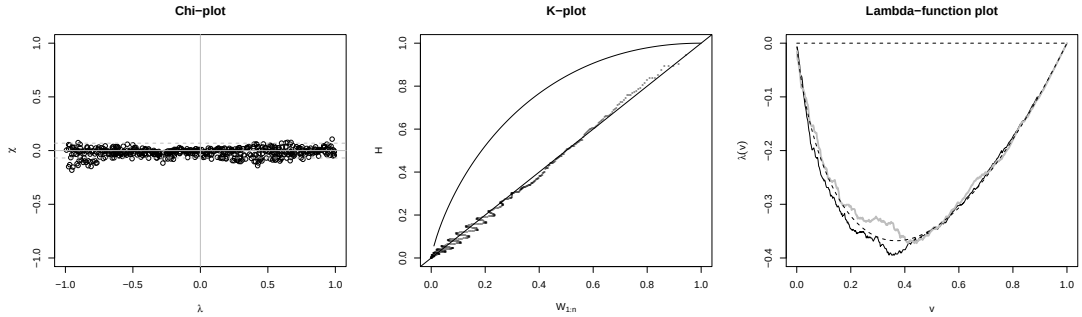

Figure 3.29: Chi-plot, K-plot and Lambda-function plot of angle dataset: edge 35/4 - The figure shows Chi-plot, K-plot and Lambda-function plot of angle dataset for edge: 35/4.

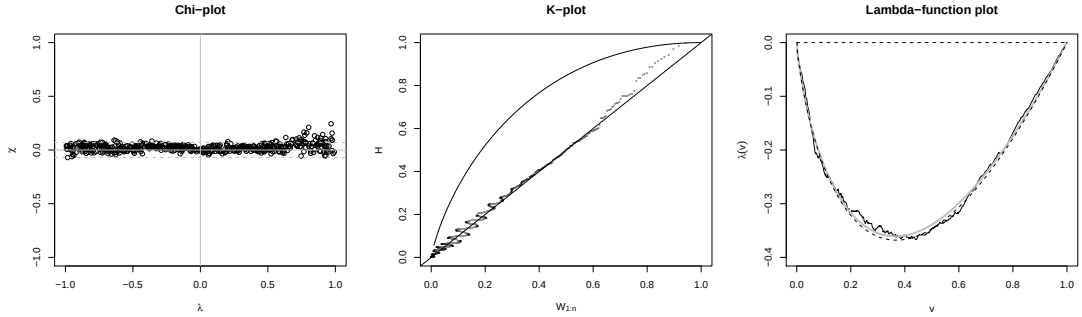

Figure 3.30: Chi-plot, K-plot and Lambda-function plot of angle dataset: edge 14/23 - The figure shows Chi-plot, K-plot and Lambda-function plot of angle dataset for edge: 14/23.

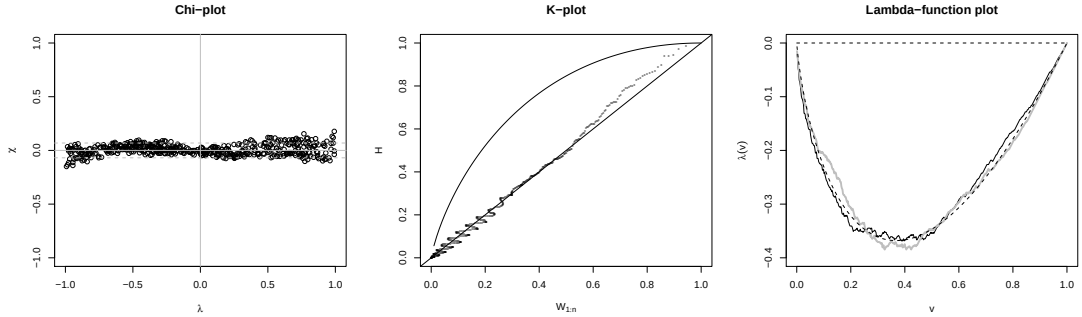

Figure 3.31: Chi-plot, K-plot and Lambda-function plot of angle dataset: edge 25/34 - The figure shows Chi-plot, K-plot and Lambda-function plot of angle dataset for edge: 25/34.

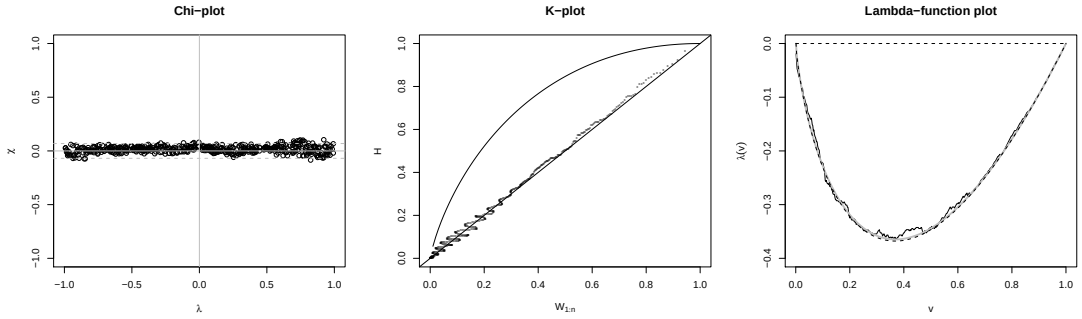

Figure 3.32: Chi-plot, K-plot and Lambda-function plot of angle dataset: edge 15/234 - The figure shows Chi-plot, K-plot and Lambda-function plot of angle dataset for edge: 15/234.

Conclusion: For angle dataset in five dimensions with sample size = 500 and interval = 5, we conclude results following.

1. When we select the observations with interval = 5 and sample size = 500, the best fit family for all trees is different (except for tree 1 and 4, the best fit family is similar: Frank). As independent test for all trees is similar: independent (see Table 3.15).

2. All three plots: Chi-plot, K-plot and Lambda-function plot from all edges and trees show that all edges are independent according to the statistical independent test from Table 3.15 (see Figure 3.23 - 3.32).

3.4.2 Amplitude Dataset

3.4.2.1 Dimension = 5

First of all, we carry out the data analysis according to the algorithms from the previous section following.

1. We examine the marginal distribution of X_1, X_2, X_3, X_4 and X_5 separately by plotting their histograms and densities. From Figure 3.16 and 3.17, the histograms and densities show that X_1, X_2, X_3, X_4 and X_5 are uniform (0,1) distributed.

2. Let $X_1 = U_1$, $X_2 = U_2$, ..., $X_5 = U_5$ be a uniform (0,1) variable.

3. We fit D-vine to U_1, U_2, U_3, U_4 and U_5 using maximum likelihood method. The fitted family, MLE of pair-copulas parameter and independent test of each edge are given in Table 3.16.

Table 3.16: The best fit family, MLE of pair-copulas parameter and independent test of each edge for D-vine of amplitude dataset when sample size = 10000 and interval = 0.

Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
1	12	Joe	1.02505395	Independent
	23	Joe	1.02483765	Independent
	34	Joe	1.02499034	Independent
	45	Joe	1.02529102	Independent
2	13/2	Gumbel	1.02485330	Independent
	24/3	Gumbel	1.02503373	Independent
	35/4	Gumbel	1.02490471	Independent
3	14/23	Student-t	0.05545252	Independent
	25/34	Student-t	0.05569520	Independent
4	15/234	Frank	0.18920396	Independent

From Table 3.16, we fit D-vine for amplitude dataset when sample size = 10000 and no interval. The results can summarize following.

- The best fit family of each edge in same tree is similar. The best fit family for tree 1, 2, 3 and 4 is Joe, Gumbel, Student-t and Frank respectively.
- When we consider the estimator of pair-copulas parameter of each edge and tree, the estimator is quite similar, for example, the estimator for the second tree is between 1.02485330 and 1.02503373, the difference is approximately 0.00018 etc. See Table 3.17 for the differences of estimator.

Table 3.17: The estimator differences of each tree of amplitude dataset when sample size = 10000 and interval = 0.

Tree	Best fit family	max.	min.	difference
1	Joe	1.02529102	1.02483765	0.00045337
2	Gumbel	1.02503373	1.02485330	0.00018043
3	Student-t	0.05569520	0.05545252	0.00024268

- For the independent test, all pair-copulas parameters of each edge and trees are independent when the test systems observations do not interval.

For amplitude dataset, we have tried to interval the observations, say, interval = 2, 5 and 10 and fit and investigate the D-vine results.

For interval = 2, the results are given in Table 3.18. We summarize following.

- The best fit family of each edge in same tree both n = 10000 and 50000 is similar. For n = 10000 and 50000, the best fit family of each edge for the first tree is

Table 3.18: The best fit family, MLE of pair-copulas parameter and independent test of each edge for D-vine of amplitude dataset when sample size = 10000, 50000 and interval = 2.

n	Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
10000	1	12	Student-t	0.05381809	Dependent
		23	Student-t	0.05376497	Dependent
		34	Student-t	0.05391102	Dependent
		45	Student-t	0.05383518	Dependent
	2	13/2	Gaussian	0.04654094	Dependent
		24/3	Gaussian	0.04651902	Dependent
		35/4	Gaussian	0.04677893	Dependent
	3	14/23	Joe	1.01360257	Independent
		25/34	Joe	1.01369486	Independent
	4	15/234	Gumbel	1.00720106	Independent
50000	1	12	Student-t	0.04503454	Dependent
		23	Student-t	0.04504611	Dependent
		34	Student-t	0.04506819	Dependent
		45	Student-t	0.04504407	Dependent
	2	13/2	Gumbel	1.02031625	Independent
		24/3	Gumbel	1.02031793	Dependent
		35/4	Gumbel	1.02035261	Dependent
	3	14/23	Gumbel	1.00948104	Independent
		25/34	Gumbel	1.00950404	Independent
	4	15/234	Gumbel	1.00897995	Independent

Table 3.19: The estimator differences of each tree of amplitude dataset when sample size = 10000, 50000 and interval = 2.

n	Tree	Best fit family	max.	min.	difference
10000	1	Student-t	0.05391102	0.05376497	0.00014605
	2	Gaussian	0.04677893	0.04651902	0.00025991
	3	Joe	1.01369486	1.01360257	0.00009229
50000	1	Student-t	0.04506819	0.04503454	0.00003365
	2	Gumbel	1.02035261	1.02031625	0.00003636
	3	Gumbel	1.00950404	1.00948104	0.000023

Student-t and Gumbel is the best fit family of all edges for tree 2, 3 and 4 when n = 50000.

- When we consider the estimator of pair-copulas parameter of each edge and sample size in same tree, the estimator is quite similar, for example, n = 50000, the estimator for the first tree is between 0.04503454 and 0.04506819, the difference is 0.00003365 etc. The differences of estimator are given in Table 3.19. Moreover, for the same best fit family in tree 1 both n = 10000 and 50000: Student-t, the estimator decreases when sample size increases.
- From Table 3.19, for n = 10000 and 50000, the differences are very small.
- For n = 10000, the independent test for pair-copulas parameter of each edge in tree 1 and 2 are similar, that is, all edges are dependent. As all edges in tree 3 and 4 are independent.
- For n = 50000, the independent test for pair-copulas parameter of each edge in tree 3 and 4 are similar, that is, all edges are independent. As all edges in tree 1 and edge 24/3 and 35/4 in tree 2 are dependent.

For interval = 5, the results are given in Table 3.20. We summarize following.

- The best fit family of each edge in same tree both n = 10000 and 50000 is similar. For n = 10000 and n = 50000, the best fit family of each edge for the first tree is Gaussian and Clayton is the best fit family of edge 15/234 for tree 4. As Joe and Gumbel are the best fit family of each edge in tree 2 and 3 when n = 10000 and 50000, respectively.
- When we consider the estimator of pair-copulas parameter of each edge and sample size in same tree, the estimator is quite similar, for example, n = 10000, the estimator for the second tree is between 1.00892862 and 1.00895327, the difference is 0.00002465 etc. The differences of estimator are given in Table 3.21. Moreover, for the same best fit family in tree 1 both n = 10000 and 50000: Gaussian, the estimator decreases when sample size increases.
- From Table 3.21, the differences are very small both n = 10000 and 50000.
- For n = 10000, the independent test for pair-copulas parameter of all edges in tree 1 are dependent. As all edges in tree 2, 3 and 4 are independent.
- For n = 50000, the independent test for pair-copulas parameter of all edges in tree 1 and edge 24/3 in tree 2 are dependent. As all edges in tree 3, 4 and edge 13/23 and 35/4 in tree 2 are independent.

For interval = 10, we have taken a variety of sample sizes, for example, n = 50, 100, 500 etc. and investigated the results. The interesting D-vine results are given in Table 3.22 - 3.25. We summarize following.

Table 3.20: The best fit family, MLE of pair-copulas parameter and independent test of each edge for D-vine of amplitude dataset when sample size = 10000, 50000 and interval = 5.

n	Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
10000	1	12	Gaussian	0.03807380	Dependent
		23	Gaussian	0.03809083	Dependent
		34	Gaussian	0.03800228	Dependent
		45	Gaussian	0.03792276	Dependent
	2	13/2	Joe	1.00893926	Independent
		24/3	Joe	1.00895327	Independent
		35/4	Joe	1.00892862	Independent
	3	14/23	Joe	1.01678585	Independent
		25/34	Joe	1.01686493	Independent
	4	15/234	Clayton	0.01513928	Independent
50000	1	12	Gaussian	0.03714331	Dependent
		23	Gaussian	0.03713724	Dependent
		34	Gaussian	0.03714251	Dependent
		45	Gaussian	0.03713840	Dependent
	2	13/2	Gumbel	1.01175764	Independent
		24/3	Gumbel	1.01175347	Dependent
		35/4	Gumbel	1.01175149	Independent
	3	14/23	Gumbel	1.01142917	Independent
		25/34	Gumbel	1.01143467	Independent
	4	15/234	Clayton	0.01628561	Independent

Table 3.21: The estimator differences of each tree of amplitude dataset when sample size = 10000, 50000 and interval = 5.

n	Tree	Best fit family	max.	min.	difference
10000	1	Gaussian	0.03809083	0.03792276	0.00016807
	2	Joe	1.00895327	1.00892862	0.00002465
	3	Joe	1.01686493	1.01678585	0.00007908
50000	1	Gaussian	0.03714331	0.03713724	0.00000607
	2	Gumbel	1.01175764	1.01175149	0.00000615
	3	Gumbel	1.01143467	1.01142917	0.0000055

1. **The best fit family**

   - For n = 50, the best fit family of each edge for the first tree is different: Clayton, Gaussian, Frank and Joe, respectively. As the second and thrid tree, Frank is the best fit family for all edges and the best fit family for edge 15/234 in tree 4 is Gaussian.
   - For n = 100, the best fit family of each edge for the second and thrid tree is similar, that is Frank. As Gaussian is the best fit family for edge 15/234 in tree 4 and for tree 1, the best fit family of edge 12 and 23 is Frank, edge 34 and 45 is Gaussian.
   - For n = 500, the best fit family of each edge for the first tree is similar, that is Clayton. As Frank is the best fit family for all edges in tree 2, 3 and 4.
   - For n = 1000, 5000, 10000 and 20000, the best fit family of each edge for the first and fourth tree is similar, that is Clayton. Moreover, when n = 10000 and 20000, the best fit family of all edges from every trees is definitely similar, for example, the best fit family for all edges in tree 1, 3 and 4 is Clayton etc.
   - For n = 30000, the best fit family of each edge for the first, second and thrid tree is similar, that is Gumbel. As Clayton is the best fit family for edge 15/234 in tree 4.
   - For n = 40000, 50000 and 70000, the best fit family of all edges from every trees is definitely similar, that is Gumbel.
   - For n = 60000, the best fit family of each edge for the first, thrid and fourth tree is similar, that is Gumbel. As Clayton is the best fit family for all edges in tree 2.

2. **The estimator of pair-copulas parameter**

   - Normally, when we consider the estimator of pair-copulas parameter of each edge and sample size in same tree, the estimator is quite similar, for example, n = 30000, the estimator for the first tree is between 1.00894512 and 1.00895967, the difference is 0.00001455 etc.

3. **The independent test**

   - For n = 50, 100 and 500, the independent test for pair-copulas parameter of most edges are dependent.
   - For n $\geq$ 1000, the independent test for pair-copulas parameter of each edge tends to be independent.
   - For n $\geq$ 10000, the independent test for pair-copulas parameter of all edges is definitely similar, especially edge 24/3 is dependent.

Table 3.22: The best fit family, MLE of pair-copulas parameter and independent test of each edge for D-vine of amplitude dataset when sample size = 50, 100, 500 and interval = 10.

n	Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
50	1	12	Clayton	0.30898248	Dependent
		23	Gaussian	0.04028315	Dependent
		34	Frank	-0.25060016	Dependent
		45	Joe	1.02703649	Independent
	2	13/2	Frank	-0.93412275	Dependent
		24/3	Frank	-0.50498392	Independent
		35/4	frank	-0.75199224	Dependent
	3	14/23	Frank	-1.03419099	Dependent
		25/34	Frank	-0.73489602	Dependent
	4	15/234	Gaussian	-0.24003994	Dependent
100	1	12	Frank	0.07930623	Independent
		23	Frank	0.09105638	Independent
		34	Gaussian	-0.01635361	Dependent
		45	Gaussian	-0.01471796	Dependent
	2	13/2	Frank	-0.73512869	Dependent
		24/3	Frank	-0.73986490	Dependent
		35/4	Frank	-0.72515772	Dependent
	3	14/23	Frank	-0.06602848	Independent
		25/34	Frank	-0.06547951	Independent
	4	15/234	Gaussian	-0.11951409	Dependent
500	1	12	Clayton	0.08246362	Dependent
		23	Clayton	0.08266453	Dependent
		34	Clayton	0.08201867	Dependent
		45	Clayton	0.08197900	Dependent
	2	13/2	Frank	-0.40256109	Dependent
		24/3	Frank	-0.39608615	Dependent
		35/4	Frank	-0.39956456	Dependent
	3	14/23	Frank	-0.09685026	Dependent
		25/34	Frank	-0.09618729	Independent
	4	15/234	Frank	-0.10467630	Dependent

Table 3.23: The best fit family, MLE of pair-copulas parameter and independent test of each edge for D-vine of amplitude dataset when sample size = 1000, 5000, 10000 and interval = 10.

n	Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
1000	1	12	Clayton	0.02712214	Dependent
		23	Clayton	0.02677438	Dependent
		34	Clayton	0.02624599	Independent
		45	Clayton	0.02609961	Independent
	2	13/2	Gumbel	1.01108277	Independent
		24/3	Gumbel	1.01162303	Dependent
		35/4	Gumbel	1.01148316	Independent
	3	14/23	Frank	-0.02235942	Independent
		25/34	Frank	-0.01948309	Independent
	4	15/234	Clayton	0.02731661	Independent
5000	1	12	Clayton	0.017472094	Independent
		23	Clayton	0.017537782	Independent
		34	Clayton	0.017514233	Independent
		45	Clayton	0.017451232	Independent
	2	13/2	Gaussain	0.006133191	Independent
		24/3	Gaussain	0.006275151	Independent
		35/4	Gaussain	0.006162226	Independent
	3	14/23	Clayton	0.033423662	Dependent
		25/34	Clayton	0.033356167	Dependent
	4	15/234	Clayton	0.015367143	Independent
10000	1	12	Clayton	0.01569222	Independent
		23	Clayton	0.01569607	Independent
		34	Clayton	0.01567268	Independent
		45	Clayton	0.01565847	Independent
	2	13/2	Gumbel	1.01203087	Independent
		24/3	Gumbel	1.01207055	Dependent
		35/4	Gumbel	1.01198581	Independent
	3	14/23	Clayton	0.02439605	Independent
		25/34	Clayton	0.02441037	Independent
	4	15/234	Clayton	0.01765788	Independent

Table 3.24: The best fit family, MLE of pair-copulas parameter and independent test of each edge for D-vine of amplitude dataset when sample size = 20000, 30000, 40000 and interval = 10.

n	Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
20000	1	12	Clayton	0.018748782	Independent
		23	Clayton	0.018749899	Independent
		34	Clayton	0.018714307	Independent
		45	Clayton	0.018712047	Independent
	2	13/2	Gumbel	1.009806708	Independent
		24/3	Gumbel	1.009830321	Dependent
		35/4	Gumbel	1.009783266	Independent
	3	14/23	Clayton	0.019979520	Independent
		25/34	Clayton	0.019978830	Independent
	4	15/234	Clayton	0.008310428	Independent
30000	1	12	Gumbel	1.00895344	Independent
		23	Gumbel	1.00895967	Independent
		34	Gumbel	1.00895126	Independent
		45	Gumbel	1.00894512	Independent
	2	13/2	Gumbel	1.00998215	Independent
		24/3	Gumbel	1.00997721	Dependent
		35/4	Gumbel	1.00998244	Independent
	3	14/23	Gumbel	1.00384583	Independent
		25/34	Gumbel	1.00384977	Independent
	4	15/234	Clayton	0.01030207	Independent
40000	1	12	Gumbel	1.009509	Independent
		23	Gumbel	1.0095911	Independent
		34	Gumbel	1.009539	Independent
		45	Gumbel	1.009540	Independent
	2	13/2	Gumbel	1.009492	Independent
		24/3	Gumbel	1.009504	Dependent
		35/4	Gumbel	1.009496	Independent
	3	14/23	Gumbel	1.006083	Independent
		25/34	Gumbel	1.006081	Independent
	4	15/234	Gumbel	1.003720	Independent

Table 3.25: The best fit family, MLE of pair-copulas parameter and independent test in each edge for D-vine of amplitude dataset when sample size = 50000, 60000, 70000 and interval = 10.

n	Tree	Pair-copulas(Edge)	Best fit family	estimator	Independent test
50000	1	12	Gumbel	1.012655	Independent
		23	Gumbel	1.012656	Independent
		34	Gumbel	1.012665	Independent
		45	Gumbel	1.012682	Independent
	2	13/2	Gumbel	1.008451	Independent
		24/3	Gumbel	1.008458	Dependent
		35/4	Gumbel	1.008461	Independent
	3	14/23	Gumbel	1.005550	Independent
		25/34	Gumbel	1.005550	Independent
	4	15/234	Gumbel	1.003901	Independent
60000	1	12	Gumbel	1.01120095	Independent
		23	Gumbel	1.01120147	Independent
		34	Gumbel	1.01120496	Independent
		45	Gumbel	1.01121002	Independent
	2	13/2	Clayton	0.01745462	Independent
		24/3	Clayton	0.01745662	Dependent
		35/4	Clayton	0.01746042	Independent
	3	14/23	Gumbel	1.00535491	Independent
		25/34	Gumbel	1.00535454	Independent
	4	15/234	Gumbel	1.00570488	Independent
70000	1	12	Gumbel	1.010039	Independent
		23	Gumbel	1.010040	Independent
		34	Gumbel	1.010045	Independent
		45	Gumbel	1.010048	Independent
	2	13/2	Gumbel	1.009466	Independent
		24/3	Gumbel	1.009472	Dependent
		35/4	Gumbel	1.009469	Independent
	3	14/23	Gumbel	1.006273	Independent
		25/34	Gumbel	1.006271	Independent
	4	15/234	Gumbel	1.006420	Independent

Conclusion: For amplitude dataset in five dimensions, we conclude results following.

1. When we select the observations without interval (interval = 0), the best fit family for all trees is different, as independent test for all trees are similar (Table 3.16).

2. When we select the observations with interval = 2 and n = 10000 and 50000, the best fit family and independent test for all trees are quite similar (except for tree 2 and 3 when n = 50000, the best fit family is different and when n = 50000, the independent test for edge 13/2 is independent, Table 3.18).

3. When we select the observations with interval = 5 and n = 10000 and 50000, the best fit family and independent test for all trees are quite similar (except for tree 2 and 3 when n = 50000, the best fit family is different and when n = 50000, the independent test for edge 24/3 is dependent) (Table 3.20).

4. When we select the observations with interval = 10 and large sample size (n $\geq$ 10000), the best fit family for all trees is quite similar, as independent test for all trees is similar. When sample size $\leq$ 5000, the best fit family and independent test for all trees are different (Table 3.22 - 3.25).

5. We found that D-vine results for amplitude dataset in five dimensions are different when we take the observations without or with interval. When we take the observations with any interval and large sample size, the D-vine results are quite similar.

3.4.2.2 Dimension = 10

For dimension = 10, we carry out the data analysis according to the algorithms as same as dimension = 5 following.

1. We examine the marginal distribution of X_1, X_2, X_3, ..., X_{10} separately by plotting their histograms and densities. The histograms and densities show that all of variables are uniform (0,1) distribution.

2. Let $X_1 = U_1$, $X_2 = U_2$, ..., $X_{10} = U_{10}$ be a uniform (0,1) variable.

3. We fit D-vine to U_1, U_2, U_3, ..., U_{10} using maximum likelihood method. The fitted family, MLE of pair-copulas parameter and independent test of each edge with skipping = 1, 2, 3, 5 are showed in Figure 3.33, 3.34, 3.35 and 3.36.

From Figure 3.33, 3.34, 3.35 and 3.36, we fit D-vine for 10 uniform (0,1) variables of amplitude dataset with interval = 1, 2, 3 and 5. The 10-dimensional D-vine has 9 trees and 45 edges and the results are following.

1. **interval = 1**
    - The best fit family for most trees is different both n = 10000 and 50000, excepting, when tree = 1, 4, 8 and 9, the best fit family is similar, that is Gaussian, Gumbel, Clayton and Clayton, respectively.
    - The estimator of pair-copulas parameter of each edge for tree 1, 4, 8 and 9 that have a similar best fit family decreases when sample size increases.
    - For the independent test for pair-copulas parameter, most edges for tree 1 - 3 both n = 10000 and 50000 are dependent. As, most edges in tree 4 - 9 tend to be independent.
2. **interval = 2**
    - The best fit family for most trees is similar both n = 10000 and 50000, that is, when tree = 1, 4, 5, 7 and 8, the best fit family is Student-t, Gumbel, Gumbel, Frank and Clayton, respectively.
    - The estimator of pair-copulas parameter of each edge for tree 1, 5, 7 and 8 that have a similar best fit family decreases when sample size increases but for tree 4, the estimator increases when sample size increases.
    - For the independent test for pair-copulas parameter, most edges for tree 1 - 2 both n = 10000 and 50000 are dependent. As, most edges in tree 3 - 9 are independent.
3. **interval = 3**
    - The best fit family for most trees is different both n = 10000 and 50000, excepting, when tree = 2, 5, 6 and 8, the best fit family is similar, that is Gumbel, Clayton, Gumbel and Frank, respectively.
    - The estimator of pair-copulas parameter of each edge for tree 2, 6 and 8 that have a similar best fit family increases when sample size increases but for tree 5, the estimator decreases when sample size increases.
    - For the independent test for pair-copulas parameter, most edges for all trees both n = 10000 and 50000 are independent.
4. **interval = 5**
    - The best fit family of each edge in the first tree is quite similar both n = 10000 and 50000, excepting, edge 78 and 89. When tree = 4, 5 and 6, the best fit family is Clayton, Frank and Frank, respectively.
    - The estimator of pair-copulas parameter of each edge for tree 5 and 6 that have a similar best fit family decreases when sample size increases but for tree 4, the estimator increases when sample size increases.
    - For the independent test for pair-copulas parameter, most edges of the first tree both n = 10000 and 50000 are dependent. As most edges of rest of trees are independent.

Conclusion: For amplitude dataset in ten dimensions, we conclude results following.

1. When we select the observations with interval = 1, 2, 3 and 5, the best fit family and independent test for some trees from both sample sizes (n =10000 and 50000) are similar (Table 3.33 - 3.36).

2. We found that D-vine results for amplitude dataset in ten dimensions are different when we take the observations with any interval.

D-Vine Amplitude Dataset, Dimension = 10, Skip = 1

n = 10000			n = 50000	
Pair-copula(Edge)	**Best fit family**	**Estimated**	**Best fit family**	**estimated**
12	Gaussian	.05849204*	Gaussian	.03517041*
23	Gaussian	.05874221*	Gaussian	.03519608*
34	Gaussian	.05841567*	Gaussian	.03514365*
45	Gaussian	.05840175*	Gaussian	.03512379*
56	Gaussian	.05827107*	Gaussian	.03512265*
67	Gaussian	.05826967*	Gaussian	.03512910*
78	Gaussian	.05821660*	Gaussian	.03506970*
89	Gaussian	.05840795*	Gaussian	.03509173*
910	Gaussian	.05841636*	Gaussian	.03512739*
13/2	Joe	1.03172500*	Gumbel	1.01710362*
24/3	Joe	1.03167175*	Gumbel	1.01709231*
35/4	Joe	1.03168780*	Gumbel	1.01709394
46/5	Joe	1.03165364*	Gumbel	1.01710141*
57/6	Joe	1.03165286	Gumbel	1.01707853
68/7	Joe	1.03158931*	Gumbel	1.01707525*
79/8	Joe	1.03158954*	Gumbel	1.01705530*
810/9	Joe	1.03161729*	Gumbel	1.01707770
14/23	Frank	.16230619*	Student-t	.03009092,29.51418*
25/34	Frank	.16139709*	Student-t	.03015072,29.49344*
36/45	Frank	.16297454*	Student-t	.03017025,29.43826*
47/56	Frank	.16247139*	Student-t	.03015866,29.45912*
58/67	Frank	.16309286*	Student-t	.03018373,29.45038*
69/78	Frank	.16203363*	Student-t	.03018228,29.46207*
710/89	Frank	.16246196	Student-t	.03014700,29.45749*
15/234	Gumbel	1.01272526	Gumbel	1.00917442
26/345	Gumbel	1.01284587	Gumbel	1.00915648
37/456	Gumbel	1.01285933	Gumbel	1.00916145
48/567	Gumbel	1.01285840*	Gumbel	1.00914927
59/678	Gumbel	1.01288600	Gumbel	1.00916503
610/789	Gumbel	1.01286531	Gumbel	1.00915822
16/2345	Gaussian	.03566179*	Clayton	.01896240
27/3456	Gaussian	.03565798*	Clayton	.01891383
38/4567	Gaussian	.03525427*	Clayton	.01890785*
49/5678	Gaussian	.03527585*	Clayton	.01890066
510/6789	Gaussian	.03530207*	Clayton	.01892215
17/23456	Joe	1.01995001	Gumbel	1.00987725
28/34567	Joe	1.01988375*	Gumbel	1.00991367
39/45678	Joe	1.01992915*	Gumbel	1.00992122*
410/56789	Joe	1.01997021	Gumbel	1.00992510
18/234567	Student-t	.01457066,29.92050	Gumbel	1.01007413
29/345678	Student-t	.01438608,29.90794	Gumbel	1.01010553
310/456789	Student-t	.01460214,29.79871	Gumbel	1.01013246
19/2345678	Clayton	.01990886	Clayton	.01723853
210/3456789	Clayton	.01994536*	Clayton	.01725404
110/23456789	Clayton	.03136980*	Clayton	.02045858

*Independent test is significant at 0.05

Figure 3.33: D-vine of amplitude dataset when sample size = 10000, 50000 and interval = 1 - The figure shows the best fit family, MLE of pair-copulas parameter and independent test in each edge for D-vine of amplitude dataset when sample size = 10000, 50000 and interval = 1.

D-Vine Ampltude Dataset, Dimension = 10, Skip = 2

n = 10000			n = 50000	
Pair-copula(Edge)	**Best fit family**	**Estimated**	**Best fit family**	**estimated**
12	Student-t	.05381809,23.15331*	Student-t	.04503454,24.27413*
23	Student-t	.05376497,23.13603*	Student-t	.04504611,24.27201*
34	Student-t	.05391102,23.11871*	Student-t	.04506819,24.26838*
45	Student-t	.05383518,23.07191*	Student-t	.04504407,24.26000*
56	Student-t	.05365904,23.03501*	Student-t	.04500606,24.25505*
67	Student-t	.05374981,23.03490*	Student-t	.04502012,24.25165*
78	Student-t	.05380266,22.99701*	Student-t	.04502526,24.24195*
89	Student-t	.05388988,22.93065*	Student-t	.04505209,24.22863*
910	Student-t	.05384318,22.95005*	Student-t	.04506301,24.23248*
13/2	Gaussian	.04654094*	Gumbel	1.02031625*
24/3	Gaussian	.04651902*	Gumbel	1.02031793*
35/4	Gaussian	.04677893*	Gumbel	1.02035261
46/5	Gaussian	.04669842*	Gumbel	1.02034899*
57/6	Gaussian	.04674105*	Gumbel	1.02036097*
68/7	Gaussian	.04659460*	Gumbel	1.02034333*
79/8	Gaussian	.04650006*	Gumbel	1.02033404*
810/9	Gaussian	.04650941*	Gumbel	1.02035017*
14/23	Joe	1.01360257	Gumbel	1.00948104
25/34	Joe	1.01369486	Gumbel	1.00950404
36/45	Joe	1.01371124	Gumbel	1.00950691
47/56	Joe	1.01369161	Gumbel	1.00951000
58/67	Joe	1.01331054	Gumbel	1.00946073
69/78	Joe	1.01331133	Gumbel	1.00946054
710/89	Joe	1.01332180	Gumbel	1.00945510
15/234	Gumbel	1.00720106	Gumbel	1.00897995
26/345	Gumbel	1.00722375	Gumbel	1.00897726
37/456	Gumbel	1.00718570	Gumbel	1.00896722
48/567	Gumbel	1.00718173	Gumbel	1.00896271
59/678	Gumbel	1.00726176	Gumbel	1.00898339
610/789	Gumbel	1.00728337	Gumbel	1.00898312
16/2345	Gumbel	1.00696406	Gumbel	1.00578257
27/3456	Gumbel	1.00697080	Gumbel	1.00577446
38/4567	Gumbel	1.00704502*	Gumbel	1.00579411*
49/5678	Gumbel	1.00706958	Gumbel	1.00579530
510/6789	Gumbel	1.00711455	Gumbel	1.00581020
17/23456	Gumbel	1.01329140	Clayton	.02401203
28/34567	Gumbel	1.01333759*	Clayton	.02402510*
39/45678	Gumbel	1.01333323*	Clayton	.02403626*
410/56789	Gumbel	1.01331347*	Clayton	.02403966
18/234567	Frank	.13108724	Frank	.11541394
29/345678	Frank	.13057981	Frank	.11552594
310/456789	Frank	.13011361	Frank	.11558571
19/2345678	Clayton	.03483291	Clayton	.01641545
210/3456789	Clayton	.03484014*	Clayton	.01642596
110/23456789	Joe	1.00261299	Clayton	.01009515

*Independent test is significant at 0.05

Figure 3.34: D-vine of amplitude dataset when sample size = 10000, 50000 and interval = 2 - The figure shows the best fit family, MLE of pair-copulas parameter and independent test in each edge for D-vine of amplitude dataset when sample size = 10000, 50000 and interval = 2.

D-Vine Amplitude Dataset, Dimension = 10, Skip = 3

	n = 10000		n = 50000	
Pair-copula(Edge)	**Best fit family**	**Estimated**	**Best fit family**	**estimated**
12	Gumbel	1.00843882	Clayton	.030626246*
23	Gumbel	1.00846988	Clayton	.030631626*
34	Gumbel	1.00850022	Clayton	.030641548
45	Gumbel	1.00846488	Clayton	.030624112
56	Gumbel	1.00847498	Clayton	.030633934
67	Gumbel	1.00847169*	Clayton	.030633467*
78	Gumbel	1.00847750	Clayton	.030635753
89	Gumbel	1.00850133	Clayton	.030635824
910	Gumbel	1.00853138	Clayton	.030630380*
13/2	Gumbel	1.00858688	Gumbel	1.010765185
24/3	Gumbel	1.00862386	Gumbel	1.010763738
35/4	Gumbel	1.00856624	Gumbel	1.010747958
46/5	Gumbel	1.00856297	Gumbel	1.010743632
57/6	Gumbel	1.00855842	Gumbel	1010747131
68/7	Gumbel	1.00855860	Gumbel	1.010745161
79/8	Gumbel	1.00856639*	Gumbel	1.010745542*
810/9	Gumbel	1.00861880	Gumbel	1.010752435
14/23	Joe	1.02175505*	Gumbel	1.011794070
25/34	Joe	1.02173660	Gumbel	1.011797337
36/45	Joe	1.02179291	Gumbel	1.011800828
47/56	Joe	1.02180011	Gumbel	1.011799016
58/67	Joe	1.02177357	Gumbel	1.011798496
69/78	Joe	1.02177921*	Gumbel	1.011798817*
710/89	Joe	1.02184795	Gumbel	1.011807790
15/234	Clayton	.02859856	Gumbel	1.009078992
26/345	Clayton	.02858242	Gumbel	1.009077212
37/456	Clayton	.02858026	Gumbel	1.009080405
48/567	Clayton	.02886638*	Gumbel	1.009080709
59/678	Clayton	.02879017	Gumbel	1.009082651
610/789	Clayton	.02882299	Gumbel	1.009093236
16/2345	Clayton	.02366748	Clayton	.019687987
27/3456	Clayton	.02366195	Clayton	.019689767
38/4567	Clayton	.02363094*	Clayton	.019694652*
49/5678	Clayton	.02373349	Clayton	.019696550
510/6789	Clayton	.02365220*	Clayton	.019672176*
17/23456	Gumbel	1.00691984	Gumbel	1.007403801
28/34567	Gumbel	1.00694328	Gumbel	1.007403294
39/45678	Gumbel	1.00695878*	Gumbel	1.007402720*
410/56789	Gumbel	1.00701028	Gumbel	1.007409879
18/234567	Gumbel	1.01212323	Gaussian	.018326970
29/345678	Gumbel	1.01214270	Gaussian	.018327269
310/456789	Gumbel	1.01207251	Gaussian	.018298298
19/2345678	Frank	.08455291	Frank	.085711968
210/3456789	Frank	.08530821*	Frank	.085792364*
110/23456789	Gaussian	-.01146150	Clayton	.008074189

*Independent test is significant at 0.05

Figure 3.35: D-vine of amplitude dataset when sample size = 10000, 50000 and interval = 3 - The figure shows the best fit family, MLE of pair-copulas parameter and independent test in each edge for D-vine of amplitude dataset when sample size = 10000, 50000 and interval = 3.

D-Vine Amplitude Dataset, Dimension = 10, Skip = 5

n = 10000			n = 50000	
Pair-copula(Edge)	**Best fit family**	**Estimated**	**Best fit family**	**Estimated**
12	Gaussian	.03807380*	Gaussian	.03714331*
23	Gaussian	.03809083*	Gaussian	.03713724*
34	Gaussian	.03800228*	Gaussian	.03714251*
45	Gaussian	.03792276*	Gaussian	.03713840*
56	Gaussian	.03798658*	Gaussian	.03715324*
67	Gaussian	.03806790*	Gaussian	.03714842*
78	Frank	.23051490*	Gaussian	.03717015*
89	Frank	.23041880*	Gaussian	.03716593*
910	Gaussian	.03822361*	Gaussian	.03714466*
13/2	Joe	1.00893926	Gumbel	1.01175764
24/3	Joe	1.00895327	Gumbel	1.01175347
35/4	Joe	1.00892862	Gumbel	1.01175149
46/5	Joe	1.00898420	Gumbel	1.01175945
57/6	Joe	1.00890508	Gumbel	1.11073889
68/7	Joe	1.00913134	Gumbel	1.01174051
79/8	Joe	1.00931328	Gumbel	1.01174336*
810/9	Joe	1.00899765	Gumbel	1.01172547
14/23	Joe	1.01678585	Gumbel	1.01142917
25/34	Joe	1.01686493	Gumbel	1.01143467
36/45	Joe	1.01685403	Gumbel	1.01143977
47/56	Joe	1.01685770	Gumbel	1.01142773
58/67	Joe	1.01703492	Gumbel	1.01145017
69/78	Joe	1.01700810	Gumbel	1.01145207
710/89	Joe	1.01697054	Gumbel	1.01146758
15/234	Clayton	.01513928	Clayton	.01628561
26/345	Clayton	.01517820	Clayton	.01628181
37/456	Clayton	.01512696	Clayton	.01626474
48/567	Clayton	.01522038	Clayton	.01628977
59/678	Clayton	.01525056	Clayton	.01629339
610/789	Clayton	.01535678	Clayton	.01629355
16/2345	Frank	.14502475	Frank	.06828997
27/3456	Frank	.14559990	Frank	.06844752
38/4567	Frank	.14713015*	Frank	.06866127*
49/5678	Frank	.14768160*	Frank	.06883048
510/6789	Frank	.14787464*	Frank	.06892126
17/23456	Frank	.09912312	Frank	.08881701
28/34567	Frank	.09853431*	Frank	.08867370
39/45678	Frank	.09879379*	Frank	.08879822*
410/56789	Frank	.09977836	Frank	.08875132
18/234567	Frank	-.01238611	Gumbel	1.00476559
29/345678	Frank	-.01244415	Gumbel	1.00476034
310/456789	Frank	-.01101804	Gumbel	1.00475864
19/2345678	Joe	1.01212913	Gumbel	1.00894782
210/3456789	Joe	1.01196651	Gumbel	1.00895121
110/23456789	Joe	1.01155571	Gaussian	.00666789

*Independent test is significant at 0.05

Figure 3.36: D-vine of amplitude dataset when sample size = 10000, 50000 and interval = 5 - The figure shows the best fit family, MLE of pair-copulas parameter and independent test in each edge for D-vine of amplitude dataset when sample size = 10000, 50000 and interval = 5.

3.5 Summary

In this chapter, we have started to study and apply copulas to the dynamics of the protein and surrounding water molecules from the sample dataset which is called test systems. We have carried out the test systems: N-variate copulas both two and five dimensions and D-vine both five and ten dimensions for angle and amplitude dataset and studied some statistical properties of the fitted model from N-variate copulas. For D-vine, we have fitted D-vine to both datasets and we have results in the following areas: the best fit family, the estimator of pair-copulas parameter estimation and the independent test. Moreover, we have also carried out graphical analysis for independent test by three graphical tools: Chi-plot, K-plot and Lambda-function plot.

For results of N-variate copulas both angle and amplitude dataset in two and five dimensions, we concluded that all copulas parameters from fitted copulas family are statistically significant. That means that X_1 and X_2 in two dimensions and X_1, X_2, X_3, X_4 and X_5 in five dimensions are positive dependent with the small coefficient value of copulas parameters.

For D-vine results both angle and amplitude dataset in five and ten dimensions, we concluded that the results are different when we take the observations without or with interval. When we take the observations with any interval and large sample size, the results are quite similar, and at the high level of tree for angle dataset, the best fit family is Gaussian.

In conclusion, the overall studies in this chapter are very useful and helpful. We understand how to apply copulas and D-vine to the dataset and we can clarify the meaning of copulas and D-vine results. For the test systems, we expect the behaviour of variables in two, five and ten dimensions should be dependent when the observations are taken without interval (we take the observations continuously). When the observations are taken with interval, the expected behaviour of variables should be less dependent when interval and sample size are larger. Both N-variate copulas and D-vine give results which confirm our expectation. Moreover, we notice that the best fit family at the high level of tree tends to be Gaussian, one theorem which may support this result is the **central limit theorem (CLT)** in statistics. Roughly, the theorem states that the distribution of sum (or average) of all samples (independent and identically distributed variables) from a population with a finite variance and sufficiently large sample size will be approximately normal (Gaussian) distribution, regardless of the underlying distribution (Hays [54]). Therefore, we can take advantage of the test systems study to help us further when we do analysis for real dataset in the next chapter.

4

Results

4.1 Details of the Data on Molecular System

For this research, we have two datasets: original and random that are provided using classical molecular dynamics (MD) simulation with explicit water molecules. Each dataset consist of dihedral angles of peptide: psi and phi and the density of water atoms (hydrogen and oxygen) at 3D grid points for 76 different delay times between 0 and 50.1 picoseconds (ps) before the transition. Also, there is important information for data analysis, that is, the centre of mass of the peptide (COM) is located at x = 14.012, y = 4.30417, z = 3.04016. Figure 4.1 shows the dialanine molecule with dihedral angles: psi and phi and Figure 4.2 shows the time frames for the time before transition (Nerukh and Karabasov [82]).

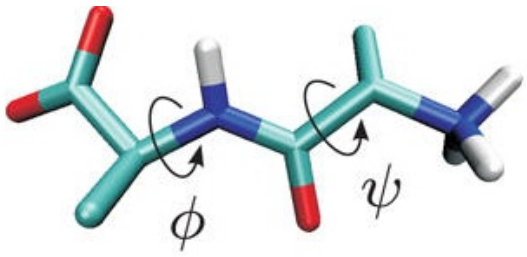

Figure 4.1: Dialanine molecule - The figure shows dialanine molecule with dihedral angles: psi and phi (Source: Nerukh and Karabasov [82], p.815).

Therefore, we have the following random variables for the analysis.

1. the value of psi,
2. the value of phi,
3. the value of densities for hydrogens at different grid points,
4. the value of densities for oxygens at different grid points.

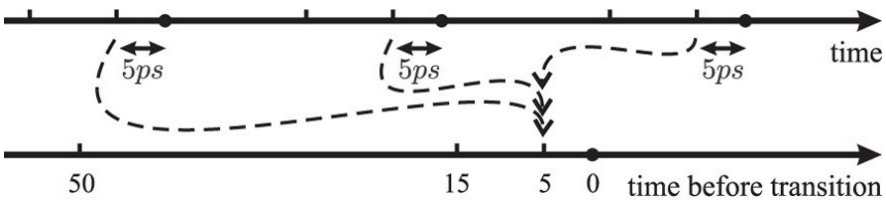

Figure 4.2: The time frames for the time before transition - The figure shows time frames for the time before transition (Source: Nerukh and Karabasov [82], p.817).

From Figure 4.2, it is collecting the time frames for the time before transition statistics. The dots on the time axes are the transition moments. In this study, we are interested in the time t in advance of the transition (before transition) which is chosen from the "time" axes.

The particular realisations of these variables will be the values at different time frames, so we have the values at different 199 and 198 time frames for the original and random dataset, respectively. Since the dihedral angles are periodic data, the sine of two angles: sin(psi) and sin(phi) are used instead of psi and phi. For the density of water atoms, we calculate the probability map for hydrogens at 0.3 ps before the transition and find the location of the maximum of these probabilities nearest to the centre of mass of the peptide. Moreover, we calculate the probability to find a hydrogen and oxygen in the whole volume within radius 4 Angstrom from the COM, also, we select several grid points of densities in space, measure the distance from those points to the peptide and analyse their behaviour. There are four grid points of hydrogen density in space that are studied following.

- X_0 is a grid point of the maximum probability of hydrogen that is located at $x = 20.25$, $y = 5.75$, $z = 1.75$ and distance from X_0 to COM is 6.532.
- X_1 is an opposite grid point of X_0 that is located at $x = 12.25$, $y = 6.75$, $z = 4.75$ and distance from X_1 to COM is 3.466.
- X_2 is a neighbor grid point of X_0 that is located at $x = 20.75$, $y = 5.75$, $z = 1.75$ and distance from X_2 to COM is 7.011.
- X_3 is a neighbor grid point of X_0 that is located at $x = 20.25$, $y = 6.25$, $z = 1.75$ and distance from X_3 to COM is 6.661.

Four grid points of hydrogen density in space and the distance from each grid point to COM are illustrated in Figure 4.3.

The main goal of this research is to study the statistical correlations between different variables and at different periods of time. For the statistical tools, we use pair-copulas and the Kendall's tau correlation, also, we apply the D-vine to study the

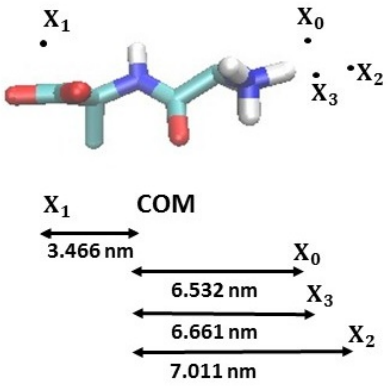

Figure 4.3: Four grid points of hydrogen density in space - The figure shows four grid points of hydrogen density in spac and the distance from each grid point to COM.

statistical correlations between variables describing molecular conformation of a peptide and the properties of water molecules surrounding the peptide both original and random dataset.

Remark.

1. **Random dataset** is the dataset that the values of psi/phi are taken at the same delay times, but the 0 delay is chosen randomly along the trajectory. This almost guarantees that there is no transition between the conformational states in this data.

2. **Original dataset** is the sequence of delay times, all aligned such that the 0 delay coincides with the moment of transition from one state, A to other state, B conformational states. Figure 4.4, left shows normalized probabilities of conformations (Ramachandran plot) formed by a 2 microseconds (μs) trajectory and right shows same probabilities emphasizing the presence of two minor conformations and the partitioning for symbolization which is designated as "A", "B", or "C" (Nerukh and Karabasov [82]).

3. The probability map is the probabilities of finding hydrogen/oxygen at all the locations on the grid points.

4. 76 different delay times before the transition are following: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.1, 2.2, 2.3, 2.5, 2.6, 2.8, 2.9, 3.1, 3.3, 3.5, 3.7, 3.9, 4.2, 4.4, 4.7, 5.0, 5.3, 5.6, 5.9, 6.3, 6.6, 7.0, 7.4, 7.9, 8.4, 8.9, 9.4, 10.0, 10.5, 11.2, 11.8, 12.5, 13.3, 14.1, 14.9, 15.8, 16.7, 17.7, 18.8, 19.9, 21.1, 22.3, 23.7, 25.1, 26.6, 28.1, 29.8, 31.6, 33.4, 35.4, 37.5, 39.8, 42.1, 44.6, 47.3 and 50.1 picoseconds (ps) as illustrated in Figure 4.5 (a) and Figure 4.5 (b) shows the natural logarithm of delay times before the transition.

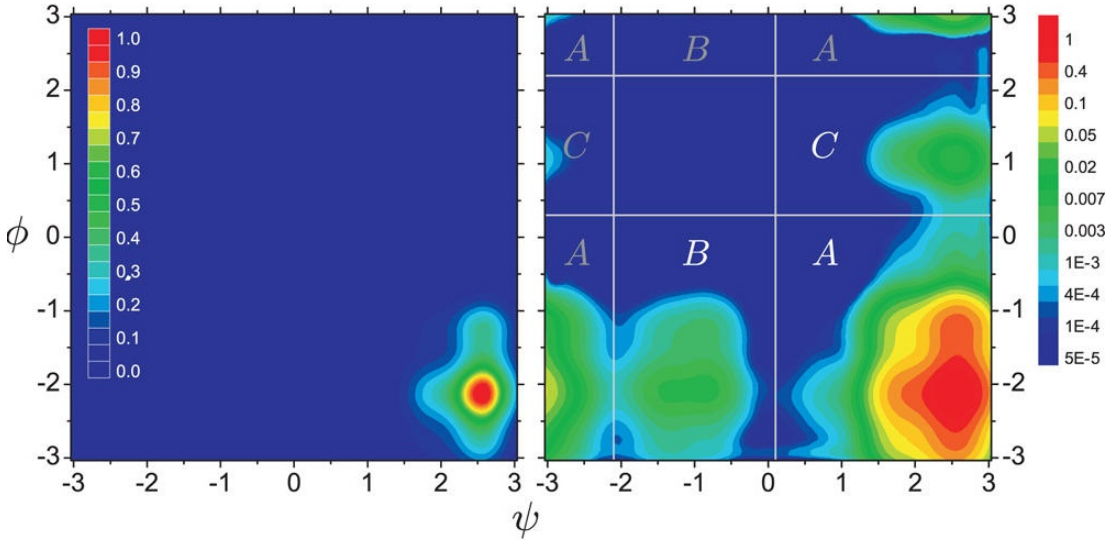

Figure 4.4: The space of conformation probabilities - The figure shows three well-separated metastable states in the space of conformation probabilities (Source: Nerukh and Karabasov [82], p.815).

For 76 delay times, we summarize the useful information of psi, sin(psi), phi, sin(phi), hydrogen and oxygen density at the maximum probability point (X_0) from delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps of original and random dataset as illustrated in Table 4.1.

Normally, checking the distribution of variables is very helpful to correlation analysis and D-vine. For correlation analysis, we will choose the correlation coefficient according to the distribution of variable. For D-vine, the expected distribution of variables should be uniform distribution on (0,1). First of all, we examine the distribution of all variables for random and original dataset by histogram as illustrated in Table 4.1. The estimated distribution of most variables for random and original dataset is similar, for example, the estimated distribution of sin(psi) and hydrogen are left and right skewed, respectively. As the estimated distribution of phi and sin(phi) for random dataset are similar: right skewed but they are different for original dataset, that is the estimated distribution of phi and sin(phi) are symmetric and left skewed, respectively. For the range of each variable is similar both random and original dataset. Figure 4.6 and 4.7 are few examples of histograms of phi at the maximum probability point (X_0) for delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for both dataset. For histograms of other variables are given in the Appendix.

For the next section, we will analyse and investigate the correlation of each variable that depends on delay time before the transition by using the correlation analysis and D-vine.

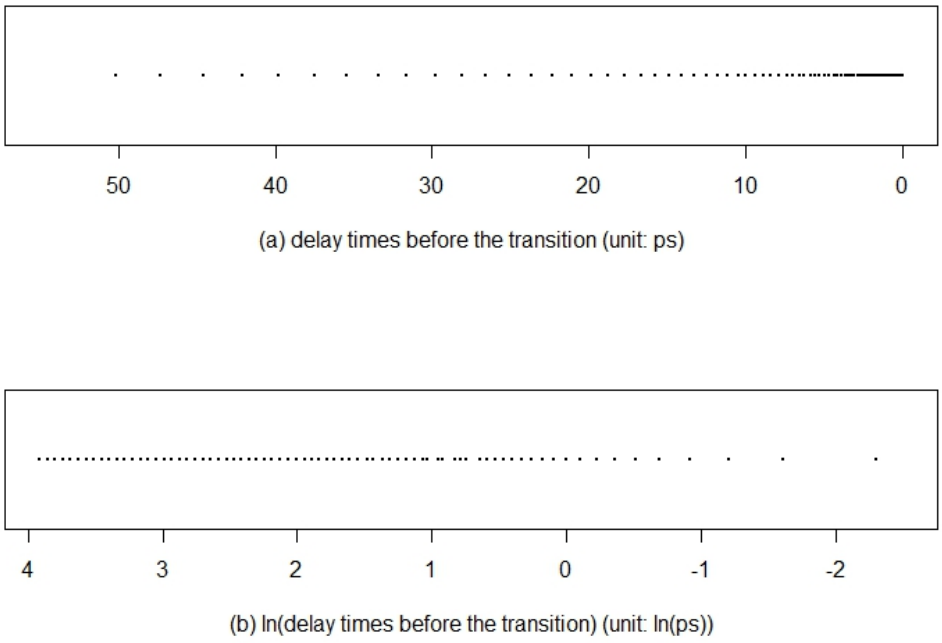

Figure 4.5: 76 different delay times before the transition - The figure shows 76 different delay times before the transition.

Table 4.1: The estimated distribution and range of psi, sin(psi), phi, sin(phi), hydrogen and oxygen density at the maximum probability point (X_0) for 76 delay times.

Item	Variable	Random dataset	Original dataset
estimated distribution	psi	left skewed	left skewed
	sin(psi)	left skewed	left skewed
	phi	right skewed	symmetric
	sin(phi)	right skewed	left skewed
	hydrogen	right skewed	right skewed
	oxygen	right skewed	right skewed
range	psi	(-3.14060, 3.14069)	(-3.14128, 3.13946)
	sin(psi)	(-0.96960, 1.00000)	(-0.99999, 1.00000)
	phi	(-3.14063, 3.14128)	(-3.14039,3.14124)
	sin(phi)	(-1.00000, 0.31567)	(-1.00000, 0.99999)
	hydrogen	(0.00000, 0.50963)	(0.00000, 0.51145)
	oxygen	(0.00007, 3.66831)	(0.00000, 3.66815)

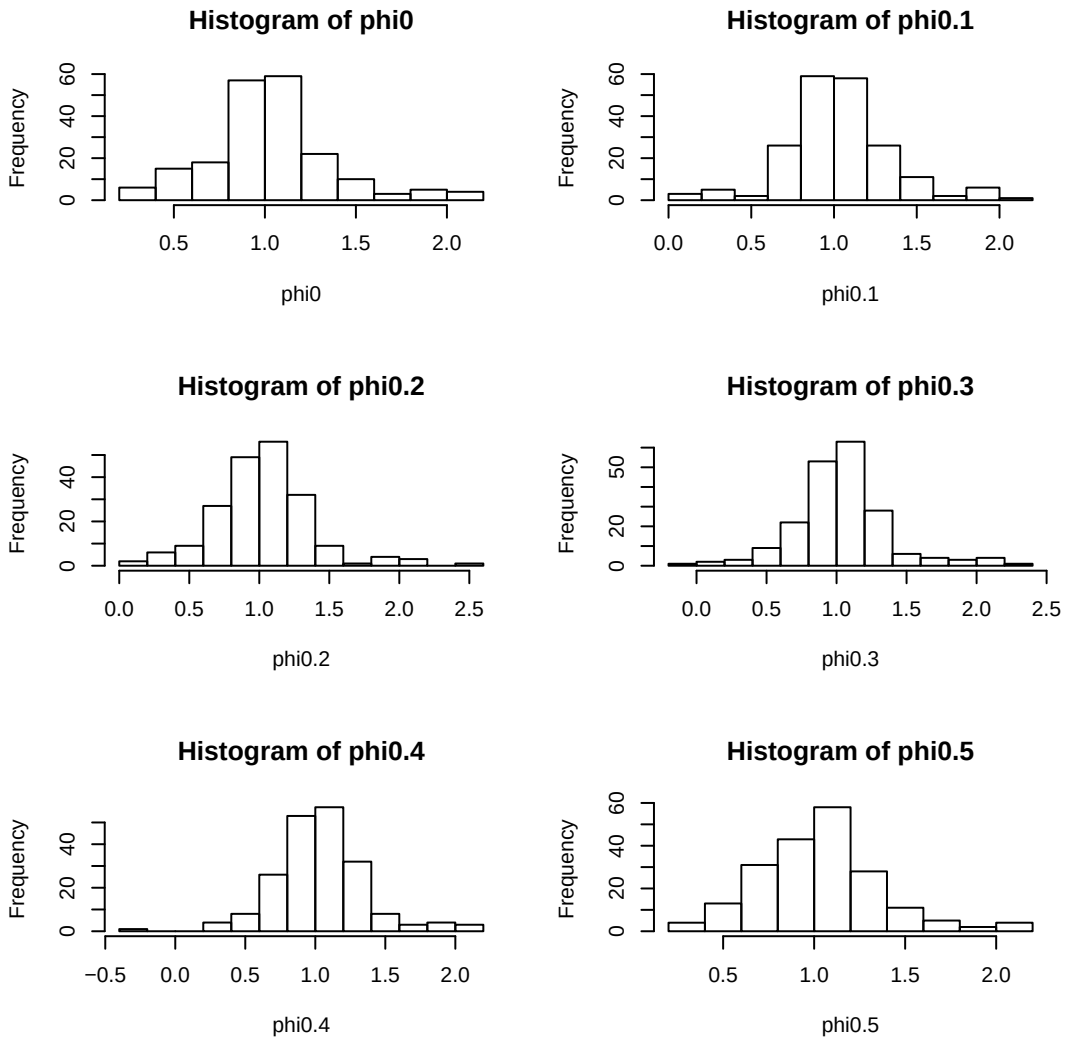

Figure 4.6: Histogram of phi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows histograms of phi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for original dataset.

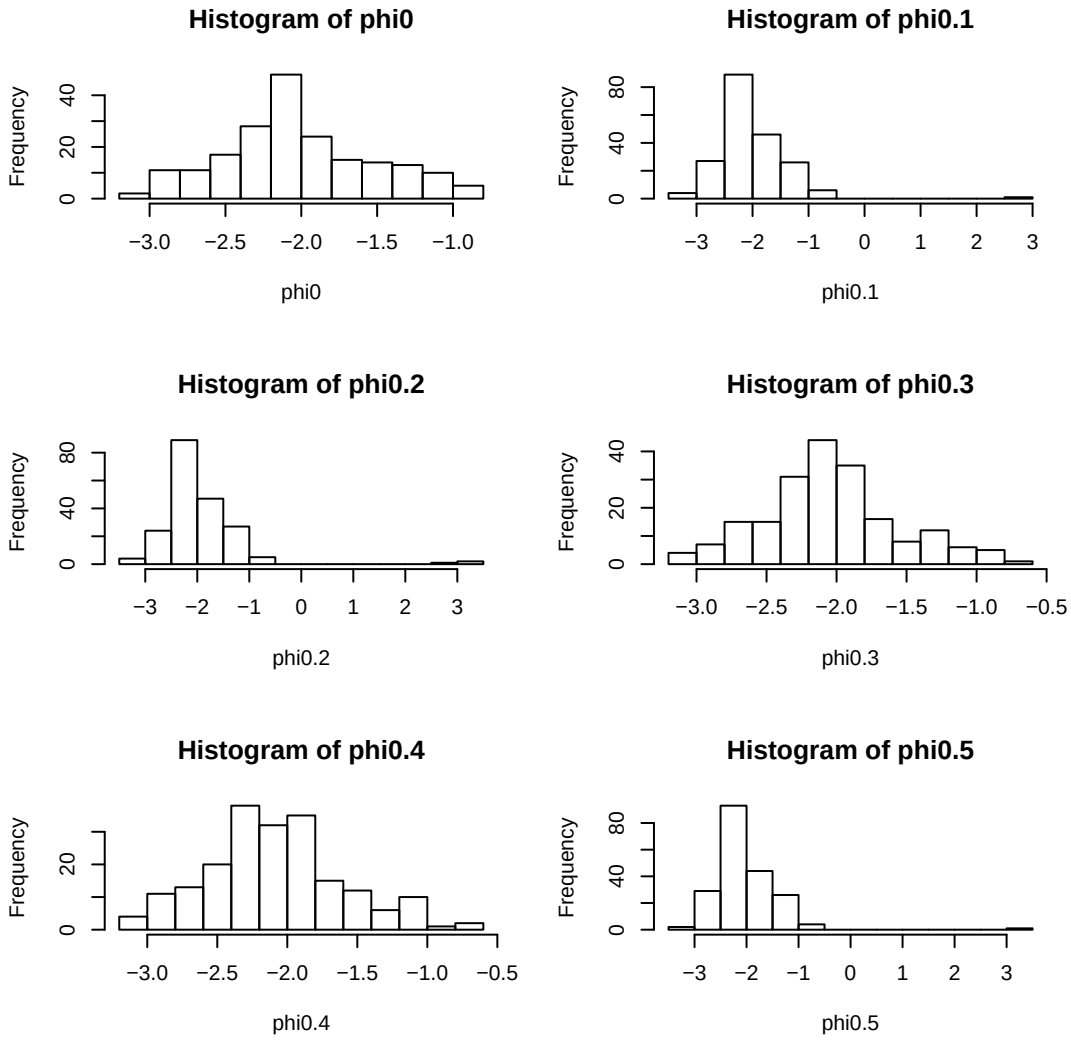

Figure 4.7: Histogram of phi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows histograms of phi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for random dataset.

4.2 Time Correlations

In this section, we carry on a statistical analysis of dihedral angles motion of a dialanine molecule with explicit water. We are interested in finding statistical correlations between the values of the angle at the moment of transition and several moments in advance of the transition between 0 and 50.1 ps.

Firstly, we examine roughly the linear relationship of psi, sin(psi), phi, sin(phi), hydrogen and oxygen density for original and random datasets by the scatter plot. Figure 4.8 and 4.9 are few examples of scatter plots that show the scatter plots of phi for both dataset at the maximum probability point (X_0) for delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps. For scatter plots of other variables are given in the Appendix. We summarize the linear relationship of psi, sin(psi), phi, sin(phi), hydrogen and oxygen density at the maximum probability point (X_0) for original and random datasets in Table 4.2 and Table 4.3 shows linear correlation of psi, sin(psi), phi, sin(phi), hydrogen and oxygen density for original dataset at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps by Pearson correlation coefficient.

Table 4.2: The linear relationship of psi, sin(psi), phi, sin(phi), hydrogen and oxygen density at the maximum probability point (X_0) for 76 delay times.

Variable	Random dataset	Original dataset
psi	not linear relationship	not linear relationship
sin(psi)	not linear relationship	not linear relationship
phi	not linear relationship	deviate from linear relationship
sin(phi)	not linear relationship	not linear relationship
hydrogen	not linear relationship	not linear relationship
oxygen	not linear relationship	not linear relationship

From Table 4.2, most variables both original and random dataset, there is no linear correlation between variables at the different delay times i and j, $i \neq j$ where $i = 50.1, 47.3, \ldots, 0.1, 0$ and $j = 50.1, 47.3, \ldots, 0.1, 0$, except, phi for original dataset, the correlation deviates from linear relationship between phi at delay time i and j. When we consider Pearson correlation coefficient of all variables at different delay times from Table 4.3, we found that most coefficients are small values which conform to scatter plots.

Table 4.3: Pearson correlation coefficient of psi, sin(psi), phi, sin(phi), hydrogen and oxygen density at the maximum probability point (X_0) for delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for original dataset.

Variable	Delay time	0	0.1	0.2	0.3	0.4	0.5
psi	0	1.0000	0.6521	0.1736	0.2378	0.2587	0.1460
	0.1	0.6521	1.0000	0.4686	0.5156	0.2263	0.1096
	0.2	0.1736	0.4686	1.0000	0.4421	0.1630	0.3180
	0.3	0.2378	0.5156	0.4421	1.0000	0.1852	0.1352
	0.4	0.2587	0.2263	0.1630	0.1852	1.0000	0.4855
	0.5	0.1460	0.1096	0.3180	0.1352	0.4855	1.0000
sin(psi)	0	1.0000	0.5009	0.4042	0.3810	0.3814	0.4767
	0.1	0.5009	1.0000	0.7538	0.6438	0.5969	0.6192
	0.2	0.4042	0.7538	1.0000	0.7267	0.5828	0.6034
	0.3	0.3811	0.6438	0.7267	1.0000	0.7687	0.6114
	0.4	0.3814	0.5969	0.5828	0.7687	1.0000	0.7435
	0.5	0.4767	0.6192	0.6034	0.6114	0.7435	1.0000
phi	0	1.0000	0.4623	0.4190	0.4058	0.3685	0.4157
	0.1	0.4623	1.0000	0.6912	0.5877	0.5000	0.54012
	0.2	0.4190	0.6913	1.0000	0.7012	0.6277	0.5758
	0.3	0.4058	0.5877	0.7012	1.0000	0.7189	0.6151
	0.4	0.3685	0.5000	0.6277	0.7189	1.0000	0.7187
	0.5	0.4157	0.5401	0.5758	0.6151	0.7187	1.0000
sin(phi)	0	1.0000	0.3866	0.3781	0.2973	0.1929	0.2574
	0.1	0.3866	1.0000	0.5722	0.4594	0.3492	0.4166
	0.2	0.3781	0.5722	1.0000	0.5675	0.5127	0.3968
	0.3	0.2973	0.4594	0.5675	1.0000	0.6617	0.4752
	0.4	0.1929	0.3492	0.5127	0.6617	1.0000	0.6012
	0.5	0.2574	0.4166	0.3968	0.4752	0.6012	1.0000
hydrogen	0	1.0000	0.2816	0.1821	0.0912	0.0165	0.0995
	0.1	0.2816	1.0000	0.5189	0.3210	0.1660	0.1607
	0.2	0.1821	0.5189	1.0000	0.4876	0.3566	0.3248
	0.3	0.0912	0.3210	0.4876	1.0000	0.5091	0.2970
	0.4	0.0165	0.1660	0.3566	0.5091	1.0000	0.3388
	0.5	0.0995	0.1607	0.3248	0.2970	0.3388	1.0000
oxygen	0	1.0000	0.4052	0.2366	0.1032	0.0576	0.0240
	0.1	0.4052	1.0000	0.6416	0.3131	0.1085	0.1590
	0.2	0.2366	0.6416	1.0000	0.5216	0.1982	0.1925
	0.3	0.1032	0.3131	0.5216	1.0000	0.5536	0.2706
	0.4	0.0576	0.1085	0.1982	0.5536	1.0000	0.4421
	0.5	0.0240	0.1590	0.1925	0.2706	0.4421	1.0000

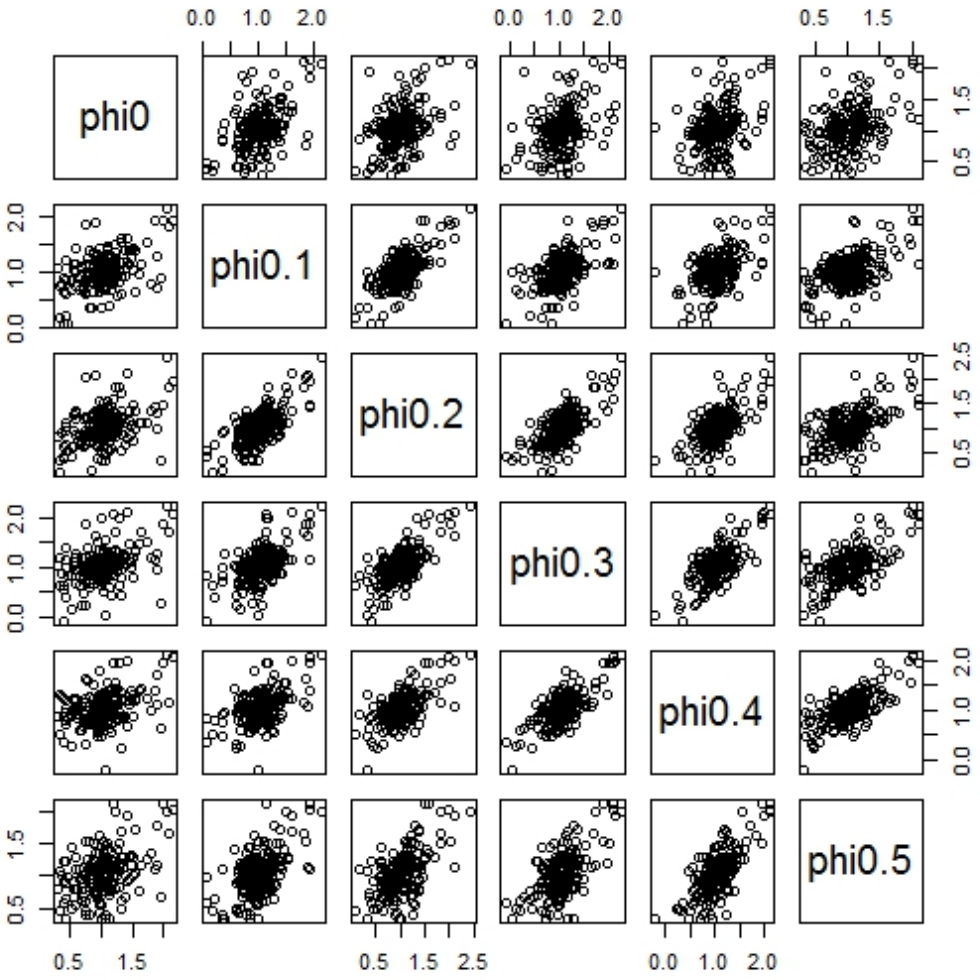

Figure 4.8: Scatter plot of phi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows scatter plot of phi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for original dataset.

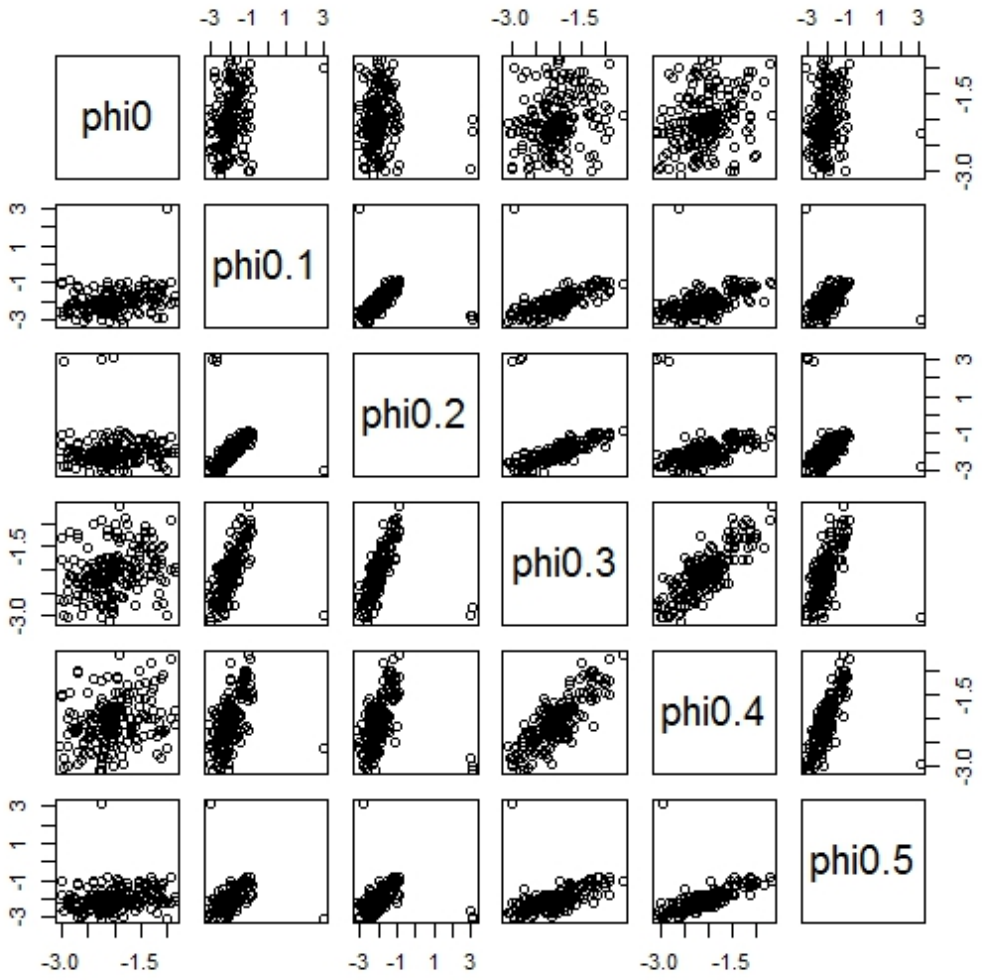

Figure 4.9: Scatter plot of phi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows scatter plot of phi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for random dataset.

Secondly, we calculate three correlation coefficients: Pearson (2.9), Spearman (2.10) and Kendall (2.11) of the dihedral angles: psi, phi, sine of dihedral angles: sin(psi), sin(phi) and water atoms: hydrogen and oxygen. The correlation matrix graph of all variables are given in Figure 4.10 - 4.18.

Regarding the results on the statistical analysis.

1. **psi, sin(psi), phi and sin(phi)**

   - For the two-point correlations of original dataset from Figure 4.10, 4.12 and 4.14, three measures of four varibles: psi, sin(psi), phi and sin(phi) are different. Since, Kendall and Spearman are non-parametric correlation coefficients that capture functional dependence including non-linear, while Pearson shows only linear dependence.
   - In contrast, for random dataset of four variables from Figure 4.11, 4.13 and 4.15, both Pearson and Spearman give the same result. As Kendall gives a slightly different result, because for random dataset there is very little correlation.
   - Therefore, for the case when there is no correlation, three correlations give the similar result, while when correlations are present, the results are different.

2. **hydrogen and oxygen density**

   - For the two-point correlations of original dataset from Figure 4.16, 4.17 and 4.18, both Pearson and Spearman give similar result. As Kendall gives the different result.
   - For random dataset of two variables from Figure 4.16, 4.17 and 4.18, both Pearson and Spearman give the same result. As Kendall gives a slightly different result.
   - Therefore, whenever there is no correlation or correlations are present, the correlations from Pearson and Spearman are quite similar, while Kendall gives the same result both original and ramdom dataset.

Regarding the molecular meaning of the statistical results.

Figure 4.10 - 4.15 show the Pearson, Spearman and Kendall correlation matrices of four variables: psi, phi, sin(psi) and sin(phi) for original and random dataset. For the correlation matrix graph, x-axis is delay time before the transition that starts from 50.1, 47.3, 44.6, ..., 0.1, 0 ps, also y-axis is delay time before the transition that starts from 0, 0.1, 0.2, ..., 47.3, 50.1 ps. Each correlation matrix is a symmetric matrix which consists of a pair correlation of variable between delay time i and j, as a main diagonal is a correlation of variable at same delay time. For the strength of relationship, it is between -1.0 and 1.0 and illustrated with a variety of colours, for example, red

colour means positive strong relationship (=1.0), black colour means negative strong relationship (= -1.0) etc.

1. The most obvious is that random and original dataset are very different because the process of obtaining dataset between random and original is different. Clearly, the correlation matrices of all variables show the different colour between random and original dataset, that means that the strength of association of variable at different delay time is different.

2. The matrices graphs of correlations show the complete picture of statistical dependence of psi(phi) at different time moments with respect to the transition moment. For example, the row starting at 50.1 shows how much the value of psi(phi) at 50.1 picoseconds before the transition depends on all the values of psi (phi) at previous time moments. While, for example, the row starting at 1.5 shows the dependence of the psi(phi) value at 1.5 picoseconds before the transition on all the values of psi (phi) at previous time moments.

3. The "random" matrices graphs show that there is no difference in dependencies at all starting times: the sequence of correlations at 0 delay is the same as the sequence at 1.5 ps or 10 ps delays. That means that there is no change in behaviour and the psi(phi) values fluctuate in the same regime.

4. The "original" matrices graphs show very different behaviour in correlations depending on the time moment before transition:

   - In advance of the transition (rows starting at 50.1 - 10 delays) the correlations are essentially the same as for the "random" dataset.
   - Starting from 10 delays and up to 1.5 ps, the correlations are much stronger and longer. The row at 1.7 ps, for example, has very strongly correlated values of psi up to 2 - 3 ps in advance (the correlation coefficient is 0.7 - 1).
   - At 1.2 - 1.5 ps, surprisingly, the correlations are very low again, almost like in the "random" dataset.
   - Before the transition, 0 - 1.1 ps the correlations are stronger than usual again, lasting for ≈0.5 ps, in contrast to ≈0.2 ps in "random" dataset. There is also anti-correlation (negative corr.:blue colour) with the data at 2 - 7 ps delays (correlation value is -0.5 - -0.4).

5. What molecular picture can be deduced from this? First of all, the peptide (we consider sin(psi)) starts "feeling" the upcoming transition at ≈10 ps in advance of the actual transition as illustrated in Figure 4.10 and 4.12. Then, from 10 ps to ≈1.5 ps before the transition, the peptide follows more or less the same conformational trajectory every time it approaches the transition: strong correlations at these times testify that. Then, at ≈1.5 ps before the transition, the trajectory jumps to a set of very different conformations: the shape of the peptide has very

little connection to what is at times ≈1.8 ps and earlier. The final approach to the transition happens along the same route again: even though the peptide starts from different shapes at 1.5 ps delay moment, it then follows the same trajectory until it reaches the next conformational state. The anti-correlation with ≈2 - 7 ps times tells us that the psi angle rotated to approximately opposite value: a half-circle rotation has happened.

To clarify those, we produce one dimension cuts of two dimensions of Spearman correlation matrix graph of sin(psi) for quantitative analysis. Figure 4.19 shows Spearman correlation graph of sin(psi) on X_0 at delay time 1.5, 2.1, 7 and 10 ps before the transition for origial dataset:

- At delay time 1.5 ps before the transition, the pair-correlations from 50.1 ps to 1.5 ps before the transition tend to be steadily increasing. The lowest and highest of pair-correlation are at 47.3 ps and 1.6 ps in advance of transition with the coefficient -0.12287 and 0.82693.
- At delay time 2.1 ps before the transition, the pair-correlations from 50.1 ps to 2.1 ps before the transition tend to be steadily increasing same as at delay time 1.5 ps. The lowest and highest of pair-correlation are at 47.3 ps and 2.2 ps in advance of transition with the coefficient -0.06534 and 0.83385.
- At delay time 7 ps before the transition, the pair-correlations from 50.1 ps to 7 ps before the transition tend to be steadily increasing same as at delay time 1.5 ps and 2.1 ps but the pair-correlation coefficients are smaller. The lowest and highest of pair-correlation are at 39.8 ps and 7.4 ps in advance of transition with the coefficient -0.07659 and 0.61639.
- At delay time 10 ps before the transition, the pair-correlations from 50.1 ps to 10 ps before the transition tend to be steadily increasing same as at delay time 7 ps but the pair-correlation coefficients are smaller. The lowest and highest of pair-correlation are at 25.1 ps and 10.5 ps in advance of transition with the coefficient -0.12476 and 0.51822.
- We hypothesise that the the conformational trajectory of peptide at any closed delay time before the transition is correlated, as at any far away delay time before the transition is un- or less correlated. Hence, the results conform to our hypothesis.

Finally, it is clear that sin(psi) shows the big differences of correlation among three measures of dependence, as hydrogen and oxygen density do not show too much differences. When we consider Pearson, Spearman and Kendall correlations, Kendall is more reasonable correlation than Pearson and Spearman for this research with three main reasons:

1. Pearson correlation is a measure of the linear correlation between two variables but our variables do not show the linear association each other as illustrated in Figure B.1 - B.10.

2. Spearman correlation is a non-parametric measure of association that can measure both linear and non-linear but the calculation is defined as same as the Pearson correlation.

3. Kendall's tau correlation is a non-parametric measure of association as same as Spearman correlation but the calculation is different which Kendall's tau correlation depends on number of concordant and disconcordant pairs of observations. Moreover, Kendall's tau correlation is applicable to copulas especially pair-copulas and D-vine.

To support using a non-parametric measure of relationship or association between two variables in this research, example of interesting studies following.

- Capéraá and Genest [19] studied a nonparametric estimation procedure for bivariate extreme values copulas and stated that Spearman is larger than Kendalls tau for positively dependent random variables.

- Khamis [65] studied measures of association: how to choose? He mentioned that the choice of the proper measure of association is based on the characteristics of each of the two variables involved.

- Lieberson [75] studied limitations in the application of non-parametric coefficients of correlation and stated that Kendall's tau and Spearman correlation are appropriate for non-linear relationship.

- Lindskog et al. [77] applied Kendalls tau for elliptical distributions and showed the relation between Kendalls tau and the linear correlation coefficient for bivariate normal distributions holds more generally for the class of elliptical distributions.

- Rupinski and Dunlap [88] studied approximating Pearson product-moment correlations from Kendalls tau and Spearmans rho. The use of Monte Carlo methods demonstrated that a formula presented by Kendall for estimating Pearson from Kendall's tau is more accurate than a formula presented by Pearson for estimating Pearson from Spearman coefficient.

- Strahan [104] studied assessing magnitude of effect from rank-order correlation coefficients. Spearman and Kendall's tau are equally powerful rank-order correlation coefficients under conditions of normality. When applied to the same data set, Kendall's tau typically is smaller in absolute value, often no more than two-thirds the size of Spearman.

- Taylor [105] studied Kendalls and Spearmans correlation coefficients in the presence of a blocking variable. Normally, Kendall's tau and Spearman correlation are two commonly used non-parametric methods of detecting associations between two variables. In this research, he considered that there are circumstances where the joint distribution of the two variables of interest (X and Y) is affected

by the value of a third variable, called a blocking variable. Therefore, a weighted sum of Kendall's tau and Spearman were used to test for associations across the blocks in this study.

From the previous correlation research, we found that Kendall's tau and Spearman correlation are well-known non-parametric measure of relationship between two variables. Kendall [63] claimed that for many practical and most theoretical points of view, kendall's tau is preferable to Spearman correlation. One reason for preferring Kendall's tau, when estimating a correlation, is that the population parameter being estimated has a simpler interpretation. As many reasons that are mentioned, therefore, we focus mainly on Kendall's tau result for this research in next section.

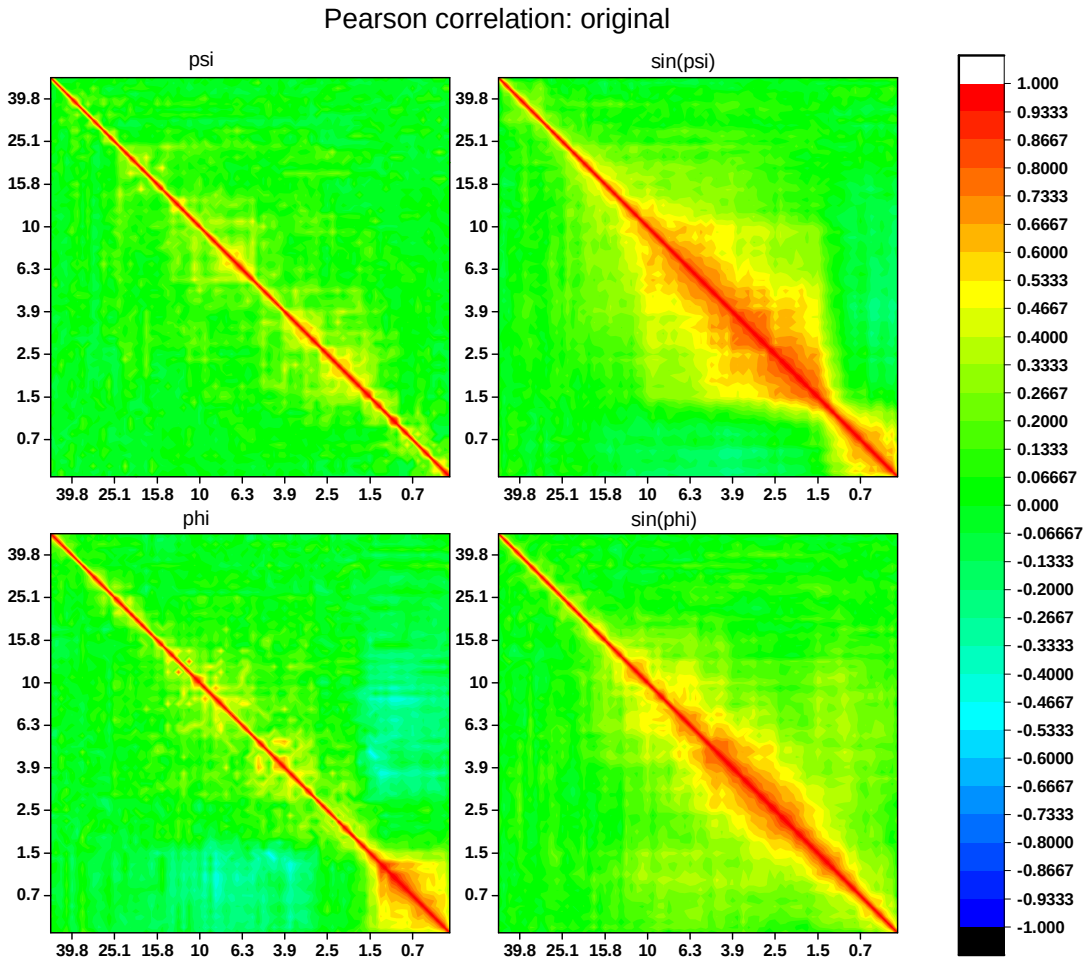

Figure 4.10: Pearson correlation of psi, phi, sin(psi) and sin(phi): original dataset - The figure shows Pearson correlation matrix graph of psi, phi, sin(psi) and sin(phi) at delay time 50.1, 47.3, . . . , 0 ps before the transition for original dataset.

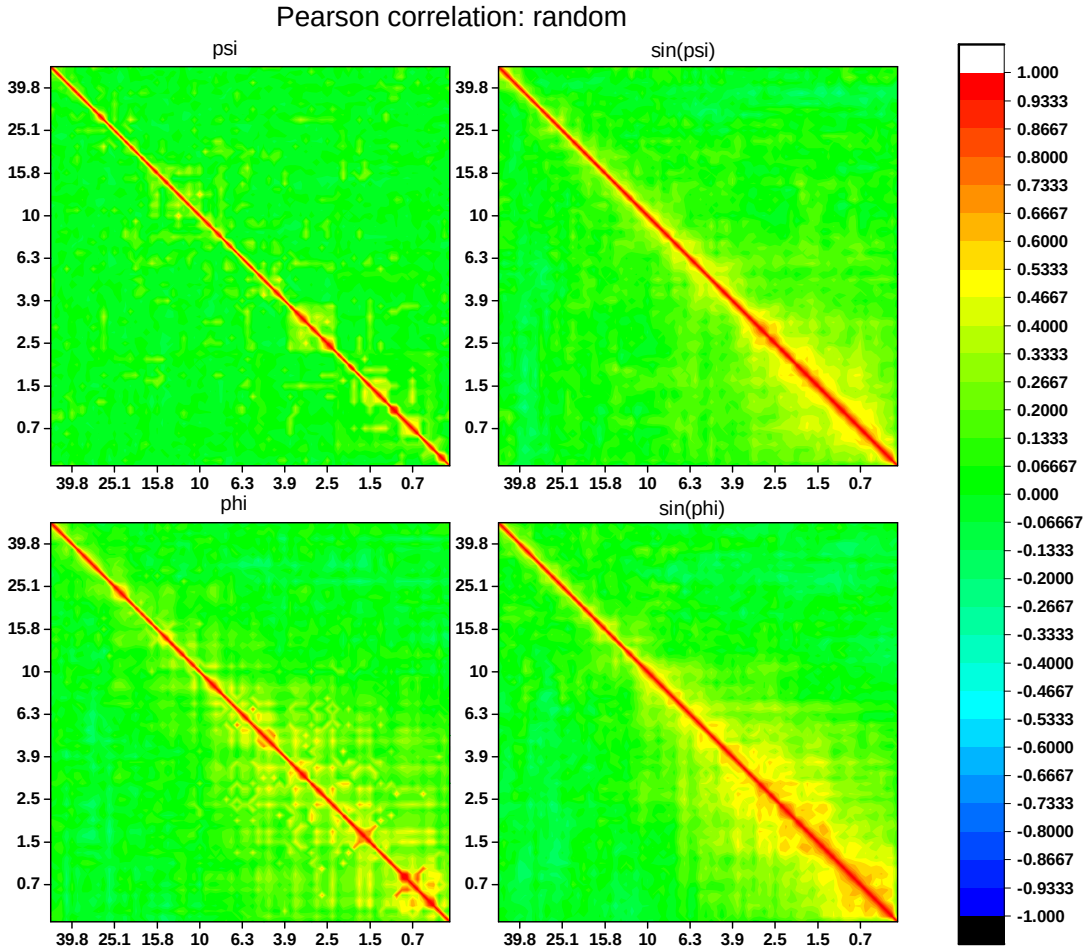

Figure 4.11: Pearson correlation of psi, phi, sin(psi) and sin(phi): random dataset - The figure shows Pearson correlation matrix graph of psi, phi, sin(psi) and sin(phi) at delay time 50.1, 47.3, . . . , 0 ps before the transition for random dataset.

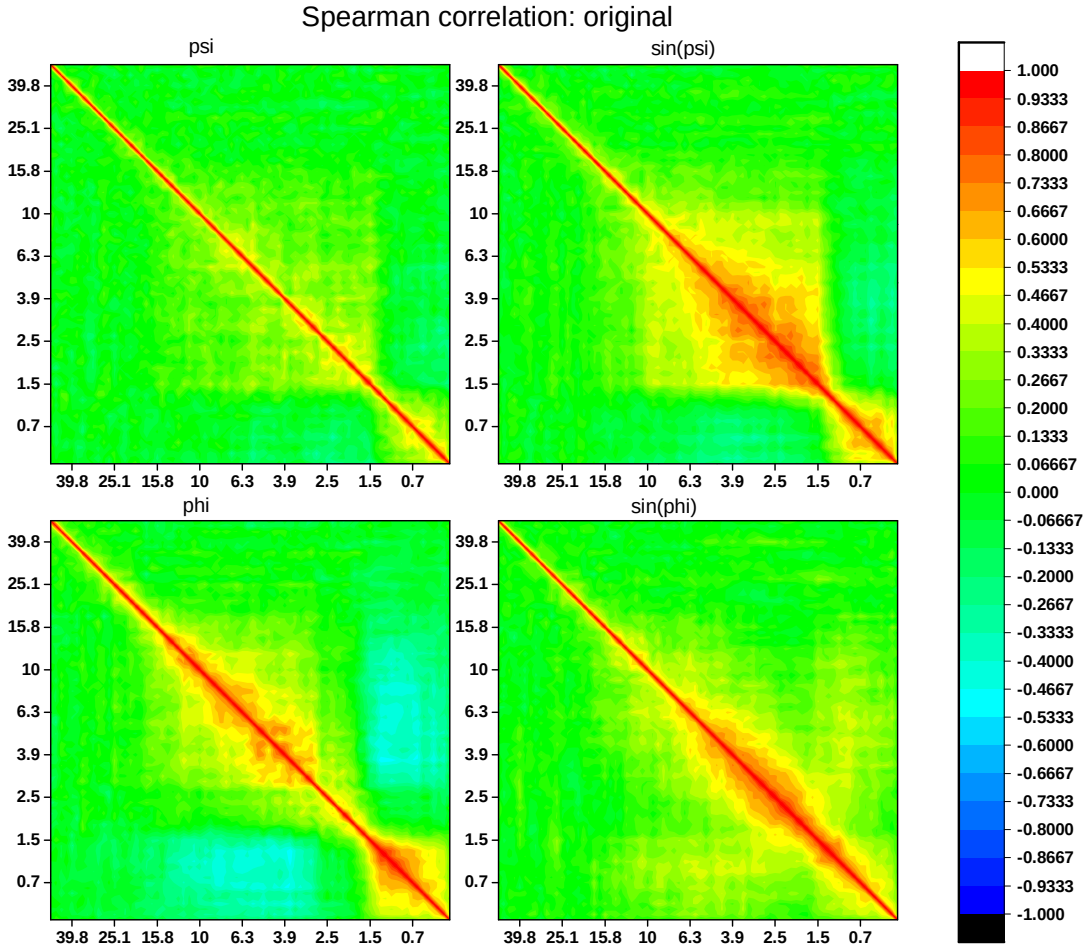

Figure 4.12: Spearman correlation of psi, phi, sin(psi) and sin(phi): original dataset - The figure shows Spearman correlation matrix graph of psi, phi, sin(psi) and sin(phi) at delay time 50.1, 47.3, . . . , 0 ps before the transition for original dataset.

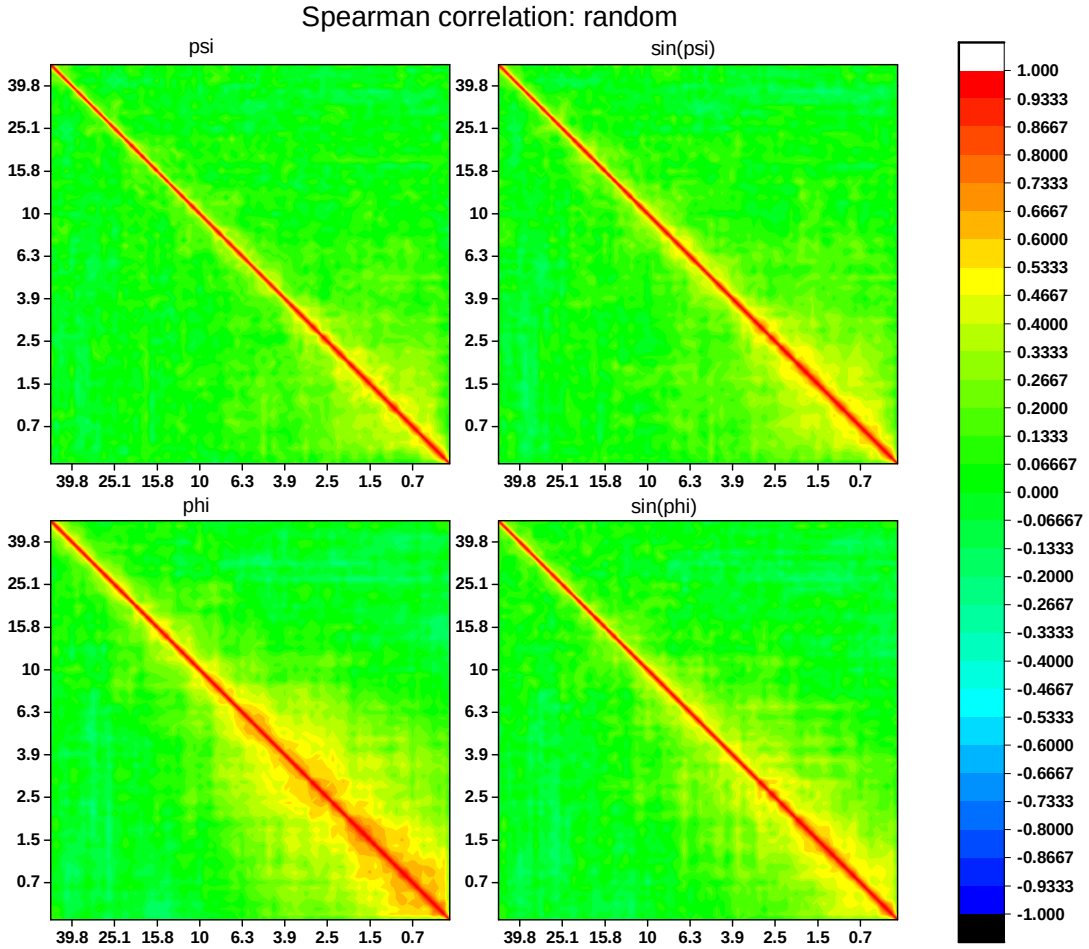

Figure 4.13: Spearman correlation of psi, phi, sin(psi) and sin(phi): random dataset - The figure shows Spearman correlation matrix graph of psi, phi, sin(psi) and sin(phi) at delay time 50.1, 47.3, . . . , 0 ps before the transition for random dataset.

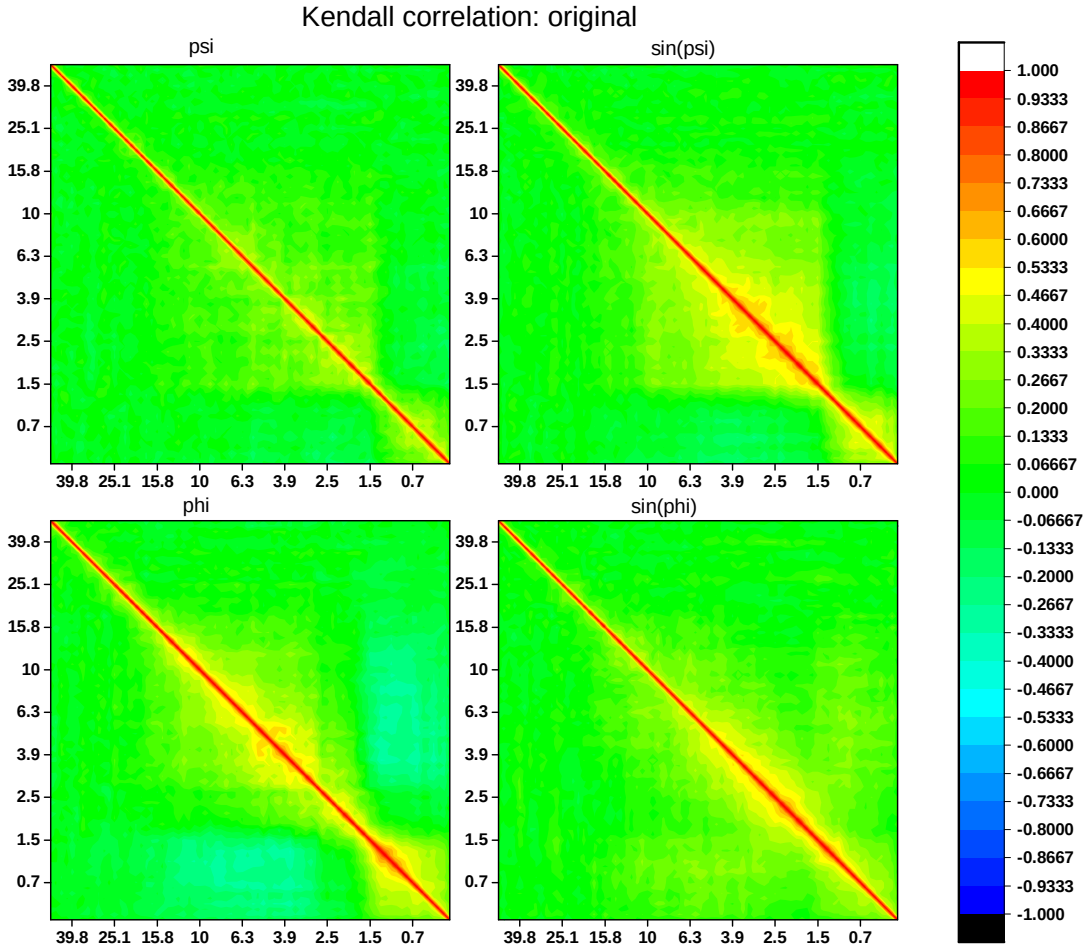

Figure 4.14: Kendall correlation of psi, phi, sin(psi) and sin(phi): original dataset - The figure shows Kendall correlation matrix graph of psi, phi, sin(psi) and sin(phi) at delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset.

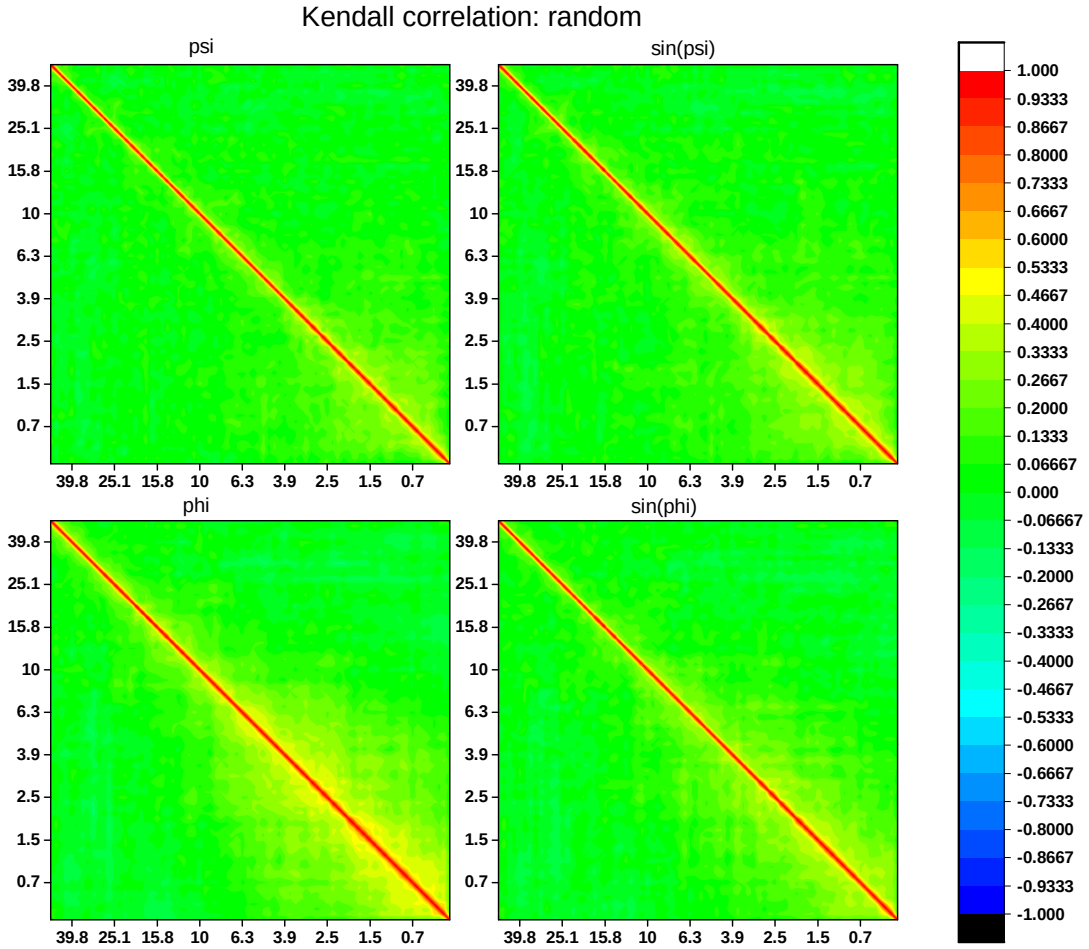

Figure 4.15: Kendall correlation of psi, phi, sin(psi) and sin(phi): random dataset - The figure shows Kendall correlation matrix graph of psi, phi, sin(psi) and sin(phi) at delay time 50.1, 47.3, . . . , 0 ps before the transition for random dataset.

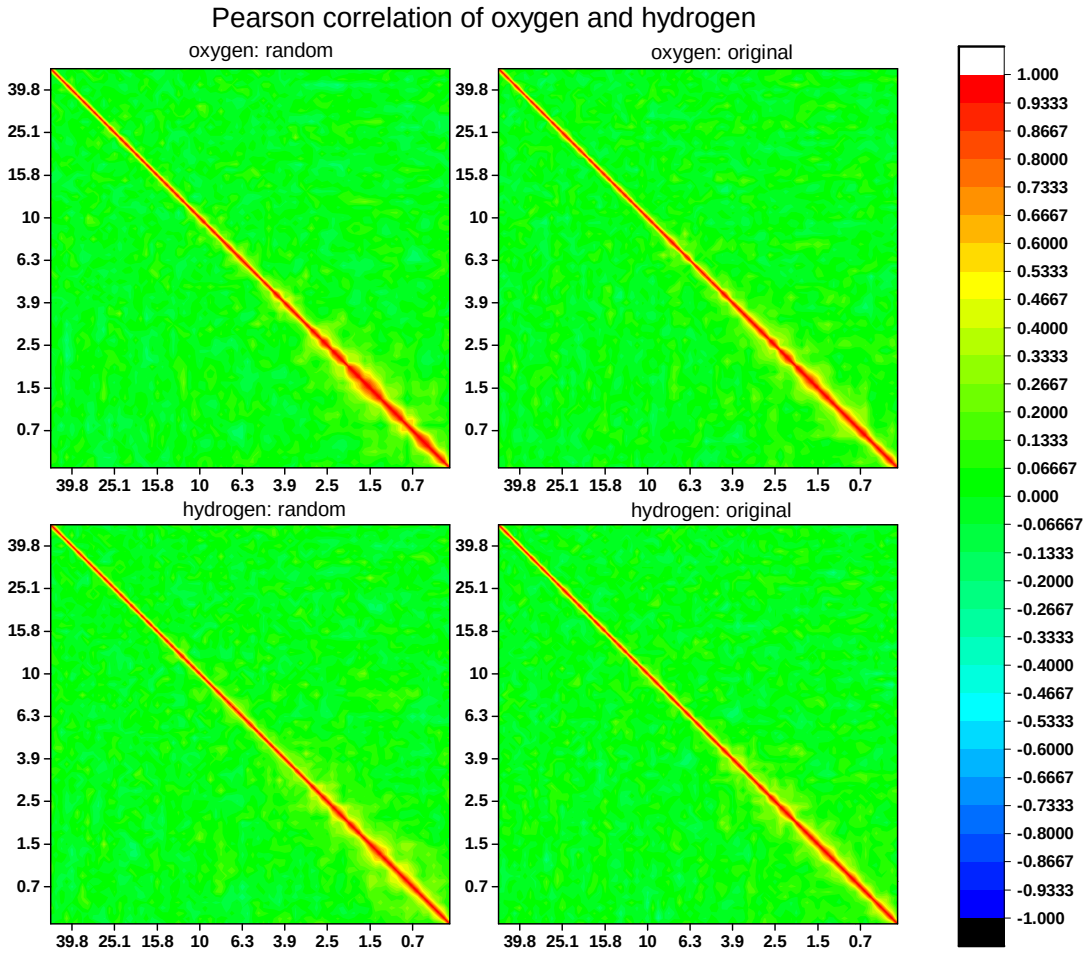

Figure 4.16: Pearson correlation of oxygen and hydrogen density: random and original dataset - The figure shows Pearson correlation matrix graph of oxygen and hydrogen density on X_0 at delay time 50.1, 47.3, ..., 0 ps before the transition for random and origial dataset.

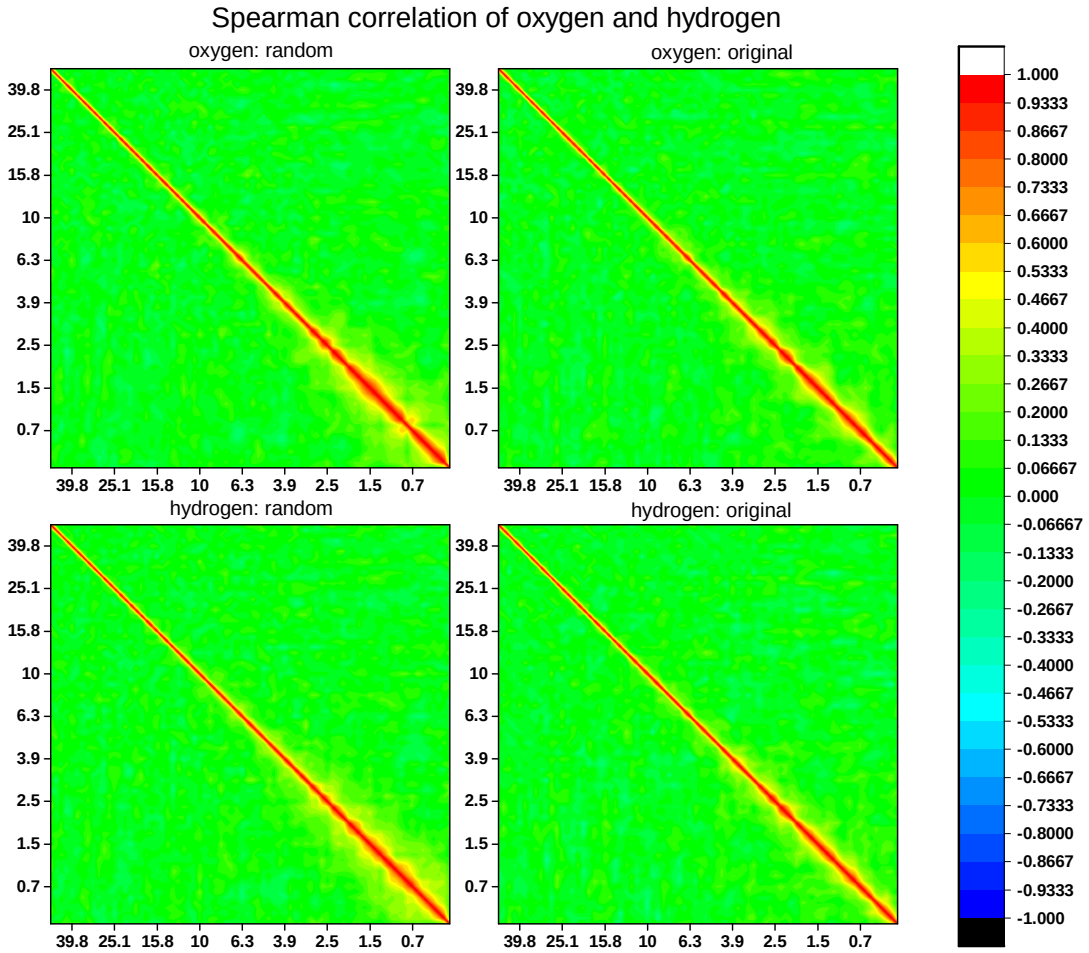

Figure 4.17: Spearman correlation of oxygen and hydrogen density: random and original dataset - The figure shows Spearman correlation matrix graph of oxygen and hydrogen density on X_0 at delay time 50.1, 47.3, ..., 0 ps before the transition for random and origial dataset.

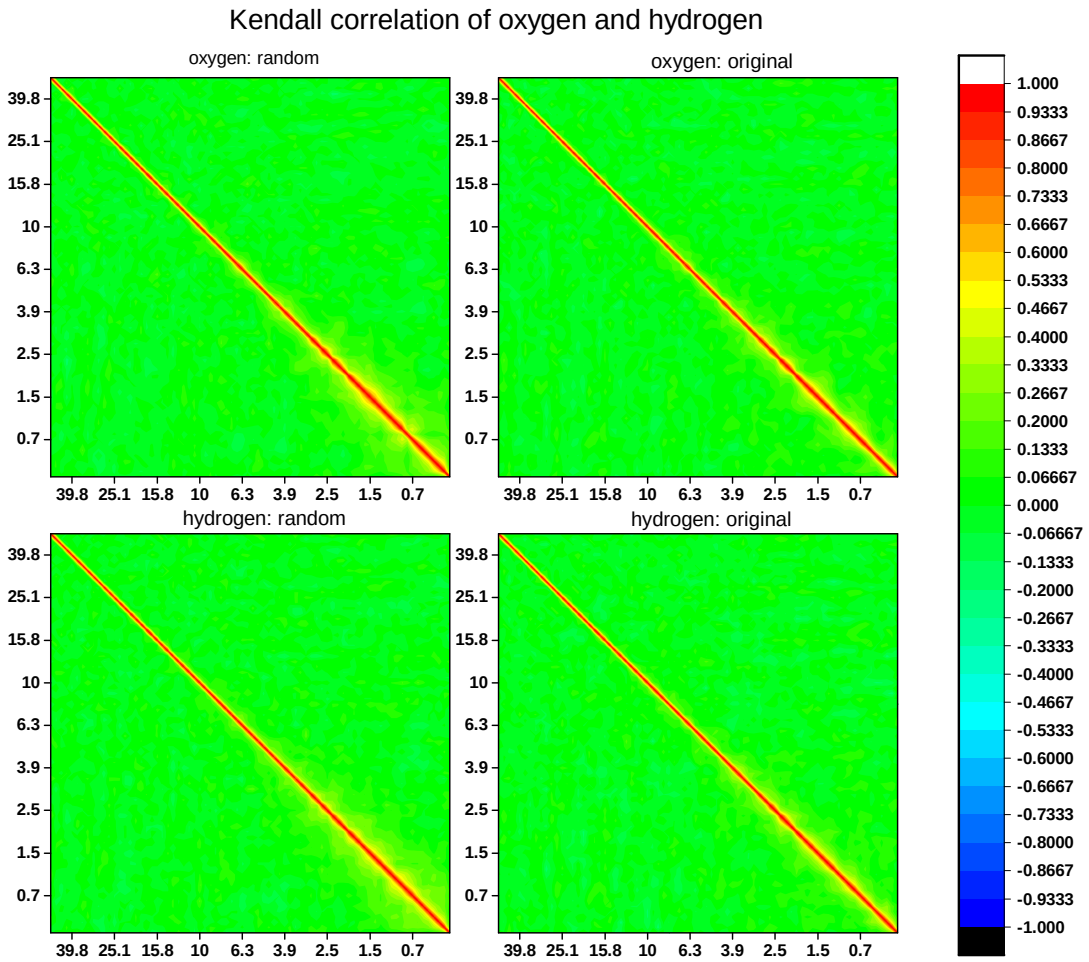

Figure 4.18: Kendall correlation of oxygen and hydrogen density: random and original dataset - The figure shows Kendall correlation matrix graph of oxygen and hydrogen density on X_0 at delay time 50.1, 47.3, ..., 0 ps before the transition for random and origial dataset.

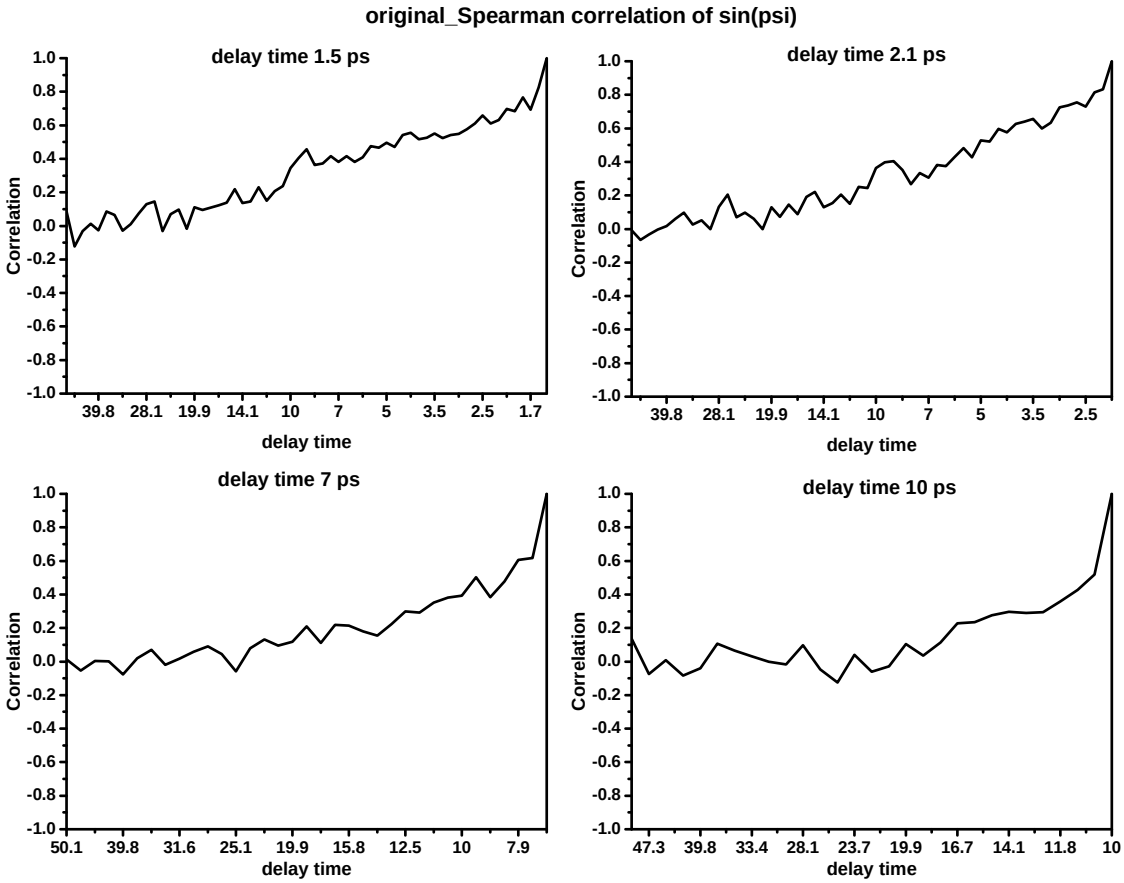

Figure 4.19: Spearman correlation of sin(psi) - one dimension: original dataset - The figure shows Spearman correlation graph of sin(psi) on X_0 at delay time 1.5, 2.1, 7 and 10 ps before the transition for origial dataset.

4.3 Conditional Correlations on X_0, X_1, X_2 and X_3

From the last section, we found that sin(psi) shows the big differences of correlation among three measures of dependence. In this section, our interest is to study conditional correlation of sin(psi) when the condition is related to hydrogen density. Therefore, we calculate the correlations of dihedral angles: sin(psi) at different delay times before transition, 50.1, 47.3, ..., 0.1, 0 under the condition of water atoms: hydrogen density on four grid points in space: X_0, X_1, X_2 and X_3 by definition of the conditional correlation (2.12) and D-vine.

Let $\sin(\psi_i)$ and $\sin(\psi_j)$ be the sine of psi at delay time i and j, respectively, where $i = j = 50.1, 47.3, \ldots, 0.1, 0$. Let δ_i be the hydrogen density at delay time i, where $i = 50.1, 47.3, \ldots, 0.1, 0$.

Therefore, the conditional correlation of $\sin(\psi_i)$ and $\sin(\psi_j)$ given δ_i is denoted by $\rho_{\psi_i\psi_j|\delta_i}$, where $i = j = 50.1, 47.3, \ldots, 0.1, 0$. From the histogram of hydrogen density both original and random dataset as Figure A.4 and A.9, we set the condition of hydrogen density as follows:

- **Condition 1:** $0.0 < \delta_i \leq 0.1$
- **Condition 2:** $0.1 < \delta_i \leq 0.2$
- **Condition 3:** $0.2 < \delta_i \leq 0.3$
- **Condition 4:** $0.3 < \delta_i \leq 0.4$

Remark. Hydrogen density at delay time i or δ_i is called the **conditioning variable**.

All hydrogen density conditions are applied for calculation of conditional correlations on X_0, X_1, X_2 and X_3 and the results are classified by four grid points of hydrogen density in space.

4.3.1 Conditional Correlations on X_0

As we know that X_0 is a grid point of the maximum probability of hydrogen atom that is located at $x = 20.25$, $y = 5.75$, $z = 1.75$ as illustrated in Figure 4.3. So the conditional correlations between $\sin(\psi_i)$ and $\sin(\psi_j)$ given δ_i on X_0 or $\rho_{\psi_i\psi_j|\delta_i}$ on X_0, we calculate three conditional correlation coefficients: Pearson, Spearman and Kendall by the definition and the conditional correlations matrix graphs both original and random dataset are given in Figure 4.20 - 4.25. For the conditional correlations by D-vine, when we fit D-vine to dataset, Kendall's tau is calculated from a bivariate copula parameter of the best fit family. Therefore, Kendall's tau conditional correlation from D-vine both original and random dataset are given in Figure 4.26 - 4.27.

Regarding the results on the statistical analysis.

1. **The definition of the conditional correlation**

   - **Condition 1:** $0.0 < \delta_i \leq 0.1$ (Figure 4.20 (a) - 4.25 (a))
   - For the two-point conditional correlations of original dataset, Pearson and Spearman correlations of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 are quite similar, as Kendall correlation is different.
   - $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 of random dataset, Pearson and Spearman correlations of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 are similar, as Kendall correlation is different.
   - For condition 1, there are lots of data according to this condition, so we investigate that the conditional correlations matrices graphs both original and random dataset are similar to the unconditional correlations in last section.
   - **Condition 2:** $0.1 < \delta_i \leq 0.2$ (Figure 4.20 (b) - 4.25 (b))
   - For the two-point conditional correlations of original dataset, three correlations of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 are different.
   - $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 of random dataset, Pearson and Spearman correlations of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 are similar, as Kendall correlation is different.
   - **Condition 3:** $0.2 < \delta_i \leq 0.3$ (Figure 4.20 (c) - 4.25 (c))
   - For the two-point conditional correlations of original dataset, Pearson and Spearman correlations of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 are quite similar, as Kendall correlation is different.
   - $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 of random dataset, Pearson and Spearman correlations of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 are similar, as Kendall correlation is a bit different.
   - **Condition 4:** $0.3 < \delta_i \leq 0.4$ (Figure 4.20 (d) - 4.25 (d))
   - For the two-point conditional correlations of original dataset, three measures of correlation of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 are quite similar.
   - $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 of random dataset, three correlations of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 are similar.
   - For condition 4, there is less data according to this condition, it means that we do not have enough data for the calculation.

2. **D-vine**

   - **Condition 1:** $0.0 < \delta_i \leq 0.1$ (Figure 4.26 (a), 4.27 (a))
   - For the two-point conditional correlations of original dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 is similar to the conditional correlations by the definition.
   - $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 of random dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 is similar to the conditional correlations by the definition.

- **Condition 2:** $0.1 < \delta_i \leq 0.2$ (Figure 4.26 (b), 4.27 (b))

- For the two-point conditional correlations of original dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 is quite similar to the conditional correlations by the definition.

- $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 of random dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 is similar to the conditional correlations by the definition.

- **Condition 3:** $0.2 < \delta_i \leq 0.3$ (Figure 4.26 (c), 4.27 (c))

- For the two-point conditional correlations of original dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 is different from the conditional correlations by the definition.

- $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 of random dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 is quite similar to the conditional correlations by the definition.

- **Condition 4:** $0.3 < \delta_i \leq 0.4$ (Figure 4.26 (d), 4.27 (d))

- For the two-point conditional correlations of original dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 is different from the conditional correlations by the definition.

- $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 of random dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 is different from the conditional correlations by the definition.

Obviously, the conditional correlations $\rho_{\psi_i\psi_j|\delta_i}$ on X_0 by D-vine can explain or show the conditional correlations matrix graphs both original and random dataset clearer than by the definition. Because D-vine gives a specific way of decomposing the probability density. The dependency structure is determined by the bivariate copulas and a nested set of trees using pair-copula. Therefore, this is an advantage which we take from D-vine to be an alternative way for the conditional correlation study in this research.

Regarding the molecular meaning of the statistical results.

1. The matrices graphs of conditional correlations show the complete picture of statistical dependence of sin(psi) at different time moments with respect to the transition moment under the hydrogen density conditions on X_0. For example, the row starting at 50.1 shows how much the value of psi(phi) at 50.1 picoseconds before the transition depends on all the values of psi (phi) at previous time moments under each condition of the hydrogen density on X_0.

2. The "random" matrices graphs show a little bit different behaviour in conditional correlations under four conditions depending on the time moment before transition:

- **Condition 1:** $0.0 < \delta_i \leq 0.1$ (Figure 4.23 (a), 4.24 (a), 4.25 (a))
- Three conditional correlations by definition and Kendall conditional correlation by D-vine show that there is no difference in conditional dependencies at all starting times: the sequence of conditional correlations at 0 delay is the same as the sequence at 1.5 ps or 10 ps delays.
- **Condition 2:** $0.1 < \delta_i \leq 0.2$ (Figure 4.23 (b), 4.24 (b), 4.25 (b))
- The conditional correlations matrices show that there is negative correlation at 39.8 - 10 ps delays.
- **Condition 3:** $0.2 < \delta_i \leq 0.3$ (Figure 4.23 (c), 4.24 (c), 4.25 (c))
- The conditional correlations matrices show that there is negative correlation at 50.1 - 2.5 ps delays.
- **Condition 4:** $0.3 < \delta_i \leq 0.4$ (Figure 4.23 (d), 4.24 (d), 4.25 (d))
- There is less data according to this condition, so we do not have enough data to analyse and summarize for this case.

3. The "original" matrices graphs show very different behaviour in conditional correlations under four conditions depending on the time moment before transition:

- **Condition 1:** $0.0 < \delta_i \leq 0.1$ (Figure 4.20 (a), 4.21 (a), 4.22 (a))
- In advance of the transition (rows starting at 50.1 - 10 delays) the conditional correlations are essentially the same as for the "random" dataset.
- Starting from 10 delays and up to 1.5 ps, the conditional correlations are much stronger and longer. The row at 1.9 ps, for example, has very strongly correlated values of psi up to 2 - 5 ps in advance (the correlation coefficient is 0.8 - 0.9).
- At 1.0 - 1.5 ps, surprisingly, the conditional correlations are very low again, almost like in the "random" dataset.
- Before the transition, 0.5 - 1.1 ps, the conditional correlations are stronger than usual again. There is negative correlation with the data at 1.5 - 12.5 ps delays (correlation value is -0.5 - -0.4).
- **Condition 2:** $0.1 < \delta_i \leq 0.2$ (Figure 4.20 (b), 4.21 (b), 4.22 (b))
- Starting from 6.3 delays and up to 1.5 ps, the conditional correlations are much stronger and longer. The row at 5 ps, for example, has very strongly correlated values of psi up to 3.9 - 3.1 ps in advance (the correlation coefficient is 0.6 - 0.9).
- At 1.5 ps, the conditional correlation is low.
- Before the transition, 0 - 1.1 ps, the conditional correlations are stronger. There is negative correlation with the data at 1.5 - 7.9 ps delays (correlation value is -0.5 - -0.3).

- **Condition 3:** $0.2 < \delta_i \leq 0.3$ (Figure 4.20 (c), 4.21 (c), 4.22 (c))

- Starting from 7.9 delays and up to 2.5, the conditional correlations are much stronger and longer. The row at 6.3 ps, for example, has very strongly correlated values of psi up to 3.1 - 3.9 ps in advance (the correlation coefficient is 0.7 - 0.9).
- At 1.1 - 1.9 ps, surprisingly, the conditional correlations are very low.
- Before the transition, 0 - 1.1 ps, the conditional correlations are stronger again. There is negative correlation with the data at 1.5 - 39.8 ps delays (correlation value is -0.5 - -0.3).

- **Condition 4:** $0.3 < \delta_i \leq 0.4$ (Figure 4.20 (d), 4.21 (d), 4.22 (d))

- There is less data according to this condition, so we do not have enough data to analyse and summarize for this case.

Summary: For the conditional correlation of sin(psi) under the hydrogen density condition on X_0, we summarize that

1. **Regarding the results on the statistical analysis**

   - By the definition of the conditional correlation, Pearson and Spearman correlations gave quite similar conditional correlation matrices graphs both original and random dataset, as Kendall correlation gave different matrices graphs under the most conditions.
   - By D-vine, Kendall correlation gave conditional correlation matrices graphs similar to the conditional correlation by the definition both original and random dataset under the most conditions.

2. **Regarding the molecular meaning of the statistical results**

   - For random dataset, the conditional correlation matrices graphs show a little bit different behaviour in conditional correlations under four conditions depending on the time moment before transition.
   - For original dataset, the conditional correlation matrices graphs show very different behaviour in conditional correlations under four conditions depending on the time moment before transition.

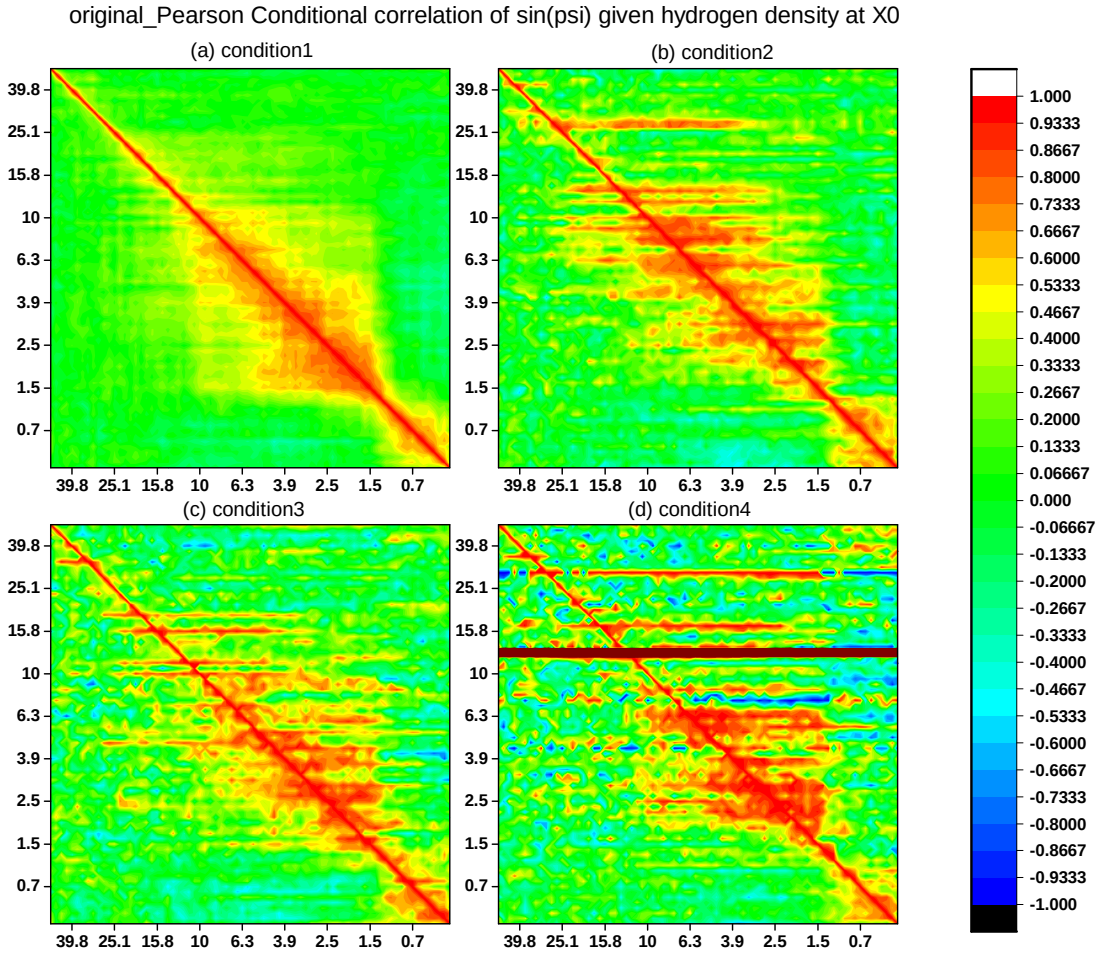

Figure 4.20: Pearson conditional correlation of sin(psi) on X_0: original dataset - The figure shows Pearson conditional correlation matrix graph of sin(psi) on X_0 and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset.

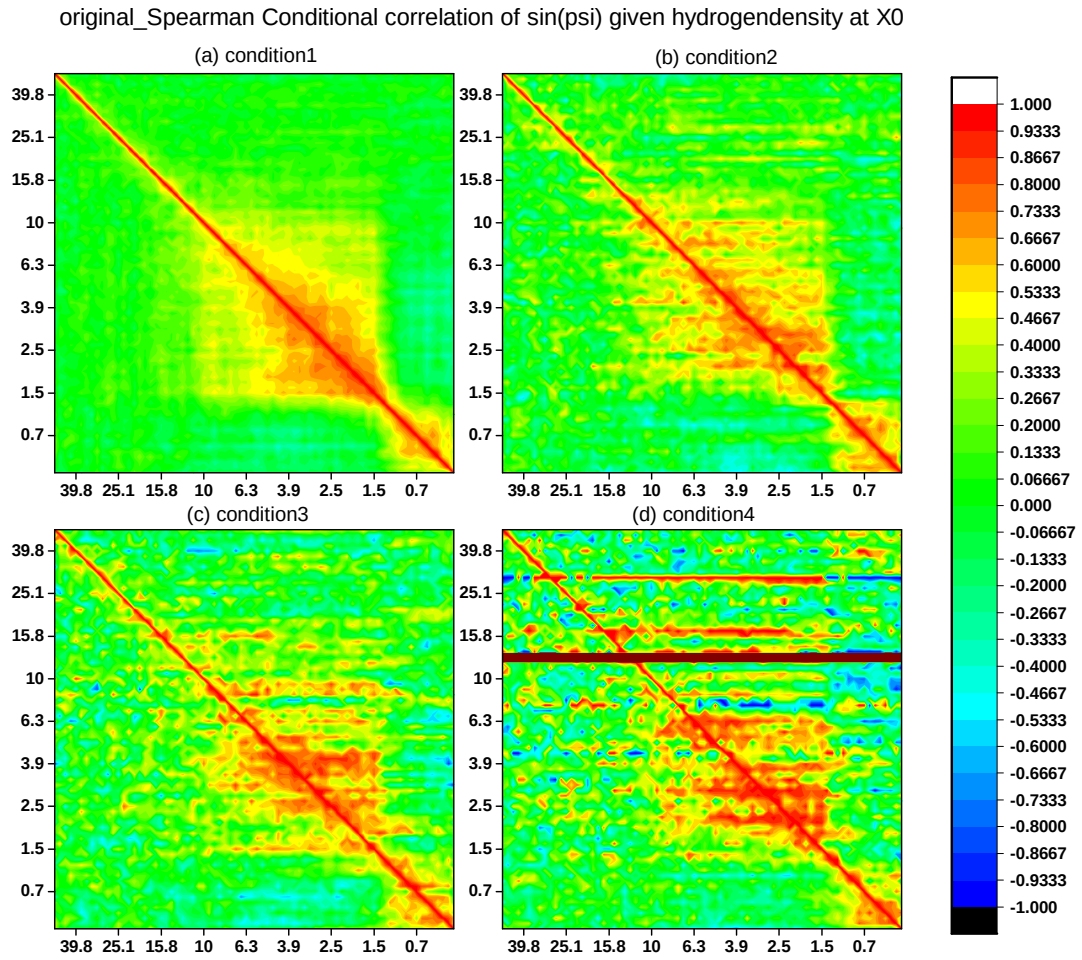

Figure 4.21: Spearman conditional correlation of sin(psi) on X_0: original dataset - The figure shows Spearman conditional correlation matrix graph of sin(psi) on X_0 and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset.

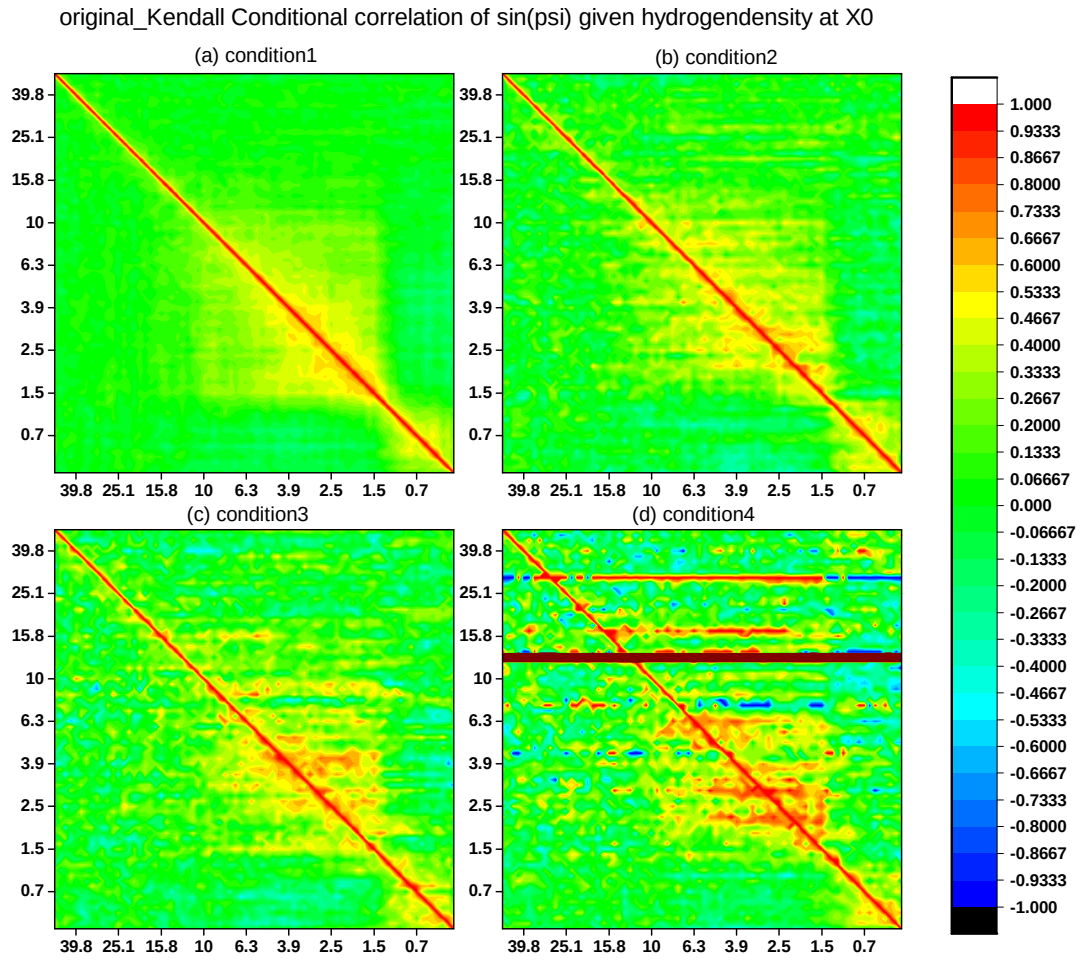

Figure 4.22: Kendall conditional correlation of sin(psi) on X_0: original dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on X_0 and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset.

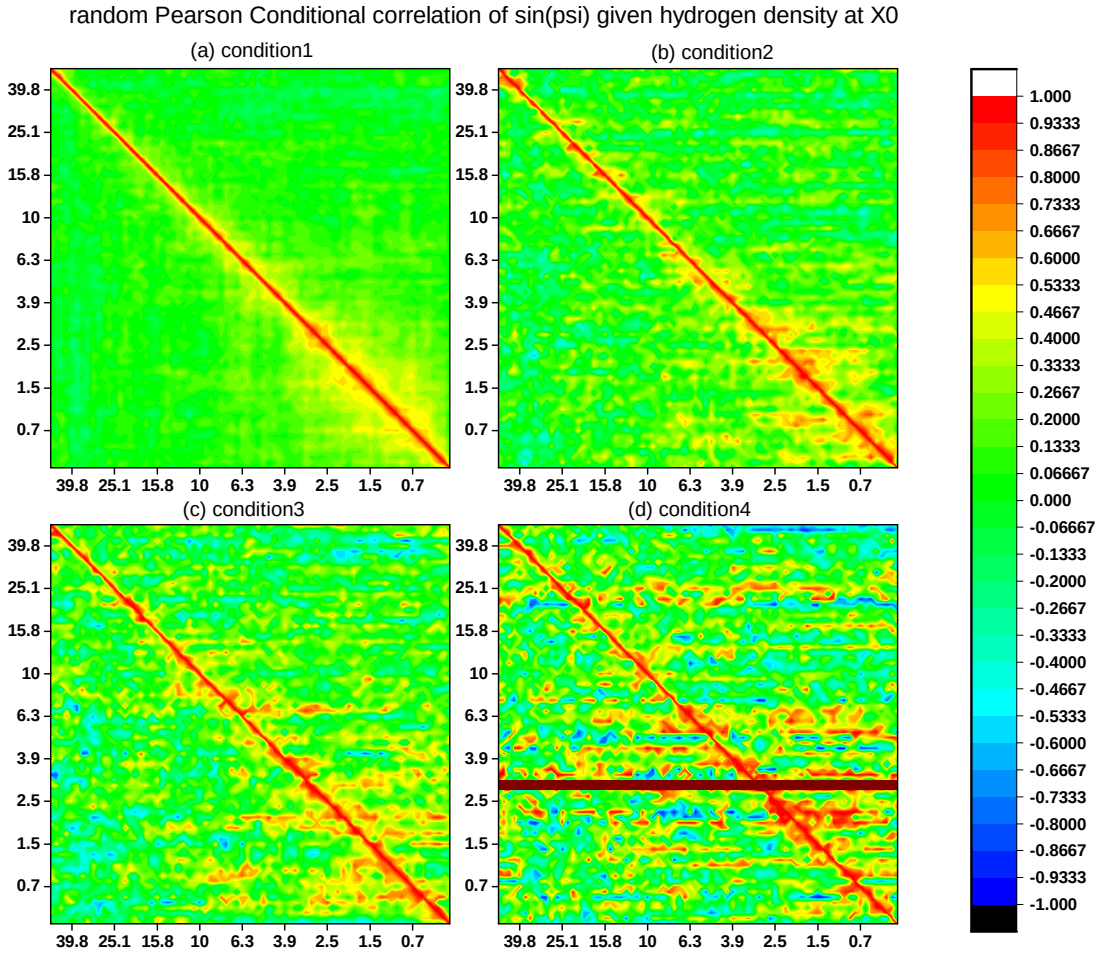

Figure 4.23: Pearson conditional correlation of sin(psi) on X_0: random dataset - The figure shows Pearson conditional correlation matrix graph of sin(psi) on X_0 and delay time 50.1, 47.3, ..., 0 ps before the transition for random dataset.

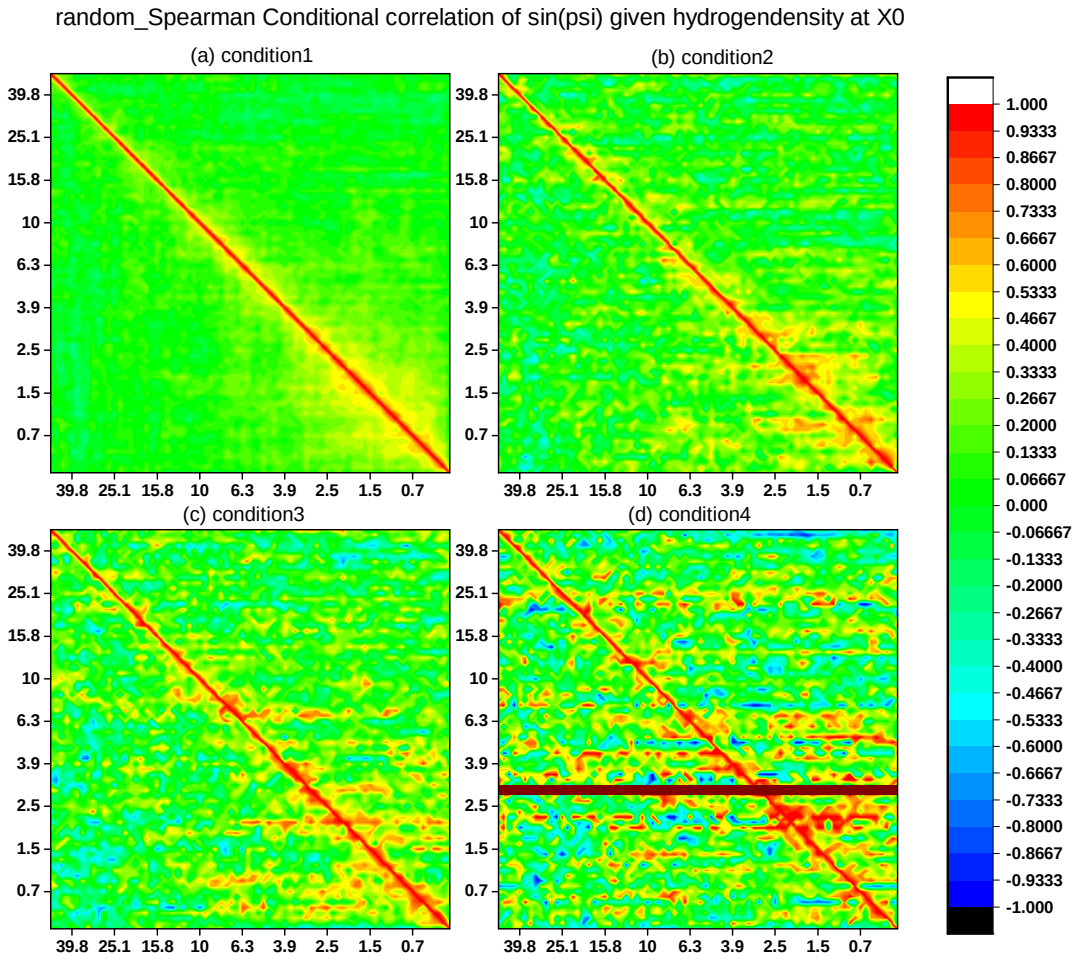

Figure 4.24: Spearman conditional correlation of sin(psi) on X_0: random dataset - The figure shows Spearman conditional correlation matrix graph of sin(psi) on X_0 and delay time 50.1, 47.3, ..., 0 ps before the transition for random dataset.

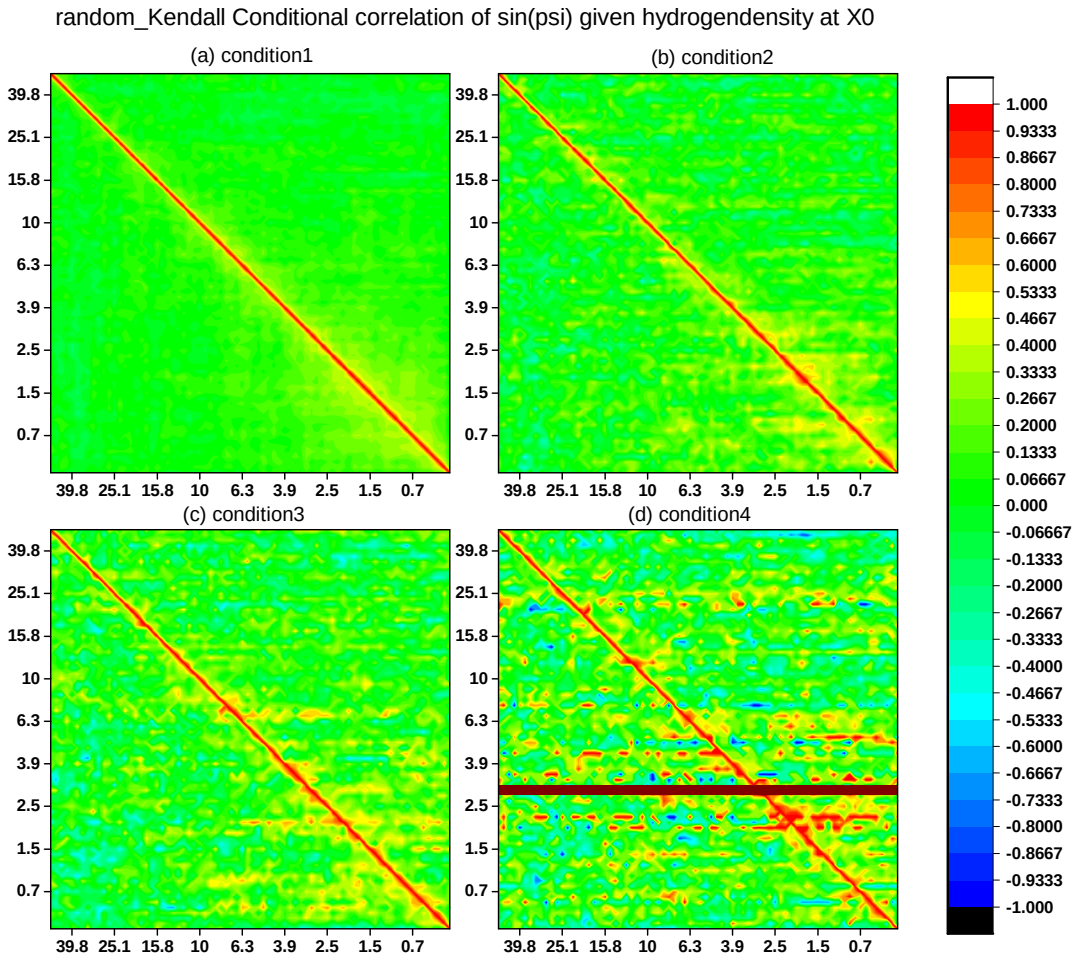

Figure 4.25: Kendall conditional correlation of sin(psi) on X_0: random dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on X_0 and delay time 50.1, 47.3, ..., 0 ps before the transition for random dataset.

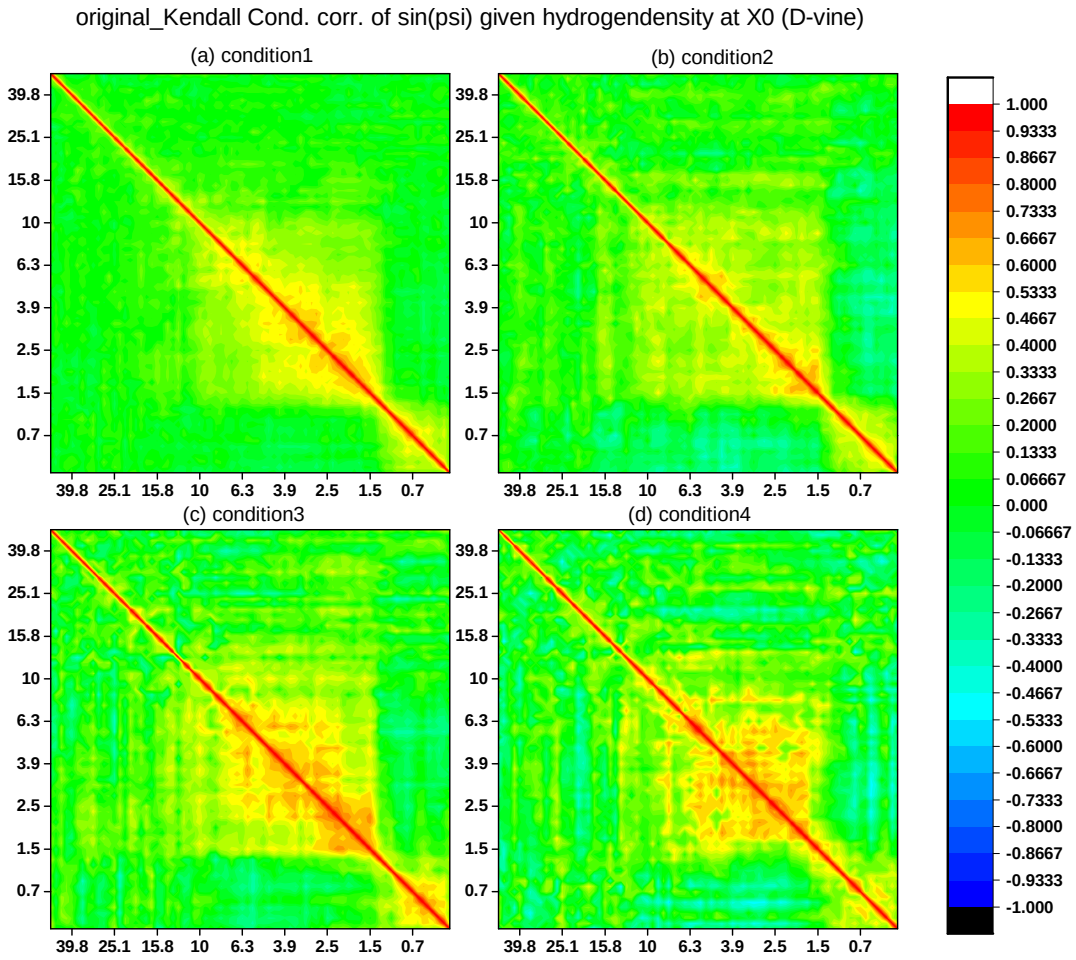

Figure 4.26: Kendall conditional correlation of sin(psi) on X_0 by D-vine: original dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on X_0 and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset by D-vine.

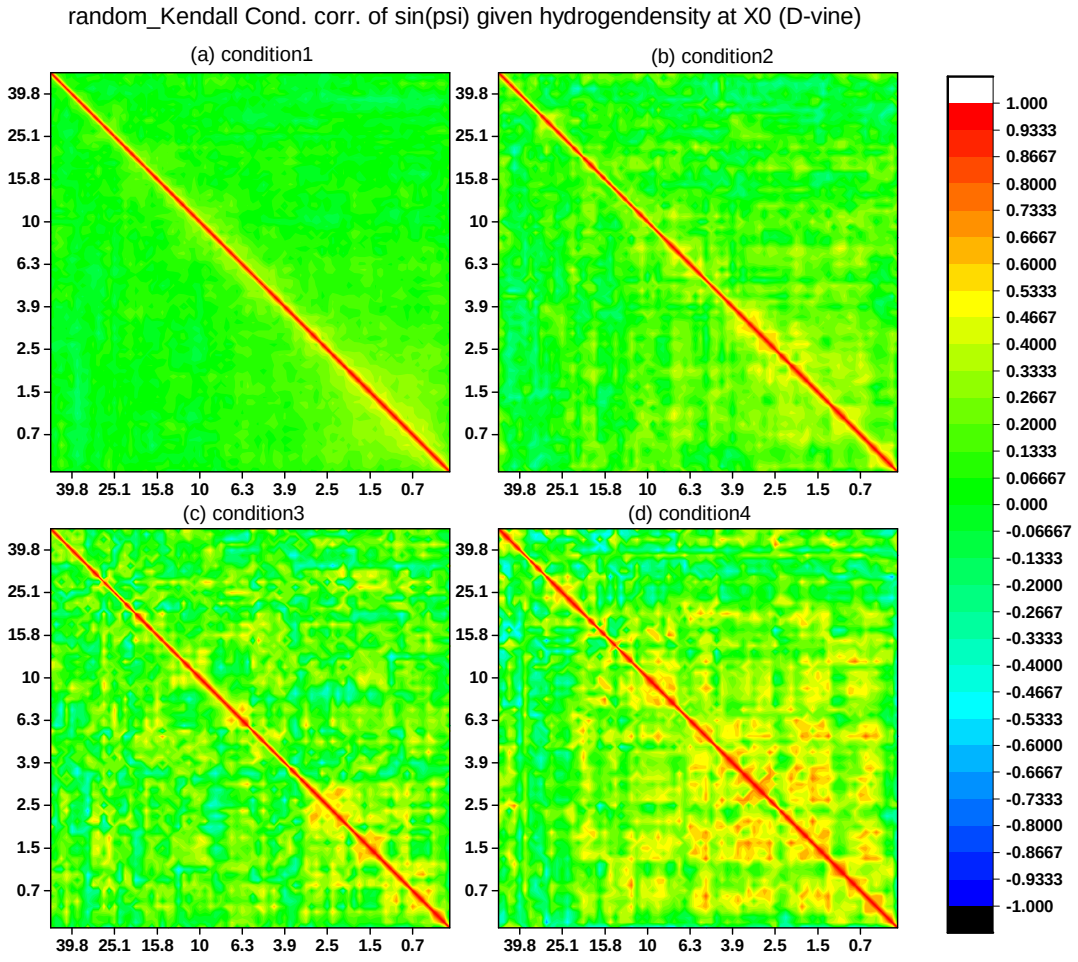

Figure 4.27: Kendall conditional correlation of sin(psi) on X_0 by D-vine: random dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on X_0 and delay time 50.1, 47.3, ..., 0 ps before the transition for random dataset by D-vine.

4.3.2 Conditional Correlations on X_1

In this subsection, we would like to study and investigate the conditional correlations between $\sin(\psi_i)$ and $\sin(\psi_j)$ given δ_i or $\rho_{\psi_i\psi_j|\delta_i}$ on X_1. From the first section, we know that X_1 is an opposite grid point of X_0 in space that is located at $x = 12.25$, $y = 6.75$, $z = 4.75$. Therefore, we calculate Kendall conditional correlation coefficient by definition of the conditional correlation (2.12) and D-vine only the original dataset and the conditional correlations matrix graphs are given in Figure 4.28 and 4.29.

Regarding the results on the statistical analysis, the results under the definition of the conditional correlation and D-vine are given in Table 4.4.

Table 4.4: The conditional correlation results of sin(psi) under the condition of hydrogen density on X_1 regarding the results on the statistical analysis under the definition of the conditional correlation and D-vine.

The definition

condition	results
1 (Figure 4.28 (a))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_1 is similar to $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0
2 (Figure 4.28 (b))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_1 is a bit different from $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0
3 (Figure 4.28 (c))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_1 is quite similar to $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0
4 (Figure 4.28 (d))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_1 is different from $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0

D-vine

condition	results
1 (Figure 4.29 (a))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_1 is similar to the cond. cor. by the definition
2 (Figure 4.29 (b))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_1 is different from the cond. cor. by the definition
3 (Figure 4.29 (c))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_1 is different from the cond. cor. by the definition
4 (Figure 4.29 (d))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_1 is different from the cond. cor. by the definition

Obviously, the conditional correlations $\rho_{\psi_i\psi_j|\delta_i}$ on X_1 by D-vine can explain or show the conditional correlations matrix graphs clearer than by the definition.

Regarding the molecular meaning of the statistical results, the results are given in Table 4.5.

Table 4.5: The conditional correlation results of sin(psi) under the condition of hydrogen density on X_1 regarding the molecular meaning of the statistical results.

condition	results
1 (Figure 4.28 (a))	- from 10 delays and up to 1.5 ps, the cond. cor. are much stronger - at 1.2 - 1.5 ps, the cond. cor. are very low - there is negative cor. with the data at 1.5 - 7.9 ps delays
2 (Figure 4.28 (b))	- from 5 delays and up to 3.1 ps, the cond. cor. are much stronger - at 1.5 ps, the cond. cor. is low - there is negative cor. with the data at 1.9 - 12.5 ps delays
3 (Figure 4.28 (c))	- from 5 delays and up to 2.5 ps, the cond. cor. are much stronger - at 1.5 - 1.9 ps, the cond. cor. are very low - there is negative cor. with the data at 1.5 - 39.8 ps delays
4 (Figure 4.28 (d))	- we do not have enough data to analyse

Summary: For the conditional correlation of sin(psi) under the hydrogen density condition on X_1, we summarize that

1. **Regarding the results on the statistical analysis**
   - By the definition of the conditional correlation, Kendall correlation gave conditional correlation matrices graphs similar to the conditional correlation on X_0 under condition 1 and 3. As Kendall correlation gave different conditional correlation matrices graphs from the conditional correlation on X_0 under condition 2 and 4.
   - By D-vine, Kendall correlation gave different conditional correlation matrices graphs from the conditional correlation by the definition under condition 2, 3 and 4.
2. **Regarding the molecular meaning of the statistical results**
   - For original dataset, the conditional correlation matrices graphs show very different behaviour in conditional correlations under four conditions depending on the time moment before transition.

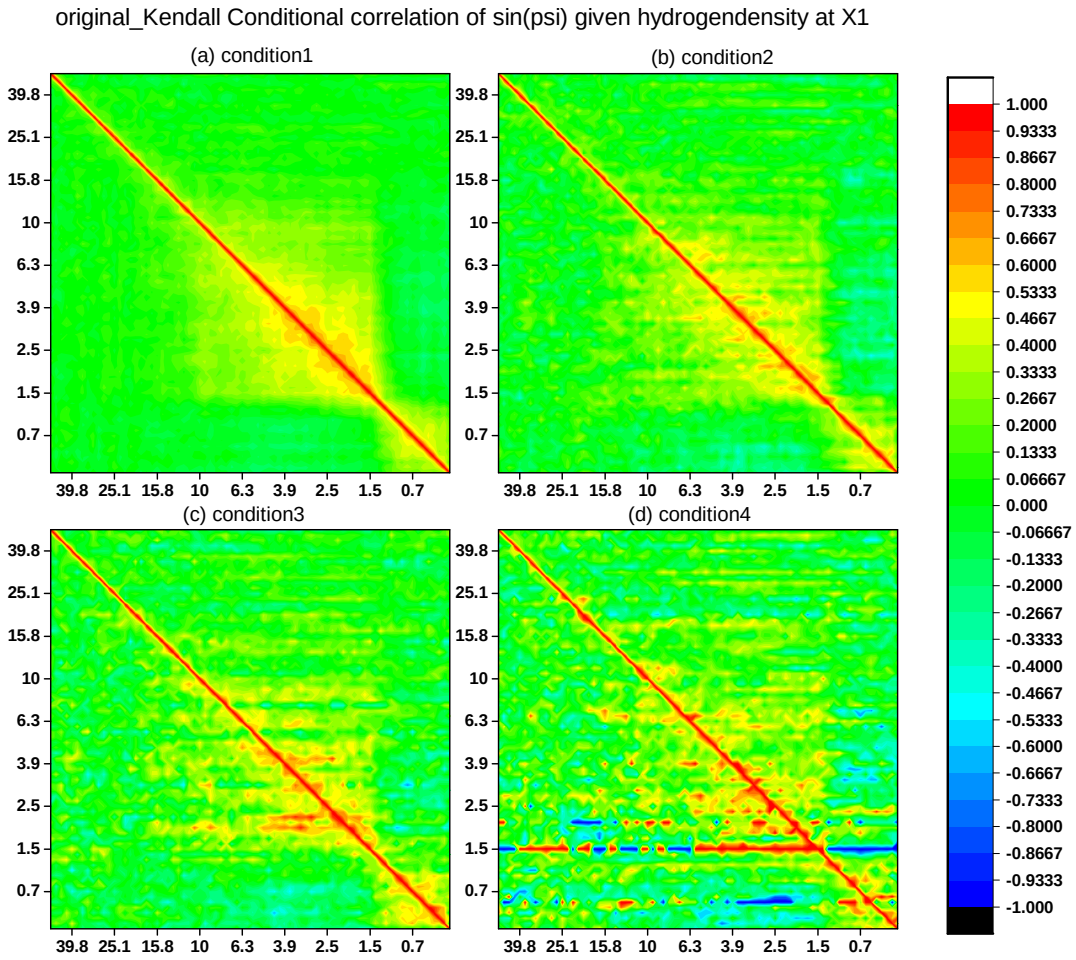

Figure 4.28: Kendall conditional correlation of sin(psi) at X_1: original dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) at X_1 and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset.

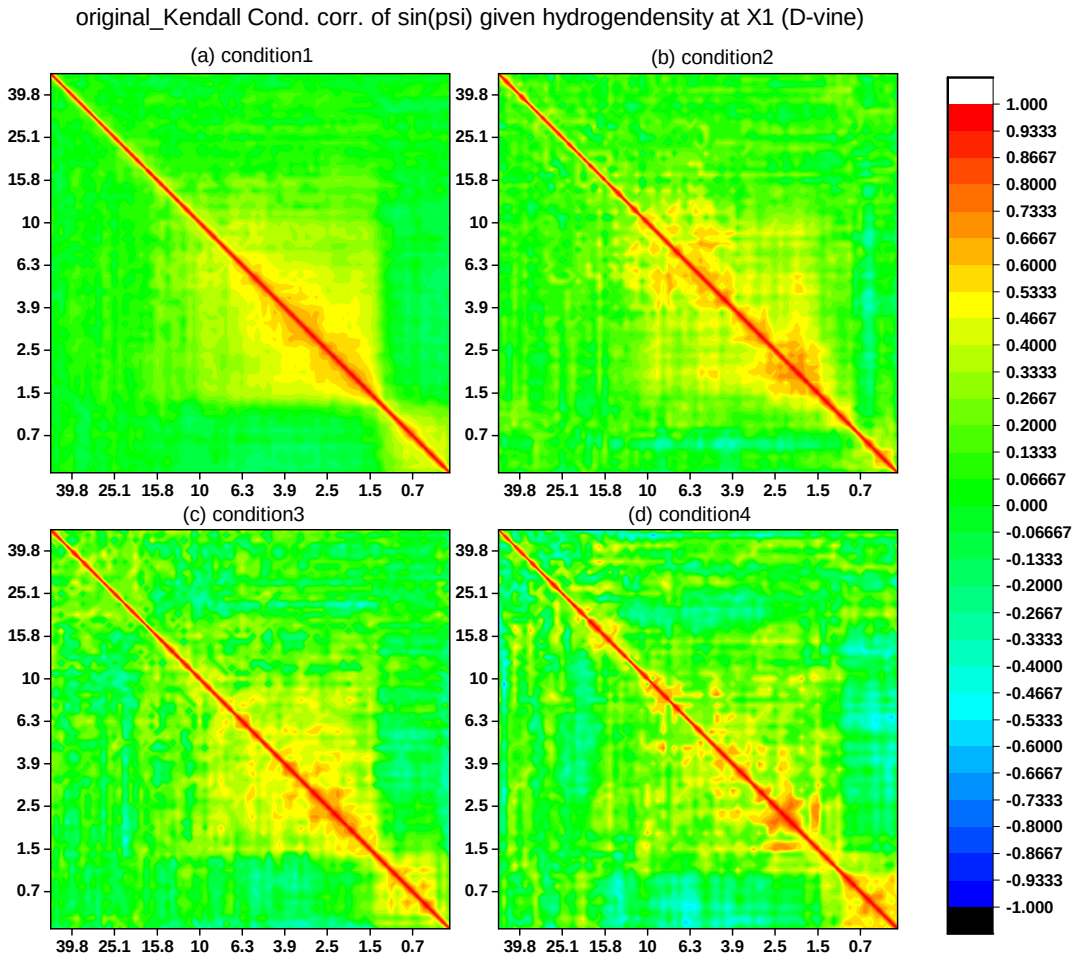

Figure 4.29: Kendall conditional correlation of sin(psi) on X_1 by D-vine: original dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on X_1 and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset by D-vine.

4.3.3 Conditional Correlations on X_2

In this subsection, we would like to study and investigate the conditional correlations between $\sin(\psi_i)$ and $\sin(\psi_j)$ given δ_i or $\rho_{\psi_i\psi_j|\delta_i}$ on X_2. From the first section, we know that X_2 is a neighbor grid point of X_0 in space that is located at $x = 20.75$, $y = 5.75$, $z = 1.75$. Therefore, we calculate Kendall conditional correlation coefficient by the definition of the conditional correlation (2.12) and D-vine only the original dataset and the conditional correlations matrix graphs are given in Figure 4.30 and 4.31.

Regarding the results on the statistical analysis, the results under the definition of the conditional correlation and D-vine are given in Table 4.6.

Table 4.6: The conditional correlation results of sin(psi) under the condition of hydrogen density on X_2 regarding the results on the statistical analysis under the definition of the conditional correlation and D-vine.

The definition

condition	results
1 (Figure 4.30 (a))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_2 is similar to $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0 and X_1
2 (Figure 4.30 (b))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_2 is similar to $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0 and a bit different on X_1
3 (Figure 4.30 (c))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_2 is different from $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0 and X_1
4 (Figure 4.30 (d))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_1 is different from $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0 and X_1

D-vine

condition	results
1 (Figure 4.31 (a))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_2 is similar to the cond. cor. by the definition
2 (Figure 4.31 (b))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_2 is different from the cond. cor. by the definition
3 (Figure 4.31 (c))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_2 is different from the cond. cor. by the definition
4 (Figure 4.31 (d))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_2 is different from the cond. cor. by the definition

Regarding the molecular meaning of the statistical results, the results are given in Table 4.7.

Table 4.7: The conditional correlation results of sin(psi) under the condition of hydrogen density on X_2 regarding the molecular meaning of the statistical results.

condition	results
1 (Figure 4.30 (a))	- from 5 delays and up to 1.5 ps, the cond. cor. are much stronger - at 1.1 - 1.5 ps, the cond. cor. are very low - there is negative cor. with the data at 1.5 - 10 ps delays
2 (Figure 4.30 (b))	- from 3.1 delays and up to 1.9 ps, the cond. cor. are much stronger - at 1.5 ps, the cond. cor. is low - there is negative cor. with the data at 1.5 - 50.1 ps delays
3 (Figure 4.30 (c))	- from 7.9 delays and up to 2.5 ps, the cond. cor. are much stronger - there is negative cor. with the data at 1.5 - 50.1 ps delays
4 (Figure 4.30 (d))	- we do not have enough data to analyse

Summary: For the conditional correlation of sin(psi) under the hydrogen density condition on X_2, we summarize that

1. **Regarding the results on the statistical analysis**
   - By the definition of the conditional correlation, Kendall correlation gave conditional correlation matrices graphs similar to the conditional correlation on X_0 and X_1 under condition 1 and 2. As Kendall correlation gave different conditional correlation matrices graphs from the conditional correlation on X_0 and X_1 under condition 3 and 4.
   - By D-vine, Kendall correlation gave different conditional correlation matrices graphs from the conditional correlation by the definition under condition 2, 3 and 4.
2. **Regarding the molecular meaning of the statistical results**
   - For original dataset, the conditional correlation matrices graphs show very different behaviour in conditional correlations under four conditions depending on the time moment before transition.

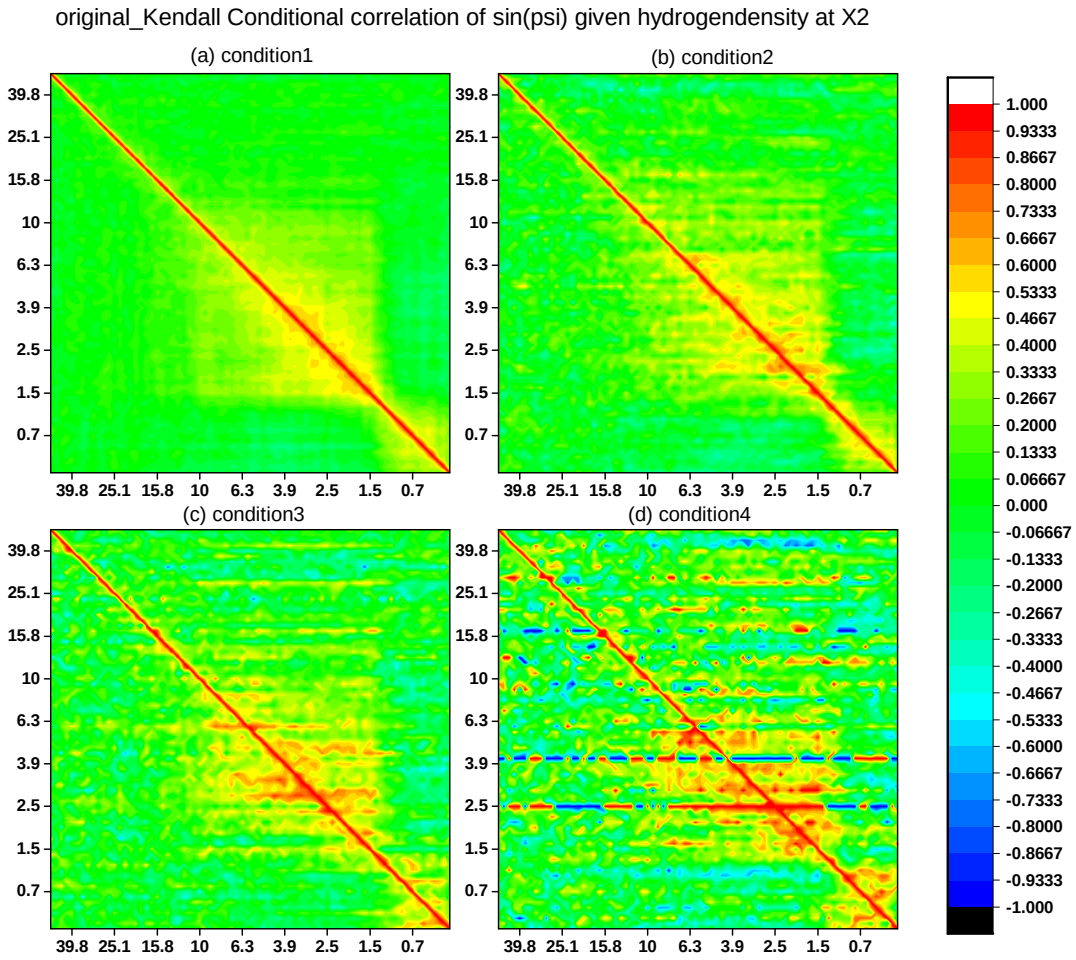

Figure 4.30: Kendall conditional correlation of sin(psi) at X_2: original dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) at X_2 and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset.

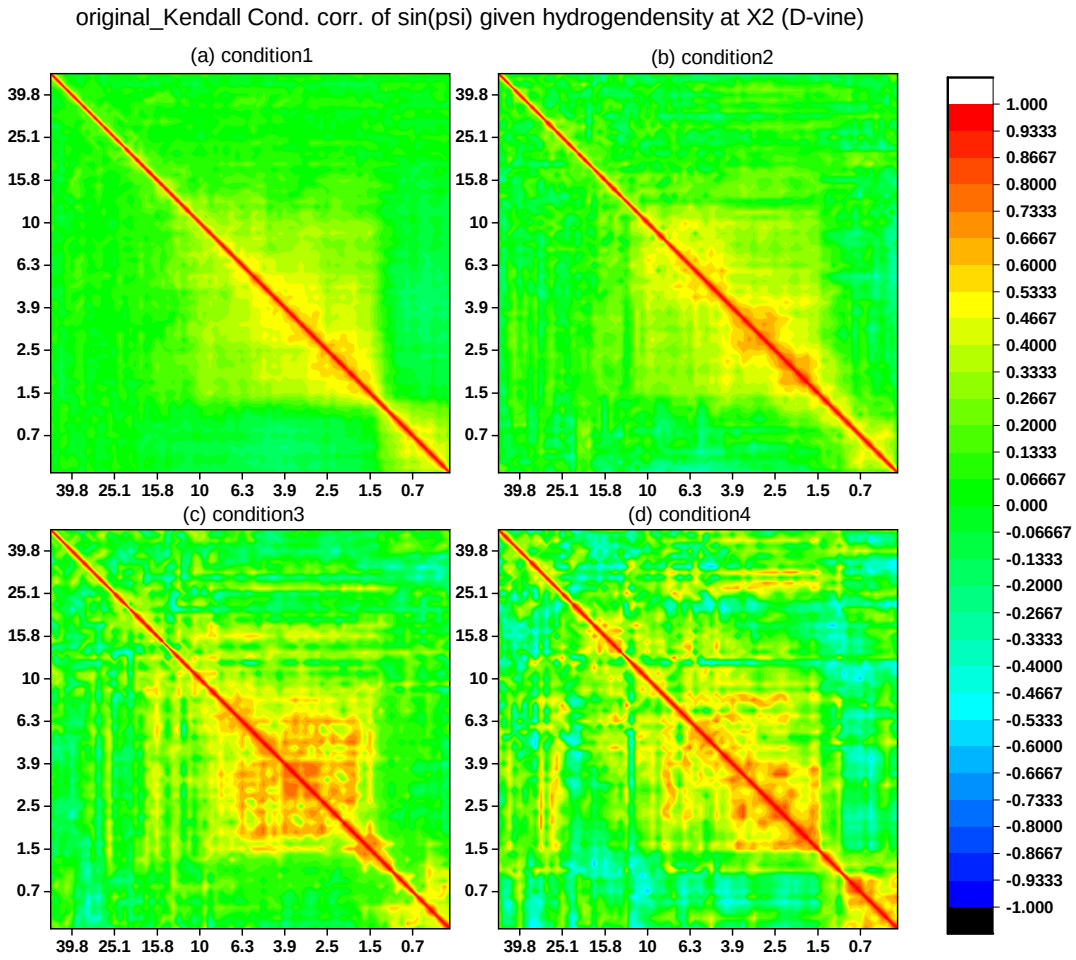

Figure 4.31: Kendall conditional correlation of sin(psi) on X_2 by D-vine: original dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on X_2 and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset by D-vine.

4.3.4 Conditional Correlations on X_3

In this subsection, we would like to study and investigate the conditional correlations between $\sin(\psi_i)$ and $\sin(\psi_j)$ given δ_i or $\rho_{\psi_i\psi_j|\delta_i}$ on X_3. From the first section, we know that X_3 is a neighbor grid point of X_0 in space that is located at $x = 20.25$, $y = 6.25$, $z = 1.75$. Therefore, we calculate Kendall conditional correlation coefficient by the definition of the conditional correlation (2.12) and D-vine only the original dataset and the conditional correlations matrix graphs are given in Figure 4.32 and 4.33.

Regarding the results on the statistical analysis, the results under the definition of the conditional correlation and D-vine are given in Table 4.8.

Table 4.8: The conditional correlation results of sin(psi) under the condition of hydrogen density on X_3 regarding the results on the statistical analysis under the definition of the conditional correlation and D-vine.

The definition

condition	results
1 (Figure 4.32 (a))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_3 is similar to $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0, X_1 and X_2
2 (Figure 4.32 (b))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_3 is a bit different from $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0, X_1 and X_2
3 (Figure 4.32 (c))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_3 is a bit different from $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0, X_1 and X_2
4 (Figure 4.32 (d))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_3 is different from $\rho_{\psi_i\psi_j\|\delta_i}$ on X_0, X_1 and X_2

D-vine

condition	results
1 (Figure 4.33 (a))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_3 is similar to the cond. cor. by the definition
2 (Figure 4.33 (b))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_3 is different from the cond. cor. by the definition
3 (Figure 4.33 (c))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_3 is different from the cond. cor. by the definition
4 (Figure 4.33 (d))	$\rho_{\psi_i\psi_j\|\delta_i}$ on X_3 is different from the cond. cor. by the definition

Regarding the molecular meaning of the statistical results, the results are given in Table 4.9.

Table 4.9: The conditional correlation results of sin(psi) under the condition of hydrogen density on X_3 regarding the molecular meaning of the statistical results.

condition	results
1 (Figure 4.32 (a))	- from 5 delays and up to 2.5 ps, the cond. cor. are much stronger - at 1.1 - 1.5 ps, the cond. cor. are very low - there is negative cor. with the data at 1.9 - 39.8 ps delays
2 (Figure 4.32 (b))	- from 5 delays and up to 1.5 ps, the cond. cor. are much stronger - at 1.1 - 1.5 ps, the cond. cor. is low - there is negative cor. with the data at 1.5 - 15.8 ps delays
3 (Figure 4.32 (c))	- from 3.9 delays and up to 1.9 ps, the cond. cor. are much stronger - there is negative cor. with the data at 1.9 - 19.9 ps delays
4 (Figure 4.32 (d))	- we do not have enough data to analyse

Summary: For the conditional correlation of sin(psi) under the hydrogen density condition on X_3, we summarize that

1. **Regarding the results on the statistical analysis**
   - By the definition of the conditional correlation, Kendall correlation gave conditional correlation matrices graphs similar to the conditional correlation on X_0, X_1 and X_2 under condition 1. As Kendall correlation gave different conditional correlation matrices graphs from the conditional correlation on X_0, X_1 and X_2 under condition 2, 3 and 4.
   - By D-vine, Kendall correlation gave different conditional correlation matrices graphs from the conditional correlation by the definition under condition 2, 3 and 4.
2. **Regarding the molecular meaning of the statistical results**
   - For original dataset, the conditional correlation matrices graphs show very different behaviour in conditional correlations under four conditions depending on the time moment before transition.

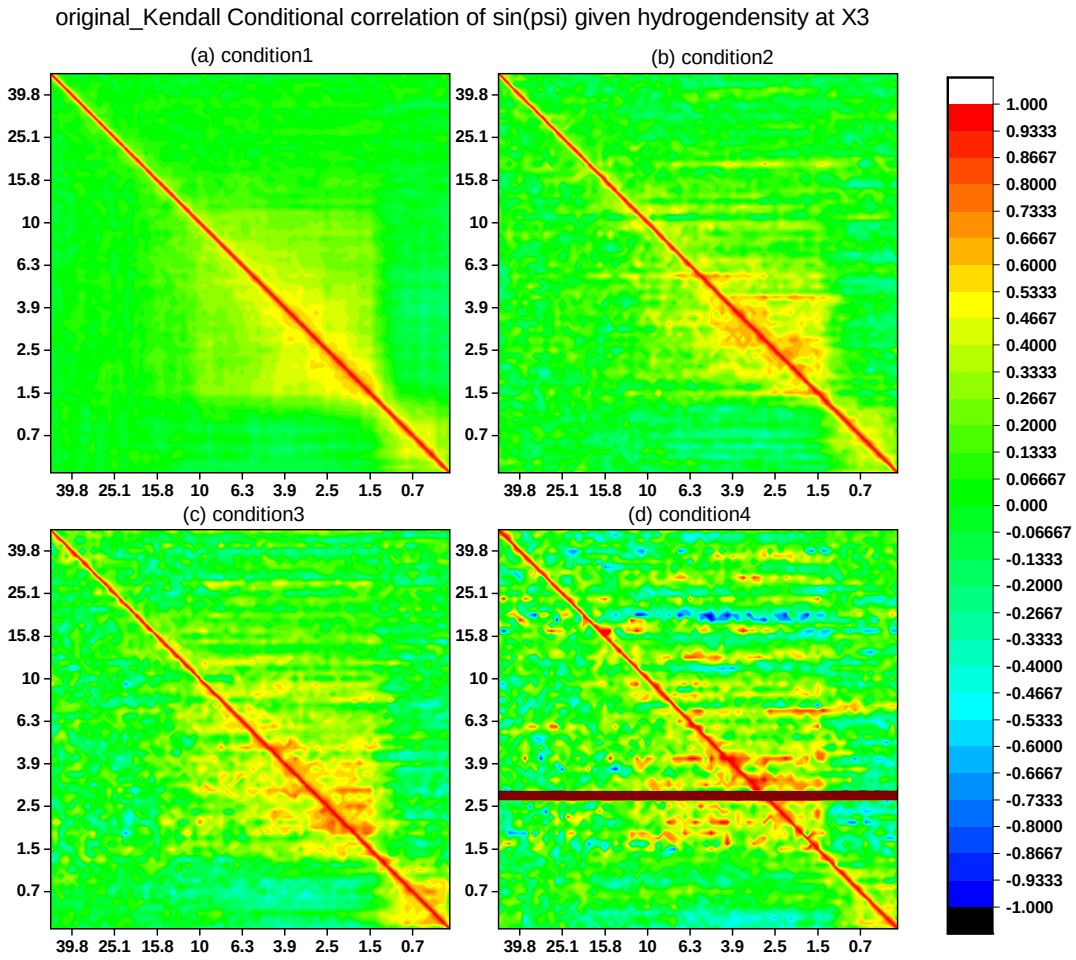

Figure 4.32: Kendall conditional correlation of sin(psi) at X_3: original dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) at X_3 and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset.

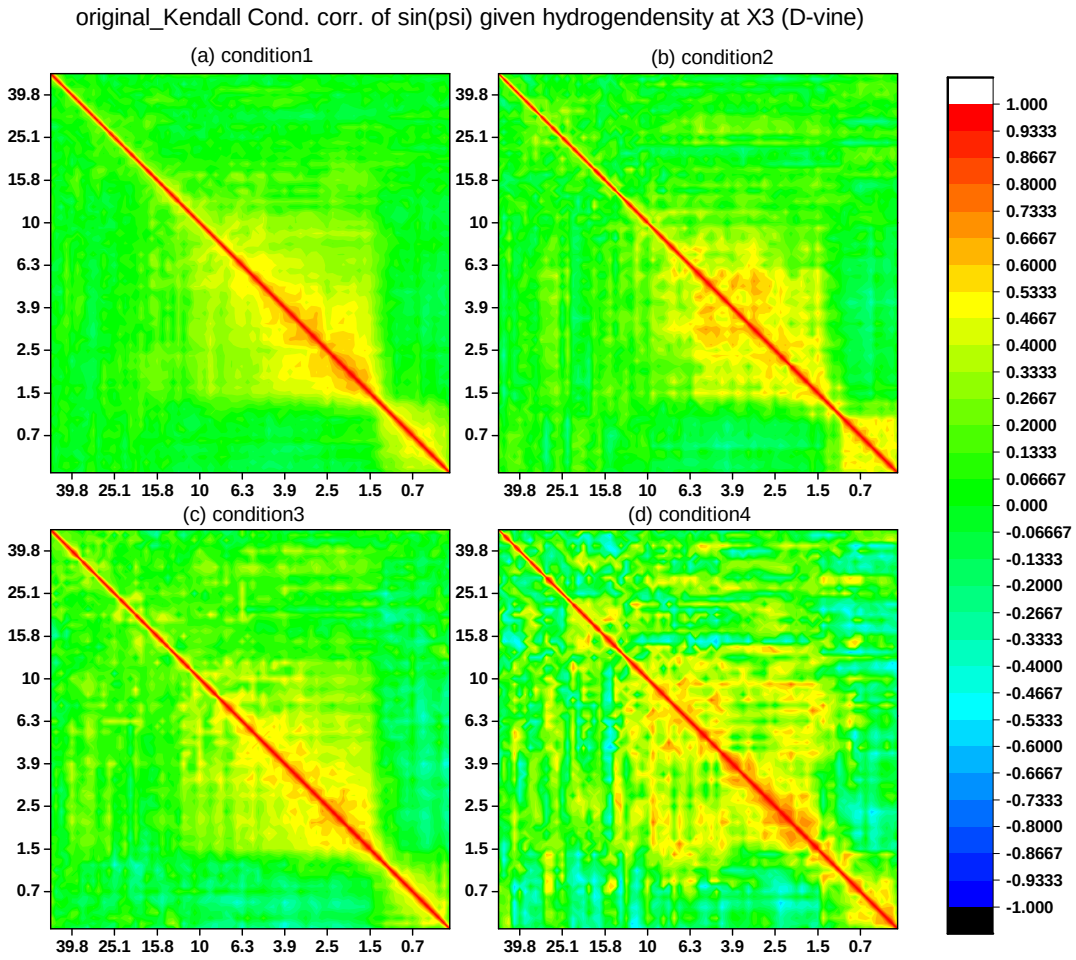

Figure 4.33: Kendall conditional correlation of sin(psi) on X_3 by D-vine: original dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on X_3 and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset by D-vine.

Conclusions: Conditional correlations on X_0, X_1, X_2 and X_3 under four conditions, we conclude that

1. X_0 is a grid point of the maximum probability of hydrogen and the distance from X_0 to the centre of mass of the peptide (COM) is 6.532.
2. X_1 is an opposite grid point of X_0 and the distance from X_1 to COM is 3.466 that is the shortest distance.
3. X_2 is a neighbor grid point of X_0 and the distance from X_2 to COM is 7.011 that is the longest distance.
4. X_3 is a neighbor grid point of X_0 and the distance from X_3 to COM is 6.661.
5. For condition 1, there are lots of data according to this condition, so we notice that the conditional correlations matrices graphs are similar to the unconditional correlations matrices graphs. As condition 4, there is less data according to this condition, so we do not have enough data for the calculation.
6. The matrices graphs of conditional correlations show the complete picture of statistical dependence of sin(psi) at different time moments with respect to the transition moment under the hydrogen density conditions on X_0, X_1, X_2 and X_3.
7. The conditional correlations matrices graphs on X_0, X_1, X_2 and X_3 by D-vine are clearer than by the definition of conditional correlation.
8. The conditional correlations matrices graphs on X_0 for random dataset show a little bit different behaviour in conditional correlations under four conditions depending on the time moment before transition.
9. The conditional correlations matrices graphs on X_0, X_1, X_2 and X_3 for original dataset show very different behaviour in conditional correlations under four conditions depending on the time moment before transition.
10. Generally, the conditional correlations matrices graphs on X_1, X_2 and X_3 for original dataset by the definition of the conditional correlation and D-vine are quite similar to the conditional correlations matrices graphs on X_0 under four conditions.

4.4 Conditional Correlations on Middle Points in Time

From the last section, we studied the conditional correlation of sin(psi) under the condition of hydrogen density on four grid points in space: X_0, X_1, X_2 and X_3 by the definition of the conditional correlation and D-vine. We found that the conditional correlation of sin(psi) under the hydrogen density on X_1 which is an opposite grid point of X_0 and X_2 and X_3 which are the neighbor grid point of X_0 are quite similar to the conditional correlation of sin(psi) under the hydrogen density on X_0. In this section, our interests are to study and investigate how the correlation of sin(psi) change in behaviour under the condition of sin(psi) on middle points in delays time. We study and calculate the correlations of dihedral angles: sin(psi) at different delay times before transition, 50.1, 47.3, ..., 0.1, 0 under the condition of dihedral angles: sin(psi) on the middle points in delay times by the defitnition of the conditional correlation (2.12) and D-vine.

Let $\sin(\psi_i)$ and $\sin(\psi_j)$ be the sine of ψ_i and ψ_j at delay time i and j where $i = j$ = 50.1, 47.3, ..., 0.1, 0. Let $\sin(\psi_{(i+j)/2})$ be the sine of $\psi_{(i+j)/2}$ at delay time $(i+j)/2$ where $i = j$ = 50.1, 47.3, ..., 0.1, 0.

Therefore, the conditional correlation of $\sin(\psi_i)$ and $\sin(\psi_j)$ given $\sin(\psi_{(i+j)/2})$ is denoted by $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ where $i = j$ = 50.1, 47.3, ..., 0.1, 0. From the histogram of $\sin(\psi)$ density both original and random dataset as Figure A.2 and A.7, we set the condition of $\sin(\psi_{(i+j)/2})$ following.

- **Condition 1:** -1.0 $\leq \sin(\psi_{(i+j)/2}) < 0.4$
- **Condition 2:** $0.4 \leq \sin(\psi_{(i+j)/2}) < 0.6$
- **Condition 3:** $0.6 \leq \sin(\psi_{(i+j)/2}) < 0.8$
- **Condition 4:** $0.8 \leq \sin(\psi_{(i+j)/2}) < 1.0$

For the conditional correlation of $\sin(\psi_i)$ and $\sin(\psi_j)$ given $\sin(\psi_{(i+j)/2})$ or $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$, we calculate Pearson, Spearman and Kendall correlations by the definition and by D-vine, we calculate only Kendall correlation. The results of the conditional correlations on middle points in time are given in Figure 4.34 - 4.41.

Regarding the results on the statistical analysis.

1. **The definition of the conditional correlation**

   - **Condition 1:** -1.0 $\leq \sin(\psi_{(i+j)/2}) < 0.4$
     (Figure 4.34 (a), 4.35 (a), 4.36 (a), 4.38 (a), 4.39 (a), 4.40 (a))
   - For the two-point conditional correlations of original dataset, Pearson, Spearman and Kendall correlations of $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points are different.
   - $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points of random dataset, Pearson and Spearman correlations are similar, as Kendall correlation is different.

- **Condition 2:** $0.4 \leq \sin(\psi_{(i+j)/2}) < 0.6$
(Figure 4.34 (b), 4.35 (b), 4.36 (b), 4.38 (b), 4.39 (b), 4.40 (b))
- For the two-point conditional correlations of original dataset, Pearson and Spearman correlations of $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points are quite similar, as Kendall correlation is different.
- $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points of random dataset, Pearson and Spearman correlations are similar, as Kendall correlation is different.
- **Condition 3:** $0.6 \leq \sin(\psi_{(i+j)/2}) < 0.8$
(Figure 4.34 (c), 4.35 (c), 4.36 (c), 4.38 (c), 4.39 (c), 4.40 (c))
- For the two-point conditional correlations of original dataset, Pearson and Spearman correlations of $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points are quite similar, as Kendall correlation is different.
- $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points of random dataset, Pearson and Spearman correlations are similar, as Kendall correlation is different.
- **Condition 4:** $0.8 \leq \sin(\psi_{(i+j)/2}) < 1.0$
(Figure 4.34 (d), 4.35 (d), 4.36 (d), 4.38 (d), 4.39 (d), 4.40 (d))
- For the two-point conditional correlations of original dataset, Pearson and Spearman correlations of $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points are quite similar, as Kendall correlation is different.
- $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points of random dataset, Pearson and Spearman correlations are similar, as Kendall correlation is different.

2. **D-vine**

- **Condition 1:** $-1.0 \leq \sin(\psi_{(i+j)/2}) < 0.4$ (Figure 4.37 (a), 4.41 (a))
- For the two-point conditional correlations of original dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points is quite similar to the conditional correlations by the definition.
- $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points of random dataset, Kendall correlation is similar to the conditional correlations by the definition.
- **Condition 2:** $0.4 \leq \sin(\psi_{(i+j)/2}) < 0.6$ (Figure 4.37 (b), 4.41 (b))
- For the two-point conditional correlations of original dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points is similar to the conditional correlations by the definition.
- $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points of random dataset, Kendall correlation is similar to the conditional correlations by the definition.
- **Condition 3:** $0.6 \leq \sin(\psi_{(i+j)/2}) < 0.8$ (Figure 4.37 (c), 4.41 (c))
- For the two-point conditional correlations of original dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points is similar to the conditional correlations by the definition.
- $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points of random dataset, Kendall correlation is similar to the conditional correlations by the definition.

- **Condition 4:** $0.8 \leq \sin(\psi_{(i+j)/2}) < 1.0$ (Figure 4.37 (d), 4.41 (d))

- For the two-point conditional correlations of original dataset, Kendall correlation of $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points is similar to the conditional correlations by the definition.
- $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points of random dataset, Kendall correlation is similar to the conditional correlations by the definition.

Obviously, the conditional correlations $\rho_{\psi_i\psi_j|\psi_{(i+j)/2}}$ on middle points by D-vine still explain or show the conditional correlations matrix graphs both original and random dataset clearer than by the definition.

Regarding the molecular meaning of the statistical results.

1. The matrices graphs of conditional correlations show the complete picture of statistical dependence of psi(phi) at different time moments with respect to the transition moment under psi(phi) on middle points. For example, the row starting at 50.1 shows how much the value of psi(phi) at 50.1 picoseconds before the transition depends on all the values of psi (phi) at previous time moments under each condition of psi(phi) on middle points.

2. The "random" matrices graphs show a little bit different behaviour in conditional correlations under four conditions depending on the time moment before transition:

   - **Condition 1:** -1.0 $\leq \sin(\psi_{(i+j)/2}) < 0.4$ (Figure 4.38 (a) - 4.41 (a))
   - The conditional correlations matrices by definition and D-vine show that there is less correlation (both positive and negative) in conditional dependencies at all starting times.
   - **Condition 2:** $0.4 \leq \sin(\psi_{(i+j)/2}) < 0.6$ (Figure 4.38 (b) - 4.41 (b))
   - The conditional correlations matrices by definition and D-vine show that there is less correlation (both positive and negative) in conditional dependencies at all starting times.
   - **Condition 3:** $0.6 \leq \sin(\psi_{(i+j)/2}) < 0.8$ (Figure 4.38 (c) - 4.41 (c))
   - The conditional correlations matrices by definition and D-vine show that there is less correlation (both positive and negative) in conditional dependencies at all starting times. The conditional correlations matrices show there is negative correlation at 50.1 - 5 ps delays.
   - **Condition 4:** $0.8 \leq \sin(\psi_{(i+j)/2}) < 1.0$ (Figure 4.38 (d) - 4.41 (d))
   - The conditional correlations matrices by definition and D-vine show that there is less correlation (both positive and negative) in conditional dependencies at all starting times. The conditional correlations matrices show there is negative correlation at 50.1 - 7.9 ps delays.

3. The "original" matrices graphs show a bit different behaviour in conditional correlations under four conditions depending on the time moment before transition:

   - **Condition 1:** -1.0 $\leq \sin(\psi_{(i+j)/2}) <$ 0.4 (Figure 4.34 (a) - 4.37 (a))
   - The conditional correlations matrices by definition and D-vine show that there is less correlation (both positive and negative) in conditional dependencies at all starting times.
   - Starting from 5 delays and up to 1.9 ps, the conditional correlations are much stronger and longer. The row at 2.5 ps, for example, has very strongly correlated values of psi up to 1.9 ps in advance (the correlation coefficient is 0.4 - 0.5).
   - There is negative correlation with the data at 1.5 - 50.1 ps delays (correlation value is -0.2 - -0.1).
   - **Condition 2:** 0.4 $\leq \sin(\psi_{(i+j)/2}) <$ 0.6 (Figure 4.34 (b) - 4.37 (b))
   - The conditional correlations matrices by definition and D-vine show that there is less correlation (both positive and negative) in conditional dependencies at all starting times.
   - There is negative correlation with the data at 1.5 - 10 ps delays (correlation value is -0.3 - -0.1).
   - **Condition 3:** 0.6 $\leq \sin(\psi_{(i+j)/2}) <$ 0.8 (Figure 4.34 (c) - 4.37 (c))
   - The conditional correlations matrices by definition and D-vine show that there is less correlation (both positive and negative) in conditional dependencies at all starting times.
   - Starting from 6.3 delays and up to 2.5, the conditional correlations are much stronger and longer (the correlation coefficient is 0.4 - 0.5).
   - There is negative correlation with the data at 1.5 - 39.8 ps delays (correlation value is -0.5 - -0.3).
   - **Condition 4:** 0.8 $\leq \sin(\psi_{(i+j)/2}) <$ 1.0 (Figure 4.34 (d) - 4.37 (d))
   - The conditional correlations matrices by definition and D-vine show that there is less correlation (both positive and negative) in conditional dependencies at all starting times.
   - Starting from 12.5 delays and up to 3.9, the conditional correlations are much stronger and longer (the correlation coefficient is 0.4 - 0.5).
   - There is negative correlation with the data at 1.5 - 39.8 ps delays (correlation value is -0.5 - -0.3).

Summary: For the conditional correlation of sin(psi) under the condition of sin(psi) on middle points in time, we summarize that

1. **Regarding the results on the statistical analysis**
    - By the definition of the conditional correlation, Pearson and Spearman correlations gave quite similar conditional correlation matrices graphs both original and random dataset, as Kendall correlation gave different matrices graphs under the most conditions.
    - By D-vine, Kendall correlation gave conditional correlation matrices graphs similar to the conditional correlation by the definition both original and random dataset under four conditions.
2. **Regarding the molecular meaning of the statistical results**
    - For random dataset, the conditional correlation matrices graphs show a little bit different behaviour in conditional correlations under four conditions depending on the time moment before transition. Under four conditions, the conditional correlations matrices graphs by definition and D-vine show that there is less correlation (both positive and negative) in conditional dependencies at all starting times.
    - For original dataset, the conditional correlation matrices graphs show a bit different behaviour in conditional correlations under four conditions depending on the time moment before transition. Under four conditions, the conditional correlations matrices graphs by definition and D-vine show that there is less correlation (both positive and negative) in conditional dependencies at all starting times.

Conclusions: Conditional correlations on middle points in time under four conditions, we conclude that

1. The matrices graphs of conditional correlations show the complete picture of statistical dependence of sin(psi) at different time moments with respect to the transition moment under the condition of sin(psi) on middle points.

2. The conditional correlations matrices graphs on middle points in time by the definition of the conditional correlation and D-vine are quite similar under four conditions both original and random dataset.

3. The conditional correlations matrices graphs on middle points in time by D-vine show the behaviour in conditional correlations that are clearer than by the definition of conditional correlation. To clarify this, we calculate the differences of Kendall conditional correlations matrices graphs of sin(psi) on middle points by the definition of correlation and D-vine under four conditions for original dataset as illustrated in Figure 4.42. From Figure 4.42, we found that the large different behaviour in conditional correlations is in condition4 because there is less data at some delays time according to this condition, so we do not have enough data for the calculation of conditional correlation by the definition. For D-vine, the conditional correlation is determined by the bivariate copulas and a nested set of trees using pair-copula.

4. The conditional correlations matrices graphs on middle points in time both original and random dataset show small different behaviour in conditional correlations under four conditions depending on the time moment before transition.

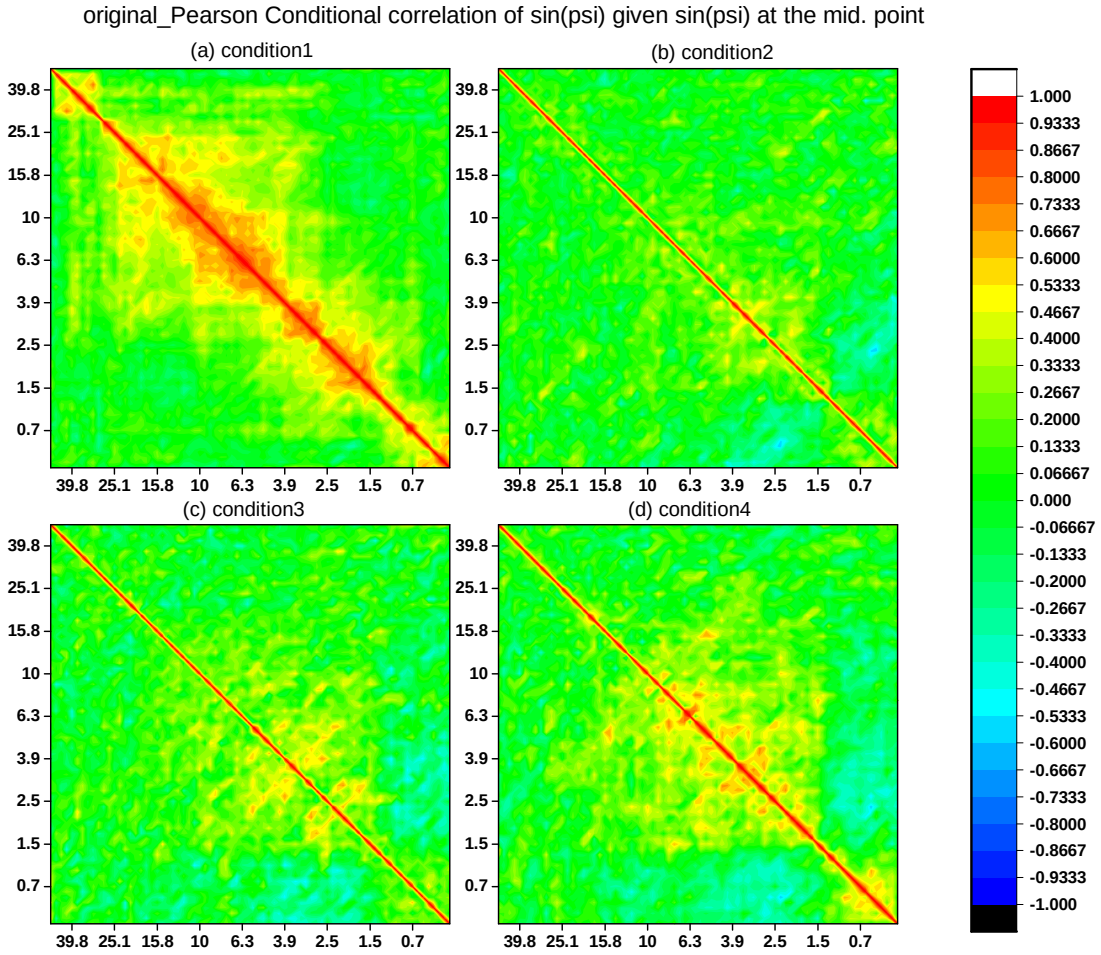

Figure 4.34: Pearson conditional correlation of sin(psi) on middle points: original dataset - The figure shows Pearson conditional correlation matrix graph of sin(psi) on middle points and delay time 50.1, 47.3, . . . , 0 ps before the transition for original dataset.

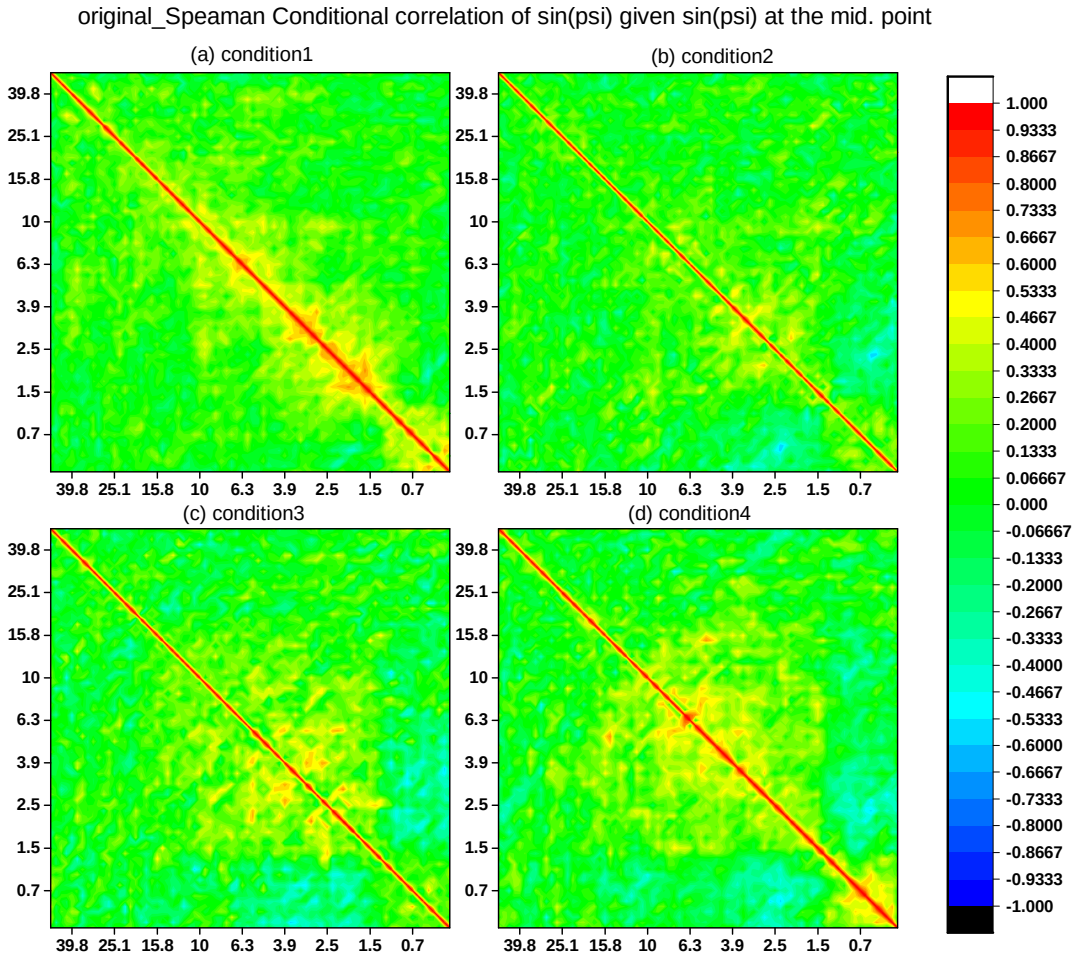

Figure 4.35: Spearman conditional correlation of sin(psi) on middle points: original dataset - The figure shows Spearman conditional correlation matrix graph of sin(psi) on middle points and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset.

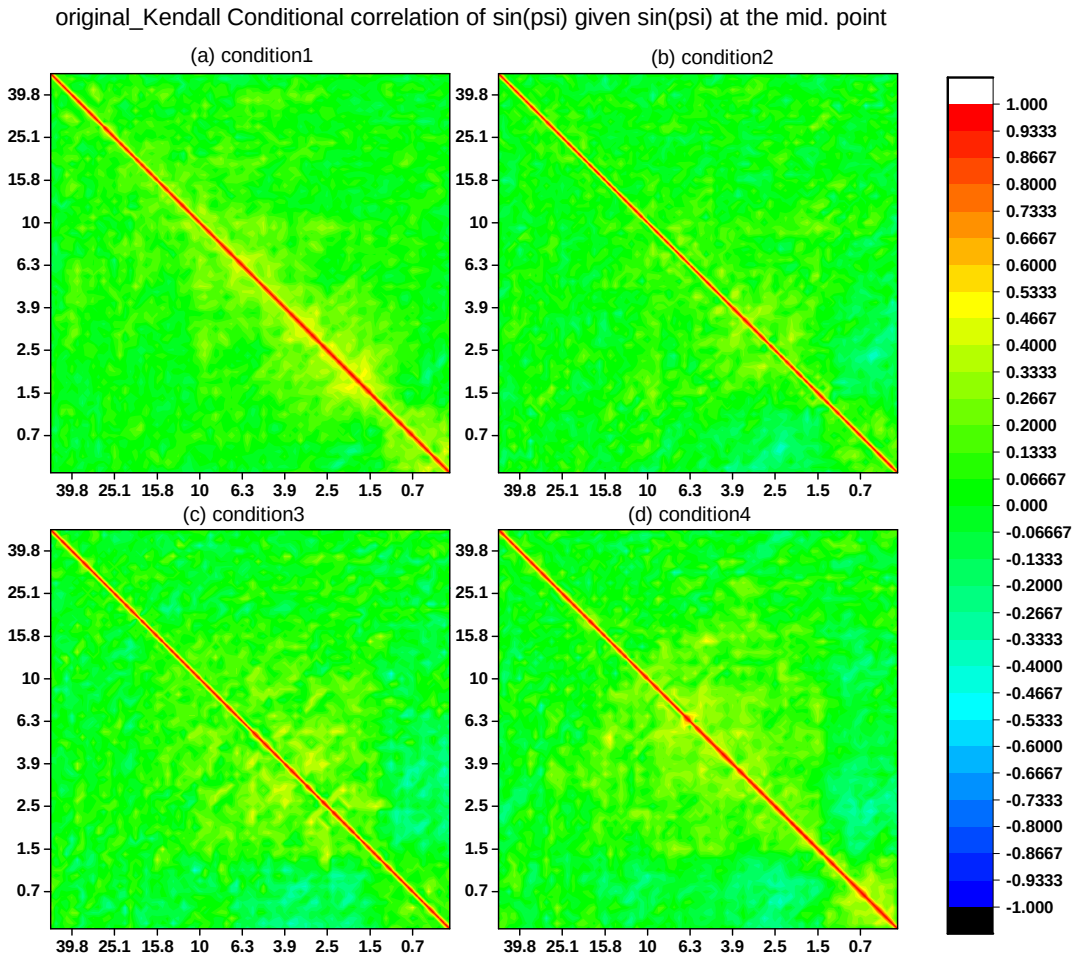

Figure 4.36: Kendall conditional correlation of sin(psi) on middle points: original dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on middle points and delay time 50.1, 47.3, . . . , 0 ps before the transition for original dataset.

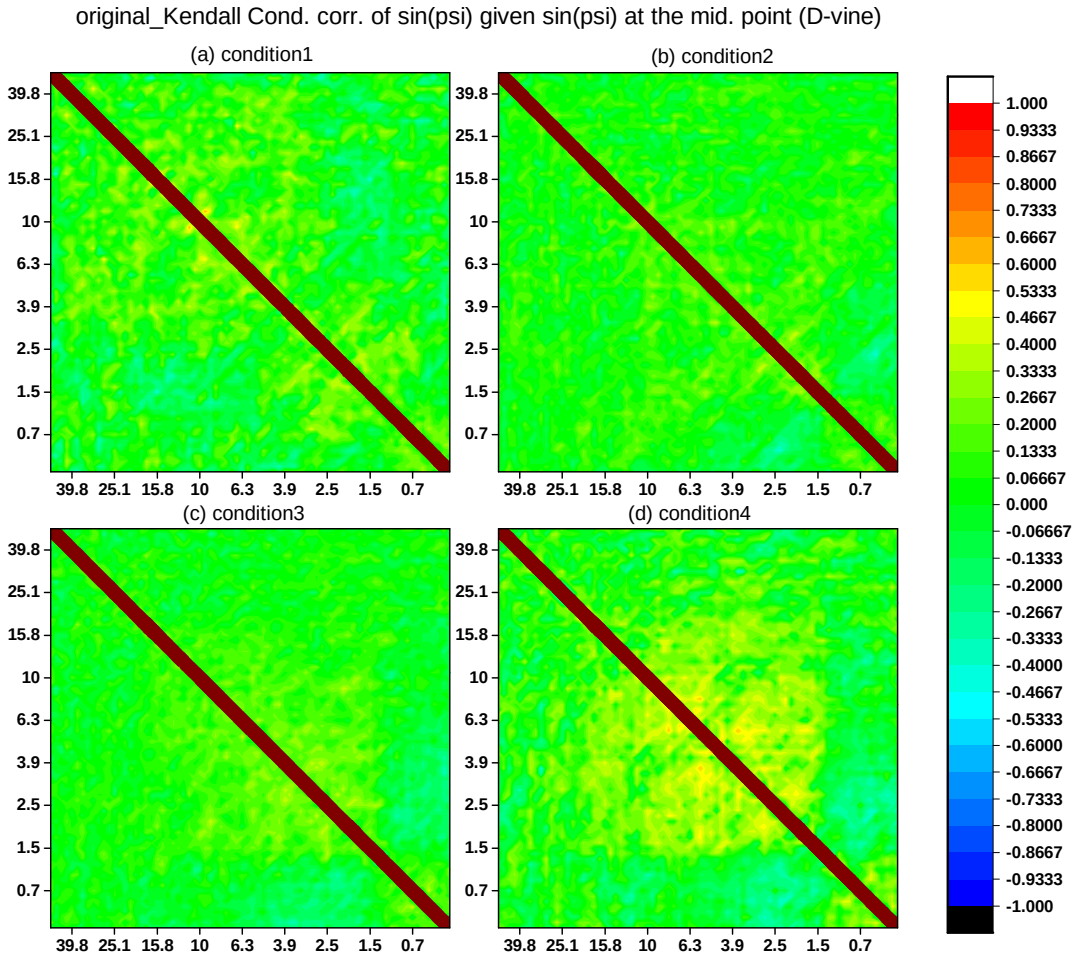

Figure 4.37: Kendall conditional correlation of sin(psi) on middle points by D-vine: original dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on middle points and delay time 50.1, 47.3, ..., 0 ps before the transition for original dataset by D-vine.

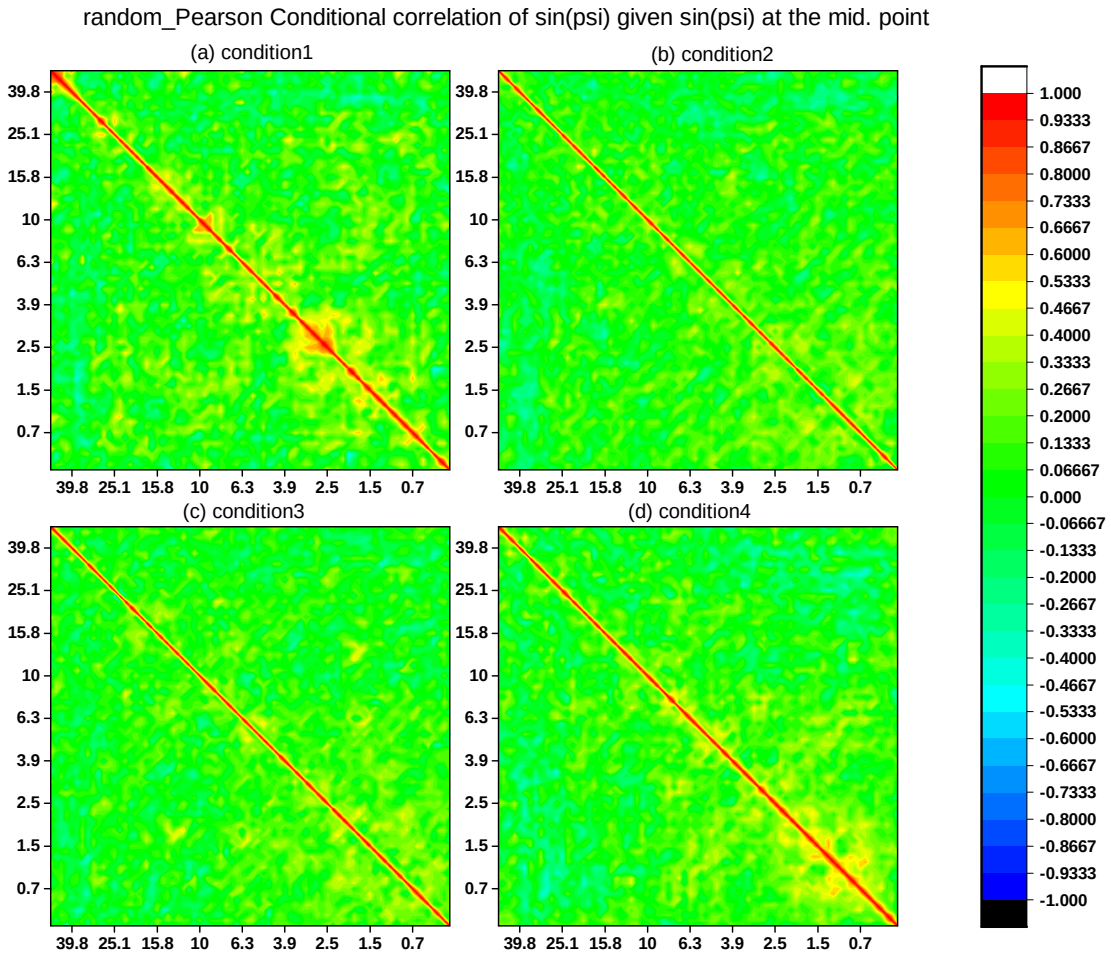

Figure 4.38: Pearson conditional correlation of sin(psi) on middle points: random dataset - The figure shows Pearson conditional correlation matrix graph of sin(psi) on middle points and delay time 50.1, 47.3, ..., 0 ps before the transition for random dataset.

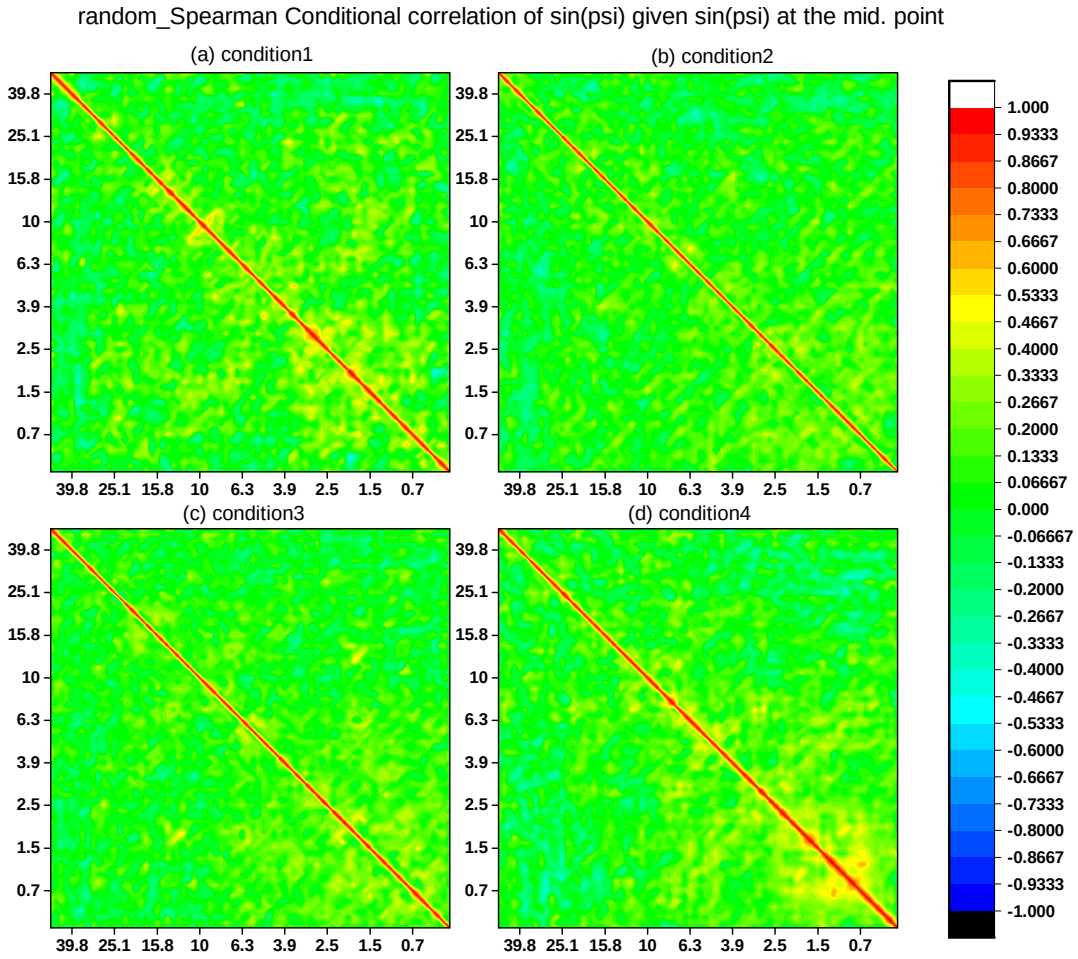

Figure 4.39: Spearman conditional correlation of sin(psi) on middle points: random dataset - The figure shows Spearman conditional correlation matrix graph of sin(psi) on middle points and delay time 50.1, 47.3, ..., 0 ps before the transition for random dataset.

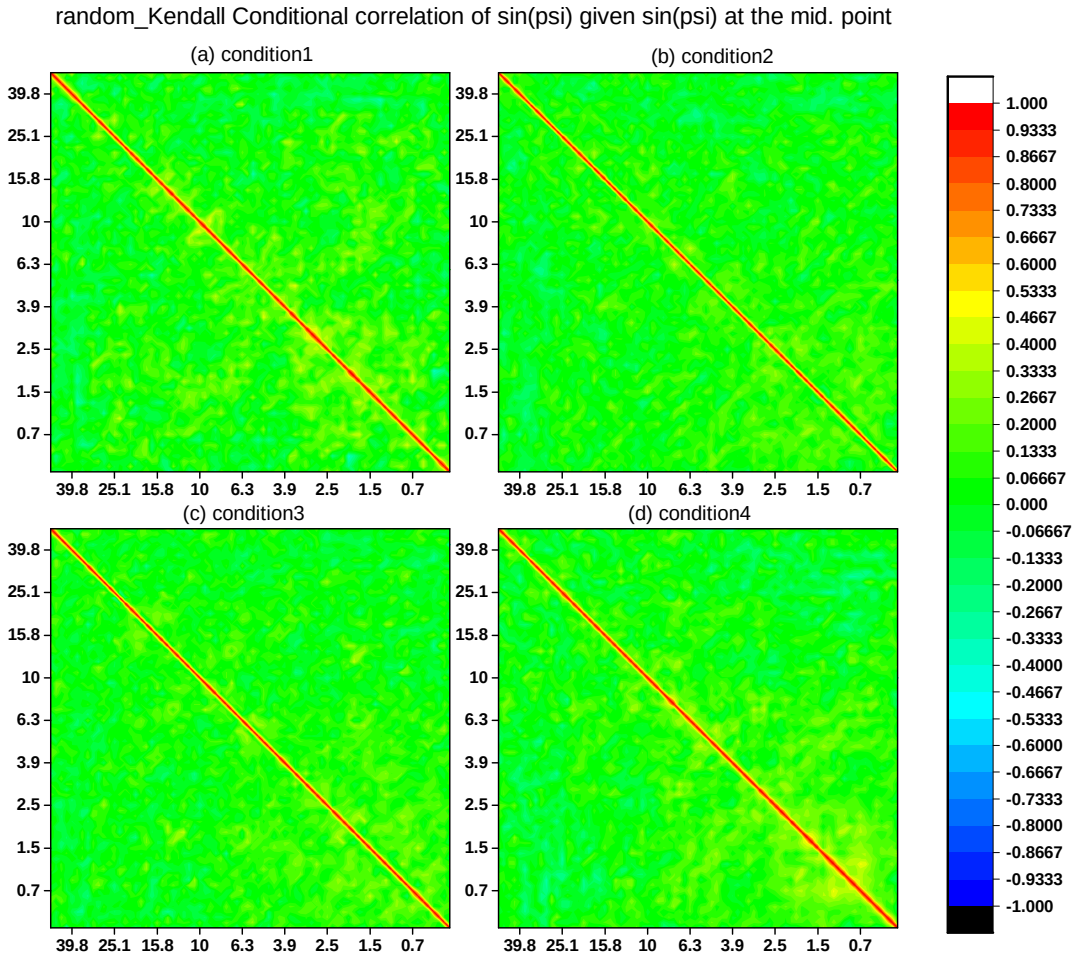

Figure 4.40: Kendall conditional correlation of sin(psi) on middle points: random dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on middle points and delay time 50.1, 47.3, ..., 0 ps before the transition for random dataset.

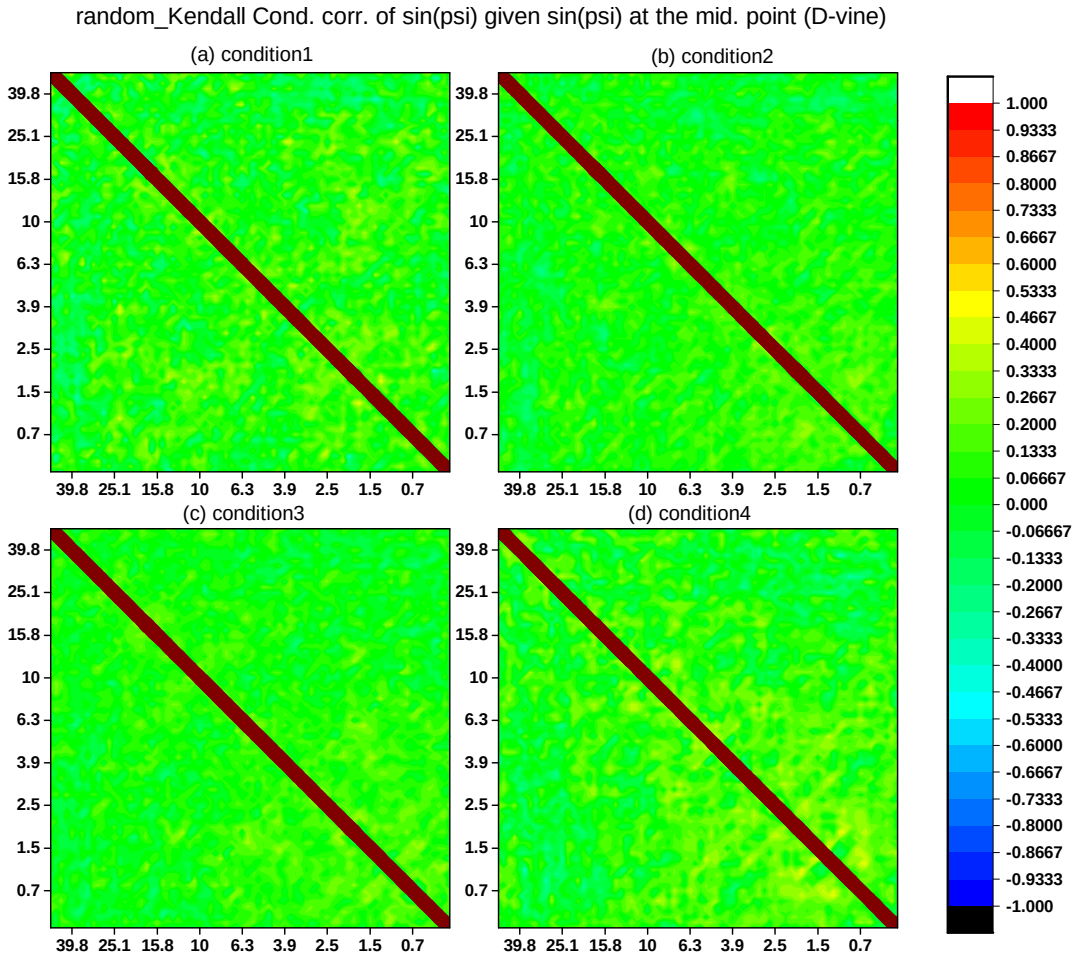

Figure 4.41: Kendall conditional correlation of sin(psi) on middle points by D-vine: random dataset - The figure shows Kendall conditional correlation matrix graph of sin(psi) on middle points and delay time 50.1, 47.3, ..., 0 ps before the transition for random dataset by D-vine.

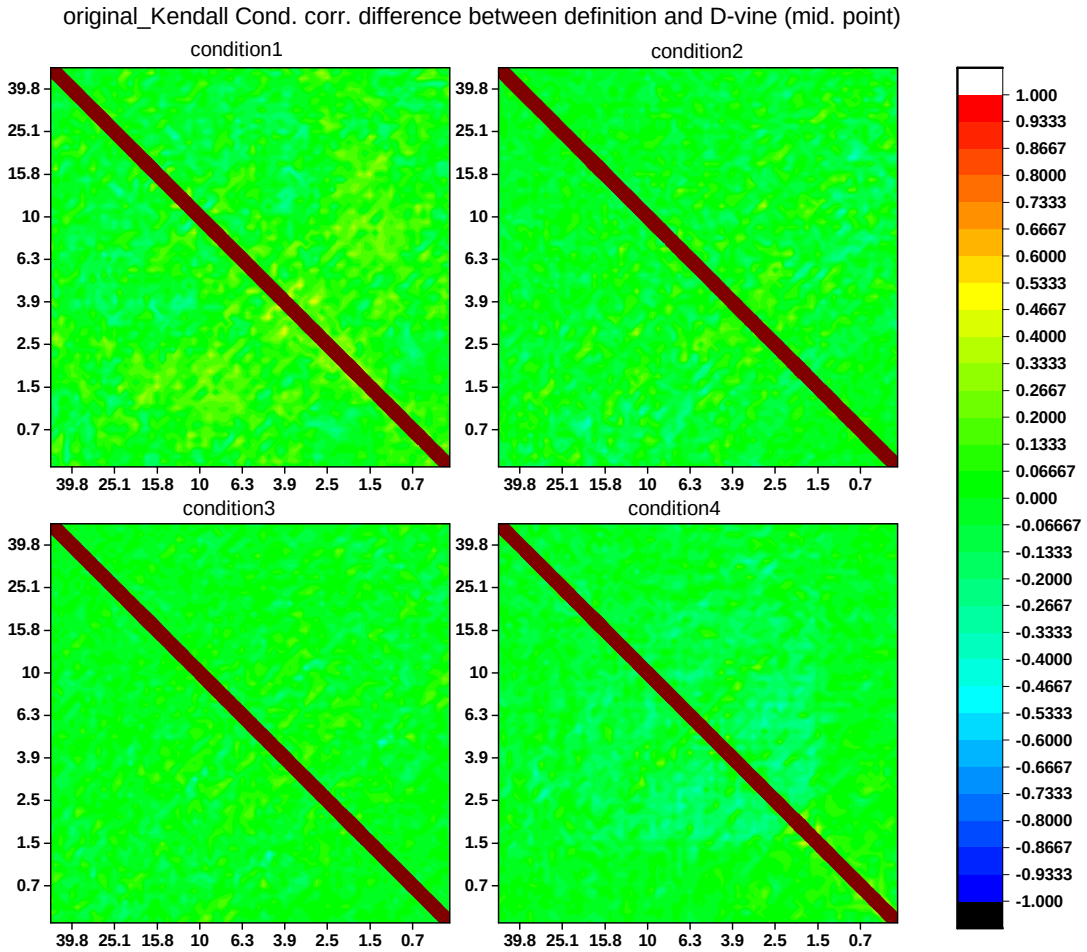

Figure 4.42: The differences of Kendall conditional correlation of sin(psi) on middle points by the definition of correlation and D-vine: original dataset - The figure shows the differences of Kendall conditional correlation matrix graph of sin(psi) on middle points and delay time 50.1, 47.3, . . . , 0 ps before the transition for original dataset by the definition of correlation and D-vine.

4.5 Grid Points

From the last two sections, we study the conditional correlations of dihedral angles: sin(psi) on some grid points of hydrogen density in space: X_0, X_1, X_2 and X_3. For this section, we focus on the density of hydrogen at 3D grid points that have 1845616 grid points in space for original dataset. Figure 4.43 shows one of protein molecule: dialanine surround with water atoms in space. We choose some grid points for analysis and the criteria for choosing the grid points is high level of hydrogen density at delay time 0, say, the hydrogen density is more than 0.2 (see Figure A.4). After, we get chosen grid points then we calculate the Kendall un- and conditional correlations of psi(phi) given the density of hydrogen on chosen grid points in space and compare both un- and conditional correlations by Fisher's transformation with large sample size ($n \geq 10$).

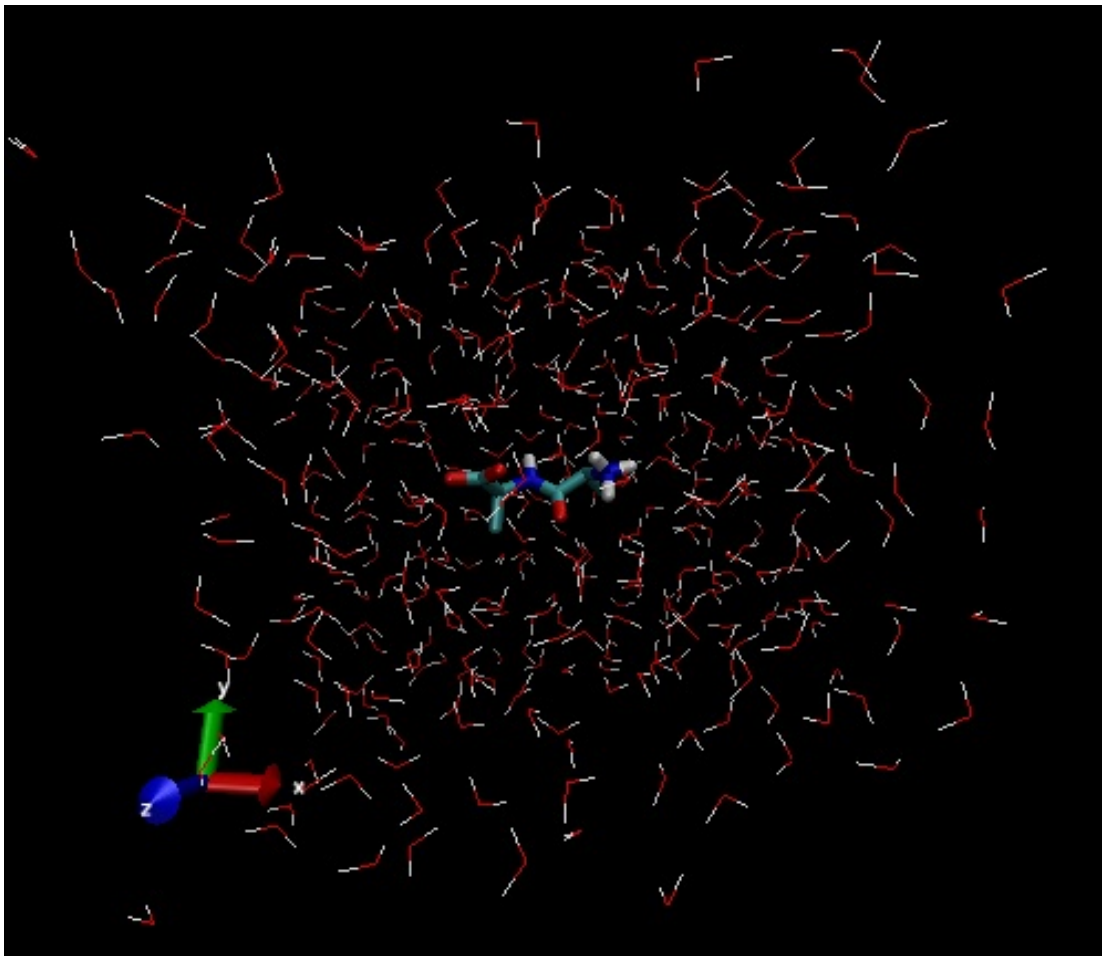

Figure 4.43: Dialanine molecule surround with water - The figure shows dialanine molecule surround with water atoms in space.

Let $\sin(\psi_i)$ and $\sin(\psi_j)$ be the sine of ψ_i and ψ_j at delay time i and j where $i = j$ = 50.1, 47.3, ..., 0.1, 0. Let $\delta_{0(k)}$ be the hydrogen density at delay time 0 on a chosen grid point k.

Therefore, the unconditional correlation of $\sin(\psi_i)$ and $\sin(\psi_j)$ is denoted by $\rho_{\psi_i\psi_j}$ and the conditional correlation of $\sin(\psi_i)$ and $\sin(\psi_j)$ given $\delta_{0(k)}$ is denoted by $\rho_{\psi_i\psi_j|\delta_{0(k)}}$ where $i = j$ = 50.1, 47.3, ..., 0.1, 0.

The procedures of calculation for this work are following.

1. To calculate the Kendall unconditional correlations of $\sin(\psi_i)$ and $\sin(\psi_j)$ or $\rho_{\psi_i\psi_j}$.
2. To calculate the Kendall conditional correlations of $\sin(\psi_i)$ and $\sin(\psi_j)$ given $\delta_{0(k)}$ or $\rho_{\psi_i\psi_j|\delta_{0(k)}}$ with the condition $\delta_{0(k)} > 0.2$ and $n \geq 10$.
3. Using Fisher's transformation for two correlations from 1 and 2,
   Let $r_1 = \rho_{\psi_i\psi_j}$ and $r_2 = \rho_{\psi_i\psi_j|\delta_{0(k)}}$, we transform r_1 and r_2 to $r_1^{'}$ and $r_2^{'}$ by
   $$r_i^{'} = (0.5)log_e\left[\frac{1 + r_i^{'}}{1 - r_i^{'}}\right].$$
4. To calculate the test statistics: z by
   $$z = \frac{r_1^{'} - r_2^{'}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}},$$
   where n_1 and n_2 are numbers of pairs of observations to calculate un- and conditional correlations, respectively $(n_1, n_2 \geq 10)$.
5. To test hypothesis about the equality of two correlations:
   $$\text{Ho: } \rho_1 = \rho_2 \quad \text{vs.} \quad \text{Ha: } \rho_1 \neq \rho_2$$
   where ρ_1 and ρ_2 are the un- and conditional correlations, respectively. The criteria to reject the null hypothesis Ho: $\rho_1 = \rho_2$ at significance level 0.05 or $\alpha = 0.05$ is $|z| > Z_{\alpha/2} = Z_{0.025} = 1.96 \approx 2.0$.

From the above procedures of calculation, they are the statistical hypothesis testing for comparing two correlations, for example, $\rho_{\psi_i\psi_j}$ and $\rho_{\psi_i\psi_j|\delta_{0(k)}}$ with the condition $\delta_{0(k)} > 0.2$ and $n \geq 10$ etc. The testing result can conclude that wheter there is the difference between two correlations. If the test result is significant, that means that two correlations are different. Normally, if there is difference between two correlations, we can do any further study, on the other hand, we do not need to do anything.

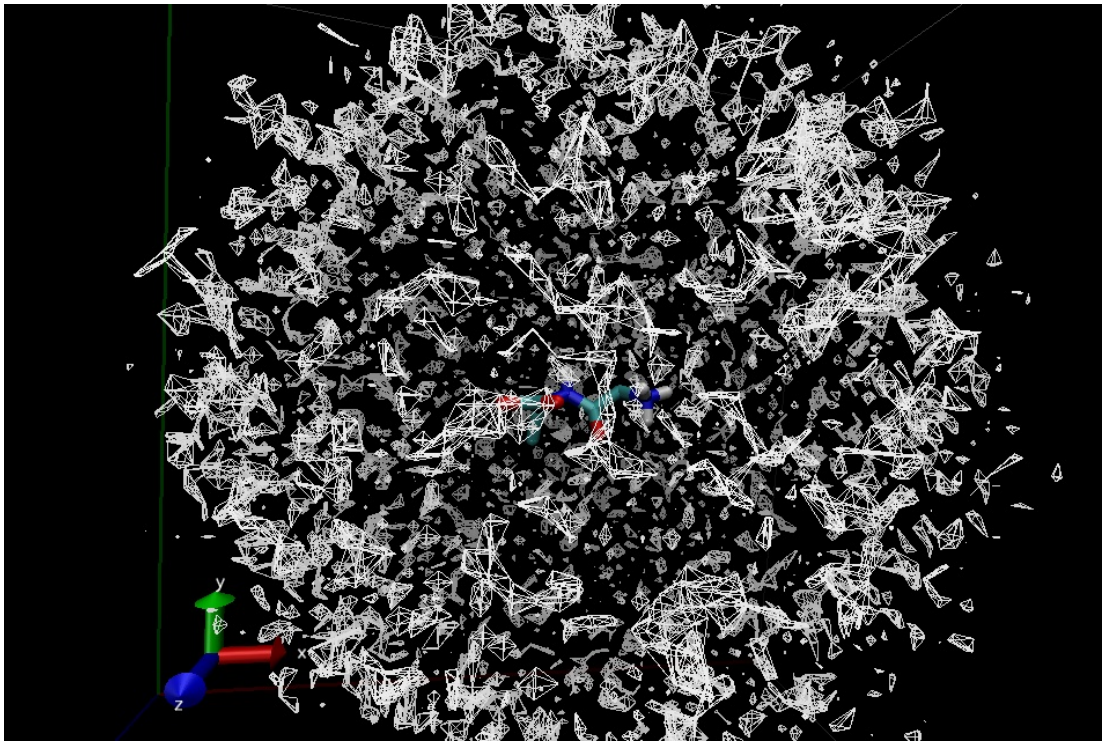

Figure 4.44: The maximum value of $|z|$ or max(abs(z)) - The figure shows the maximum value of $|z|$ or max(abs(z)) of all grid points in space when isovalue = 3.3754.

In this study, for each chosen grid point of the hydrogen density, we have 76×76 symmetric matrix of z values. We choose the maximum value of $|z|$ or max(abs(z)) to be a representative of all z values that quantifies the difference and we are interested to identify a pair of delay time that makes the largest difference. The one method to identify is checking the highest number of times that a pair have max(abs(z)). Figure 4.44 shows the maximum value of $|z|$ or max(abs(z)) of all grid points in space.

The results for hypothesis testing about the equality of two correlations: $\rho_{\psi_i\psi_j}$ and $\rho_{\psi_i\psi_j|\delta_{0(k)}}$ are following.

1. The range of the maximum value of $|z|$ or max(abs(z)) of all grid points in space is 0 - 6.47336.

2. In general, the hypothesis testing concludes that the unconditional correlation is different from conditional correlation on all grid points in space or $\rho_{\psi_i\psi_j} \neq \rho_{\psi_i\psi_j|\delta_{0(k)}}$.

3. The pair of delays time that makes the largest difference is (2.3, 2.5).

Remark. The maximum value of $|z|$ or max(abs(z)) is equal to 0 may come from:

- there is no density of hydrogen at that grid point, or
- the density of hydrogen does not satisfy the condition (the density is more than 0.2 and $n \geq 10$).

Conclusions: In this section, we extend to study the conditional correlation of sin(psi) under the condition of hydrogen density on some selected grid points to all grid points in space and compare between the un- and conditional correlations. We found that there is difference between the un- and conditional correlations. Also, we did a preliminary analysis to identify a pair of delay time that makes the largest difference from the highest number of times that a pair have the maximum value of absolute value of test statistic (max(abs(z))). This is an alternative study, for more detail study, it will be a future work.

5

Conclusions

5.1 Introduction

Clearly, the methods of statistics can apply to study molecular system dynamics. In this thesis, we particularly focus on statistical dependencies in (i) the dynamics of the protein and surrounding water molecules and (ii) collective long space- and time-range correlations in the molecular trajectories. Kendall's tau correlation is the well-known nonparametric correlation that is applied to measure the association of the dihedral angles of peptide and water atoms both un- and conditional correlations at delay time from 0 to 50.1 ps before transition. Moreover, we also apply the D-vine which is one of the ways to study the statistical correlations between variables describing molecular conformation of a peptide and the properties of water molecules surrounding the peptide. D-vine is a graphical tool to give a specific way of decomposing the probability density into bivariate copulas, so called pair-copulas. The dependency structure is determined by the bivariate copulas and a nested set of trees using pair-copulas. Main goal of the thesis is to study statistical properties of liquid protein-water molecular system dynamics.

Test systems were carried out to study copulas and D-vine for liquid protein-water molecular system dynamics and investigate the statistical dependencies of the dynamics. Angle and amplitude dataset with dimension 2 and 5 were analysed by N-variate copulas, as D-vine was applied for both datasets with dimension 5 and 10. Copulas analysis for two and five dimensions were performed by taking the same sample sizes both angle and amplitude datasets, say, 10000. D-vine analysis for five and ten dimensions were performed by taking a variety of sample sizes (n) both angle and amplitude datasets, for example, n = 10000, 50000. Furthermore, we skipped the observations to investigate the differences of results, for example interval = 0, 1, 2 and we also carried out both analytical and graphical analysis.

The main results on molecular system were presented in four sections:

- **Time correlations**

  We carried on a statistical analysis of dihedral angles motion of a dialanine molecule with explicit water. We calculated three correlation coefficients: Pearson, Spearman and Kendall of the dihedral angles: psi, phi, sine of dihedral angles: sin(psi), sin(phi) and water atoms: hydrogen and oxygen at the moment of transition and several moments in advance of the transition between 0 and 50.1 ps (Figure 4.10 - **??**).

- **Conditional correlations on X_0, X_1, X_2 and X_3**

  We studied and calculated the correlations of dihedral angles: psi at different delay times before transition, 50.1, 47.3, ..., 0.1, 0 under four conditions of water atoms: hydrogen density on four grid points in space: X_0, X_1, X_2 and X_3 by the definition of the conditional correlation and D-vine (Figure 4.20 - 4.25, 4.26 - 4.33, Table 4.4 - 4.9).

- **Conditional correlations on middle points in time**

  We studied and calculated the correlations of dihedral angles: psi at different delay times before transition, 50.1, 47.3, ..., 0.1, 0 under the condition of dihedral angles: psi on the middle points in delay times by the definition of the conditional correlation and D-vine (Figure 4.34 - 4.41).

- **Grid points**

  We studied and focused on the density of hydrogen at 3D grid points in space. Some grid points were chosen by the high level of hydrogen density at delay time 0, say, the hydrogen density is more than 0.2. We calculated the Kendall un- and conditional correlations of psi(phi) given the density of hydrogen on chosen grid points in space and compare both un- and conditional correlations by Fisher's transformation with large sample size ($n \geq 10$) (Figure 4.43, 4.44).

5.2 Conclusions

For this thesis, the simulated trajectories were provided from an external source, and we have done the statistical analysis of them. In Chapter 3, two datasets: the angle and amplitude came from the process of converting the continuous atomic velocity (coordinate) of the oxygen and hydrogen atoms of one of the water molecules that was used as a 3-dimensional signal. At the locations where the velocity pierces the xy plane, the points of a 2-dimensional map were generated and used as the original continuous signal for analysis. The symmetry of the 2-dimensional set of points can be further illustrated by tranforming the data to the polar coordinates (x,y) $\longrightarrow$ (ρ,ϕ) as illustrated in Figure 3.1. Both datasets, we exhibited the copula and D-vine analysis through test systems.

1. **N-variate Copulas**

   - **Angle dataset**
     For two and five dimensions (variables), the copulas results were similar. We concluded that all copulas parameters from fitted copulas family are statistically significant. That is X_1 and X_2 (two dimensions) and X_1, X_2, X_3, X_4 and X_5 (five dimensions) are positive correlated with a little value of copulas parameters (Table 3.2, 3.4).
   - **Amplitude dataset**
     For two and five dimensions (variables), the copulas results were similar. We concluded that all copulas parameters from fitted copulas family are statistically significant. That is X_1 and X_2 (two dimensions) and X_1, X_2, X_3, X_4 and X_5 (five dimensions) are positive correlated with a little value of copulas parameters (Table 3.6, 3.8).

2. **D-vine**

   - **Angle dataset**
     For five dimensions, the D-vine results which has 4 trees and 10 edges were different when we took the observations without or with interval. When we took the observations with any interval and large sample size, the D-vine results were similar and at the high level of tree, the best fit family was Gaussian (Table 3.9 - 3.13).

     For ten dimensions, D-vine results which has 9 trees and 45 edges were different when we took the observations with interval same as in five dimensions. When we took the observations with high values interval, for example interval = 3 and 5 etc. and n = 10000 and 50000, the D-vine results were quite similar and at the high level of tree, the best fit family was Gaussian (Table 3.19 - 3.22).
   - **Amplitude dataset**
     For five dimensions, the D-vine results were different when we took the observations without or with interval. When we took the observations with any interval and large sample size, the D-vine results were quite similar for each tree and edge (Table 3.16 - 3.25).

     For ten dimensions, D-vine results were different when we took the observations with any interval (Table 3.33 - 3.22).

In Chapter 4, two datasets: original and random were provided using classical molecular dynamics (MD) simulation with explicit water molecules. Each dataset consist of dihedral angles of peptide: psi and phi and the density of water atoms (hydrogen and oxygen) at 3D grid points for 76 different delay times between 0 and 50.1 picoseconds (ps) before the transition.

1. **Time correlations**

   We calculated three measures of correlation and plotted the correlation matrices graphs for dihedral angles: psi, sin(psi), phi, sin(phi) and water atoms: hydrogen and oxygen density at the maximum probability point (X_0) both original and random dataset. The graphs of all variables showed the complete picture of statistical dependence of dihedral angles and water atoms at different time moments with respect to the transition moment, the graphs also showed the different behaviour between random and original dataset. From the matrices graphs of correlations, we found that Pearson and Spearman correlations gave quite similar results, as Kendall correlation was different. Clearly, sin(psi) showed the big differences of correlation among three measures of dependence, as hydrogen and oxygen density did not show too much differences (Figure 4.10 - 4.18).

2. **Conditional correlations on X_0, X_1, X_2 and X_3**

   The conditional correlations on X_0, X_1, X_2 and X_3 were calculated by the definition of the conditional correlation and D-vine. Both methods gave quite similar correlation matrices but the conditional correlations matrices graphs from D-vine were clearer than the definition. When the condition of hydrogen density changed, there was very different behaviour in correlations depending on the time moment before transition. Generally, the conditional correlations on X_1, X_2 and X_3 for original dataset by the definition of the conditional correlation and D-vine were quite similar to the conditional correlations on X_0 under the conditions (Figure 4.20 - 4.25, 4.26 - 4.33, Table 4.4 - 4.9).

3. **Conditional correlations on middle points in time**

   The conditional correlations on middle points in time were also calculated by the definition of the conditional correlation and D-vine. Both methods still gave similar correlation matrices but the conditional correlations matrices graphs from D-vine were clearer than the definition. When the condition of psi changed, there was small difference behaviour in correlations depending on the time moment before transition. The conditional correlations matrices graphs on middle points in time both original and random dataset showed small difference behaviour in conditional correlations under four conditions depending on the time moment before transition (Figure 4.34 - 4.41).

4. **Grid points**

   We calculated the Kendall un- and conditional correlations of dihedral angles: sin(psi) given the density of hydrogen at delay time 0 (the hydrogen density is more than 0.2) on chosen 3D grid points in space for original dataset and compared un- and conditional correlations by Fisher's transformation with large sample size ($n \geq 10$). We found that there was difference between un- and conditional correlation, then we did the preliminary study to quantify the difference. In addition to identify a pair of delay time that makes the largest difference, we chose a pair of delay time that has the highest number of times of max(abs(z)).

5.3 Novelty

From thesis objectives in Chapter 1, we have achieved all thesis objectives:

1. to study time correlations of liquid protein-water molecular system dynamics.

   For objective 1, we did the correlation analysis for dihedral angles and water atoms and make the correlation matrices graphs to show the picture of statistical dependence of dihedral angles and water atoms at different time moments with respect to the transition moment.

2. to study conditional correlations of liquid protein-water molecular system dynamics on selected points.

   For objective 2, the study of conditional correlations of liquid protein-water molecular system dynamics on selected points was classified in two groups:

   (a) We studied and calculated the conditional correlations of dihedral angles: sin(psi) under the condition of hydrogen density on selected points in space: X_0, X_1, X_2 and X_3 by the definition of the conditional correlation and D-vine. We got the results showed the different behaviour of dihedral angles when the conditions changed.

   (b) We studied and calculated the conditional correlations of dihedral angles: sin(psi) under the condition of itself on middle points in time by the definition of the conditional correlation and D-vine. We got the results showed the different behaviour of dihedral angles when the conditions changed.

3. to study conditional correlations of liquid protein-water molecular system dynamics on grid points.

   For objective 3, we extended the study of conditional correlations of liquid protein-water molecular system dynamics on selected points to all grid points in space. Therefore, we calculated the Kendall un- and conditional correlations of dihedral angles: sin(psi) under the condition of hydrogen density on all grid points by the definition of the conditional correlation.

4. to compare un- and conditional correlations of liquid protein-water molecular system dynamics.

   For objective 4, we compared the un- and conditional correlations of liquid protein-water molecular system dynamics on all grid points in space by Fisher's transformation with large sample size ($n \geq 10$).

Regarding the molecular meaning of the statistical results for time correlations both original (Figure 4.10, 4.12, 4.14) and random (Figure 4.11, 4.13, 4.15) dataset.

(a) The most obvious is that random and original dataset are very different because the process of obtaining dataset between random and original is different. Clearly, the correlation matrices of all variables show the different colour between random and original dataset, that means that the strength of association of variable at different delay time is different.

(b) The matrices graphs of correlations show the complete picture of statistical dependence of psi(phi) at different time moments with respect to the transition moment. For example, the row starting at 50.1 shows how much the value of psi(phi) at 50.1 picoseconds before the transition depends on all the values of psi (phi) at previous time moments. While, for example, the row starting at 1.5 shows the dependence of the psi(phi) value at 1.5 picoseconds before the transition on all the values of psi (phi) at previous time moments.

(c) The "random" matrices graphs show that there is no difference in dependencies at all starting times: the sequence of correlations at 0 delay is the same as the sequence at 1.5 ps or 10 ps delays. That means that there is no change in behaviour and the psi(phi) values fluctuate in the same regime.

(d) The "original" matrices graphs show very different behaviour in correlations depending on the time moment before transition:

- In advance of the transition (rows starting at 50.1 - 10 delays) the correlations are essentially the same as for the "random" dataset.
- Starting from 10 delays and up to 1.5 ps, the correlations are much stronger and longer. The row at 1.7 ps, for example, has very strongly correlated values of psi up to 2 - 3 ps in advance (the correlation coefficient is 0.7 - 1).
- At 1.2 - 1.5 ps, surprisingly, the correlations are very low again, almost like in the "random" dataset.
- Before the transition, 0 - 1.1 ps the correlations are stronger than usual again, lasting for ≈0.5 ps, in contrast to ≈0.2 ps in "random" dataset. There is also anti-correlation (negative corr.:blue colour) with the data at 2 - 7 ps delays (correlation value is -0.5 - -0.4).

(e) What molecular picture can be deduced from this? First of all, the peptide (we consider sin(psi)) starts "feeling" the upcoming transition at ≈10 ps in advance of the actual transition as illustrated in Figure 4.10 and 4.12. Then, from 10 ps to ≈1.5 ps before the transition, the peptide follows more or less the same conformational trajectory every time it approaches the transition: strong correlations at these times testify that. Then, at ≈1.5 ps before the transition, the trajectory jumps to a set of very different conformations: the shape of the peptide has very little connection to what is at times ≈1.8 ps

and earlier. The final approach to the transition happens along the same route again: even though the peptide starts from different shapes at 1.5 ps delay moment, it then follows the same trajectory until it reaches the next conformational state. The anti-correlation with ≈2 - 7 ps times tells us that the psi angle rotated to approximately opposite value: a half-circle rotation has happened.

To clarify that, Figure 4.19 shows Spearman correlation graph of sin(psi) on X_0 at delay time 1.5, 2.1, 7 and 10 ps before the transition for origial dataset:

- At delay time 1.5 ps before the transition, the pair-correlations from 50.1 ps to 1.5 ps before the transition tend to be steadily increasing. The lowest and highest of pair-correlation are at 47.3 ps and 1.6 ps in advance of transition with the coefficient -0.12287 and 0.82693.
- At delay time 2.1 ps before the transition, the pair-correlations from 50.1 ps to 2.1 ps before the transition tend to be steadily increasing same as at delay time 1.5 ps. The lowest and highest of pair-correlation are at 47.3 ps and 2.2 ps in advance of transition with the coefficient -0.06534 and 0.83385.
- At delay time 7 ps before the transition, the pair-correlations from 50.1 ps to 7 ps before the transition tend to be steadily increasing same as at delay time 1.5 ps and 2.1 ps but the pair-correlation coefficients are smaller. The lowest and highest of pair-correlation are at 39.8 ps and 7.4 ps in advance of transition with the coefficient -0.07659 and 0.61639.
- At delay time 10 ps before the transition, the pair-correlations from 50.1 ps to 10 ps before the transition tend to be steadily increasing same as at delay time 7 ps but the pair-correlation coefficients are smaller. The lowest and highest of pair-correlation are at 25.1 ps and 10.5 ps in advance of transition with the coefficient -0.12476 and 0.51822.
- We hypothesise that the the conformational trajectory of peptide at any closed delay time before the transition is correlated, as at any far away delay time before the transition is un- or less correlated. Hence, the results conform to our hypothesis.

Appendix A

Histogram

A.1 Original Dataset

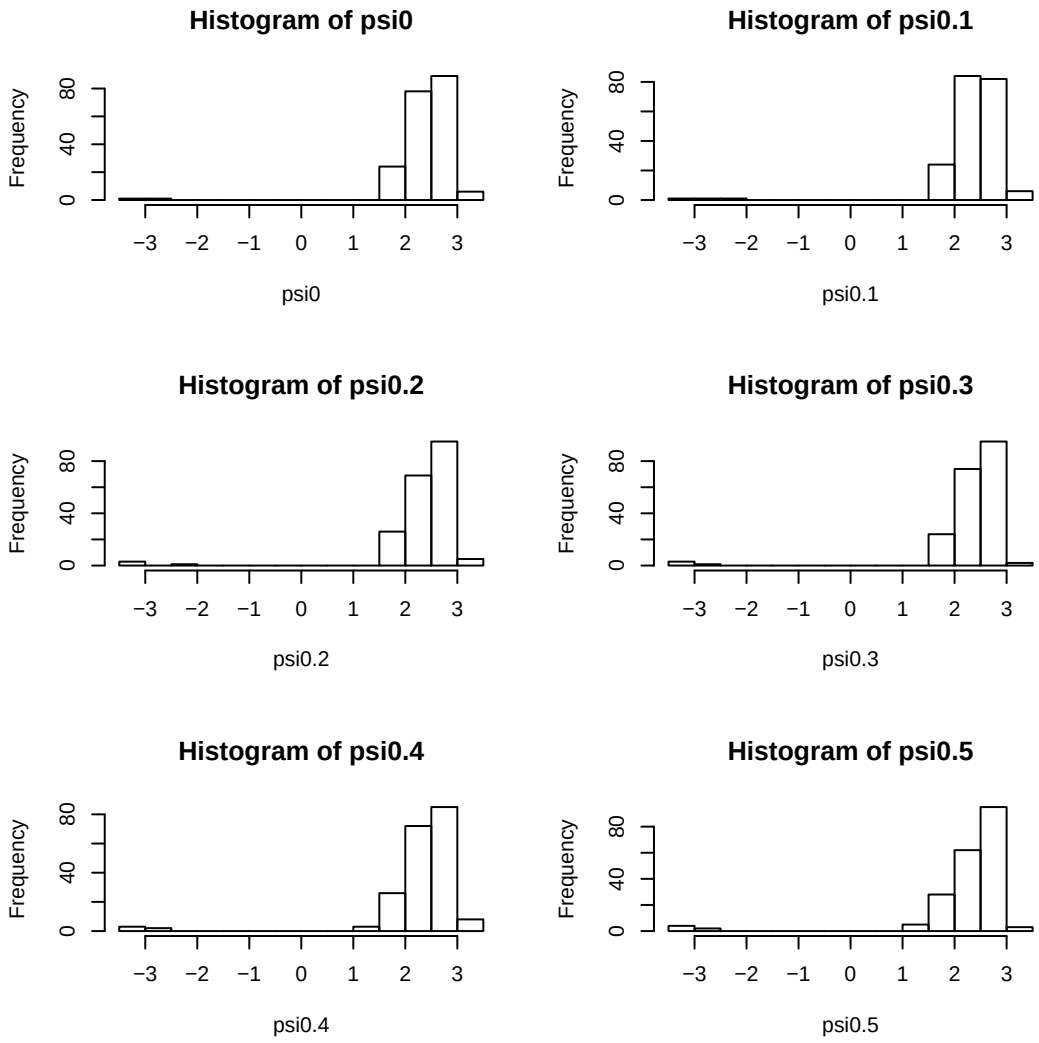

Figure A.1: Histogram of psi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows histograms of psi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for original dataset.

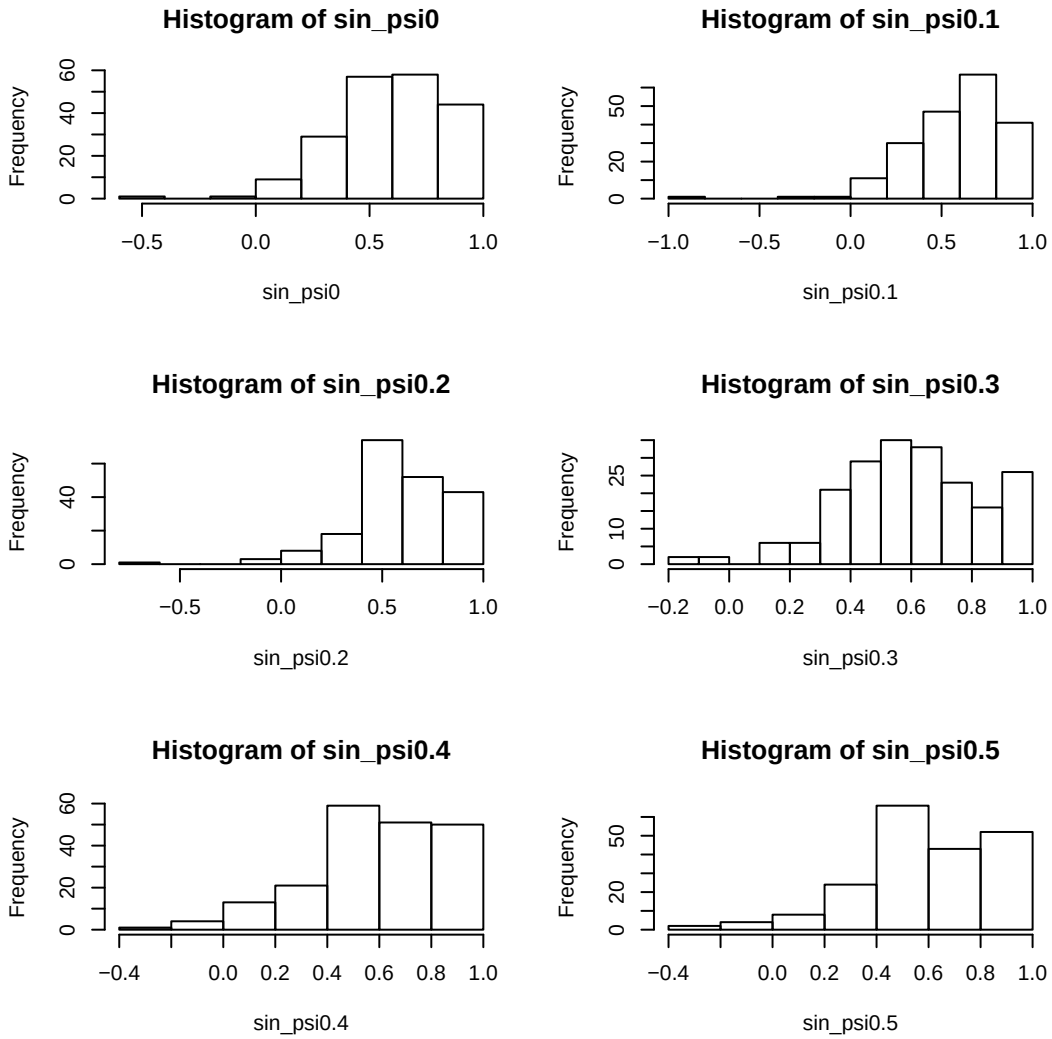

Figure A.2: Histogram of sin(psi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows histograms of sin(psi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for original dataset.

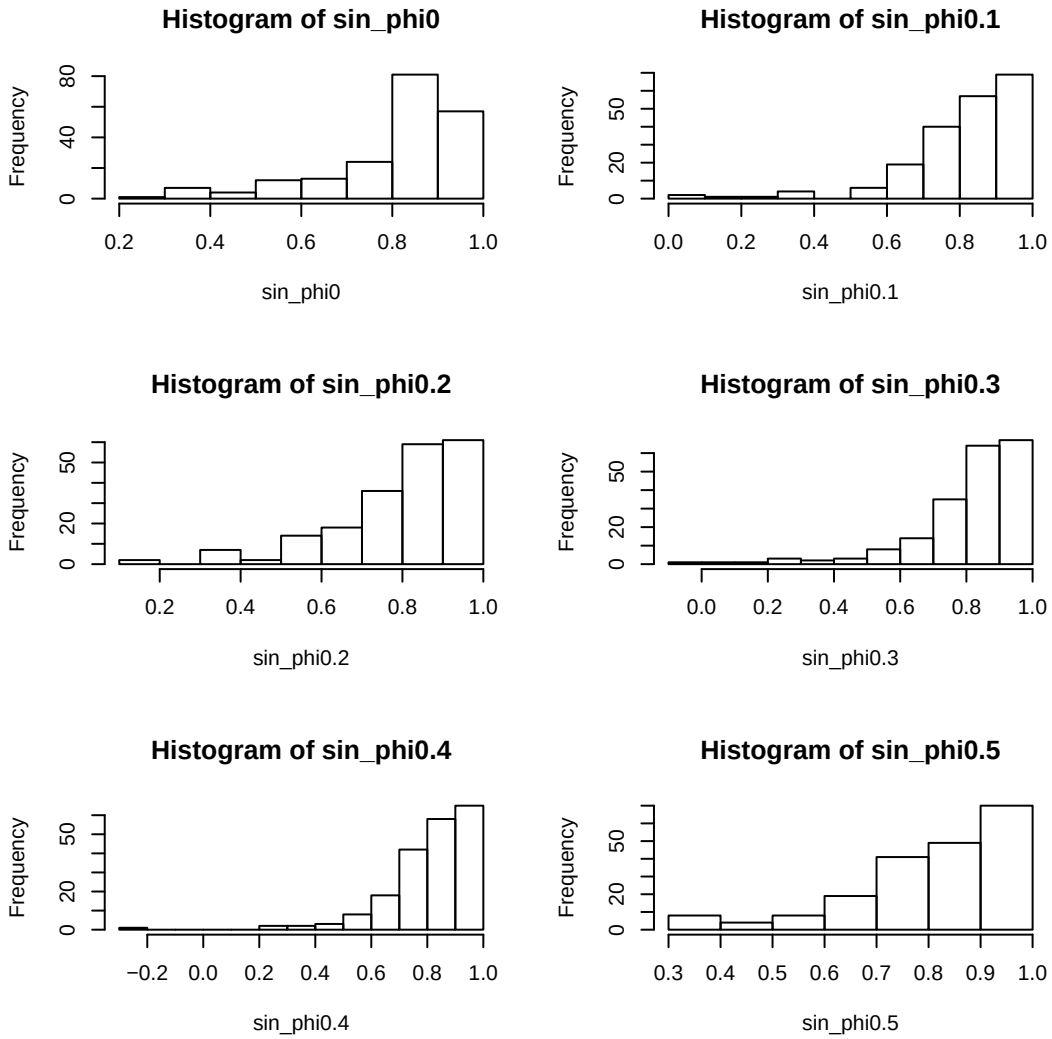

Figure A.3: Histogram of sin(phi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows histograms of sin(phi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for original dataset.

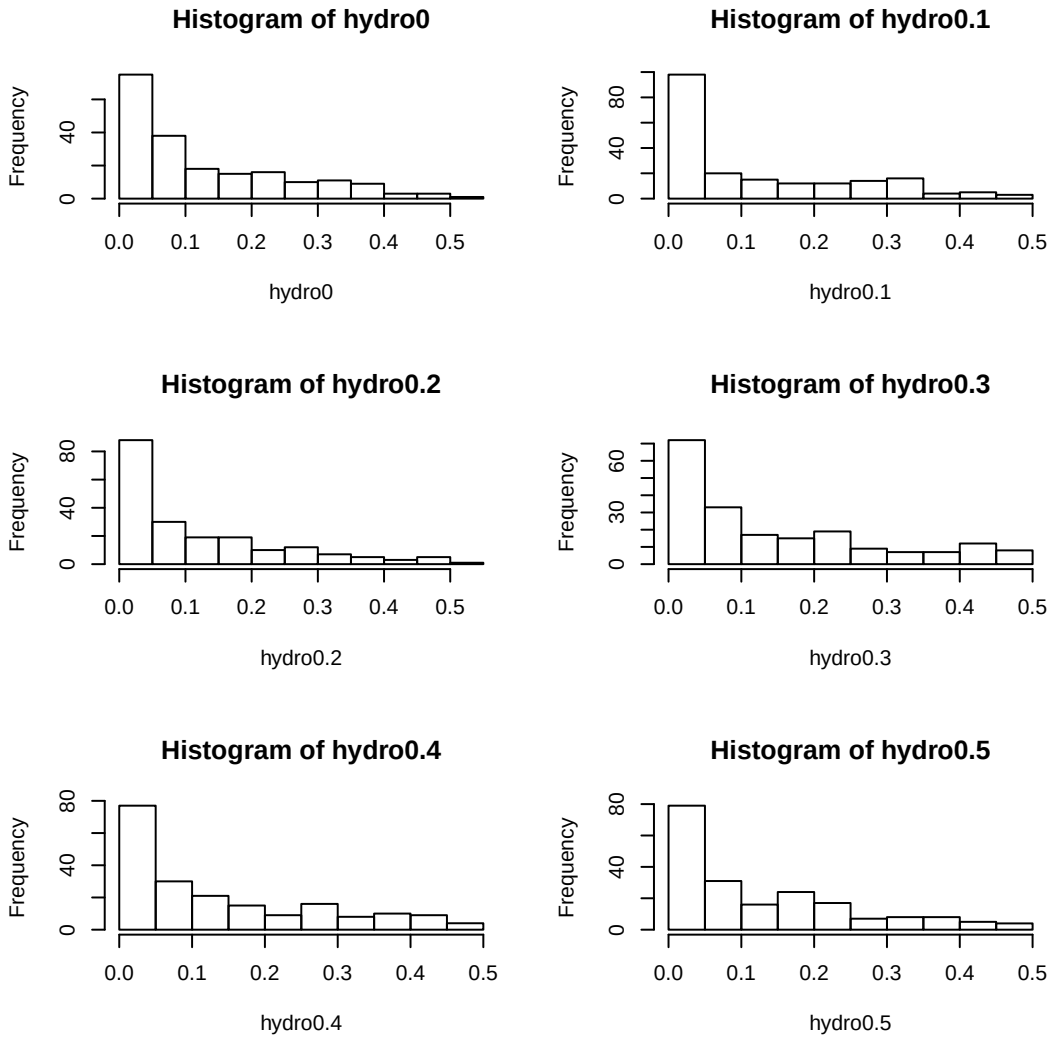

Figure A.4: Histogram of hydrogen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows histograms of hydrogen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for original dataset at X_0.

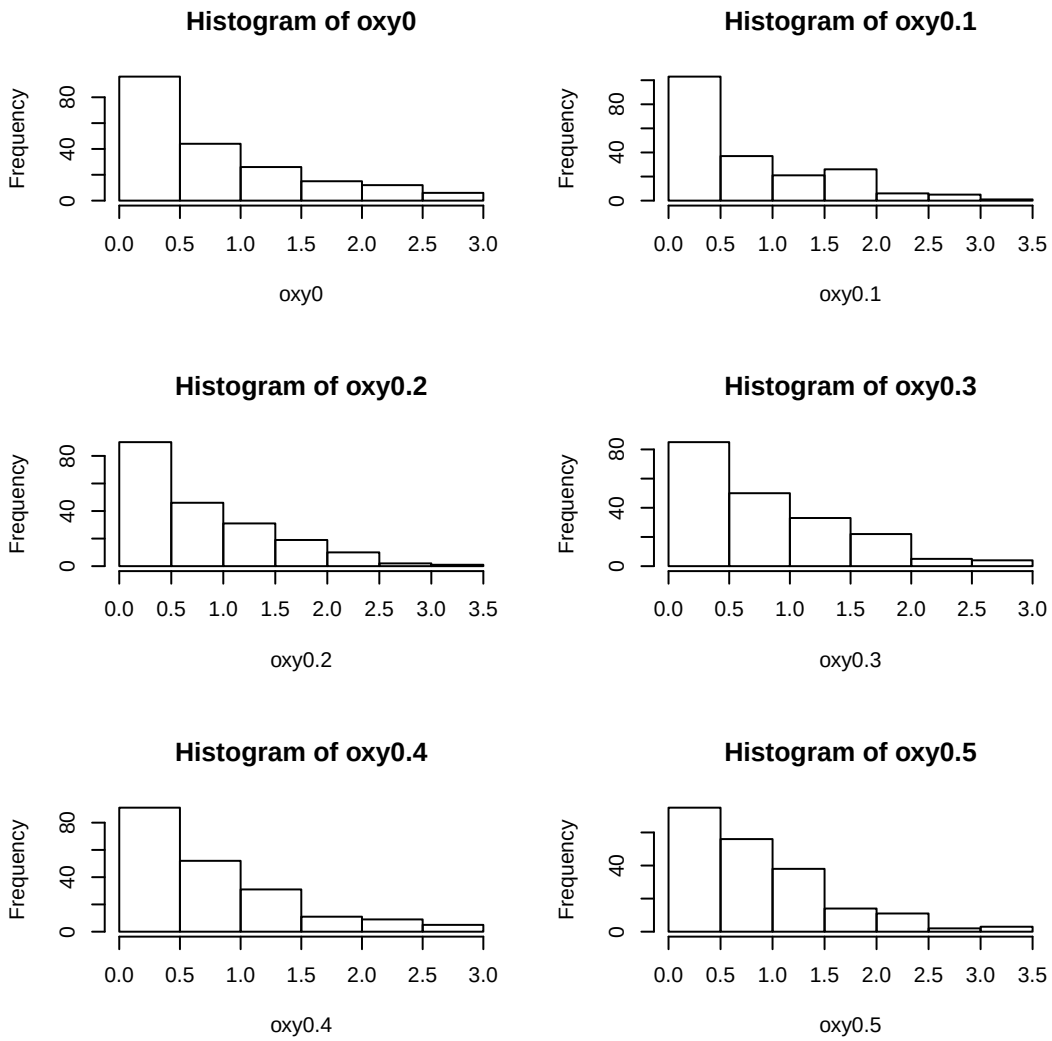

Figure A.5: Histogram of oxygen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows histograms of oxygen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for original dataset at X_0.

A.2 Random Dataset

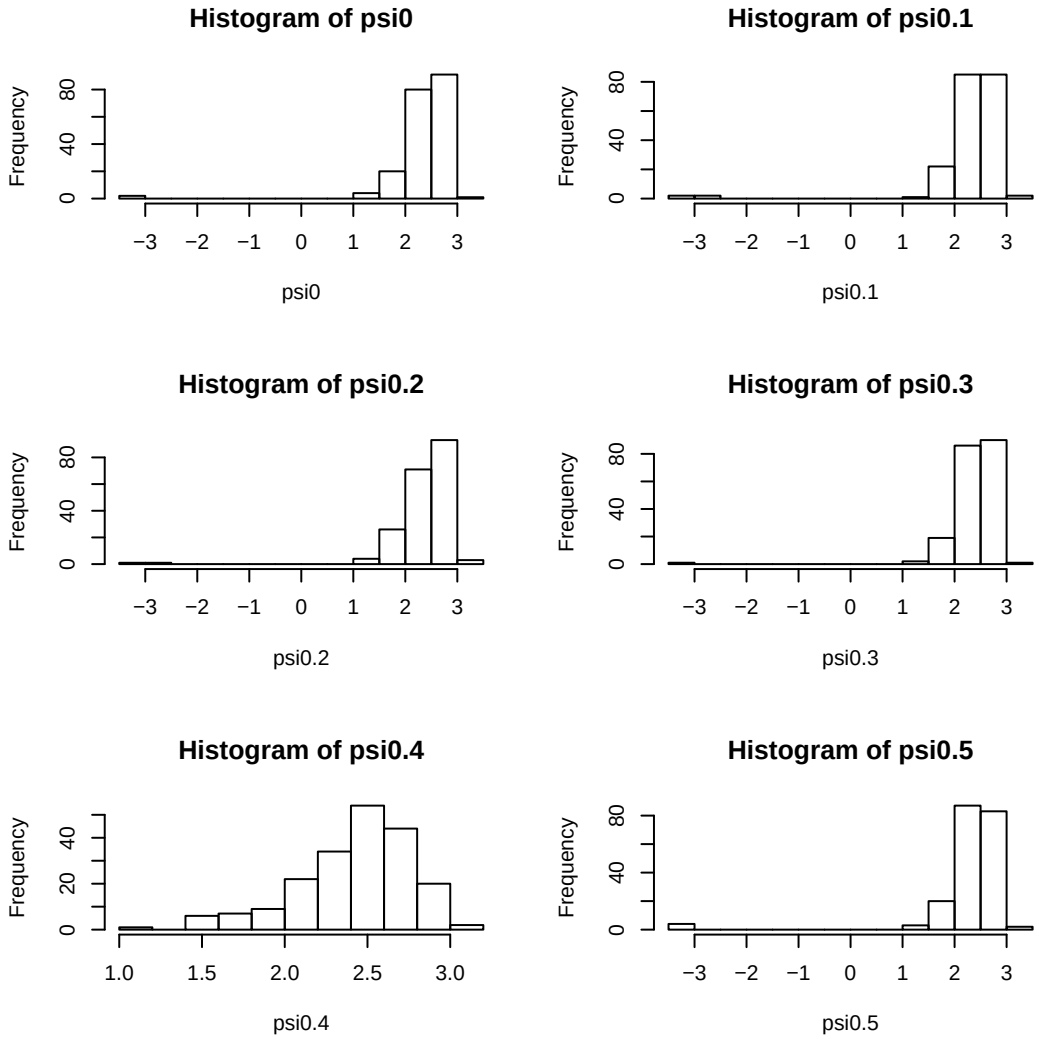

Figure A.6: Histogram of psi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows histograms of psi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for random dataset.

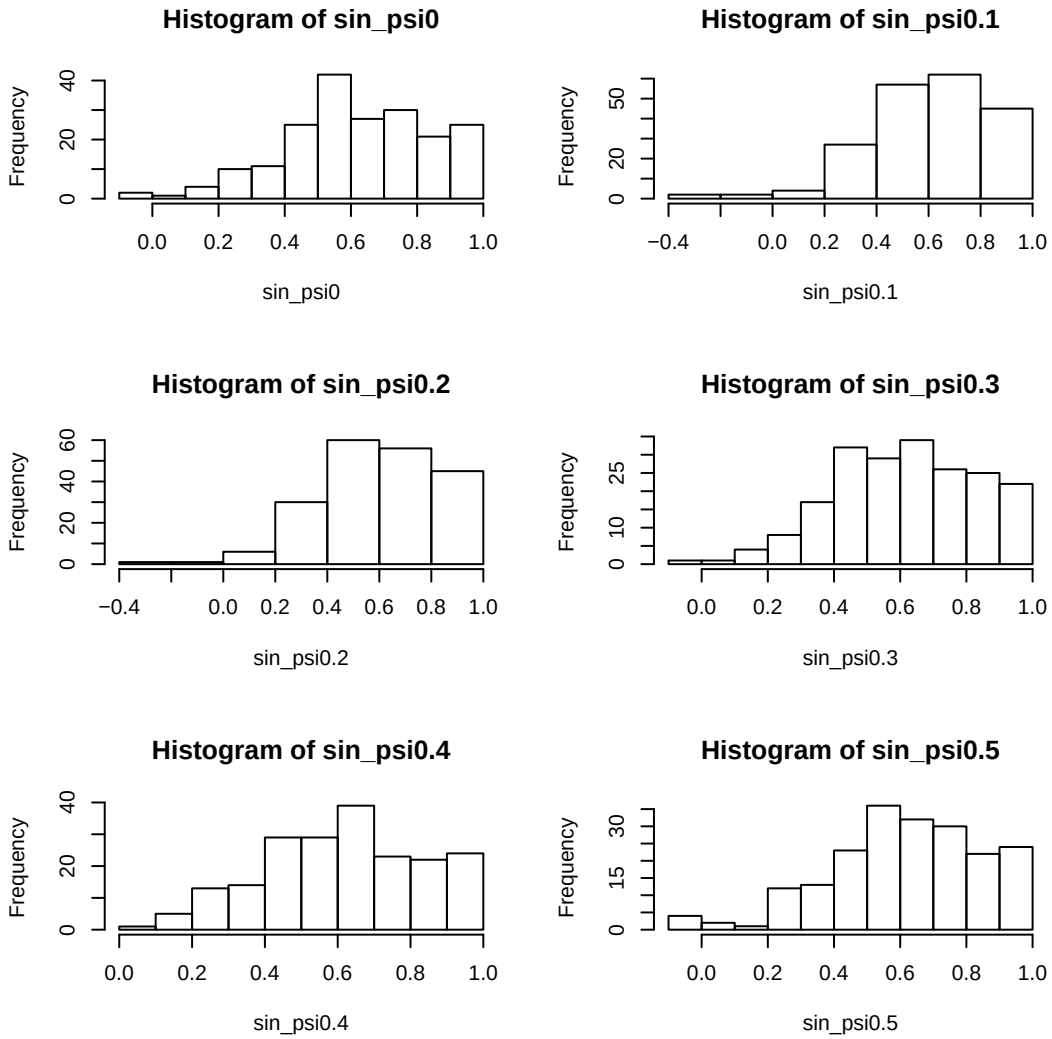

Figure A.7: Histogram of sin(psi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows histograms of sin(psi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for random dataset.

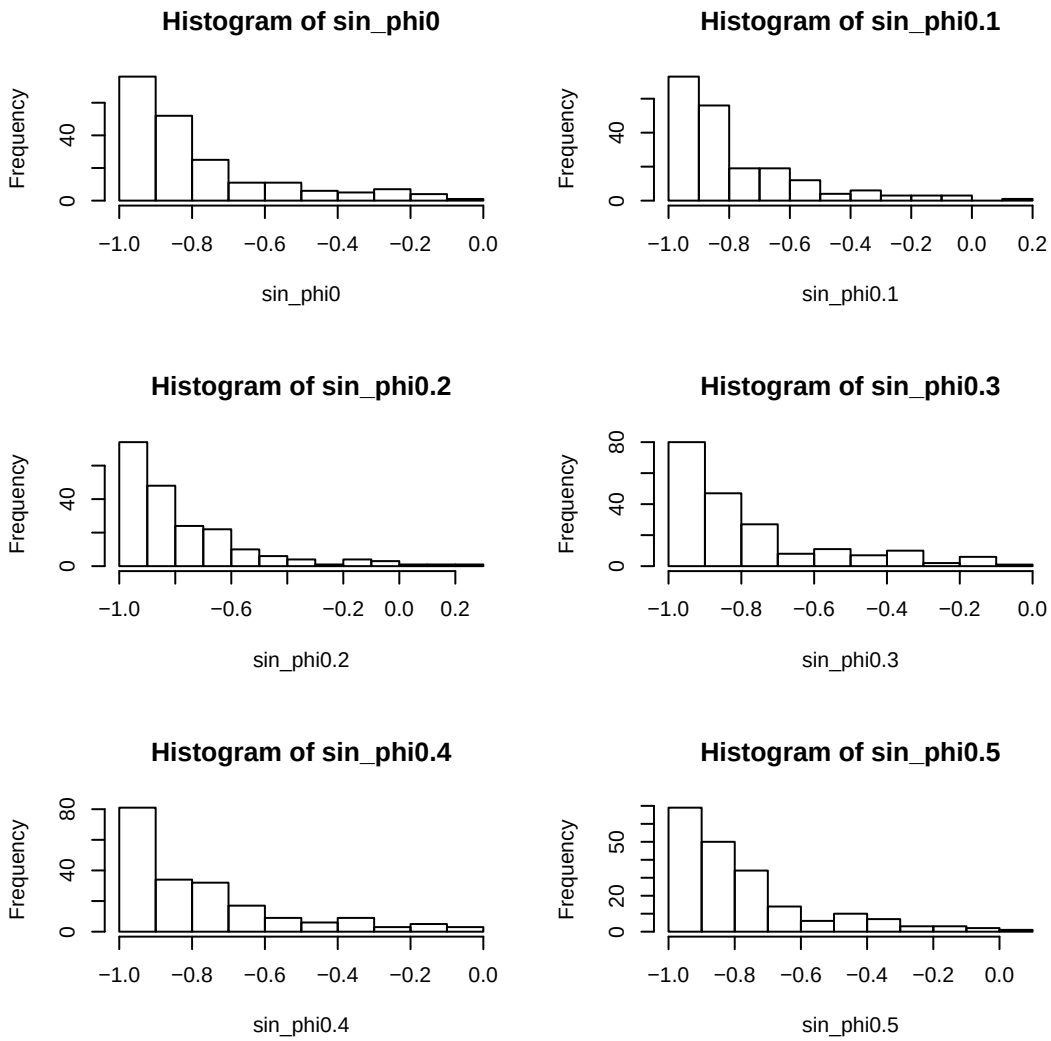

Figure A.8: Histogram of sin(phi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows histograms of sin(phi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for random dataset.

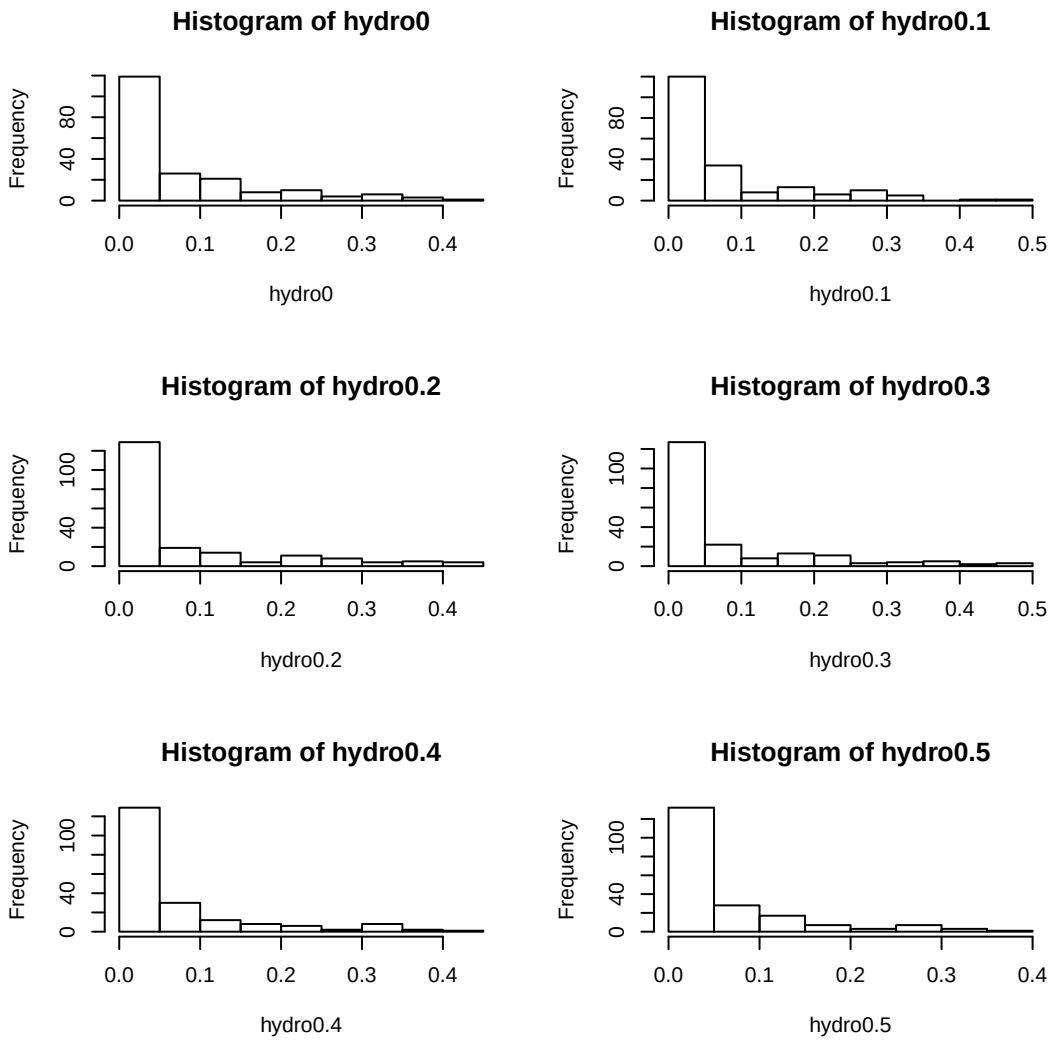

Figure A.9: Histogram of hydrogen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows histograms of hydrogen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for random dataset at X_0.

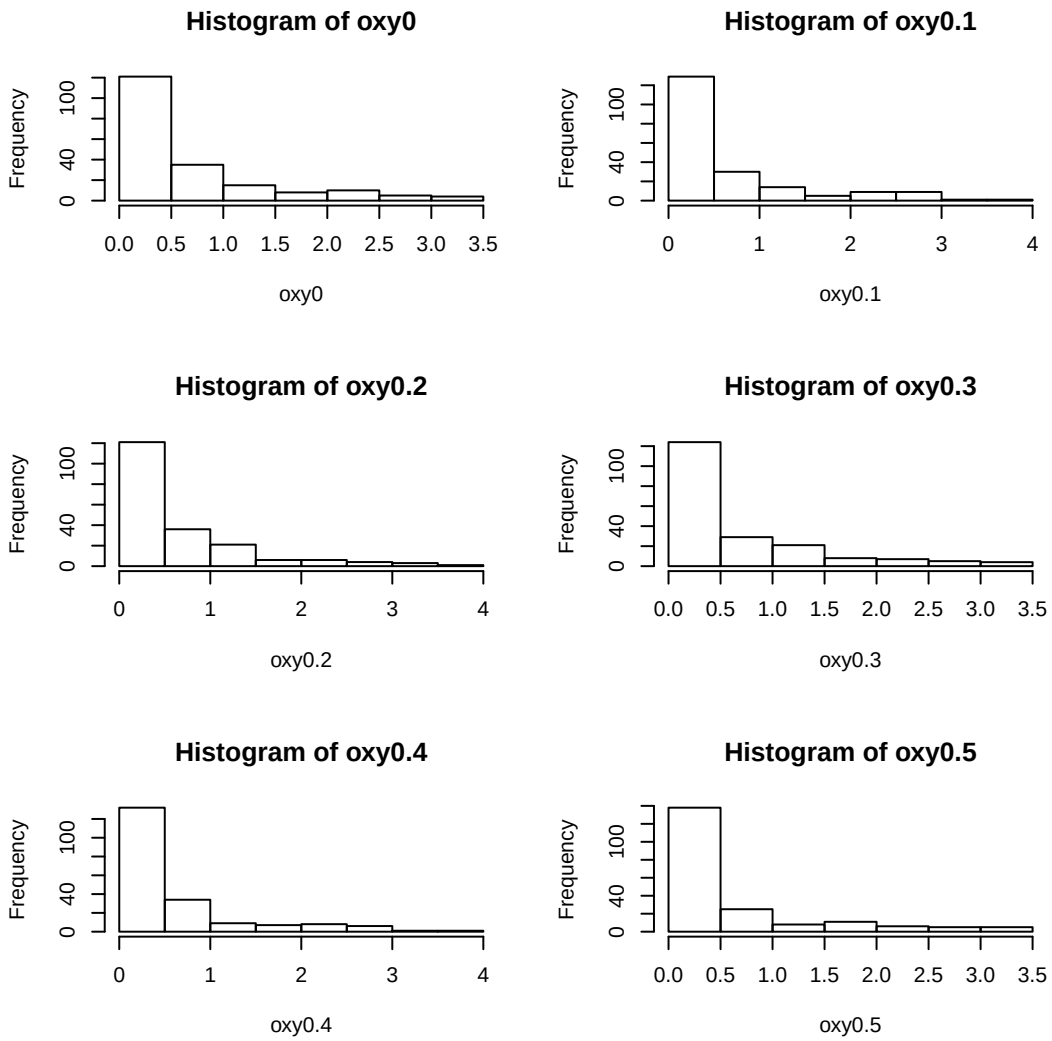

Figure A.10: Histogram of oxygen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows histograms of oxygen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps for random dataset at X_0.

Appendix B

Scatter Plot

B.1 Original Dataset

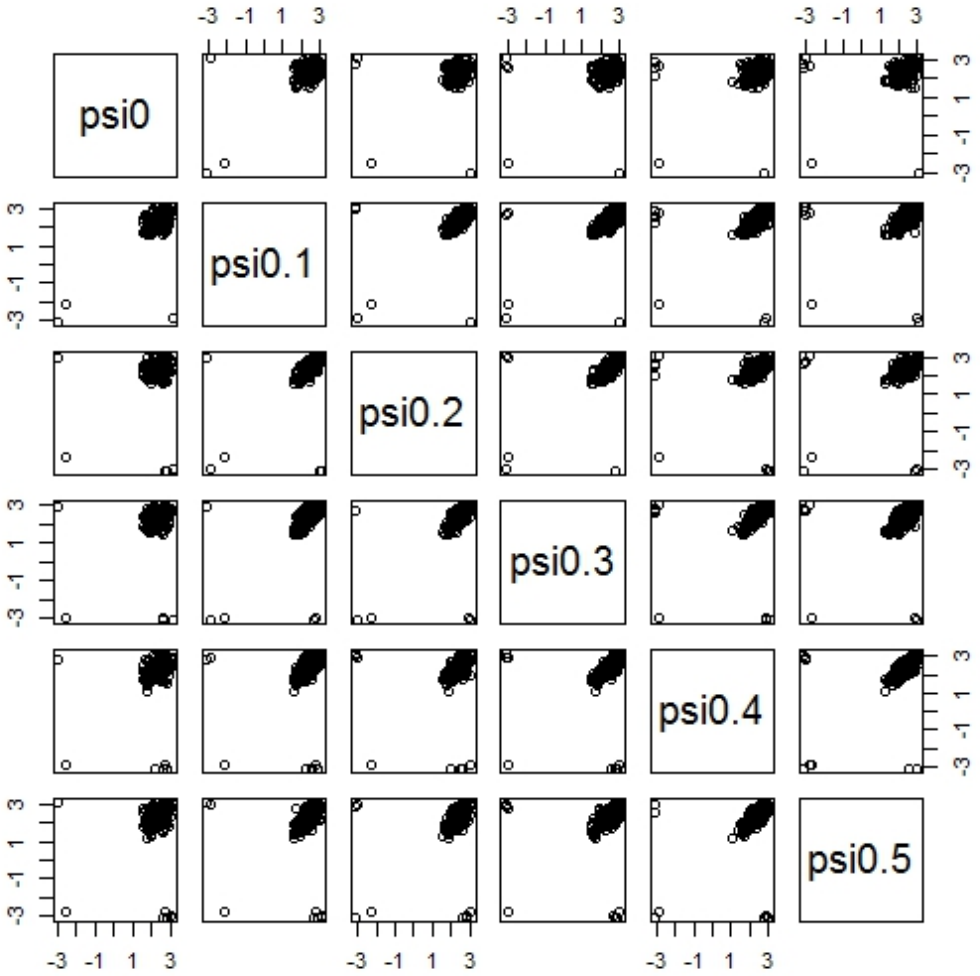

Figure B.1: Scatter plot of psi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows scatter plot of psi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for original dataset.

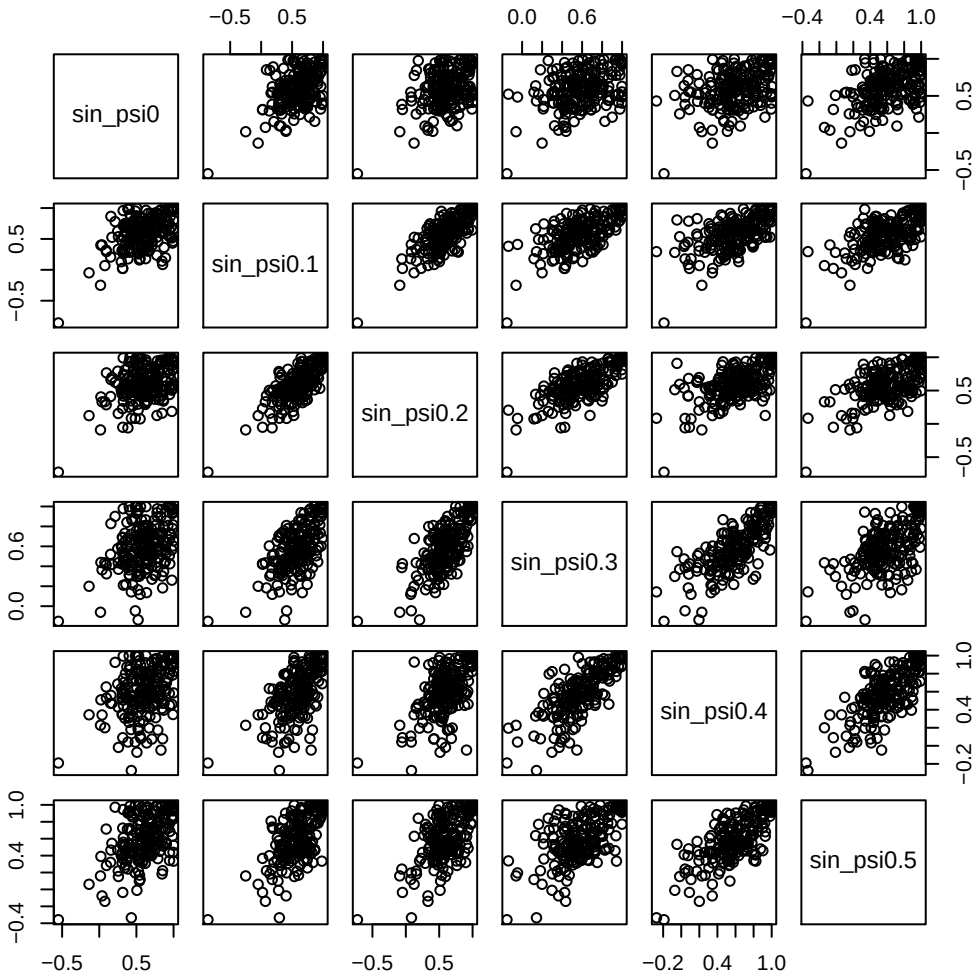

Figure B.2: Scatter plot of sin(psi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows scatter plot of sin(psi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for original dataset.

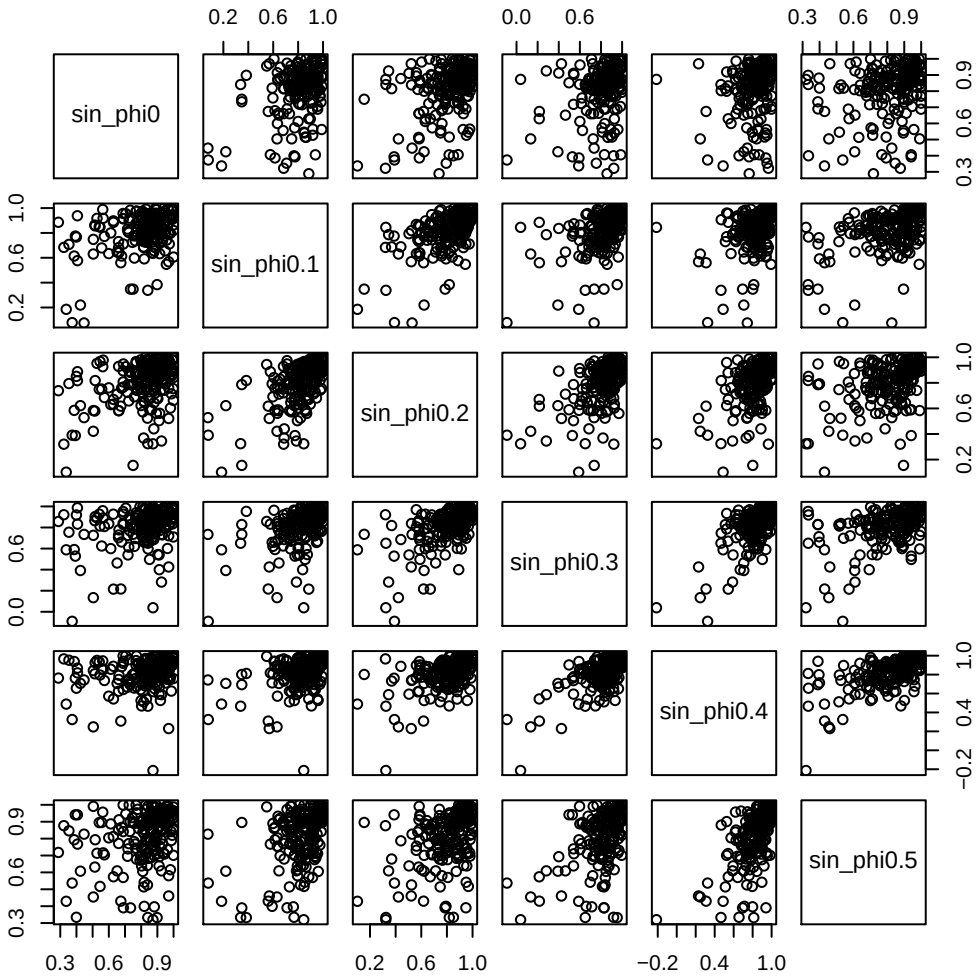

Figure B.3: Scatter plot of sin(phi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows scatter plot of sin(phi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for original dataset.

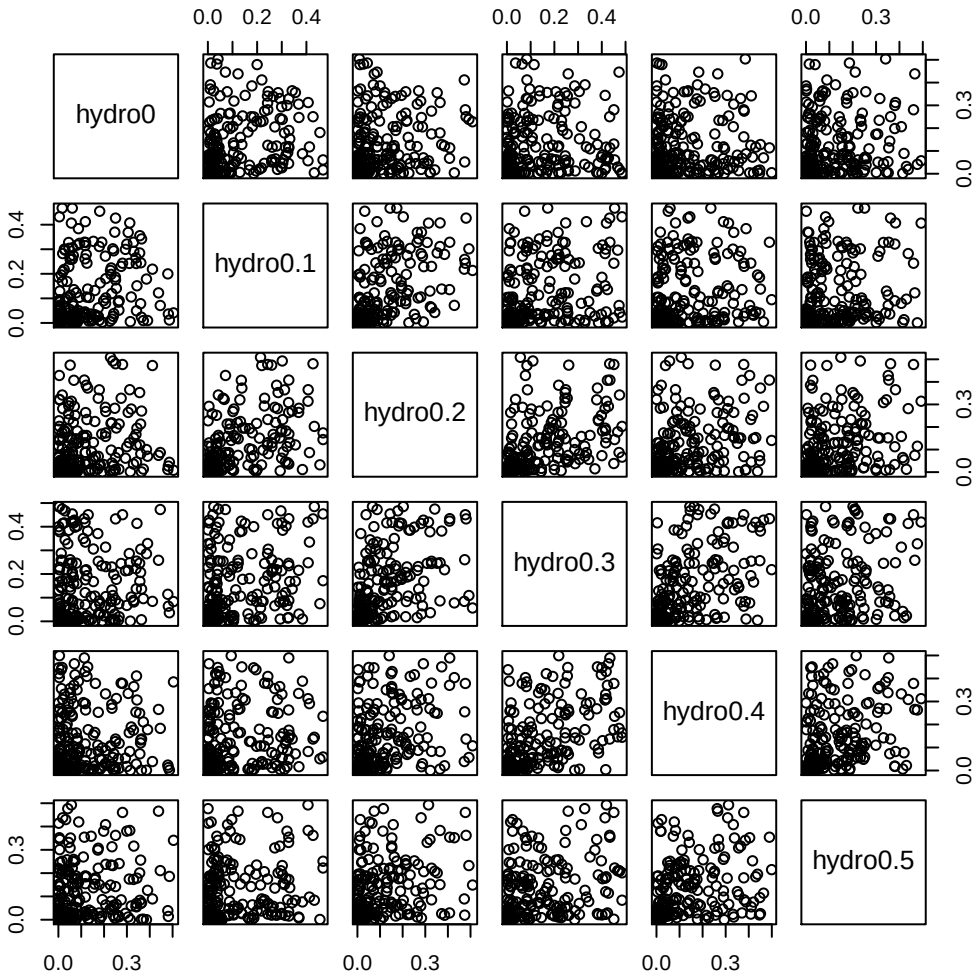

Figure B.4: Scatter plot of hydrogen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows scatter plot of hydrogen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for original dataset.

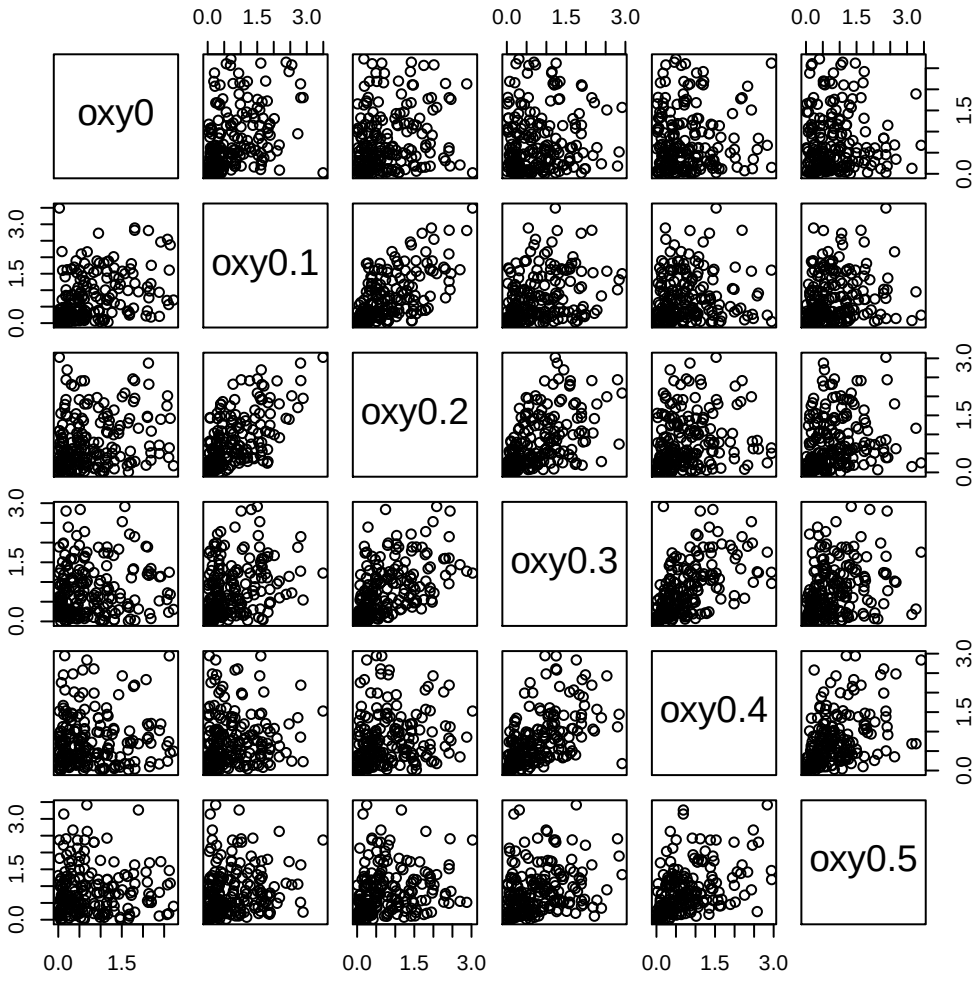

Figure B.5: Scatter plot of oxygen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: original dataset - The figure shows scatter plot of oxygen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for original dataset.

B.2 Random Dataset

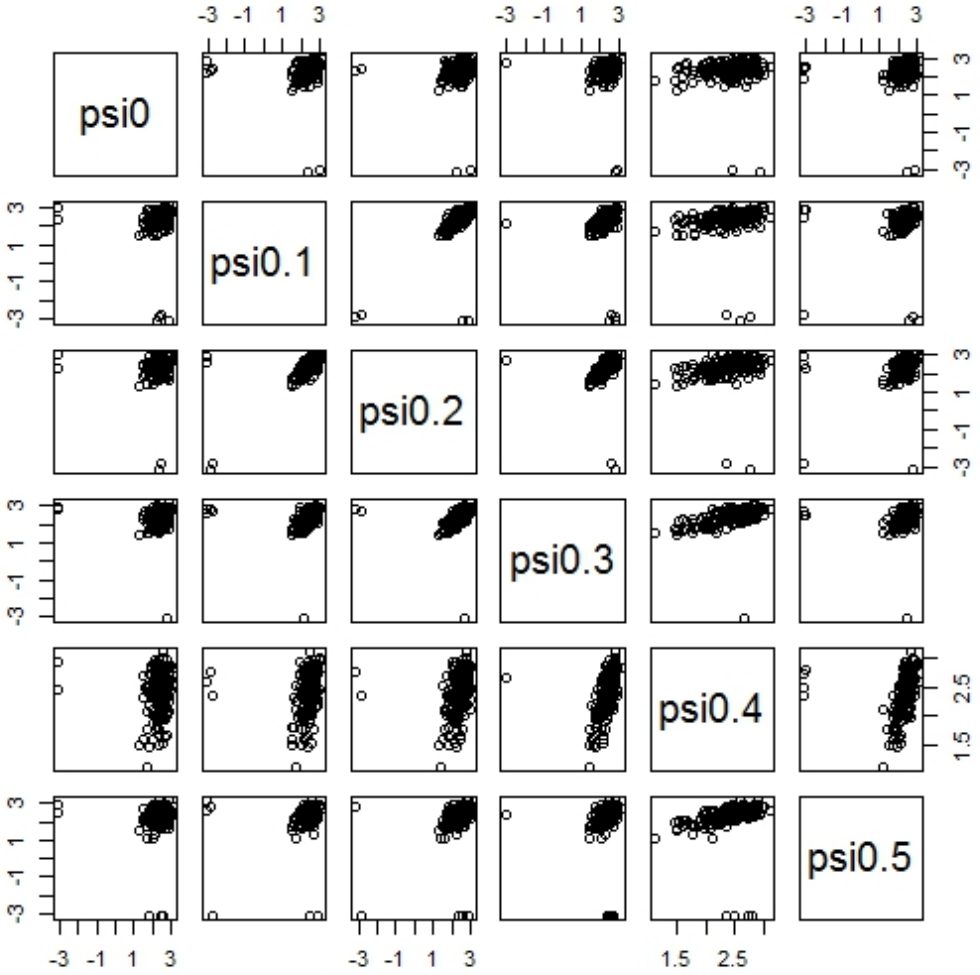

Figure B.6: Scatter plot of psi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows scatter plot of psi at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for random dataset.

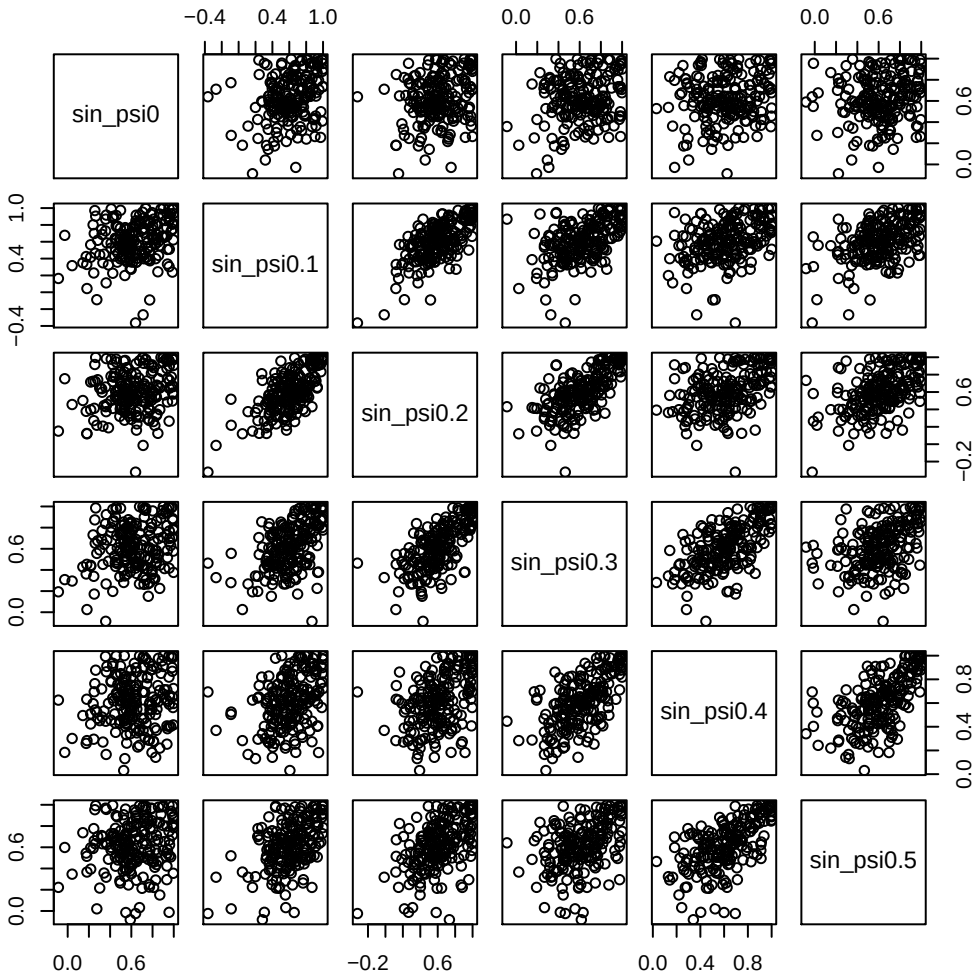

Figure B.7: Scatter plot of sin(psi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows scatter plot of sin(psi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for random dataset.

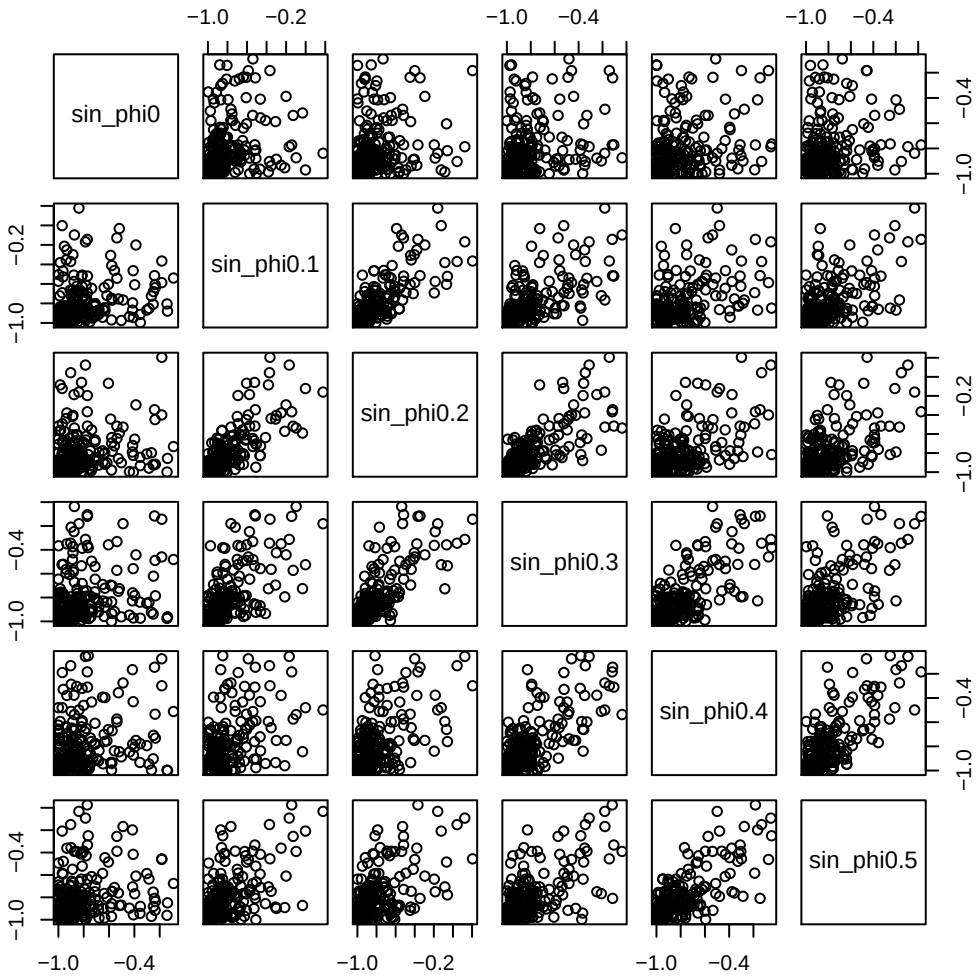

Figure B.8: Scatter plot of sin(phi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows scatter plot of sin(phi) at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for random dataset.

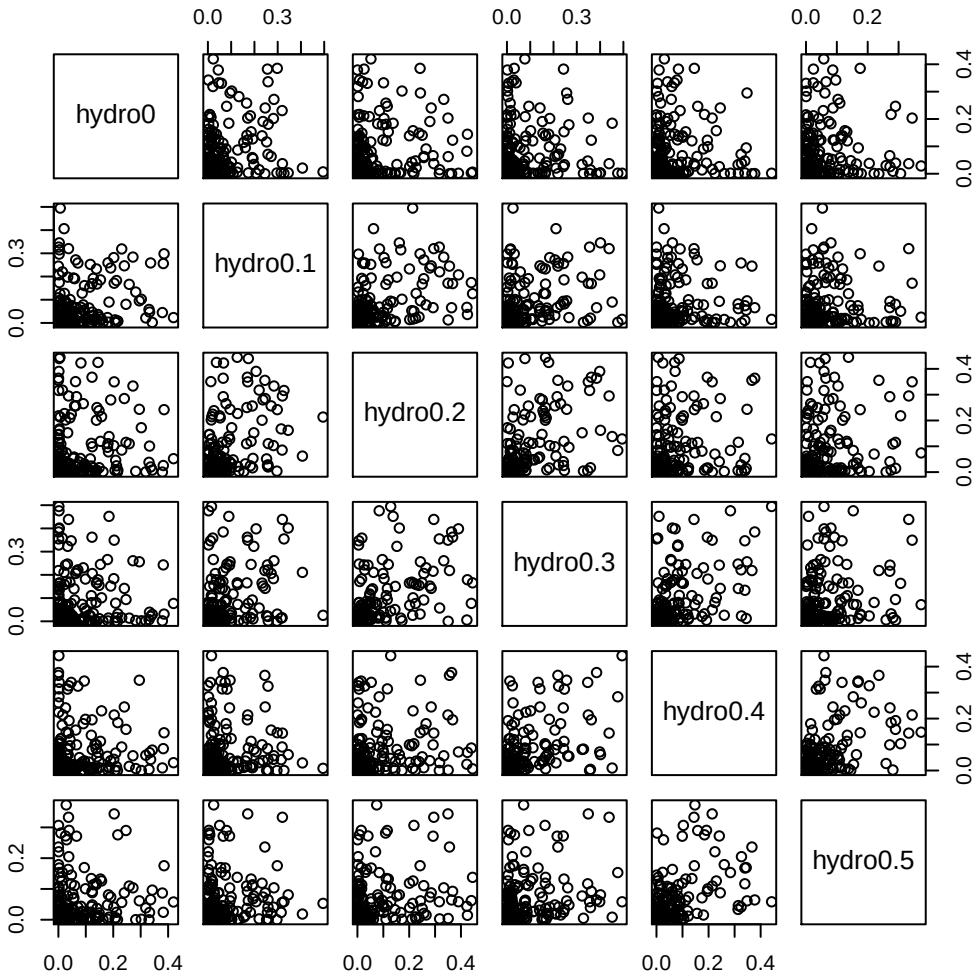

Figure B.9: Scatter plot of hydrogen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows scatter plot of hydrogen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for random dataset.

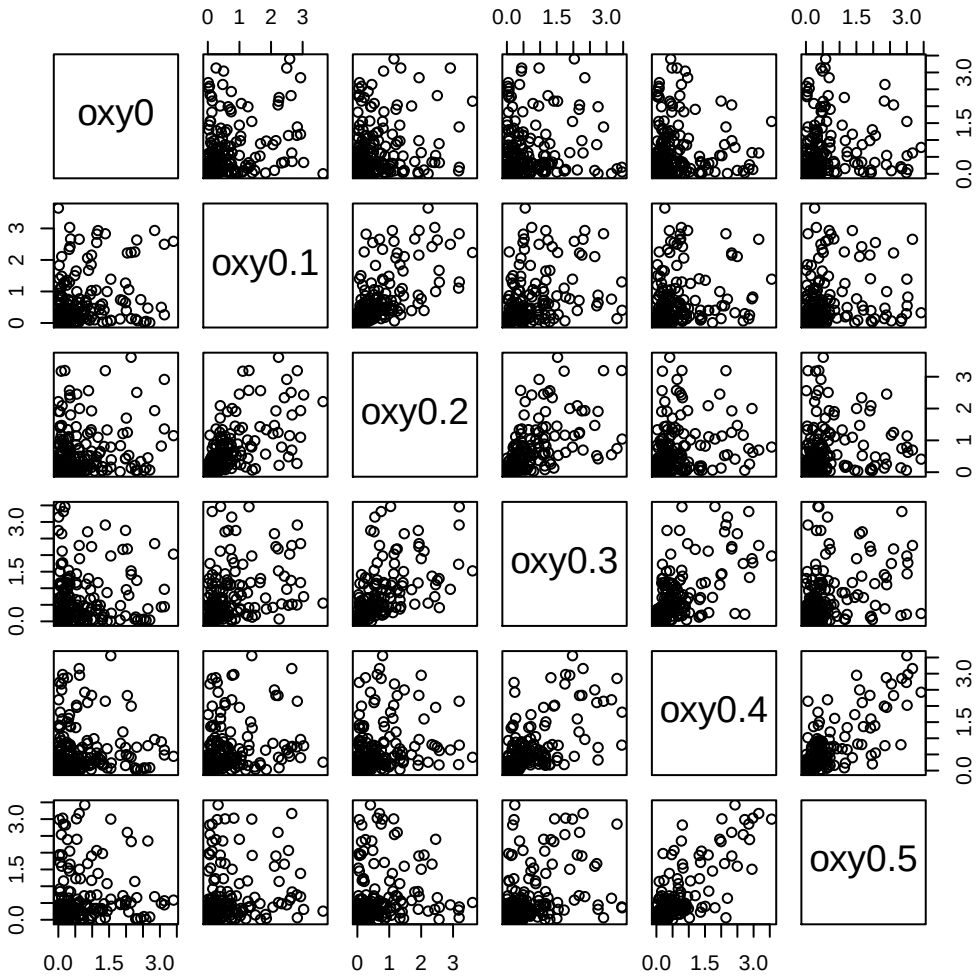

Figure B.10: Scatter plot of oxygen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps: random dataset - The figure shows scatter plot of oxygen density at delay time 0, 0.1, 0.2, 0.3, 0.4 and 0.5 ps before the transition for random dataset.

Appendix C

R Codes

The R codes that are used for analysis in Chapter 3: Test Systems and Chapter 4: Results are listed in this section as follows:

C.1 N-Variate Copulas

C.1.1 Angle Dataset

- Dimension = 2

```
##dimension=2
a210000<-read.table("E:/Dmitry/Copulas/angledataset/data/
   a210000.txt",header=T)
attach(a210000)
############################
##Histogram&Density
par(mfrow=c(2,2))
hist(x1)
hist(x2)
plot(density(x1),type="l",xlab="x1",main="Density")
plot(density(x2),type="l",xlab="x2",main="Density")
############################
##Scatter Plot Matrix
pairs(a210000)
############################
##Correlation
cor(a210000)
############################
##FIT A BETA DISTRIBUTION for x1
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
```

```
fitx1=vglm(x1 ~ 1, beta.ab, trace = TRUE)
coef(fitx1, matrix = TRUE)
Coef(fitx1)
summary(fitx1)

alpha1 = Coef(fitx1)[1]
beta1  = Coef(fitx1)[2]
y1 = seq(0,1,0.001)
dfBeta= pbeta(y1, shape1=alpha1, shape2=beta1)
dBeta = dbeta(y1, shape1=alpha1, shape2=beta1)
par(mfrow=c(1, 2))
plot(y1, dfBeta, cex=0.3, xlab="dist function")
plot(y1, dBeta,  cex=0.3, xlab="density function")
##############################
##FIT A BETA DISTRIBUTION for x2
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx2=vglm(x2 ~ 1, beta.ab, trace = TRUE)
coef(fitx2, matrix = TRUE)
Coef(fitx2)
summary(fitx2)

alpha2 = Coef(fitx2)[1]
beta2  = Coef(fitx2)[2]
y2 = seq(0,1,0.001)
dfBeta= pbeta(y2, shape1=alpha2, shape2=beta2)
dBeta = dbeta(y2, shape1=alpha2, shape2=beta2)
par(mfrow=c(1, 2))
plot(y2, dfBeta, cex=0.3, xlab="dist function")
plot(y2, dBeta,  cex=0.3, xlab="density function")
##############################
##PROBABILITY INTEGRAL TRANSFORM
u1= pbeta(x1, shape1=alpha1, shape2=beta1)
par(mfrow=c(1, 2))
hist(u1, main="", xlab="Histogram of Transformed x1")
Fn <- ecdf(x1)
lines(y1, Fn(y1), lty=2)
plot(Fn(y1),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u2= pbeta(x2, shape1=alpha2, shape2=beta2)
par(mfrow=c(1, 2))
hist(u2, main="", xlab="Histogram of Transformed x2")
```

```
Fn <- ecdf(x2)
lines(y2, Fn(y2), lty=2)
plot(Fn(y2),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

plot(u1,u2, cex=0.5, xlim=c(-0.1,1.1), ylim=c(-0.1,1.1),
     xlab="Transformed x1", ylab="Transformed x2")
cor(u1,u2, method="spearman")
################################
##FIT NORMAL'S COPULA

uu = cbind(u1,u2)
normal.cop <- ellipCopula("normal", dim = 2, dispstr="un")
(fit.ml <- fitCopula(normal.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
normal.cop <- ellipCopula("normal", param= param, dim = 2)
spearmansRho(normal.cop)
################################
##FIT T'S COPULA

uu = cbind(u1,u2)
t.cop <- ellipCopula("t", dim = 2, dispstr="un")
(fit.ml <- fitCopula(t.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
#t.cop <- ellipCopula("t", param= param, dim = 2)
#spearmansRho(t.cop)
################################
##FIT FRANK'S COPULA

uu = cbind(u1,u2)
frank.cop <- archmCopula("frank", dim = 2)
(fit.ml <- fitCopula(frank.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
frank.cop <- archmCopula("frank", param= param, dim = 2)
spearmansRho(frank.cop)
################################
##FIT CLAYTON'S COPULA

uu = cbind(u1,u2)
clayton.cop <- archmCopula("clayton", dim = 2)
```

```
(fit.ml <- fitCopula(clayton.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
clayton.cop <- archmCopula("clayton", param= param, dim = 2)
spearmansRho(clayton.cop)
################################
```

- Dimension = 5

```
##dimension=5
a510000<-read.table("E:/Dmitry/Copulas/angledataset/data/
   a510000.txt",header=T)
attach(a510000)
############################
##Histogram
par(mfrow=c(3,2))
hist(x1)
hist(x2)
hist(x3)
hist(x4)
hist(x5)
##Density
par(mfrow=c(3,2))
plot(density(x1),type="l",xlab="x1",main="Density")
plot(density(x2),type="l",xlab="x2",main="Density")
plot(density(x3),type="l",xlab="x3",main="Density")
plot(density(x4),type="l",xlab="x4",main="Density")
plot(density(x5),type="l",xlab="x5",main="Density")
##Scatter Plot Matrix
pairs(a510000)
##Correlation
cor(a510000)
####################
##FIT A BETA DISTRIBUTION for x1
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx1=vglm(x1 ~ 1, beta.ab, trace = TRUE)
coef(fitx1, matrix = TRUE)
Coef(fitx1)
summary(fitx1)

alpha1 = Coef(fitx1)[1]
beta1  = Coef(fitx1)[2]
```

```
y1 = seq(0,1,0.001)
dfBeta= pbeta(y1, shape1=alpha1, shape2=beta1)
dBeta = dbeta(y1, shape1=alpha1, shape2=beta1)
par(mfrow=c(1, 2))
plot(y1, dfBeta, cex=0.3, xlab="dist function")
plot(y1, dBeta,  cex=0.3, xlab="density function")
##############################
##FIT A BETA DISTRIBUTION for x2
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx2=vglm(x2 ~ 1, beta.ab, trace = TRUE)
coef(fitx2, matrix = TRUE)
Coef(fitx2)
summary(fitx2)

alpha2 = Coef(fitx2)[1]
beta2  = Coef(fitx2)[2]
y2 = seq(0,1,0.001)
dfBeta= pbeta(y2, shape1=alpha2, shape2=beta2)
dBeta = dbeta(y2, shape1=alpha2, shape2=beta2)
par(mfrow=c(1, 2))
plot(y2, dfBeta, cex=0.3, xlab="dist function")
plot(y2, dBeta,  cex=0.3, xlab="density function")
##############################
##FIT A BETA DISTRIBUTION for x3
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx3=vglm(x3 ~ 1, beta.ab, trace = TRUE)
coef(fitx3, matrix = TRUE)
Coef(fitx3)
summary(fitx3)

alpha3 = Coef(fitx3)[1]
beta3  = Coef(fitx3)[2]
y3 = seq(0,1,0.001)
dfBeta= pbeta(y3, shape1=alpha3, shape2=beta3)
dBeta = dbeta(y3, shape1=alpha3, shape2=beta3)
par(mfrow=c(1, 2))
plot(y3, dfBeta, cex=0.3, xlab="dist function")
plot(y3, dBeta,  cex=0.3, xlab="density function")
##############################
##FIT A BETA DISTRIBUTION for x4
library(VGAM)
```

```
beta.ab(lshape1="identity", lshape2 = "identity")
fitx4=vglm(x4 ~ 1, beta.ab, trace = TRUE)
coef(fitx4, matrix = TRUE)
Coef(fitx4)
summary(fitx4)

alpha4 = Coef(fitx4)[1]
beta4  = Coef(fitx4)[2]
y4 = seq(0,1,0.001)
dfBeta= pbeta(y4, shape1=alpha4, shape2=beta4)
dBeta = dbeta(y4, shape1=alpha4, shape2=beta4)
par(mfrow=c(1, 2))
plot(y4, dfBeta, cex=0.3, xlab="dist function")
plot(y4, dBeta,  cex=0.3, xlab="density function")
################################
##FIT A BETA DISTRIBUTION for x5
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx5=vglm(x5 ~ 1, beta.ab, trace = TRUE)
coef(fitx5, matrix = TRUE)
Coef(fitx5)
summary(fitx5)

alpha5 = Coef(fitx5)[1]
beta5  = Coef(fitx5)[2]
y5 = seq(0,1,0.001)
dfBeta= pbeta(y5, shape1=alpha5, shape2=beta5)
dBeta = dbeta(y5, shape1=alpha5, shape2=beta5)
par(mfrow=c(1, 2))
plot(y5, dfBeta, cex=0.3, xlab="dist function")
plot(y5, dBeta,  cex=0.3, xlab="density function")
################################
##PROBABILITY INTEGRAL TRANSFORM
u1= pbeta(x1, shape1=alpha1, shape2=beta1)
par(mfrow=c(1, 2))
hist(u1, main="", xlab="Histogram of Transformed x1")
Fn <- ecdf(x1)
lines(y1, Fn(y1), lty=2)
plot(Fn(y1),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u2= pbeta(x2, shape1=alpha2, shape2=beta2)
par(mfrow=c(1, 2))
```

```
hist(u2, main="", xlab="Histogram of Transformed x2")
Fn <- ecdf(x2)
lines(y2, Fn(y2), lty=2)
plot(Fn(y2),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u3= pbeta(x3, shape1=alpha3, shape2=beta3)
par(mfrow=c(1, 2))
hist(u3, main="", xlab="Histogram of Transformed x3")
Fn <- ecdf(x3)
lines(y3, Fn(y3), lty=2)
plot(Fn(y3),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u4= pbeta(x4, shape1=alpha4, shape2=beta4)
par(mfrow=c(1, 2))
hist(u4, main="", xlab="Histogram of Transformed x4")
Fn <- ecdf(x4)
lines(y4, Fn(y4), lty=2)
plot(Fn(y4),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u5= pbeta(x5, shape1=alpha5, shape2=beta5)
par(mfrow=c(1, 2))
hist(u5, main="", xlab="Histogram of Transformed x5")
Fn <- ecdf(x5)
lines(y5, Fn(y5), lty=2)
plot(Fn(y5),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)
################################
##FIT FRANK'S COPULA

uu = cbind(u1,u2,u3,u4,u5)
frank.cop <- archmCopula("frank", dim = 5)
(fit.ml <- fitCopula(frank.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
frank.cop <- archmCopula("frank", param= param, dim = 5)
spearmansRho(frank.cop)
################################
##FIT CLAYTON'S COPULA

uu = cbind(u1,u2,u3,u4,u5)
```

```
clayton.cop <- archmCopula("clayton", dim = 5)
(fit.ml <- fitCopula(clayton.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
clayton.cop <- archmCopula("clayton", param= param, dim = 5)
spearmansRho(clayton.cop)
################################
##FIT Gumbel'S COPULA

uu = cbind(u1,u2,u3,u4,u5)
gumbel.cop <- archmCopula("gumbel", dim = 5)
(fit.ml <- fitCopula(gumbel.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
gumbel.cop <- archmCopula("gumbel", param= param, dim = 5)
spearmansRho(gumbel.cop)
################################
```

C.1.2 Amplitude Dataset

- Dimension = 2

```
##dimension=2
a210000<-read.table("E:/Dmitry/Copulas/amplitudedataset/data/
   a210000.txt",header=T)
attach(a210000)
############################
##Histogram&Density
par(mfrow=c(2,2))
hist(x1)
hist(x2)
plot(density(x1),type="l",xlab="x1",main="Density")
plot(density(x2),type="l",xlab="x2",main="Density")
############################
##Scatter Plot Matrix
pairs(a210000)
############################
##Correlation
cor(a210000)
############################
##Plots of distribution function and density function
y = seq(0,1,0.001)
dfUniform = punif(y,0,1)
dUniform = dunif(y,0,1)
```

```
par(mfrow=c(1, 2))
plot(y, dfUniform, cex=0.3, xlab="dist function")
plot(y, dUniform,  cex=0.3, xlab="density function")
##############################
##PROBABILITY INTEGRAL TRANSFORM
u1= punif(x1,0,1)
par(mfrow=c(1, 2))
hist(u1, main="", xlab="Histogram of Transformed x1")
Fn <- ecdf(x1)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u2= punif(x2,0,1)
par(mfrow=c(1, 2))
hist(u2, main="", xlab="Histogram of Transformed x2")
Fn <- ecdf(x2)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

plot(u1,u2, cex=0.5, xlim=c(-0.1,1.1), ylim=c(-0.1,1.1),
     xlab="Transformed x1", ylab="Transformed x2")
cor(u1,u2, method="spearman")
##############################
##FIT NORMAL'S COPULA

uu = cbind(u1,u2)
normal.cop <- ellipCopula("normal", dim = 2, dispstr="un")
(fit.ml <- fitCopula(normal.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
normal.cop <- ellipCopula("normal", param= param, dim = 2)
spearmansRho(normal.cop)
##############################
##FIT T'S COPULA

uu = cbind(u1,u2)
t.cop <- ellipCopula("t", dim = 2, dispstr="un")
(fit.ml <- fitCopula(t.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
#t.cop <- ellipCopula("t", param= param, dim = 2)
```

```
#spearmansRho(t.cop)
################################
##FIT FRANK'S COPULA

uu = cbind(u1,u2)
frank.cop <- archmCopula("frank", dim = 2)
(fit.ml <- fitCopula(frank.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
frank.cop <- archmCopula("frank", param= param, dim = 2)
spearmansRho(frank.cop)
################################
##FIT CLAYTON'S COPULA

uu = cbind(u1,u2)
clayton.cop <- archmCopula("clayton", dim = 2)
(fit.ml <- fitCopula(clayton.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
clayton.cop <- archmCopula("clayton", param= param, dim = 2)
spearmansRho(clayton.cop)
################################
```

- Dimension = 5

```
##dimension=5
a510000<-read.table("E:/Dmitry/Copulas/amplitudedataset/data/
 a510000.txt",header=T)
attach(a510000)
################################
##Histogram
par(mfrow=c(3,2))
hist(x1)
hist(x2)
hist(x3)
hist(x4)
hist(x5)
##Density
par(mfrow=c(3,2))
plot(density(x1),type="l",xlab="x1",main="Density")
plot(density(x2),type="l",xlab="x2",main="Density")
plot(density(x3),type="l",xlab="x3",main="Density")
plot(density(x4),type="l",xlab="x4",main="Density")
```

```
plot(density(x5),type="l",xlab="x5",main="Density")
##Scatter Plot Matrix
pairs(a510000)
##Correlation
cor(a510000)
####################
##Plots of distribution function and density function
y = seq(0,1,0.001)
dfUniform = punif(y,0,1)
dUniform = dunif(y,0,1)
par(mfrow=c(1, 2))
plot(y, dfUniform, cex=0.3, xlab="dist function")
plot(y, dUniform,  cex=0.3, xlab="density function")
############################
##PROBABILITY INTEGRAL TRANSFORM
u1= punif(x1,0,1)
par(mfrow=c(1, 2))
hist(u1, main="", xlab="Histogram of Transformed x1")
Fn <- ecdf(x1)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u2= punif(x2,0,1)
par(mfrow=c(1, 2))
hist(u2, main="", xlab="Histogram of Transformed x2")
Fn <- ecdf(x2)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u3= punif(x3,0,1)
par(mfrow=c(1, 2))
hist(u3, main="", xlab="Histogram of Transformed x3")
Fn <- ecdf(x3)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u4= punif(x4,0,1)
par(mfrow=c(1, 2))
hist(u4, main="", xlab="Histogram of Transformed x4")
Fn <- ecdf(x4)
```

```
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u5= punif(x5,0,1)
par(mfrow=c(1, 2))
hist(u5, main="", xlab="Histogram of Transformed x5")
Fn <- ecdf(x5)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)
########################
##FIT NORMAL'S COPULA

uu = cbind(u1,u2,u3,u4,u5)
normal.cop <- ellipCopula("normal", dim = 5, dispstr="un")
(fit.ml <- fitCopula(normal.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
normal.cop <- ellipCopula("normal", param= param, dim = 5)
spearmansRho(normal.cop)
###############################
##FIT T'S COPULA

uu = cbind(u1,u2,u3,u4,u5)
t.cop <- ellipCopula("t", dim = 5, dispstr="un")
(fit.ml <- fitCopula(t.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
#t.cop <- ellipCopula("t", param= param, dim = 5)
#spearmansRho(t.cop)
###############################
##FIT FRANK'S COPULA

uu = cbind(u1,u2,u3,u4,u5)
frank.cop <- archmCopula("frank", dim = 5)
(fit.ml <- fitCopula(frank.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
frank.cop <- archmCopula("frank", param= param, dim = 5)
spearmansRho(frank.cop)
###############################
##FIT CLAYTON'S COPULA
```

```
uu = cbind(u1,u2,u3,u4,u5)
clayton.cop <- archmCopula("clayton", dim = 5)
(fit.ml <- fitCopula(clayton.cop, uu, method="ml"))
summary(fit.ml)
(param = fit.ml@estimate)
clayton.cop <- archmCopula("clayton", param= param, dim = 5)
spearmansRho(clayton.cop)
################################
```

C.2 D-vine

C.2.1 Angle Dataset

- Dimension = 5

```
##dimension=5
##Histogram
par(mfrow=c(3,2))
hist(x1)
hist(x2)
hist(x3)
hist(x4)
hist(x5)
##Density
par(mfrow=c(3,2))
plot(density(x1),type="l",xlab="x1",main="Density")
plot(density(x2),type="l",xlab="x2",main="Density")
plot(density(x3),type="l",xlab="x3",main="Density")
plot(density(x4),type="l",xlab="x4",main="Density")
plot(density(x5),type="l",xlab="x5",main="Density")
##Scatter Plot Matrix
pairs(phid5skip5)
##Correlation
cor(phid5skip5)
####################
##FIT A BETA DISTRIBUTION for x1
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx1=vglm(x1 ~ 1, beta.ab, trace = TRUE)
coef(fitx1, matrix = TRUE)
Coef(fitx1)
summary(fitx1)
```

```
alpha1 = Coef(fitx1)[1]
beta1  = Coef(fitx1)[2]
y1 = seq(0,1,0.001)
dfBeta= pbeta(y1, shape1=alpha1, shape2=beta1)
dBeta = dbeta(y1, shape1=alpha1, shape2=beta1)
par(mfrow=c(1, 2))
plot(y1, dfBeta, cex=0.3, xlab="dist function")
plot(y1, dBeta,  cex=0.3, xlab="density function")
##############################
##FIT A BETA DISTRIBUTION for x2
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx2=vglm(x2 ~ 1, beta.ab, trace = TRUE)
coef(fitx2, matrix = TRUE)
Coef(fitx2)
summary(fitx2)

alpha2 = Coef(fitx2)[1]
beta2  = Coef(fitx2)[2]
y2 = seq(0,1,0.001)
dfBeta= pbeta(y2, shape1=alpha2, shape2=beta2)
dBeta = dbeta(y2, shape1=alpha2, shape2=beta2)
par(mfrow=c(1, 2))
plot(y2, dfBeta, cex=0.3, xlab="dist function")
plot(y2, dBeta,  cex=0.3, xlab="density function")
##############################
##FIT A BETA DISTRIBUTION for x3
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx3=vglm(x3 ~ 1, beta.ab, trace = TRUE)
coef(fitx3, matrix = TRUE)
Coef(fitx3)
summary(fitx3)

alpha3 = Coef(fitx3)[1]
beta3  = Coef(fitx3)[2]
y3 = seq(0,1,0.001)
dfBeta= pbeta(y3, shape1=alpha3, shape2=beta3)
dBeta = dbeta(y3, shape1=alpha3, shape2=beta3)
par(mfrow=c(1, 2))
plot(y3, dfBeta, cex=0.3, xlab="dist function")
plot(y3, dBeta,  cex=0.3, xlab="density function")
##############################
```

```
##FIT A BETA DISTRIBUTION for x4
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx4=vglm(x4 ~ 1, beta.ab, trace = TRUE)
coef(fitx4, matrix = TRUE)
Coef(fitx4)
summary(fitx4)

alpha4 = Coef(fitx4)[1]
beta4  = Coef(fitx4)[2]
y4 = seq(0,1,0.001)
dfBeta= pbeta(y4, shape1=alpha4, shape2=beta4)
dBeta = dbeta(y4, shape1=alpha4, shape2=beta4)
par(mfrow=c(1, 2))
plot(y4, dfBeta, cex=0.3, xlab="dist function")
plot(y4, dBeta,  cex=0.3, xlab="density function")
################################
##FIT A BETA DISTRIBUTION for x5
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx5=vglm(x5 ~ 1, beta.ab, trace = TRUE)
coef(fitx5, matrix = TRUE)
Coef(fitx5)
summary(fitx5)

alpha5 = Coef(fitx5)[1]
beta5  = Coef(fitx5)[2]
y5 = seq(0,1,0.001)
dfBeta= pbeta(y5, shape1=alpha5, shape2=beta5)
dBeta = dbeta(y5, shape1=alpha5, shape2=beta5)
par(mfrow=c(1, 2))
plot(y5, dfBeta, cex=0.3, xlab="dist function")
plot(y5, dBeta,  cex=0.3, xlab="density function")
################################
##PROBABILITY INTEGRAL TRANSFORM
u1= pbeta(x1, shape1=alpha1, shape2=beta1)
par(mfrow=c(1, 2))
hist(u1, main="", xlab="Histogram of Transformed x1")
Fn <- ecdf(x1)
lines(y1, Fn(y1), lty=2)
plot(Fn(y1),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)
```

```
u2= pbeta(x2, shape1=alpha2, shape2=beta2)
par(mfrow=c(1, 2))
hist(u2, main="", xlab="Histogram of Transformed x2")
Fn <- ecdf(x2)
lines(y2, Fn(y2), lty=2)
plot(Fn(y2),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u3= pbeta(x3, shape1=alpha3, shape2=beta3)
par(mfrow=c(1, 2))
hist(u3, main="", xlab="Histogram of Transformed x3")
Fn <- ecdf(x3)
lines(y3, Fn(y3), lty=2)
plot(Fn(y3),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u4= pbeta(x4, shape1=alpha4, shape2=beta4)
par(mfrow=c(1, 2))
hist(u4, main="", xlab="Histogram of Transformed x4")
Fn <- ecdf(x4)
lines(y4, Fn(y4), lty=2)
plot(Fn(y4),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u5= pbeta(x5, shape1=alpha5, shape2=beta5)
par(mfrow=c(1, 2))
hist(u5, main="", xlab="Histogram of Transformed x5")
Fn <- ecdf(x5)
lines(y5, Fn(y5), lty=2)
plot(Fn(y5),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)
##############################

uu = cbind(u1,u2,u3,u4,u5)

# select the D-vine families and parameters
dvine = CDVineCopSelect(uu,c(1:6),type="DVine")
dvine

# simulate from a bivariate copula
# pair: (1,2) with Joe pair-copulas
fam12 = 6
par12 = 1.28124153
```

```
simdat12 = BiCopSim(100,fam12,par12)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat12[,1],simdat12[,2])

# create chi-plots
BiCopChiPlot(simdat12[,1],simdat12[,2],xlim=c(-1,1),ylim=c(-1,1),
main="General chi-plot")

# create K-plots
BiCopKPlot(simdat12[,1],simdat12[,2],main="Joe copula")

# pair: (2,3) with Joe pair-copulas
fam23 = 6
par23 = 1.27992139
simdat23 = BiCopSim(100,fam23,par23)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat23[,1],simdat23[,2])

# pair: (3,4) with Joe pair-copulas
fam34 = 6
par34 = 1.27405092
simdat34 = BiCopSim(100,fam34,par34)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat34[,1],simdat34[,2])

# pair: (4,5) with Gaussian pair-copulas
fam45 = 1
par45 = -0.04920439
simdat45 = BiCopSim(100,fam45,par45)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat45[,1],simdat45[,2])

# pair: 13..2 with Gaussian pair-copulas
fam13..2 = 1
par13..2 = 0.23208254
```

```
simdat13..2 = BiCopSim(100,fam13..2,par13..2)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat13..2[,1],simdat13..2[,2])

# pair: 24..3 with Frank pair-copulas
fam24..3 = 5
par24..3 = 1.59238616
simdat24..3 = BiCopSim(100,fam24..3,par24..3)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat24..3[,1],simdat24..3[,2])

# pair: 35..4 with Frank pair-copulas
fam35..4 = 5
par35..4 = 1.72711827
simdat35..4 = BiCopSim(100,fam35..4,par35..4)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat35..4[,1],simdat35..4[,2])

# pair: 14..23 with Gaussian pair-copulas
fam14..23 = 1
par14..23 = -0.07249118
simdat14..23 = BiCopSim(100,fam14..23,par14..23)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat14..23[,1],simdat14..23[,2])

# pair: 25..34 with Gaussian pair-copulas
fam25..34 = 1
par25..34 = -0.04660377
simdat25..34 = BiCopSim(100,fam25..34,par25..34)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat25..34[,1],simdat25..34[,2])

# pair: 15..234 with Frank pair-copulas
```

```
fam15..234 = 5
par15..234 = 1.66364214
simdat15..234 = BiCopSim(100,fam15..234,par15..234)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat15..234[,1],simdat15..234[,2])
```

- Dimension = 10

```
##dimension=10
##Histogram
par(mfrow=c(3,2))
hist(x1)
hist(x2)
hist(x3)
hist(x4)
hist(x5)
hist(x6)
par(mfrow=c(2,2))
hist(x7)
hist(x8)
hist(x9)
hist(x10)
##Density
par(mfrow=c(3,2))
plot(density(x1),type="l",xlab="x1",main="Density")
plot(density(x2),type="l",xlab="x2",main="Density")
plot(density(x3),type="l",xlab="x3",main="Density")
plot(density(x4),type="l",xlab="x4",main="Density")
plot(density(x5),type="l",xlab="x5",main="Density")
plot(density(x6),type="l",xlab="x6",main="Density")
par(mfrow=c(2,2))
plot(density(x7),type="l",xlab="x7",main="Density")
plot(density(x8),type="l",xlab="x8",main="Density")
plot(density(x9),type="l",xlab="x9",main="Density")
plot(density(x10),type="l",xlab="x10",main="Density")
##Scatter Plot Matrix
pairs(n10000)
##Correlation
cor(n10000)
####################
##FIT A BETA DISTRIBUTION for x1
```

```
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx1=vglm(x1 ~ 1, beta.ab, trace = TRUE)
coef(fitx1, matrix = TRUE)
Coef(fitx1)
summary(fitx1)

alpha1 = Coef(fitx1)[1]
beta1  = Coef(fitx1)[2]
y1 = seq(0,1,0.001)
dfBeta= pbeta(y1, shape1=alpha1, shape2=beta1)
dBeta = dbeta(y1, shape1=alpha1, shape2=beta1)
par(mfrow=c(1, 2))
plot(y1, dfBeta, cex=0.3, xlab="dist function")
plot(y1, dBeta,  cex=0.3, xlab="density function")
##############################
##FIT A BETA DISTRIBUTION for x2
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx2=vglm(x2 ~ 1, beta.ab, trace = TRUE)
coef(fitx2, matrix = TRUE)
Coef(fitx2)
summary(fitx2)

alpha2 = Coef(fitx2)[1]
beta2  = Coef(fitx2)[2]
y2 = seq(0,1,0.001)
dfBeta= pbeta(y2, shape1=alpha2, shape2=beta2)
dBeta = dbeta(y2, shape1=alpha2, shape2=beta2)
par(mfrow=c(1, 2))
plot(y2, dfBeta, cex=0.3, xlab="dist function")
plot(y2, dBeta,  cex=0.3, xlab="density function")
##############################
##FIT A BETA DISTRIBUTION for x3
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx3=vglm(x3 ~ 1, beta.ab, trace = TRUE)
coef(fitx3, matrix = TRUE)
Coef(fitx3)
summary(fitx3)

alpha3 = Coef(fitx3)[1]
beta3  = Coef(fitx3)[2]
```

```
y3 = seq(0,1,0.001)
dfBeta= pbeta(y3, shape1=alpha3, shape2=beta3)
dBeta = dbeta(y3, shape1=alpha3, shape2=beta3)
par(mfrow=c(1, 2))
plot(y3, dfBeta, cex=0.3, xlab="dist function")
plot(y3, dBeta,  cex=0.3, xlab="density function")
################################
##FIT A BETA DISTRIBUTION for x4
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx4=vglm(x4 ~ 1, beta.ab, trace = TRUE)
coef(fitx4, matrix = TRUE)
Coef(fitx4)
summary(fitx4)

alpha4 = Coef(fitx4)[1]
beta4  = Coef(fitx4)[2]
y4 = seq(0,1,0.001)
dfBeta= pbeta(y4, shape1=alpha4, shape2=beta4)
dBeta = dbeta(y4, shape1=alpha4, shape2=beta4)
par(mfrow=c(1, 2))
plot(y4, dfBeta, cex=0.3, xlab="dist function")
plot(y4, dBeta,  cex=0.3, xlab="density function")
################################
##FIT A BETA DISTRIBUTION for x5
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx5=vglm(x5 ~ 1, beta.ab, trace = TRUE)
coef(fitx5, matrix = TRUE)
Coef(fitx5)
summary(fitx5)

alpha5 = Coef(fitx5)[1]
beta5  = Coef(fitx5)[2]
y5 = seq(0,1,0.001)
dfBeta= pbeta(y5, shape1=alpha5, shape2=beta5)
dBeta = dbeta(y5, shape1=alpha5, shape2=beta5)
par(mfrow=c(1, 2))
plot(y5, dfBeta, cex=0.3, xlab="dist function")
plot(y5, dBeta,  cex=0.3, xlab="density function")
################################
##FIT A BETA DISTRIBUTION for x6
library(VGAM)
```

```
beta.ab(lshape1="identity", lshape2 = "identity")
fitx6=vglm(x6 ~ 1, beta.ab, trace = TRUE)
coef(fitx6, matrix = TRUE)
Coef(fitx6)
summary(fitx6)

alpha6 = Coef(fitx6)[1]
beta6  = Coef(fitx6)[2]
y6 = seq(0,1,0.001)
dfBeta= pbeta(y6, shape1=alpha6, shape2=beta6)
dBeta = dbeta(y6, shape1=alpha6, shape2=beta6)
par(mfrow=c(1, 2))
plot(y6, dfBeta, cex=0.3, xlab="dist function")
plot(y6, dBeta,  cex=0.3, xlab="density function")
################################
##FIT A BETA DISTRIBUTION for x7
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx7=vglm(x7 ~ 1, beta.ab, trace = TRUE)
coef(fitx7, matrix = TRUE)
Coef(fitx7)
summary(fitx7)

alpha7 = Coef(fitx7)[1]
beta7  = Coef(fitx7)[2]
y7 = seq(0,1,0.001)
dfBeta= pbeta(y7, shape1=alpha7, shape2=beta7)
dBeta = dbeta(y7, shape1=alpha7, shape2=beta7)
par(mfrow=c(1, 2))
plot(y7, dfBeta, cex=0.3, xlab="dist function")
plot(y7, dBeta,  cex=0.3, xlab="density function")
################################
##FIT A BETA DISTRIBUTION for x8
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx8=vglm(x8 ~ 1, beta.ab, trace = TRUE)
coef(fitx8, matrix = TRUE)
Coef(fitx8)
summary(fitx8)

alpha8 = Coef(fitx8)[1]
beta8  = Coef(fitx8)[2]
y8 = seq(0,1,0.001)
```

```
dfBeta= pbeta(y8, shape1=alpha8, shape2=beta8)
dBeta = dbeta(y8, shape1=alpha8, shape2=beta8)
par(mfrow=c(1, 2))
plot(y8, dfBeta, cex=0.3, xlab="dist function")
plot(y8, dBeta,  cex=0.3, xlab="density function")
################################
##FIT A BETA DISTRIBUTION for x9
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx9=vglm(x9 ~ 1, beta.ab, trace = TRUE)
coef(fitx9, matrix = TRUE)
Coef(fitx9)
summary(fitx9)

alpha9 = Coef(fitx9)[1]
beta9  = Coef(fitx9)[2]
y9 = seq(0,1,0.001)
dfBeta= pbeta(y9, shape1=alpha9, shape2=beta9)
dBeta = dbeta(y9, shape1=alpha9, shape2=beta9)
par(mfrow=c(1, 2))
plot(y9, dfBeta, cex=0.3, xlab="dist function")
plot(y9, dBeta,  cex=0.3, xlab="density function")
################################
##FIT A BETA DISTRIBUTION for x10
library(VGAM)
beta.ab(lshape1="identity", lshape2 = "identity")
fitx10=vglm(x10 ~ 1, beta.ab, trace = TRUE)
coef(fitx10, matrix = TRUE)
Coef(fitx10)
summary(fitx10)

alpha10 = Coef(fitx10)[1]
beta10  = Coef(fitx10)[2]
y10 = seq(0,1,0.001)
dfBeta= pbeta(y10, shape1=alpha10, shape2=beta10)
dBeta = dbeta(y10, shape1=alpha10, shape2=beta10)
par(mfrow=c(1, 2))
plot(y10, dfBeta, cex=0.3, xlab="dist function")
plot(y10, dBeta,  cex=0.3, xlab="density function")
################################
##PROBABILITY INTEGRAL TRANSFORM
u1= pbeta(x1, shape1=alpha1, shape2=beta1)
par(mfrow=c(1, 2))
```

```
hist(u1, main="", xlab="Histogram of Transformed x1")
Fn <- ecdf(x1)
lines(y1, Fn(y1), lty=2)
plot(Fn(y1),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u2= pbeta(x2, shape1=alpha2, shape2=beta2)
par(mfrow=c(1, 2))
hist(u2, main="", xlab="Histogram of Transformed x2")
Fn <- ecdf(x2)
lines(y2, Fn(y2), lty=2)
plot(Fn(y2),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u3= pbeta(x3, shape1=alpha3, shape2=beta3)
par(mfrow=c(1, 2))
hist(u3, main="", xlab="Histogram of Transformed x3")
Fn <- ecdf(x3)
lines(y3, Fn(y3), lty=2)
plot(Fn(y3),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u4= pbeta(x4, shape1=alpha4, shape2=beta4)
par(mfrow=c(1, 2))
hist(u4, main="", xlab="Histogram of Transformed x4")
Fn <- ecdf(x4)
lines(y4, Fn(y4), lty=2)
plot(Fn(y4),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u5= pbeta(x5, shape1=alpha5, shape2=beta5)
par(mfrow=c(1, 2))
hist(u5, main="", xlab="Histogram of Transformed x5")
Fn <- ecdf(x5)
lines(y5, Fn(y5), lty=2)
plot(Fn(y5),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u6= pbeta(x6, shape1=alpha6, shape2=beta6)
par(mfrow=c(1, 2))
hist(u6, main="", xlab="Histogram of Transformed x6")
Fn <- ecdf(x6)
lines(y6, Fn(y6), lty=2)
```

```
plot(Fn(y6),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u7= pbeta(x7, shape1=alpha7, shape2=beta7)
par(mfrow=c(1, 2))
hist(u7, main="", xlab="Histogram of Transformed x7")
Fn <- ecdf(x7)
lines(y7, Fn(y7), lty=2)
plot(Fn(y7),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u8= pbeta(x8, shape1=alpha8, shape2=beta8)
par(mfrow=c(1, 2))
hist(u8, main="", xlab="Histogram of Transformed x8")
Fn <- ecdf(x8)
lines(y8, Fn(y8), lty=2)
plot(Fn(y8),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u9= pbeta(x9, shape1=alpha9, shape2=beta9)
par(mfrow=c(1, 2))
hist(u9, main="", xlab="Histogram of Transformed x9")
Fn <- ecdf(x9)
lines(y9, Fn(y9), lty=2)
plot(Fn(y9),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u10= pbeta(x10, shape1=alpha10, shape2=beta10)
par(mfrow=c(1, 2))
hist(u10, main="", xlab="Histogram of Transformed x10")
Fn <- ecdf(x10)
lines(y10, Fn(y10), lty=2)
plot(Fn(y10),dfBeta, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)
################################

uu = cbind(u1,u2,u3,u4,u5,u6,u7,u8,u9,u10)

# select the D-vine families and parameters
dvine = CDVineCopSelect(uu,c(1:6),type="DVine")
dvine

# simulate from a bivariate copula
```

```
# pair: (1,2) with Gumbel pair-copulas
fam12 = 4
par12 = 1.10620051
n = 10000
simdat12 = BiCopSim(n,fam12,par12)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat12[,1],simdat12[,2])

# pair: (2,3) with Gumbel pair-copulas
fam23 = 4
par23 = 1.10632430
n = 10000
simdat23 = BiCopSim(n,fam23,par23)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat23[,1],simdat23[,2])

# pair: (3,4) with Gumbel pair-copulas
fam34 = 4
par34 = 1.10629388
n = 10000
simdat34 = BiCopSim(n,fam34,par34)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat34[,1],simdat34[,2])

# pair: (4,5) with Gumbel pair-copulas
fam45 = 4
par45 = 1.10634391
n = 10000
simdat45 = BiCopSim(n,fam45,par45)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat45[,1],simdat45[,2])

# pair: (5,6) with Gumbel pair-copulas
fam56 = 4
par56 = 1.10638950
```

```
n = 10000
simdat56 = BiCopSim(n,fam56,par56)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat56[,1],simdat56[,2])

# pair: (6,7) with Gumbel pair-copulas
fam67 = 4
par67 = 1.10640111
n = 10000
simdat67 = BiCopSim(n,fam67,par67)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat67[,1],simdat67[,2])

# pair: (7,8) with Gumbel pair-copulas
fam78 = 4
par78 = 1.10632263
n = 10000
simdat78 = BiCopSim(n,fam78,par78)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat78[,1],simdat78[,2])

# pair: (8,9) with Gumbel pair-copulas
fam89 = 4
par89 = 1.10634497
n = 10000
simdat89 = BiCopSim(n,fam89,par89)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat89[,1],simdat89[,2])

# pair: (9,10) with Gumbel pair-copulas
fam910 = 4
par910 = 1.10617025
n = 10000
simdat910 = BiCopSim(n,fam910,par910)
```

```
# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat910[,1],simdat910[,2])

# pair: 13..2 with Gaussian pair-copulas
fam13..2 = 1
par13..2 = 0.07998308
n = 10000
simdat13..2 = BiCopSim(n,fam13..2,par13..2)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat13..2[,1],simdat13..2[,2])

# pair: 24..3 with Gaussian pair-copulas
fam24..3 = 1
par24..3 = 0.07979491
n = 10000
simdat24..3 = BiCopSim(n,fam24..3,par24..3)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat24..3[,1],simdat24..3[,2])

# pair: 35..4 with Gaussian pair-copulas
fam35..4 = 1
par35..4 = 0.07978876
n = 10000
simdat35..4 = BiCopSim(n,fam35..4,par35..4)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat35..4[,1],simdat35..4[,2])

# pair: 46..5 with Gaussian pair-copulas
fam46..5 = 1
par46..5 = 0.07986084
n = 10000
simdat46..5 = BiCopSim(n,fam46..5,par46..5)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat46..5[,1],simdat46..5[,2])
```

```
# pair: 57..6 with Gaussian pair-copulas
fam57..6 = 1
par57..6 = 0.07993519
n = 10000
simdat57..6 = BiCopSim(n,fam57..6,par57..6)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat57..6[,1],simdat57..6[,2])

# pair: 68..7 with Gaussian pair-copulas
fam68..7 = 1
par68..7 = 0.07986309
n = 10000
simdat68..7 = BiCopSim(n,fam68..7,par68..7)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat68..7[,1],simdat68..7[,2])

# pair: 79..8 with Gaussian pair-copulas
fam79..8 = 1
par79..8 = 0.08004171
n = 10000
simdat79..8 = BiCopSim(n,fam79..8,par79..8)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat79..8[,1],simdat79..8[,2])

# pair: 810..9 with Gaussian pair-copulas
fam810..9 = 1
par810..9 = 0.08037421
n = 10000
simdat810..9 = BiCopSim(n,fam810..9,par810..9)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat810..9[,1],simdat810..9[,2])


# pair: 14..23 with Gaussian pair-copulas
```

```
fam14..23 = 1
par14..23 = 0.04525652
n = 10000
simdat14..23 = BiCopSim(n,fam14..23,par14..23)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat14..23[,1],simdat14..23[,2])

# pair: 25..34 with Gaussian pair-copulas
fam25..34 = 1
par25..34 = 0.04513136
n = 10000
simdat25..34 = BiCopSim(n,fam25..34,par25..34)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat25..34[,1],simdat25..34[,2])

# pair: 36..45 with Gaussian pair-copulas
fam36..45 = 1
par36..45 = 0.04505036
n = 10000
simdat36..45 = BiCopSim(n,fam36..45,par36..45)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat36..45[,1],simdat36..45[,2])

# pair: 47..56 with Gaussian pair-copulas
fam47..56 = 1
par47..56 = 0.04510724
n = 10000
simdat47..56 = BiCopSim(n,fam47..56,par47..56)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat47..56[,1],simdat47..56[,2])

# pair: 58..67 with Gaussian pair-copulas
fam58..67 = 1
par58..67 = 0.04506394
n = 10000
```

```
simdat58..67 = BiCopSim(n,fam58..67,par58..67)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat58..67[,1],simdat58..67[,2])

# pair: 69..78 with Gaussian pair-copulas
fam69..78 = 1
par69..78 = 0.04513326
n = 10000
simdat69..78 = BiCopSim(n,fam69..78,par69..78)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat69..78[,1],simdat69..78[,2])

# pair: 710..89 with Gaussian pair-copulas
fam710..89 = 1
par710..89 = 0.04527113
n = 10000
simdat710..89 = BiCopSim(n,fam710..89,par710..89)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat710..89[,1],simdat710..89[,2])


# pair: 15..234 with Frank pair-copulas
fam15..234 = 5
par15..234 = 0.27416759
n = 10000
simdat15..234 = BiCopSim(n,fam15..234,par15..234)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat15..234[,1],simdat15..234[,2])

# pair: 26..345 with Frank pair-copulas
fam26..345 = 5
par26..345 = 0.27429153
n = 10000
simdat26..345 = BiCopSim(n,fam26..345,par26..345)
```

```
# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat26..345[,1],simdat26..345[,2])

# pair: 37..456 with Frank pair-copulas
fam37..456 = 5
par37..456 = 0.27312874
n = 10000
simdat37..456 = BiCopSim(n,fam37..456,par37..456)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat37..456[,1],simdat37..456[,2])

# pair: 48..567 with Frank pair-copulas
fam48..567 = 5
par48..567 = 0.27198212
n = 10000
simdat48..567 = BiCopSim(n,fam48..567,par48..567)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat48..567[,1],simdat48..567[,2])

# pair: 59..678 with Frank pair-copulas
fam59..678 = 5
par59..678 = 0.27173976
n = 10000
simdat59..678 = BiCopSim(n,fam59..678,par59..678)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat59..678[,1],simdat59..678[,2])

# pair: 610..789 with Frank pair-copulas
fam610..789 = 5
par610..789 = 0.27159698
n = 10000
simdat610..789 = BiCopSim(n,fam610..789,par610..789)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat610..789[,1],simdat610..789[,2])
```

```
# pair: 16..2345 with Gumbel pair-copulas
fam16..2345 = 4
par16..2345 = 1.01457067
n = 10000
simdat16..2345 = BiCopSim(n,fam16..2345,par16..2345)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat16..2345[,1],simdat16..2345[,2])

# pair: 27..3456 with Gumbel pair-copulas
fam27..3456 = 4
par27..3456 = 1.01460183
n = 10000
simdat27..3456 = BiCopSim(n,fam27..3456,par27..3456)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat27..3456[,1],simdat27..3456[,2])

# pair: 38..4567 with Gumbel pair-copulas
fam38..4567 = 4
par38..4567 = 1.01461597
n = 10000
simdat38..4567 = BiCopSim(n,fam38..4567,par38..4567)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat38..4567[,1],simdat38..4567[,2])

# pair: 49..5678 with Gumbel pair-copulas
fam49..5678 = 4
par49..5678 = 1.01463313
n = 10000
simdat49..5678 = BiCopSim(n,fam49..5678,par49..5678)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat49..5678[,1],simdat49..5678[,2])

# pair: 510..6789 with Gumbel pair-copulas
fam510..6789 = 4
```

```
par510..6789 = 1.01460252
n = 10000
simdat510..6789 = BiCopSim(n,fam510..6789,par510..6789)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat510..6789[,1],simdat510..6789[,2])

# pair: 17.23456 with Gaussian pair-copulas
fam17..23456 = 1
par17..23456 = 0.04862847
n = 10000
simdat17..23456 = BiCopSim(n,fam17..23456,par17..23456)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat17..23456[,1],simdat17..23456[,2])

# pair: 28..34567 with Gaussian pair-copulas
fam28..34567 = 1
par28..34567 = 0.04883379
n = 10000
simdat28..34567 = BiCopSim(n,fam28..34567,par28..34567)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat28..34567[,1],simdat28..34567[,2])

# pair: 39..45678 with Gaussian pair-copulas
fam39..45678 = 1
par39..45678 = 0.04895098
n = 10000
simdat39..45678 = BiCopSim(n,fam39..45678,par39..45678)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat39..45678[,1],simdat39..45678[,2])

# pair: 410..56789 with Gaussian pair-copulas
fam410..56789 = 1
par410..56789 = 0.04870248
n = 10000
simdat410..56789 = BiCopSim(n,fam410..56789,par410..56789)
```

```
# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat410..56789[,1],simdat410..56789[,2])

# pair: 18..234567 with Gaussian pair-copulas
fam18..234567 = 1
par18..234567 = 0.02542544
n = 10000
simdat18..234567 = BiCopSim(n,fam18..234567,par18..234567)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat18..234567[,1],simdat18..234567[,2])

# pair: 29..345678 with Gaussian pair-copulas
fam29..345678 = 1
par29..345678 = 0.02516106
n = 10000
simdat29..345678 = BiCopSim(n,fam29..345678,par29..345678)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat29..345678[,1],simdat29..345678[,2])

# pair: 310..456789 with Gaussian pair-copulas
fam310..456789 = 1
par310..456789 = 0.02550702
n = 10000
simdat310..456789 = BiCopSim(n,fam310..456789,par310..456789)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat310..456789[,1],simdat310..456789[,2])

# pair: 19..2345678 with Frank pair-copulas
fam19..2345678 = 5
par19..2345678 = 0.11288487
n = 10000
simdat19..2345678 = BiCopSim(n,fam19..2345678,par19..2345678)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
```

```
BiCopIndTest(simdat19..2345678[,1],simdat19..2345678[,2])

# pair: 210..3456789 with Frank pair-copulas
fam210..3456789 = 5
par210..3456789 = 0.11242082
n = 10000
simdat210..3456789 = BiCopSim(n,fam210..3456789,par210..3456789)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat210..3456789[,1],simdat210..3456789[,2])

# pair: 110..23456789 with Frank pair-copulas
fam110..23456789 = 5
par110..23456789 = 0.16939812
n = 10000
simdat110..23456789 = BiCopSim(n,fam110..23456789,par110..23456789)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat110..23456789[,1],simdat110..23456789[,2])
```

C.2.2 Amplitude Dataset

- Dimension = 5

```
##dimension=5
##Histogram
par(mfrow=c(3,2))
hist(x1)
hist(x2)
hist(x3)
hist(x4)
hist(x5)
##Density
par(mfrow=c(3,2))
plot(density(x1),type="l",xlab="x1",main="Density")
plot(density(x2),type="l",xlab="x2",main="Density")
plot(density(x3),type="l",xlab="x3",main="Density")
plot(density(x4),type="l",xlab="x4",main="Density")
plot(density(x5),type="l",xlab="x5",main="Density")
##Scatter Plot Matrix
pairs(n10000)
```

```
##Correlation
cor(n10000)
####################
##Plots of distribution function and density function
y = seq(0,1,0.001)
dfUniform = punif(y,0,1)
dUniform = dunif(y,0,1)
par(mfrow=c(1, 2))
plot(y, dfUniform, cex=0.3, xlab="dist function")
plot(y, dUniform,  cex=0.3, xlab="density function")
############################
##PROBABILITY INTEGRAL TRANSFORM
u1= punif(x1,0,1)
par(mfrow=c(1, 2))
hist(u1, main="", xlab="Histogram of Transformed x1")
Fn <- ecdf(x1)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u2= punif(x2,0,1)
par(mfrow=c(1, 2))
hist(u2, main="", xlab="Histogram of Transformed x2")
Fn <- ecdf(x2)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u3= punif(x3,0,1)
par(mfrow=c(1, 2))
hist(u3, main="", xlab="Histogram of Transformed x3")
Fn <- ecdf(x3)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u4= punif(x4,0,1)
par(mfrow=c(1, 2))
hist(u4, main="", xlab="Histogram of Transformed x4")
Fn <- ecdf(x4)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)
```

```
u5= punif(x5,0,1)
par(mfrow=c(1, 2))
hist(u5, main="", xlab="Histogram of Transformed x5")
Fn <- ecdf(x5)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)
########################

uu = cbind(u1,u2,u3,u4,u5)

# select the D-vine families and parameters
dvine = CDVineCopSelect(uu,c(1:6),type="DVine")
dvine

# simulate from a bivariate copula
# pair: (1,2) with Gaussian pair-copulas
fam12 = 1
par12 = 0.05849204
n = 10000
simdat12 = BiCopSim(n,fam12,par12)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat12[,1],simdat12[,2])

# create chi-plot, K-plot and lambda-function plots
dev.new(width=16,height=5)
par(mfrow=c(1,3))
BiCopChiPlot(simdat12[,1],simdat12[,2],xlim=c(-1,1),ylim=c(-1,1),
main="General chi-plot")
BiCopKPlot(simdat12[,1],simdat12[,2],main="Gaussian copula")
BiCopLambda(simdat12[,1],simdat12[,2],family=fam12,par=par12)

# pair: (2,3) with Gaussian pair-copulas
fam23 = 1
par23 = 0.05874221
n = 10000
simdat23 = BiCopSim(n,fam23,par23)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
```

```
BiCopIndTest(simdat23[,1],simdat23[,2])

# pair: (3,4) with Gaussian pair-copulas
fam34 = 1
par34 = 0.05841567
n = 10000
simdat34 = BiCopSim(n,fam34,par34)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat34[,1],simdat34[,2])

# pair: (4,5) with Gaussian pair-copulas
fam45 = 1
par45 = 0.05840175
n = 10000
simdat45 = BiCopSim(n,fam45,par45)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat45[,1],simdat45[,2])

# pair: 13..2 with Joe pair-copulas
fam13..2 = 6
par13..2 = 1.03172500
n = 10000
simdat13..2 = BiCopSim(n,fam13..2,par13..2)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat13..2[,1],simdat13..2[,2])

# pair: 24..3 with Joe pair-copulas
fam24..3 = 6
par24..3 = 1.03167175
n = 10000
simdat24..3 = BiCopSim(n,fam24..3,par24..3)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat24..3[,1],simdat24..3[,2])

# pair: 35..4 with Joe pair-copulas
```

```
fam35..4 = 6
par35..4 = 1.03168780
n = 10000
simdat35..4 = BiCopSim(n,fam35..4,par35..4)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat35..4[,1],simdat35..4[,2])

# pair: 14..23 with Frank pair-copulas
fam14..23 = 5
par14..23 = 0.16230619
n = 10000
simdat14..23 = BiCopSim(n,fam14..23,par14..23)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat14..23[,1],simdat14..23[,2])

# pair: 25..34 with Frank pair-copulas
fam25..34 = 5
par25..34 = 0.16139709
n = 10000
simdat25..34 = BiCopSim(n,fam25..34,par25..34)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat25..34[,1],simdat25..34[,2])

# pair: 15..234 with Gumbel pair-copulas
fam15..234 = 4
par15..234 = 1.01272526
n = 10000
simdat15..234 = BiCopSim(n,fam15..234,par15..234)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat15..234[,1],simdat15..234[,2])
```

- Dimension = 10

```
##dimension=10
##Histogram
```

```
par(mfrow=c(3,2))
hist(x1)
hist(x2)
hist(x3)
hist(x4)
hist(x5)
hist(x6)
par(mfrow=c(2,2))
hist(x7)
hist(x8)
hist(x9)
hist(x10)
##Density
par(mfrow=c(3,2))
plot(density(x1),type="l",xlab="x1",main="Density")
plot(density(x2),type="l",xlab="x2",main="Density")
plot(density(x3),type="l",xlab="x3",main="Density")
plot(density(x4),type="l",xlab="x4",main="Density")
plot(density(x5),type="l",xlab="x5",main="Density")
plot(density(x6),type="l",xlab="x6",main="Density")
par(mfrow=c(2,2))
plot(density(x7),type="l",xlab="x7",main="Density")
plot(density(x8),type="l",xlab="x8",main="Density")
plot(density(x9),type="l",xlab="x9",main="Density")
plot(density(x10),type="l",xlab="x10",main="Density")
##Scatter Plot Matrix
pairs(n10000)
##Correlation
cor(n10000)
####################
##Plots of distribution function and density function
y = seq(0,1,0.001)
dfUniform = punif(y,0,1)
dUniform = dunif(y,0,1)
par(mfrow=c(1, 2))
plot(y, dfUniform, cex=0.3, xlab="dist function")
plot(y, dUniform,  cex=0.3, xlab="density function")
############################
##PROBABILITY INTEGRAL TRANSFORM
u1= punif(x1,0,1)
par(mfrow=c(1, 2))
hist(u1, main="", xlab="Histogram of Transformed x1")
Fn <- ecdf(x1)
```

```
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u2= punif(x2,0,1)
par(mfrow=c(1, 2))
hist(u2, main="", xlab="Histogram of Transformed x2")
Fn <- ecdf(x2)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u3= punif(x3,0,1)
par(mfrow=c(1, 2))
hist(u3, main="", xlab="Histogram of Transformed x3")
Fn <- ecdf(x3)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u4= punif(x4,0,1)
par(mfrow=c(1, 2))
hist(u4, main="", xlab="Histogram of Transformed x4")
Fn <- ecdf(x4)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u5= punif(x5,0,1)
par(mfrow=c(1, 2))
hist(u5, main="", xlab="Histogram of Transformed x5")
Fn <- ecdf(x5)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u6= punif(x6,0,1)
par(mfrow=c(1, 2))
hist(u6, main="", xlab="Histogram of Transformed x6")
Fn <- ecdf(x6)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)
```

```
u7= punif(x7,0,1)
par(mfrow=c(1, 2))
hist(u7, main="", xlab="Histogram of Transformed x7")
Fn <- ecdf(x7)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u8= punif(x8,0,1)
par(mfrow=c(1, 2))
hist(u8, main="", xlab="Histogram of Transformed x8")
Fn <- ecdf(x8)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u9= punif(x9,0,1)
par(mfrow=c(1, 2))
hist(u9, main="", xlab="Histogram of Transformed x9")
Fn <- ecdf(x9)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)

u10= punif(x10,0,1)
par(mfrow=c(1, 2))
hist(u10, main="", xlab="Histogram of Transformed x10")
Fn <- ecdf(x10)
lines(y, Fn(y), lty=2)
plot(Fn(y),dfUniform, cex=0.3, xlab="Empirical DF")
abline(a=0,b=1)
########################

uu = cbind(u1,u2,u3,u4,u5,u6,u7,u8,u9,u10)

# select the D-vine families and parameters
dvine = CDVineCopSelect(uu,c(1:6),type="DVine")
dvine

# simulate from a bivariate copula
# pair: (1,2) with Gaussian pair-copulas
fam12 = 1
```

```
par12 = 0.05849204
n = 10000
simdat12 = BiCopSim(n,fam12,par12)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat12[,1],simdat12[,2])

# pair: (2,3) with Gaussian pair-copulas
fam23 = 1
par23 = 0.05874221
n = 10000
simdat23 = BiCopSim(n,fam23,par23)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat23[,1],simdat23[,2])

# pair: (3,4) with Gaussian pair-copulas
fam34 = 1
par34 = 0.05841567
n = 10000
simdat34 = BiCopSim(n,fam34,par34)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat34[,1],simdat34[,2])

# pair: (4,5) with Gaussian pair-copulas
fam45 = 1
par45 = 0.05840175
n = 10000
simdat45 = BiCopSim(n,fam45,par45)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat45[,1],simdat45[,2])

# pair: (5,6) with Gaussian pair-copulas
fam56 = 1
par56 = 0.05827107
n = 10000
simdat56 = BiCopSim(n,fam56,par56)
```

```
# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat56[,1],simdat56[,2])

# pair: (6,7) with Gaussian pair-copulas
fam67 = 1
par67 = 0.05826967
n = 10000
simdat67 = BiCopSim(n,fam67,par67)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat67[,1],simdat67[,2])

# pair: (7,8) with Gaussian pair-copulas
fam78 = 1
par78 = 0.05821660
n = 10000
simdat78 = BiCopSim(n,fam78,par78)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat78[,1],simdat78[,2])

# pair: (8,9) with Gaussian pair-copulas
fam89 = 1
par89 = 0.05840795
n = 10000
simdat89 = BiCopSim(n,fam89,par89)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat89[,1],simdat89[,2])

# pair: (9,10) with Gaussian pair-copulas
fam910 = 1
par910 = 0.05841636
n = 10000
simdat910 = BiCopSim(n,fam910,par910)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
```

```
BiCopIndTest(simdat910[,1],simdat910[,2])

# pair: 13..2 with Joe pair-copulas
fam13..2 = 6
par13..2 = 1.03172500
n = 10000
simdat13..2 = BiCopSim(n,fam13..2,par13..2)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat13..2[,1],simdat13..2[,2])

# pair: 24..3 with Joe pair-copulas
fam24..3 = 6
par24..3 = 1.03167175
n = 10000
simdat24..3 = BiCopSim(n,fam24..3,par24..3)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat24..3[,1],simdat24..3[,2])

# pair: 35..4 with Joe pair-copulas
fam35..4 = 6
par35..4 = 1.03168780
n = 10000
simdat35..4 = BiCopSim(n,fam35..4,par35..4)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat35..4[,1],simdat35..4[,2])

# pair: 46..5 with Joe pair-copulas
fam46..5 = 6
par46..5 = 1.03165364
n = 10000
simdat46..5 = BiCopSim(n,fam46..5,par46..5)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat46..5[,1],simdat46..5[,2])

# pair: 57..6 with Joe pair-copulas
```

```
fam57..6 = 6
par57..6 = 1.03165286
n = 10000
simdat57..6 = BiCopSim(n,fam57..6,par57..6)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat57..6[,1],simdat57..6[,2])

# pair: 68..7 with Joe pair-copulas
fam68..7 = 6
par68..7 = 1.03158931
n = 10000
simdat68..7 = BiCopSim(n,fam68..7,par68..7)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat68..7[,1],simdat68..7[,2])

# pair: 79..8 with Joe pair-copulas
fam79..8 = 6
par79..8 = 1.03158954
n = 10000
simdat79..8 = BiCopSim(n,fam79..8,par79..8)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat79..8[,1],simdat79..8[,2])

# pair: 810..9 with Joe pair-copulas
fam810..9 = 6
par810..9 = 1.03161729
n = 10000
simdat810..9 = BiCopSim(n,fam810..9,par810..9)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat810..9[,1],simdat810..9[,2])

# pair: 14..23 with Frank pair-copulas
fam14..23 = 5
par14..23 = 0.16230619
n = 10000
```

```
simdat14..23 = BiCopSim(n,fam14..23,par14..23)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat14..23[,1],simdat14..23[,2])

# pair: 25..34 with Frank pair-copulas
fam25..34 = 5
par25..34 = 0.16139709
n = 10000
simdat25..34 = BiCopSim(n,fam25..34,par25..34)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat25..34[,1],simdat25..34[,2])

# pair: 36..45 with Frank pair-copulas
fam36..45 = 5
par36..45 = 0.16297454
n = 10000
simdat36..45 = BiCopSim(n,fam36..45,par36..45)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat36..45[,1],simdat36..45[,2])

# pair: 47..56 with Frank pair-copulas
fam47..56 = 5
par47..56 = 0.16247139
n = 10000
simdat47..56 = BiCopSim(n,fam47..56,par47..56)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat47..56[,1],simdat47..56[,2])

# pair: 58..67 with Frank pair-copulas
fam58..67 = 5
par58..67 = 0.16309286
n = 10000
simdat58..67 = BiCopSim(n,fam58..67,par58..67)

# BiCopIndTest Independence test for bivariate copula data
```

```
# perform the asymptotic independence test
BiCopIndTest(simdat58..67[,1],simdat58..67[,2])

# pair: 69..78 with Frank pair-copulas
fam69..78 = 5
par69..78 = 0.16203363
n = 10000
simdat69..78 = BiCopSim(n,fam69..78,par69..78)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat69..78[,1],simdat69..78[,2])

# pair: 710..89 with Frank pair-copulas
fam710..89 = 5
par710..89 = 0.16246196
n = 10000
simdat710..89 = BiCopSim(n,fam710..89,par710..89)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat710..89[,1],simdat710..89[,2])

# pair: 15..234 with Gumbel pair-copulas
fam15..234 = 4
par15..234 = 1.01272526
n = 10000
simdat15..234 = BiCopSim(n,fam15..234,par15..234)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat15..234[,1],simdat15..234[,2])

# pair: 26..345 with Gumbel pair-copulas
fam26..345 = 4
par26..345 = 1.01284587
n = 10000
simdat26..345 = BiCopSim(n,fam26..345,par26..345)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat26..345[,1],simdat26..345[,2])
```

```
# pair: 37..456 with Gumbel pair-copulas
fam37..456 = 4
par37..456 = 1.01285933
n = 10000
simdat37..456 = BiCopSim(n,fam37..456,par37..456)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat37..456[,1],simdat37..456[,2])

# pair: 48..567 with Gumbel pair-copulas
fam48..567 = 4
par48..567 = 1.01285840
n = 10000
simdat48..567 = BiCopSim(n,fam48..567,par48..567)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat48..567[,1],simdat48..567[,2])

# pair: 59..678 with Gumbel pair-copulas
fam59..678 = 4
par59..678 = 1.01288600
n = 10000
simdat59..678 = BiCopSim(n,fam59..678,par59..678)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat59..678[,1],simdat59..678[,2])

# pair: 610..789 with Gumbel pair-copulas
fam610..789 = 4
par610..789 = 1.01286531
n = 10000
simdat610..789 = BiCopSim(n,fam610..789,par610..789)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat610..789[,1],simdat610..789[,2])

# pair: 16..2345 with Gaussian pair-copulas
fam16..2345 = 1
par16..2345 = 0.03566179
```

```
n = 10000
simdat16..2345 = BiCopSim(n,fam16..2345,par16..2345)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat16..2345[,1],simdat16..2345[,2])

# pair: 27..3456 with Gaussian pair-copulas
fam27..3456 = 1
par27..3456 = 0.03565798
n = 10000
simdat27..3456 = BiCopSim(n,fam27..3456,par27..3456)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat27..3456[,1],simdat27..3456[,2])

# pair: 38..4567 with Gaussian pair-copulas
fam38..4567 = 1
par38..4567 = 0.03525427
n = 10000
simdat38..4567 = BiCopSim(n,fam38..4567,par38..4567)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat38..4567[,1],simdat38..4567[,2])

# pair: 49..5678 with Gaussian pair-copulas
fam49..5678 = 1
par49..5678 = 0.03527585
n = 10000
simdat49..5678 = BiCopSim(n,fam49..5678,par49..5678)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat49..5678[,1],simdat49..5678[,2])

# pair: 510..6789 with Gaussian pair-copulas
fam510..6789 = 1
par510..6789 = 0.03530207
n = 10000
simdat510..6789 = BiCopSim(n,fam510..6789,par510..6789)
```

```
# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat510..6789[,1],simdat510..6789[,2])

# pair: 17.23456 with Joe pair-copulas
fam17..23456 = 6
par17..23456 = 1.01995001
n = 10000
simdat17..23456 = BiCopSim(n,fam17..23456,par17..23456)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat17..23456[,1],simdat17..23456[,2])

# pair: 28..34567 with Joe pair-copulas
fam28..34567 = 6
par28..34567 = 1.01988375
n = 10000
simdat28..34567 = BiCopSim(n,fam28..34567,par28..34567)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat28..34567[,1],simdat28..34567[,2])

# pair: 39..45678 with Joe pair-copulas
fam39..45678 = 6
par39..45678 = 1.01992915
n = 10000
simdat39..45678 = BiCopSim(n,fam39..45678,par39..45678)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat39..45678[,1],simdat39..45678[,2])

# pair: 410..56789 with Joe pair-copulas
fam410..56789 = 6
par410..56789 = 1.01997021
n = 10000
simdat410..56789 = BiCopSim(n,fam410..56789,par410..56789)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat410..56789[,1],simdat410..56789[,2])
```

```
# pair: 18..234567 with Student-t pair-copulas
fam18..234567 = 2
par18..234567 = 0.01457066
par2 = 29.92050
n = 10000
simdat18..234567 = BiCopSim(n,fam18..234567,par18..234567,par2)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat18..234567[,1],simdat18..234567[,2])

# pair: 29..345678 with Student-t pair-copulas
fam29..345678 = 2
par29..345678 = 0.01438608
par2 = 29.90794
n = 10000
simdat29..345678 = BiCopSim(n,fam29..345678,par29..345678,par2)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat29..345678[,1],simdat29..345678[,2])

# pair: 310..456789 with Student-t pair-copulas
fam310..456789 = 2
par310..456789 = 0.01460214
par2 = 29.79871
n = 10000
simdat310..456789 = BiCopSim(n,fam310..456789,par310..456789,par2)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat310..456789[,1],simdat310..456789[,2])

# pair: 19..2345678 with Clayton pair-copulas
fam19..2345678 = 3
par19..2345678 = 0.01990886
n = 10000
simdat19..2345678 = BiCopSim(n,fam19..2345678,par19..2345678)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat19..2345678[,1],simdat19..2345678[,2])
```

```
# pair: 210..3456789 with Clayton pair-copulas
fam210..3456789 = 3
par210..3456789 = 0.01994536
n = 10000
simdat210..3456789 = BiCopSim(n,fam210..3456789,par210..3456789)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat210..3456789[,1],simdat210..3456789[,2])

# pair: 110..23456789 with Clayton pair-copulas
fam110..23456789 = 3
par110..23456789 = 0.03136980
n = 10000
simdat110..23456789 = BiCopSim(n,fam110..23456789,par110..23456789)

# BiCopIndTest Independence test for bivariate copula data
# perform the asymptotic independence test
BiCopIndTest(simdat110..23456789[,1],simdat110..23456789[,2])
```

C.3 Comparing Correlations

C.3.1 Original Dataset

```
d=cbind(mat[,6928:6929],mat[,6997:7011],mat[,7013:7018],mat[,7027:7050],
  mat[,7117:7134],mat[,7136:7139],mat[,7150:7174], mat[,7241:7246])

uncon=array(0,dim=c(76,76,100))
cc=array(0,dim=c(76,76,100))
p1=array(0,dim=c(76,76,100))
p2=array(0,dim=c(76,76,100))
z1=array(0,dim=c(76,76,100))
r_uncon=array(0,dim=c(76,76,100))
r_cc=array(0,dim=c(76,76,100))
max_abs_z1=matrix(0,100,1)

for (i in 1:76)
{
for (j in 1:76)
{
for (k in 1:100)
{
```

```
x=sinpsi[,i]
y=sinpsi[,j]
z=d[,k]
xy=cbind(x,y)
xyz=cbind(x,y,z)
uncon[i,j,k]=cor(x,y,method="kendall")
cc[i,j,k]=cor(x[z > 0.2], y[z > 0.2],method="kendall")
p1[i,j,k]=dim(xy)[1]
p2[i,j,k]=dim(xyz[which(xyz[,3] > 0.2),])[1]
r_uncon[i,j,k] = 0.5*log((1 + uncon[i,j,k])/(1 - uncon[i,j,k]))
r_cc[i,j,k] = 0.5*log((1 + cc[i,j,k])/(1 - cc[i,j,k]))
z1[i,j,k] = (r_uncon[i,j,k] - r_cc[i,j,k])/sqrt((1/(p1[i,j,k] - 3)) +
                            (1/(p2[i,j,k] - 3)))
z1[is.na(z1)] <- 0
z1[is.infinite(z1)] <- 0
max_abs_z1[k]=max(abs(z1[,,k]),na.rm=TRUE)
}
}
}
hydro0_5_10=max_abs_z1
n_hydro0_5_10=cbind(p2[76,76,],hydro0_5_10)
```

C.3.2 Random Dataset

```
d=cbind(mat[,2022:2037],mat[,2039:2042],mat[,2129:2164],mat[,2245:2247],
              mat[,2251:2286],mat[,2367:2369],mat[,2371:2372])

uncon=array(0,dim=c(76,76,100))
cc=array(0,dim=c(76,76,100))
p1=array(0,dim=c(76,76,100))
p2=array(0,dim=c(76,76,100))
z1=array(0,dim=c(76,76,100))
r_uncon=array(0,dim=c(76,76,100))
r_cc=array(0,dim=c(76,76,100))
max_abs_z1=matrix(0,100,1)

for (i in 1:76)
{
for (j in 1:76)
{
for (k in 1:100)
{
x=sinpsi[,i]
y=sinpsi[,j]
```

```
z=d[,k]
xy=cbind(x,y)
xyz=cbind(x,y,z)
uncon[i,j,k]=cor(x,y,method="kendall")
cc[i,j,k]=cor(x[z > 0.2], y[z > 0.2],method="kendall")
p1[i,j,k]=dim(xy)[1]
p2[i,j,k]=dim(xyz[which(xyz[,3] > 0.2),])[1]
r_uncon[i,j,k] = 0.5*log((1 + uncon[i,j,k])/(1 - uncon[i,j,k]))
r_cc[i,j,k] = 0.5*log((1 + cc[i,j,k])/(1 - cc[i,j,k]))
z1[i,j,k] = (r_uncon[i,j,k] - r_cc[i,j,k])/sqrt((1/(p1[i,j,k] - 3)) +
                          (1/(p2[i,j,k] - 3)))
z1[is.na(z1)] <- 0
z1[is.infinite(z1)] <- 0
max_abs_z1[k]=max(abs(z1[,,k]),na.rm=TRUE)
}
}
}
hydro0_6_3=max_abs_z1
n_hydro0_6_3=cbind(p2[76,76,],hydro0_6_3)
```

References

[1] Aas, K. (2004). Modelling the dependence structure of financial assets: A survey of four copulas. Technical Report. Norwegian Computing Center, Oslo. 18

[2] Aas, K., Czado, C., Frigessi, A. and Bakken, H. (2009). Pair-copula constructions of multiple dependence. Insurance: Mathematics and Economics, 44, 182-198. 29, 30

[3] Abdi, H. (2007). Kendall rank correlation. In Salkind, N.J. Encyclopedia of Measurement and Statistics. Thousand Oaks (CA). Sage. 37, 38

[4] Akaike, H. (1973). "Information theory and an extension of the maximum Likelihood principle" in Proceedings of the second international symposium on information theory, edited by Petrov, B. N. and Csaki, F., pp. 267-281. Akademiai Kiado, Budapest. 47

[5] Anscombe, F. J. (1973). Graphs in statistical analysis. The American Statistician, 27, 17-21. 33

[6] Atamas, N., Bardik, V., Bannikova, A., Grishina, O., Lugovskoi, E., Lavoryk, S., Makogonenko, Y., Korolovych, V., Nerukh, D. and Paschenko, V. (2017). The effect of water dynamics on conformation changes of albumin in pre-denaturation state: photon correlation spectroscopy and simulation. Journal of Molecular Liquids, http://dx.doi.org/10.1016/j.molliq.2017.01.017. 6

[7] Baba, K., Ritei Shibata, R. and Sibuya, M. (2004). Partial correlation and conditional correlation as measure of conditional independence. Australian & New Zealand Journal of Statistics, 46, 657-664. 39

[8] Balakrishnan, N. and Lai, C. D. (2009). Continuous bivariate distributions (second edition). Springer, New York. 43

[9] Barbe, P., Genest, C., Ghoudi, K., and Rémillard, B. (1996). On Kendalls process. Journal of Multivariate Analysis, 58, 197-229. 38

[10] Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. Annals of Mathematics and Artificial Intelligence, 32, 245-268. 31

[11] Bedford, T. and Cooke, R. M. (2002). Vines: a new graphical model for dependent random variables. The Annals of Statistics, 30, 1031-1068. 30

[12] Berg, D. and Aas, K. (2009). Models for construction of higher-dimensional dependence: a comparison study. European Journal of Finance, 15, 639-659. 18

[13] Blum, P., Dias, A. and Embrechts, P. (2002). The art of dependence modelling: the latest advances in correlation analysis. In Alternative Risk Strategies. Risk Books, London. 37

[14] Bobée, B., and Ashkar, F. (1991). The gamma family and derived distributions applied in hydrology, Water Resources, Littleton, Colo. 43

[15] Borkowf, C. B. (2002). Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearmans rank correlation. Computational Statistics and Data Analysis, 39, 271-286. 37

[16] Brechmann, C. E. and Czado, C. (2012). Risk management with high-dimensional vine copulas: an analysis of the Euro Stoxx 50. Submitted for publication. 18

[17] Brechmann, C. E., Czado, C. and Aas, K. (2012). Truncated regular vines in high dimensions with applications to financial data. Canadian Journal of Statistics, 40, 68-85. 18

[18] Brechmann, C. E. and Schepsmeier, U. (2013). CDVine: modeling dependence with C- and D- vine copulas in R. Journal of Statistical Software, 52, 1-27. 32

[19] Capéraá, P., and Genest, C. (1993). Spearmans rho is larger than Kendalls tau for positively dependent random variables. Journal of Nonparametric Statistics, 2, 183-194. 108

[20] Capéraá, P., Fougéres, A. -L., and Genest, C. (1997). A nonparametric estimation procedure for bivariate extreme value copulas. Biometrika, 84, 567-577. 29

[21] Cherubini, U., Luciano, E., and Vecchiato, W. (2004). Copula methods in finance, Wiley, New York. 16

[22] Chen P. Y. and Popovich P. M. (2002). Correlation: parametric and nonparametric measures. Sage Publications, Inc. California. 37

[23] Chollete, L., Heinen, A. and Valdesogo, A. (2009). Modeling international financial returns with a multivariate regime switching copula. Journal of Financial Econometrics, 7, 437-480. 19

[24] Cuendet, M. A., Weinstein, H. and LeVine, M. V. (2016). The allostery landscape: Quantifying thermodynamic couplings in biomolecular systems. Journal of Chemical Theory and Computation, 12, 5758-5767. 14

[25] Czado, C. (2010). "Pair-copula constructions of multivariate copulas" in Copula theory and its applications, edited by Jaworski, P., Durante, F., Härdle, W. and Rychlik, T., Springer-Verlag, Berlin. 29

[26] Czado, C., Schepsmeier, U. and Min, A. (2012). Maximum likelihood estimation of mixed C-vines with application to exchange rates. Statistical Modelling, 12, 229-255. 19, 30

[27] Demarta, S. and McNeil, A. J. (2004). The t copula and related copulas. Technical report, ETH, Zurich. 27

[28] De Melo Mendes, B. V., Mendes Semeraro, M. and Câmara Leal, R. P. (2010). Pair-copulas modeling in finance. Financial Markets and Portfolio Management, 24, 193-213. 20

[29] Devroye, L. (1986). Nonuniform random variate generation, Springer, New York. 43

[30] Dißmann, J., Brechmann, C. E., Czado, C. and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. Computational Statistics & Data Analysis, 59, 52-69. 20

[31] Echenique, P. (2007). Introduction to protein folding for physicists. Contemporary Physics, 48, 81-108. 1, 2, 3, 4, 5

[32] Embrechts, P., McNeil, A. J., and Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. Risk management:Value at risk and beyond (Cambridge, 1998), Cambridge University Press, Cambridge, U.K., 176-223. 38

[33] Embrechts, P., Lindskog, F. and McNeil, A. J. (2003). "Modelling dependence with copulas and applications to risk management" in Handbook of heavy tailed distributions in finance, edited by Rachev, S. T., Elsevier, North-Holland. 20, 38

[34] Favre, A. C., El Adlouni, S., Perreault, L., Thiémonge, N., and Bobée, B. (2004). Multivariate hydrological frequency analysis using copulas. Water Resources Research, 40, 1-12. 21

[35] Fermanian, J. D. (2005). Goodness-of-fit tests for copulas. Journal of Multivariate Analysis, 95, 119-152. 43

[36] Fieller, E. C., Hartley, H. O. and Pearson, E. S. (1957). Tests for rank correlation coefficients. Biometrika, 44, 470-481. 38

[37] Fischer, M., Köck, C., Schlüter, S. and Weigert, F. (2009). An empirical analysis of multivariate copula models. Quantitative Finance, 9, 839-854. 20

[38] Fisher, N. I. and Switzer, P. (1985). Chi-plots for assessing dependence. Biometrika, 72, 253-265. 33

[39] Fisher, N. I. and Switzer, P. (2001). Graphical assessment of dependence: Is a picture worth 100 tests?. The American Statistician, 55, 233-239. 33

[40] Frees, E. W., and Valdez, E. A. (1998). Understanding relationships using copulas. The North American Actuarial Journal, 2, 1-25. 16

[41] Genest, C. (1987). Franks family of bivariate distributions. Biometrika, 74, 549-555. 28

[42] Genest, C. and Boies, J. C. (2003). Detecting dependence with Kendall plots. The American Statistician, 57, 275-284. 34

[43] 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. 33, 34

[44] Genest, C., and MacKay, J. (1986) The joy of copulas: Bivariate distributions with uniform marginals. The American Statistician, 40, 280-283. 27

[45] Genest, C., Ghoudi, K., and Rivest, L. P. (1998). Discussion of understanding relationships using copulas by E. W. Frees and E. A. Valdez. The North American Actuarial Journal, 2, 143-149. 17

[46] 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. 48

[47] Genest, C. and Rémillard, B. (2004). Tests of independence and randomness based on the empirical copula process. Sociedad de Estadística e Investigación Operativa Test, 13, 335-369. 48

[48] Genest, C. and Rivest, L. P. (1993). Statistical inference procedures for bivariate Archimedean copulas. Journal of the American Statistical Association, 88, 1034-1043. 36, 48

[49] Genest, C., and Rivest, L. P. (2001). On the multivariate probability integral transformation. Statistics and Probability Letters, 53, 391-399. 48

[50] Genest, C., and Verret, F. (2005). Locally most powerful rank tests of independence for copula models. Journal of Nonparametric Statistics, 17, 521-539. 48

[51] Gibbons, J. D. (1985). Nonparametric methods for quantitative analysis (second edition). Columbus, OH: American Sciences Press, Inc. 37

[52] Gibbons J. D. (1993). Nonparametric Measures of Association. Sage Publications, Inc., California. 37

[53] Gijbels, I., and Mielniczuk J. (1990). Estimating the density of a copula function. Communications in Statistics - Theory and Methods, 19, 445-464. 25

[54] Hays, W. L. (1973). Statistics. Holt Rinehart & Winston, New York. 37, 93

[55] Henson, R. K. and Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA Task Force report and current trends. Journal of Research and Development in Education, 33, 285-296. 58

[56] Hobæk Haff, I. (2013). Parameter estimation for pair-copula constructions. Bernoulli, 19, 462-491. 29

[57] Hobæk Haff, I., Aas, K. and Frigessi, A. (2010). On the simplified pair-copula construction - simply useful or too simplistic? Journal of Multivariate Analysis, 101, 1296-1310. 29

[58] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. The Annals of Mathematical Stattistics, 19, 293-325. 38

[59] Joe, H. (1997). Multivariate models and dependence concepts. Chapman and Hall, London. 16

[60] Johnson, N., and Kotz, S. (1972). Distributions in Statistics: continuous multivariate distributions. Wiley, New York. 48

[61] Johnson, N., Kotz, S. and Balakrishnan, N. (1994). Continuous univariate distributions (second edition). John Wiley & Sons, Inc, New York. 48

[62] Kendall, M. G. (1938). A new measure of rank correlation. Biometrika 30, 81-93. 37

[63] Kendall, M. G. (1970). Rank correlation methods (fourth edition). Charles Griffin & Company, London. 37, 109

[64] Kendall, M. G. and Stuart, A. (1979). Handbook of statistics. Charles Griffin & Company, London. 37

[65] Khamis, H. (2008). Measures of association: how to choose? Journal of Diagnostic Medical Sonography, 24, 155-162. 108

[66] Kimeldorf, G., and Sampson, A. R. (1975). Uniform representations of bivariate distributions. Communications in Statistics - Theory and Methods, 4, 617-627. 43

[67] Kotz, S., Balakrishnan, N. and Johnson, N. (2000). Continuous multivariate distributions: models and applications (Volume 1, 2nd ed.). John Wiley & Sons, Inc, New York. 48

[68] Knapp, T. R. (1998). Comments on the statistical significance testing articles. Research in Schools, 5, 39-41. 58

[69] Kruskal, W. H. (1958). Ordinal measures of association. Journal of the American Statistical Association, 53, 814-861. 37

[70] Kullback, S. (1959). Information theory and statistics. John Wiley & Sons, New York. (Republished in 1997 by Dover Publications, Inc). 37

[71] Kurowicka, D. and Joe, H. (editors) (2011). Dependence modeling vine copula handbook. World Scientific Publishing Co. Pte. Ltd, Singapore. 30

[72] Lehmann, E. L. (1966). Some concepts of dependence. The Annals of Mathematical Statistics, 37, 1137-1153. 37

[73] Lehmann, E. L. (1975). Nonparametrics: statistical methods based on ranks. Holden-Day Inc., San Francisco. 37

[74] Lewis-Beck, M. S. (1995). Data analysis: an introduction. Sage Publications, Inc., California. 37

[75] Lieberson, S. (1964). Limitations in the application of non-parametric coefficients of correlation. American Sociological Review, 29, 744-746. 108

[76] Liebetrau, A. M. (1976). Measures of association. Beverly Hills and London: Sage Publications, Inc. 37

[77] Lindskog, F., McNeil, A. and Schmock, U. (2001). A note on Kendalls tau for elliptical distributions. ETH preprint. 108

[78] Mari, D. D. and Kotz, S. (2004). Correlation and dependence. Imperial College Press, London. 37

[79] Marshall, A. (1996). "Copulas, marginals and joint distributions" in Distributions with fixed marginals and related topics, edited by Rüschendorff, L., Schweizer, B. and Taylor, M., pp. 213-222. Institute of Mathematical Statistics, Hayward, California. 24

[80] Nelsen, B. R. (2006). An introduction to copulas (second edition). Springer Science + Business Media, Inc., New York. 16, 25, 27

[81] Nerukh, D. (2012). Non-Markov state model of peptide dynamics. Journal of Molecular Liquids, 176, 65-70. 7

[82] Nerukh, D. and Karabasov, S. (2013). Water-Peptide dynamics during conformational transitions. The Journal of Physical Chemistry Letters, 4, 815-819. 12, 94, 95, 96, 97

[83] Nerukh, D., Ryabov, V. and Taiji, M. (2009). Molecular phase space transport in water: Non-stationary random walk model. Physica, A 388, 4719-4726. 8

[84] Nerukh, D., Ryabov, V. and Glen, R. C. (2008). Complex temporal patterns in molecular dynamics: A direct measure of the phase-space exploration by the trajectory at macroscopic time scales. Physical Review, E 77, 036225, 1-11. 9

[85] R Development Core Team. (2015), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 42

[86] Roberts, D. M. and Kunst, R. E. (1990). A case against continuing use of the Spearman formula for rank-order correlation. Psychological Reports, 66, 339-349. 37

[87] Rosenthal, R. (1994). "Parametric measures of effect size" in the Handbook of Research Synthesis, edited by Cooper, H. & Hedges, L. V., pp. 213-244. Russell Sage Foundation, New York. 58

[88] Rupinski, M. T. and Dunlap, W. P. (1996). Approximating Pearson product-moment correlations from Kendalls tau and Spearmans rho. Educational and Psychological Measurement, 56, 419-429. 108

[89] Ryabov, V. and Nerukh, D. (2011). Quantifying long time memory in phase space trajectories of molecular liquids. Journal of Molecular Liquids, 159, 99-104. 9

[90] Ryabov, V. and Nerukh, D. (2011). Computational mechanics of molecular systems: Quantifying highdimensional dynamics by distribution of Poincaré recurrence times. American Institute of Physics, 21, 037113-1-9. 10, 41

[91] Samara, B., and Randles, R. H. (1988). A test for correlation based on Kendalls tau. Communications in Statistics - Theory and Methods, 17, 3191-3205. 38

[92] Scarsini, M. (1984). On measures of concordance. Stochastica, 8, 201-218. 38

[93] Schepsmeier, U. (2010). Maximum likelihood estimation of C-vine pair-copula constructions based on bivariate copulas from different families. Diploma thesis, Technische Universitaet Muenchen. 32, 36

[94] Schepsmeier, U. and Brechmann, E. C. (2014). Statistical inference of C- and D-vine copulas: Package CDVine. The Comprehensive R Archive Network for the R programming language (CRAN). 32, 33, 35, 36

[95] Schirmacher, D. and Schirmacher, E. (2008). Multivariate dependence modeling using pair-copulas. Society of Actuaries, Illinois, USA. 29

[96] Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. Annals of Statistics, 9, 879-885. 37

[97] Sheskin, D. J. (2007). Handbook of Parametric and Nonparametric Statistical Procedures. 4th ed. Boca Raton, FL: Chapman & Hall/CRC. 37

[98] Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. McGrawHill. New York. 37

[99] Siegel, S. (1957). Nonparametric Statistics. The American Statistician, 11, 13-19. 37

[100] Sklar, A. (1996). Random variables, distribution functions, and copulas - a personal look backward and forward in Distributions with Fixed Marginals and Related Topics, edited by Rüschendorff, L., Schweizer, B. and Taylor, M., pp. 1-14. Institute of Mathematical Statistics, Hayward, California. 24

[101] Sprent, P. and Smeeton, N. C. (2007). Applied nonparametic statistical methods (fourth edition). Boca Raton, FL: Chapman & Hall/CRC. 37

[102] Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251. 39

[103] Stewart, D. W. (2000). Testing statistical significance testing: Some observations of an agnostic. Educational and Psychological Measurement, 60, 685-690.

[104] Strahan, R. F. (1982). Assessing magnitude of effect from rank-order correlation coefficients. Educational and Psychological Measurement, 42, 763-765. 108

[105] Taylor, J. M. G. (1987) Kendalls and Spearmans correlation coefficients in the presence of a blocking variable. Biometrics, 43, 409-416. 108

[106] Taylor, H. M. and Karlin, S. (1984). An introduction to stochastic modeling. Academic Press, New York. 38

[107] Trivedi, P. K. and Zimmer, D. M. (2005). Copula modeling: an introduction for practitioners. Foundations and Trends® in Econometrics, 1, 1-111. 24

[108] Valle, L. D. (2014). Official statistics data integration using copulas. Quality Technology & Quantitative Management 11, 111-131. 24

[109] Venter, G. G. (2002). Tails of copulas. Proceedings of the Casualty Actuarial Society, 89, 68113. 29, 38

[110] Walker, D. A. (2003). Converting Kendall's Tau for correlational or meta-analytic analyses. Journal of Modern Applied Statistical Methods, 2, 525-530. 38

[111] Yan, J. (2007). Enjoy the joy of copulas: with a package copula. Journal of Statistical Software, 21, 1-21. 32, 42