Technische Universität München

Department of Mathematics

Master's Thesis

Vine Copula and mixture model based analysis of the Sachs data

Florian Kössinger

Supervisor:	Prof. Claudia Czado, PhD
Advisors:	Prof. Claudia Czado, PhD
	Özge Sahin
Submission Date:	04.02.2022

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich, 04.02.2022

Acknowledgements

First, I would like to thank Prof. Claudia Czado and Özge Sahin for their great guidance while the whole process of writing this thesis. Their suggestions and feedback were of great help to me.

Furthermore, in the sense of my agreement to the funding guidelines of April 27, 2016, I would also like to thank the Hanns Seidel Stiftung, which supported me over the last years through a scholarship and was a great help to me through ideational as well as financial support.

Contents

Chapter 1

Introduction

In this master's thesis we use different approaches and methods to analyze the Sachs dataset (Sachs et al. (2005)). The Sachs dataset consists of 14 measurements of human T-cells. In each of these experiments, external influence was applied to certain phosphorylated proteins and phospholipids. In addition, these variables influence each other, making graphical models a useful approach for the analysis. Given the generation of the data from 14 different experiments, methods based on mixture densities are also a promising approach we are pursuing.

In the next chapter we will introduce the theoretical background on which the work of the following chapters is based. In the third chapter, we will introduce the Sachs dataset and and briefly explain the biochemical effects behind the data. We also explain the necessary preprocessing we need to apply to the data before the following chapters. In the fourth chapter, we will start by modeling the entire data set consisting of all experiments as pooled data. For this purpose, we first use a multivariate Gaussian approach, then work with a vine copula model and finally with D-vine regression models introduced by Kraus and Czado (2017). In the fifth chapter, we first use Gaussian mixture models and later vine copula mixture models introduced by Sahin and Czado (2021) to find substructures in the Sachs dataset. Especially the results of the vine copula mixture models are analyzed in greater detail, as they are a major part of this thesis. In the sixth chapter, we focus on the causal analysis of the data. In every experiment and thus in every single observation, external influence was applied to the measurements. It is therefore a particularly exciting idea to be able to build models where all external influences are removed. To do this, we fit D-Vine regression models on certain subsets where specific variables were not influenced. Since Gaussian models are commonly used in previous research, we allow different marginals and copulas and discuss the Gaussian results to them of non-Gaussian copula based D-vine models.

Chapter 2

Theoretical background

2.1 Marginal distributions

We would like to start by defining all marginal distributions used in this thesis and giving their probability density functions *f*. Except for the multivariate normal distribution, all distributions are univariate. To get a better understanding of the distributions, we will additionally plot the densities of all univariate distributions except from the χ^2 distribution (which has only one parameter, and which we do not use in VCMM but for tests) with mean = 2 and variance = 3.

Definition 2.1 (Univariate normal distribution)**.** *A random variable $X \in \mathbb{R}$ follows the univariate normal distribution, if its probability density function in x is*

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{2.1}$$

with mean $\mu \in \mathbb{R}$ and variance $\sigma \in \mathbb{R}^+$. Then we denote $X \sim \mathcal{N}(\mu, \sigma^2)$. The special case of $\mathcal{N}(0,1)$ - i.e. $\mu = 0$ and $\sigma = 1$ - is called the standard normal distribution. The distribution function of $\mathcal{N}(0,1)$ is denoted as $\Phi(x)$ and the density as $\phi(x)$.

Definition 2.2 (Multivariate normal distribution)**.** *A random vector $\boldsymbol{X} = (X_1, ..., X_d)^T \in \mathbb{R}^d$ follows the multivariate normal distribution, if the probability density function of $\boldsymbol{X}$ in $\boldsymbol{x}$ is*

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{2.2}$$

with mean vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_d)^T \in \mathbb{R}^d$ and variance matrix $\Sigma \in \mathbb{R}^{d \times d}$ with $\det \Sigma \neq 0$. Then we denote $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$.

Definition 2.3 (Gamma distribution)**.** *A random variable $X \in \mathbb{R}^+$ is gamma distributed, if its probability density function in x is given by*

$$f(x; a, \lambda) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x} \tag{2.3}$$

for parameters $a \in \mathbb{R}$ and $\lambda \in \mathbb{R}$. The gamma-function Γ and more detailed information about the gamma distribution can be found in Czado et al. (2011).

Definition 2.4 (χ^2-distribution)**.** *A random variable $X \in \mathbb{R}^+$ is χ^2-distributed, if its probability density function in x is given by*

$$f(x; k) = \frac{x^{k/2-1} e^{-x/2}}{2^{k/2}\Gamma(k/2)} \tag{2.4}$$

with $k \in \mathbb{N}$ degrees of freedom. The χ^2-distribution is a special case of the gamma distribution, as one can set $a = \frac{k}{2}$ and $\lambda = \frac{1}{2}$.

Definition 2.5 (Student's t distribution)**.** *A random variable $X \in \mathbb{R}$ follows the Student's t distribution, if its probability density function in x is given by*

$$f(x; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} + \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{2.5}$$

with $\nu \in \mathbb{R}^+$ degrees of freedom. In this definition the expected value is $\mathbb{E}[X] = 0$. By transforming $x = (\tilde{x} - \mu)$ in the density function a location parameter can be added.

Definition 2.6 (Logistic distribution)**.** *A random variable $X \in \mathbb{R}$ follows a logistic distribution, if its probability density function in x is given by*

$$f(x; \mu, s) = \frac{e^{-(x-\mu)/s}}{s\left(1 + e^{-(x-\mu)/s}\right)^2} \tag{2.6}$$

with mean $\mu \in \mathbb{R}$ and scale parameter $s \in \mathbb{R}^+$.

Definition 2.7 (Log-normal distribution)**.** *A random variable $X \in \mathbb{R}$ follows a log-normal distribution, if the random variable $Y = \ln(X)$ is normal distributed. This leads to a probability density function of X in x*

$$f(x; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right) \tag{2.7}$$

with parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

Definition 2.8 (Log-logistic distribution)**.** *A random variable $X \in \mathbb{R}$ follows a log-logistic distribution, if the random variable $Y = \ln(X)$ follows a logistic distribution. This leads to a probability density function of X in x*

$$f(x;\alpha,\beta) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1+(x/\alpha)^{\beta})^2} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right) \tag{2.8}$$

with scale parameter $\alpha \in \mathbb{R}^+$ and shape parameter $\beta \in \mathbb{R}^+$.

Definition 2.9 (Skew Student-t distribution)**.** *According to Fernandez et al. (1996) a random variable $X \in \mathbb{R}$ is skew t distributed, if its probability density function in x is given by*

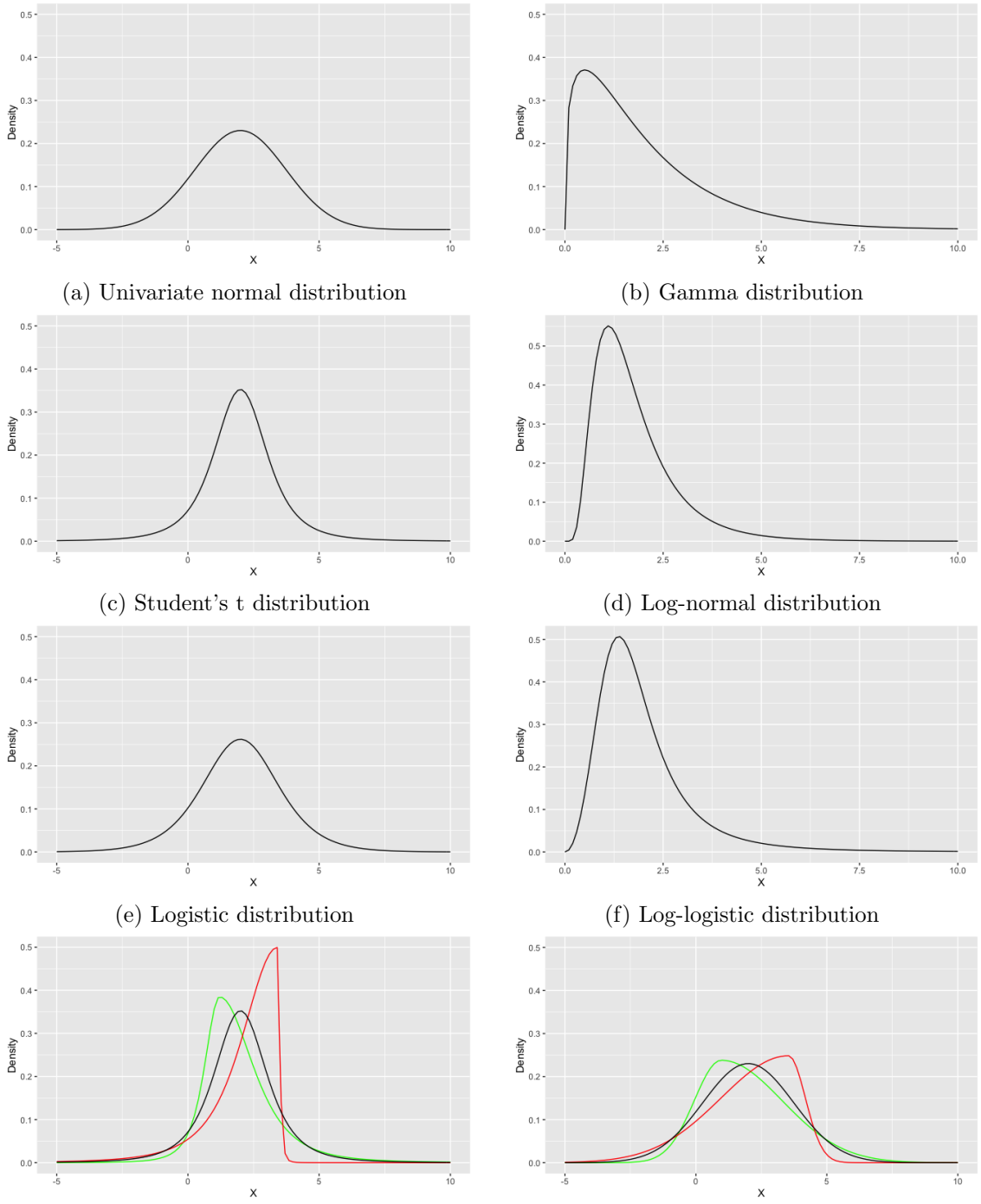
$$\begin{aligned} f(x;\mu,\tau^2,\nu,\gamma) = 2\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\frac{\tau}{\gamma+\frac{1}{\gamma}} \\ \left[1+\frac{\tau^2(x-\mu)^2}{\nu}\left\{\gamma^2\mathbb{I}_{(-\infty,0)}(x-\mu)+\frac{1}{\gamma^2}\mathbb{I}_{[0,\infty)}(x-\mu)\right\}\right]^{\frac{1-\nu}{2}} \end{aligned} \tag{2.9}$$

with location parameter $\mu \in \mathbb{R}$, scale parameter $\tau \in \mathbb{R}^+$, degrees of freedom $\nu \in \mathbb{R}^+$ (i.e. the shape parameter) and skewness parameter $\gamma \in \mathbb{R}^+$. There are also other definitions of the skew students t distribution that use fewer parameters (for example in Azzalini et al. (2013)). However, we use this skew t distribution with four parameters in the models later.

Definition 2.10 (Skew normal distribution)**.** *According to Azzalini et al. (2013) a random variable $X \in \mathbb{R}$ is skew t distributed, if its probability density function in x is given by*

$$f(x;\mu,\tau^2,\alpha) = \frac{2}{\tau}\phi\left(\frac{x-\mu}{\tau}\right)\Phi\left(\alpha\frac{x-\mu}{\tau}\right) \tag{2.10}$$

with location parameter $\mu \in \mathbb{R}$, scale parameter $\tau \in \mathbb{R}^+$, and slant parameter $\alpha \in \mathbb{R}$. ϕ and Φ are the density and distribution function of the univariate standard normal distribution.



(a) Univariate normal distribution

(b) Gamma distribution

(c) Student's t distribution

(d) Log-normal distribution

(e) Logistic distribution

(f) Log-logistic distribution

(g) Skew Student-t distribution

(h) Skew normal distribution

Figure 2.1: Density functions of different distributions with $\mathbb{E}[X] = 2$ and $\mathrm{Var}[X] = 3$. The red and green lines in the skew Student-t distribution have parameters $\gamma = 0.25$ (red) and $\gamma = 1.5$ (green) and in the skew normal distribution $\alpha = 0.5$ (red) and $\alpha = 1.5$ (green).

2.2 Copulas

In this chapter we introduce the basics of copulas, which are a good way to model multivariate distributions. If not stated differently the following definitions and concepts can be found in Czado (2019).

2.2.1 Introduction to copulas

Definition 2.11 (Probability integral transform (PIT)). *Let X be a continuous random variable following a distribution F and let x be an observed value of X. Then $u := F(x)$ is called the probability integral transform (PIT) at x.*

Furthermore the random variable $U := F(X)$ follows a uniform distribution, as it holds for every $u \in [0,1]$:

$$\mathbb{P}(U \leq u) = \mathbb{P}(F(X) \leq u) = \mathbb{P}\left(X \leq F^{-1}(u)\right) = F\left(F^{-1}(u)\right) = u$$

Definition 2.12 (x-scale, u-scale, z-scale). *Let $\boldsymbol{X} = (X_1, ..., X_d)^T \in \mathbb{R}^d$ be random vector following a multivariate distribution F. Furthermore let F_j be the corresponding marginal distribution functions for $j = 1, ..., d$ and let Φ be the cumulative distribution function of the standard normal distribution $\mathcal{N}(0,1)$. We can create the random vectors $\boldsymbol{U} = (U_1, ..., U_d)^T := (F_1(X_1), ..., F_d(X_d))^T$ and $\boldsymbol{Z} = (Z_1, ..., Z_d)^T := (\Phi^{-1}(U_1), ..., \Phi^{-1}(U_d))^T$, s.t. for all $j = 1, ..., d$, U_j is uniform and Z_j is standard normal distributed.*

If $\boldsymbol{x}$ is now an observed value of $\boldsymbol{X}$, then we analyze this observation with $\boldsymbol{x}$ on x-scale, with $\boldsymbol{u}$ on u-scale and with $\boldsymbol{z}$ on z-scale.

Definition 2.13 (Copula and copula density). *A d-dimensional copula C is a multivariate distribution function*

$$C : [0,1]^d \rightarrow [0,1]$$

with uniformly distributed marginals. By partial differentiation we can obtain the copula density c for $\boldsymbol{u} \in [0,1]^d$:

$$c(u_1, ..., u_d) = \frac{\partial^d}{\partial_1, ..., \partial_d} C(u_1, ..., u_d) \tag{2.11}$$

Theorem 2.14 (Sklar's Theorem). *It was proven by Sklar (1959), that for a $\boldsymbol{X} = (X_1, ..., X_d)^T \in \mathbb{R}^d$ random vector following a multivariate distribution F with marginal*

distribution functions F_j for $j = 1, ..., d$, the joint distribution function can be expressed as

$$F(x_1, ..., x_d) = C\left(F_1(x_1), ..., F_d(x_d)\right) \tag{2.12}$$

for some d-dimensional copula C. Let c be the copula density of C, then the density f of F is given by

$$f(x_1, ..., x_d) = c\left(F_1(x_1), ..., F_d(x_d)\right) f_1(x_1)...f_d(x_d). \tag{2.13}$$

It also holds for the inverse: the copula C for a multivariate distribution F with marginal distribution functions F_j for $j = 1, ..., d$ can be expressed as

$$C(u_1, ..., u_d) = F\left(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)\right) \tag{2.14}$$

and the copula density c is given by

$$c(u_1, ..., u_d) = \frac{f\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right)}{f_1(F_1^{-1}(u_1))...f_d(F_d^{-1}(u_d))}. \tag{2.15}$$

In later chapters, we will measure the dependence of variables more than once using Kendall's τ, which was developed by Kendall (1938). Kendall's τ has the advantage over other dependence measures that it is rank-based and thus invariant to monotone transformations of the marginals.

Definition 2.15 (Kendall's tau)**.** *Let X_1 and X_2 be two continuous random variables and let (X_{11}, X_{12}) and (X_{21}, X_{22}) be independent and identically distributed copies of (X_1, X_2). The Kendall's τ between (X_1, X_2) is defined as*

$$\tau(X_1, X_2) = \mathbb{P}((X_{11} - X_{21})(X_{12} - X_{22}) > 0) - \mathbb{P}((X_{11} - X_{21})(X_{12} - X_{22}) < 0)$$

Theorem 2.16 (Kendall's tau expressed in terms of the copula)**.** *Let (X_1, X_2) be two continuous random variables following the joint distribution F and the marginal distributions F_1 and F_2 respectively. Let for these the copula C be defined as in Equation 2.14. Then Kendall's τ can be expressed as*

$$\tau(X_1, X_2) = 4 \int_{[0,1]^2} C(u_1, u_2) dC(u_1, u_2) - 1$$

The proof for this theorem can be found in Czado (2019).

Definition 2.17 (Upper and lower tail dependence)**.** *Let (X_1, X_2) be two continuous random variables following the joint distribution F and the marginal distributions F_1 and*

F_2 respectively. Let for these the copula C be defined as in Equation 2.14. Then we call

$$\lambda^{upper} = \lim_{t\to 1^-} \mathbb{P}\left(X_2 > F_2^{-1}(t) | X_1 > F_1^{-1}(t)\right) = \lim_{t\to 1^-} \frac{1-2t+C(t,t)}{1-t}$$

the uppertail dependence coefficient and

$$\lambda^{lower} = \lim_{t\to 0^+} \mathbb{P}\left(X_2 \leq F_2^{-1}(t) | X_1 \leq F_1^{-1}(t)\right) = \lim_{t\to 0^+} \frac{C(t,t)}{t}$$

the lower tail dependence coefficient.

Definition 2.18 (Rotated copulas)**.** *Let $c(\cdot,\cdot)$ be the density of a bivariate copula. The densities of its (counterclockwise) rotations are given by*

$$90°: \quad c_{90}(u_1,u_2) = c(1-u_2,u_1)$$

$$180°: \quad c_{180}(u_1,u_2) = c(1-u_1,1-u_2)$$

$$270°: \quad c_{270}(u_1,u_2) = c(u_2,1-u_1)$$

Theorem 2.19 (Conditional copula)**.** *Let C be a copula. Then the conditional copula is given by*

$$C_{1,...,m|m+1,...,d}(u_1,...,u_m|v_{m+1},...,v_d) = \frac{\partial^{d-m}}{\partial_{m+1},...,\partial_d} C(u_1,...,u_d)|_{u_{m+1}=v_{m+1},...,u_d=v_d}.$$

Proof: With Equation (2.11) we can show

$$\begin{aligned} C_{1,...,m|m+1,...,d}(u_1,...,u_m|v_{m+1},...,v_d) &= \int_0^{u_1} ... \int_0^{u_m} c(v_1,...,v_d)dv_1...dv_m \\ &= \int_0^{u_1} ... \int_0^{u_m} \frac{\partial^d}{\partial_1,...,\partial_d} C(v_1,...,v_m,v_{m+1},...,v_d)dv_1...dv_m \\ &= \frac{\partial^{d-m}}{\partial_{m+1},...,\partial_d} C(u_1,...,u_d)|_{u_{m+1}=v_{m+1},...,u_d=v_d}. \end{aligned} \tag{2.16}$$

2.2.2 Bivariate copulas

In this subchapter we want to introduce the bivariate copulas, which will be used directly in the following chapters, or which will be used to construct multivariate copulas using vine tree structures.

Example 2.20 (Bivariate independence copula)**.** *Let U_1 and U_2 be independent random variables following a uniform distribution on $[0,1]$. Then the bivariate independence copula is defined as*

$$C(u_1, u_2) = \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2) = u_1 u_2. \tag{2.17}$$

The following bivariate Gaussian and bivariate Student's t copula are elliptical copulas. With Equation (2.14) these copulas can be easily constructed.

Example 2.21 (Bivariate Gaussian copula)**.** *Let Φ_R be the bivariate standard normal distribution function with zero mean vector, unit variance and correlation ρ. Then the bivariate Gaussian copula is given by*

$$C(u_1, u_2; R) = \Phi_\rho\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right). \tag{2.18}$$

Example 2.22 (Bivariate Student's t copula)**.** *Let $t(\cdot, \cdot; \nu, \rho)$ be the density of the bivariate Student's t distribution function with ν degrees of freedom, zero mean, and correlation ρ. Let T_ν be the univariate Student's t distribution function with ν degrees of freedom and zero mean and t_ν be its density. Then the bivariate Student's t copula is given by*

$$\begin{aligned} C(u_1, u_2; \nu, \rho) &= \int_0^{u_1} \int_0^{u_2} \frac{t(T_\nu^{-1}(v_1), T_\nu^{-1}(v_2); \nu, \rho)}{t_\nu(T_\nu^{-1}(v_1)) t_\nu(T_\nu^{-1}(v_2))} dv_1 dv_2 \\ &= \int_{-\infty}^{T_\nu^{-1}(u_1)} \int_{-\infty}^{T_\nu^{-1}(u_2)} t(x_1, x_2; \nu, \rho) dx_1 dx_2 \end{aligned} \tag{2.19}$$

Definition 2.23 (Bivariate Archimedean copula)**.** *Let $\phi : [0,1] \to [0,\infty]$ be a continuous, strictly monotone decreasing, and convex function with $\phi(1) = 0$. Then*

$$C(u_1, u_2) = \phi^{[-1]}(\phi(u_1) + \phi(u_1))$$

is with the pseudo inverse function $\phi^{[-1]} : [0,\infty] \to [0,1]$ a copula.

$$\phi^{[-1]}(t) := \begin{cases} \phi^{-1}(t) & , 0 \leq t \leq \phi(0) \\ 0 & , \phi(0) \leq t \leq \infty \end{cases}$$

We call C a bivariate Archimedean copula with generator ϕ.

Example 2.24 (Bivariate Clayton copula)**.** *The bivariate Clayton copula is defined as*

$$C(u_1, u_2; \delta) = (u_1^{-\delta} + u_2^{-\delta} - 1)^{-\frac{1}{\delta}}$$

with dependence parameter $0 < \delta < \infty$.

Example 2.25 (Bivariate Gumbel copula)**.** *The bivariate Gumbel copula is defined as*

$$C(u_1, u_2; \delta) = \exp\left(-\left((-\log u_1)^\delta + (-\log u_2)^\delta\right)^{-\frac{1}{\delta}}\right)$$

with dependence parameter $\delta \geq 1$.

Example 2.26 (Bivariate Frank copula)**.** *The bivariate Frank copula is defined as*

$$C(u_1, u_2; \delta) = -\frac{1}{\delta}\log\left(1 - \frac{(1 - e^{-\delta u_1})(1 - e^{-\delta u_2})}{(1 - e^{-\delta})}\right)$$

with dependence parameter $\delta \in [-\infty, \infty] \setminus \{0\}$.

Example 2.27 (Bivariate Joe copula)**.** *The bivariate Joe copula is defined as*

$$C(u_1, u_2; \delta) = 1 - \left((1 - u_1)^\delta + (1 - u_2)^\delta - (1 - u_1)^\delta(1 - u_2)^\delta\right)^{-\frac{1}{\delta}}$$

with dependence parameter $\delta \geq 1$.

Example 2.28 (Bivariate BB1 copula)**.** *Assume the function $\eta(s) = \eta_{\delta,\theta}^{(BB1)}(s) = (1 + s^{\frac{1}{\delta}})^{-\frac{1}{\theta}}$. Then the bivariate BB1 copula is defined as*

$$C(u_1, u_2; \delta, \theta) = \eta(\eta^{-1}(u_1) + \eta^{-1}(u_2))$$

with parameters $\delta \geq 1$ and $\theta > 0$.

Example 2.29 (Bivariate BB7 copula)**.** *Assume the function $\eta(s) = \eta_{\delta,\theta}^{(BB7)}(s) = 1 - (1 - (1 + s)^{-\frac{1}{\delta}})^{\frac{1}{\theta}}$. Then the bivariate BB7 copula is defined as*

$$C(u_1, u_2; \delta, \theta) = \eta(\eta^{-1}(u_1) + \eta^{-1}(u_2))$$

with parameters $\delta > 0$ and $\theta \geq 1$.

2.2.3 Vine copulas

Definition 2.30 (Graph). *A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a pair of a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$.*

Definition 2.31 (Path and Cycle). *Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| \geq 2$. A path in $\mathcal{G}$ is a sequence of (at least two) nodes $X_1, ..., X_d \in \mathcal{V}$, such that there is an edge $(X_i, X_{i+1}) \in \mathcal{E}$ for all $i = 1, ..., d-1$. A cycle is a path $X_1, ..., X_d$ with $X_1 = X_d$.*

Definition 2.32 (Tree). *Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. If any two nodes $X, Y \in \mathcal{V}$ are connected by a unique path, then we call $\mathcal{G}$ a tree.*

Definition 2.33 (Regular (R-) vine tree sequence). *The seqence of trees $\mathcal{T} = (T_1, ..., T_{d-1})$ is a regular vine tree sequence on d elements if:*

(i) $T_j = (\mathcal{V}_j, \mathcal{E}_j)$ *is a tree for all* $j = 1, ..., d-1$.

(ii) $T_1 = (\mathcal{V}_1, \mathcal{E}_1)$ *has node set* $\mathcal{V}_1 = \{1, ..., d\}$.

(iii) For $j \geq 2$, $T_j = (\mathcal{V}_j, \mathcal{E}_j)$ *has node set* $\mathcal{V}_j = \mathcal{E}_{j-1}$.

(iv) For all $j = 2, ..., d-1$, *it holds* $|a \cap b| = 1$, *if* $\{a, b\} \in \mathcal{E}_j$.

Example 2.34 (Four dimensional R-vine tree sequence). *The following seqence of trees $\mathcal{T} = (T_1, ..., T_{d-1})$ is a R-vine tree sequence:*

$$
\begin{aligned}
T_1: \quad & \mathcal{V}_1 = \{1,2,3,4\} \\
& \mathcal{E}_1 = \{\{1,2\},\{2,3\},\{3,4\}\} \\
T_2: \quad & \mathcal{V}_2 = \{\{1,2\},\{2,3\},\{3,4\}\} \\
& \mathcal{E}_2 = \{\{\{1,2\},\{2,3\}\},\{\{2,3\},\{3,4\}\}\} \\
T_3: \quad & \mathcal{V}_3 = \{\{\{1,2\},\{2,3\}\},\{\{2,3\},\{3,4\}\}\} \\
& \mathcal{E}_3 = \{\{\{\{1,2\},\{2,3\}\},\{\{2,3\},\{3,4\}\}\}\}
\end{aligned}
$$

Definition 2.35 (Complete union, conditioning set and conditioned sets). *The complete union of an edge $e \in \mathcal{E}_i$ is defined by*

$$A_e = \{j \in \mathcal{V}_1 | \exists e_1 \in \mathcal{E}_1, ..., e_i \in \mathcal{E}_i \text{ s.t. } j \in e_1 \in ... \in e_i \in e\}.$$

The conditioning set of an edge $e = \{a, b\}$ is given by

$$D_e = A_a \cap A_b,$$

and the conditioned sets of are given by

$$\mathcal{C}_{e,a} = A_a \setminus D_e \text{ and } \mathcal{C}_{e,b} = A_b \setminus D_e.$$

Definition 2.36 (Pair copula)**.** *Let $e = \{a, b\} \in \mathcal{E}_i$ be an edge. Then we abbreviate the copula $C_{\mathcal{C}_{e,a}\mathcal{C}_{e,b};D_e}$ with C_e and its density $c_{\mathcal{C}_{e,a}\mathcal{C}_{e,b};D_e}$ with c_e. We call C_e a pair copula.*

Definition 2.37 (C- and D-vine tree sequence)**.** *We call a R-vine seqence $\mathcal{T} = (T_1, ..., T_{d-1})$*

- *C-vine sequence, if in each tree T_i there is a node $r \in \mathcal{V}_i$ (which we call root node), s.t. $|\{e \in \mathcal{E}_i | r \in e\}| = d - i$.*
- *D-vine sequence, if we have $|\{e \in \mathcal{E}_i | n \in e\}| \leq 2$ for each node $n \in \mathcal{V}_i$.*

Definition 2.38 (R-vine distribution)**.** *Consider a d-dimensional random vector $\boldsymbol{X} = (X_1, ..., X_d)$ with distribution F. It has a regular vine distribution, if there is a triplet $(\mathcal{F}, \mathcal{T}, \mathcal{B})$, for which the following holds:*

(i) $\mathcal{F} = (F_1, ..., F_d)$ is a vector of continuous invertible marginal distribution functions. For each $i = 1, ..., d$, F_i is the marginal distribution of X_i.

(ii) $\mathcal{T} = (T_1, ..., T_{d-1})$ is a R-vine tree sequence.

(iii) $\mathcal{B} = \{C_e | e \; in \mathcal{E}_i, i = 1, ..., d-1\}$ is a set of pair copulas.

(iv) $C_e \in \mathcal{B}$, which is the copula associated with the conditional distribution of $X_{\mathcal{C}_{e,a}}$ and $X_{\mathcal{C}_{e,b}}$ given $\boldsymbol{X}_{D_e} = \boldsymbol{x}_{D_e}$, does not depend on the specific value of $\boldsymbol{x}_{D_e}$.

Theorem 2.39 (Existence of a R-vine distribution)**.** *Let $(\mathcal{F}, \mathcal{T}, \mathcal{B})$ be a triplet , which fullfilles (i)-(iii) of Definition 2.38. Then there is a unique d dimensional distribution F with density*

$$\begin{aligned} f_{1,...,d}(x_1, ..., x_d) &= f_1(x_1) \cdot ... \cdot f_d(x_d) \\ &\cdot \prod_{i=1}^{d-1} \prod_{e \in \mathcal{E}_i} c_{\mathcal{C}_{e,a}\mathcal{C}_{e,b};D_e} \left(F_{\mathcal{C}_{e,a}|D_e}\left(x_{\mathcal{C}_{e,a}} | \boldsymbol{x}_{D_e}\right), F_{\mathcal{C}_{e,b}|D_e}\left(x_{\mathcal{C}_{e,b}} | \boldsymbol{x}_{D_e}\right)\right), \end{aligned} \tag{2.20}$$

s.t. for each $e = \{a, b\} \in \mathcal{E}_i$ *the distribution function of* $X_{\mathcal{C}_{e,a}}$ *and* $X_{\mathcal{C}_{e,b}}$ *given* $\boldsymbol{X}_{D_e} = \boldsymbol{x}_{D_e}$ *is given by*

$$F_{\mathcal{C}_{e,a}\mathcal{C}_{e,b};D_e}\left(x_{\mathcal{C}_{e,a}}, x_{\mathcal{C}_{e,b}}|\boldsymbol{x}_{D_e}\right) = C_{\mathcal{C}_{e,a}\mathcal{C}_{e,b};D_e}\left(F_{\mathcal{C}_{e,a}|D_e}\left(x_{\mathcal{C}_{e,a}}|\boldsymbol{x}_{D_e}\right), F_{\mathcal{C}_{e,b}|D_e}\left(x_{\mathcal{C}_{e,b}}|\boldsymbol{x}_{D_e}\right)\right).$$

The proof for Theorem (2.39) can be found in Bedford and Cooke (2002).

2.3 Clustering

2.3.1 Mixture models

Definition 2.40 (Mixture model)**.** *Assume a d-dimensional random vector* $\boldsymbol{X} = (X_1, ..., X_d)^T$ *following a distribution with density*

$$f(\boldsymbol{x}; \boldsymbol{\eta}) = \sum_{j=1}^{g} \pi_j f_j(\boldsymbol{x}; \boldsymbol{\psi}_j) \tag{2.21}$$

in a realisation $\boldsymbol{x} = (x_1, ..., x_d)^T$*, where* $g \in \mathbb{N}$*,* $\pi_j \in (0, 1)$ *for* $j = 1, ..., g$ *and* $\sum_{j=1}^{g} \pi_j = 1$*.*

Then we call g *the number of components,* π_j *the mixture weights (or mixing proportions),* $f_j(\boldsymbol{x}; \boldsymbol{\psi}_j)$ *the density of the jth component and* f *a mixture density. The parameter* $\boldsymbol{\eta} = (\eta_1, ..., \eta_1)^T$ *with* $\eta_j = (\pi_j, \boldsymbol{\psi}_j)^T$ *for* $j = 1, ..., g$ *contains all model parameters.*

The parameters and the densities of the different components of mixture densities can be defined in may different ways. In the following chapters, we will often use the vine copula mixture model and the Gaussian mixture model. In the vine copula mixture model the densities of the components are densities of R-vine distributions, i.e. they are of the form given in Equation (2.20). In a vine copula mixture model the parameter $\boldsymbol{\psi}_j$ contain both the copula parameters and the marginal parameters of the jth component. In the Gaussian mixture model the densities of the components are multivariate Gaussian distributions as defined in Equation (2.2) i.e. it holds $f_j(\mathbf{x}; \boldsymbol{\psi}_j) = \mathcal{N}_d(\mathbf{x}; \mu_j, \Sigma_j)$. At this point we should familiarize ourselves with the parameterization of the covariance matrices Σ_j in multivariate normal distributions. We can write the covariance matrix of a component j in the form $\Sigma_j = \lambda_j D_j A_j D_j^T$, where $\lambda_j = \det(\Sigma_j)^{\frac{1}{d}} \in \mathbb{R}$, $D_j \in \mathbb{R}^{d \times d}$ is the matrix of eigenvectors of Σ_j and $A_j \in \mathbb{R}^{d \times d}$ is a diagonal matrix. λ_j controls the volume, A_j the shape and D_j the orientation. Now one can estimate λ_j, A_j and D_j for each cluster separately or use the same for all clusters, which means less computational effort and fewer parameters, but also less flexibility. This results in various characterizations, which are listed in Table 2.1. More details on the parameterizations of GMMs can be found in Celeux and Govaert (1995). More details on the vine copula mixture models can be found in Sahin and Czado (2021).

Identfier	Parameterizations	Distribution	Volume	Shape	Orientation
EII	λI	Spherical	equal	equal	
VII	$\lambda_j I$	Spherical	variable	equal	
EEI	λA	Diagonal	equal	equal	coordinate axes
VEI	$\lambda_j A$	Diagonal	variable	equal	coordinate axes
EVI	λA_j	Diagonal	equal	variable	coordinate axes
VVI	$\lambda_j A_j$	Diagonal	variable	variable	coordinate axes
EEE	$\lambda D A D^T$	Ellipsoidal	equal	equal	equal
EEV	$\lambda D_j A D_j^T$	Ellipsoidal	equal	equal	variable
VEV	$\lambda_j D_j A D_j^T$	Ellipsoidal	variable	equal	variable
VVV	$\lambda_j D_j A_j D_j^T$	Ellipsoidal	variable	variable	variable

Table 2.1: Parameterisations of the within-component covariance matrix Σ_j. (Source: Scrucca L. et al. 2016, p.8)

2.3.2 Model selection

After we have fitted different models with the data, we want to compare these. In this subsection we will therefore discuss different ways of finding the best model. At first we consider the Kullback-Leibler information criterion (KLIC) of Kullback and Leibler (1951), the Akaike information criterion (AIC) of Akaike (1973) and the Bayesian information criterion (BIC) of Schwarz (1978). Subsequently, we deal with approaches from test theory, which are the likelihood ratio test and the Vuong test of Vuong (1989).

Definition 2.41 (*Kullback-Leibler information criterion*)**.** *Let $\boldsymbol{X}$ be a random vector following the true distribution with density h_0 and $\mathbb{E}_0$ the expected value with respect to the true distribution $h_0(\boldsymbol{x})$. $f(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}})$ in the approximation to $h_0(\boldsymbol{x})$ with the estimated parameter $\hat{\boldsymbol{\theta}}$.*

$$\mathrm{KLIC}(h_0, f, \hat{\boldsymbol{\theta}}) = \mathbb{E}_0\left[\log h_0(\boldsymbol{X})\right] - \mathbb{E}_0\left[\log f(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}})\right] \tag{2.22}$$

The KLIC measures the "distance" between the true distribution and our modeled distribution. The obvious goal is to minimize this distance. As the first term is constant, we can reduce this approach to maximizing $\mathbb{E}_0\left[\log f(\mathbf{X} \mid \hat{\boldsymbol{\theta}})\right]$. Since the true distribution is unknown we can not compute expected value, but approximate the it with the observed data. This leads to the method of maximum likelihood estimation (MLE).

Since the maximum likelihood estimator only considers the fit and does not take into account any other information about the model or the data, comparison criteria like the AIC and the BIC were established. Both consist of the log-likelihood and a penalty term, which penalizes the number of parameters in the model to prevent overfitting. The BIC also uses the sample size of the dataset.

Definition 2.42 (*Akaike information criterion*)**.** *Let $f(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}})$ be the density of our model with the estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ to the observations $\boldsymbol{x}_i$, $i = 1, ..., n$. k is the number of parameters i.e. the dimension of θ. Then the Akaike information criterion is given by*

$$\mathrm{AIC} = -2\sum_{i=1}^{n} \log f(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}}) + 2k \tag{2.23}$$

Definition 2.43 (*Bayesian information criterion*)**.** *Let $f(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}})$ be the density of our model with the estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ to the observations $\boldsymbol{x}_i$, $i = 1, ..., n$. k is the number of parameters i.e. the dimension of $\boldsymbol{\theta}$. Then the Bayesian information criterion is given by*

$$\mathrm{BIC} = -2\sum_{i=1}^{n} \log f(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}}) + \log(n)k \tag{2.24}$$

Clearly one minimizes the AIC or BIC to find the best-fitting parameter $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$.

Besides the straight approaches of simply choosing the model with the highest likelihood or the smallest AIC or BIC, one could compare models by testing whether the advantage that one model has over the other is significant. For this reason, we would now like to study the Likelihood Ratio Test (LRT) following McLachlan (1987) and the Vuong test following Vuong (1989).

Likelihood Ratio Test Let Θ be a parameter space and $\mathbf{X}$ be a random vector with density functions $f(\cdot \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$. Furthermore let $\Theta_0 \subset \Theta$ and $\Theta_1 = \Theta \backslash \Theta_0$ be subspaces. Then the LRT tests the hypothesis

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \in \Theta_1.$$

Therefore we compute the maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ under $\boldsymbol{\theta} \in \Theta$ and $\hat{\boldsymbol{\theta}}_{\mathbf{0}}$ for $\boldsymbol{\theta}$ under $\boldsymbol{\theta} \in \Theta_0$ and the test statistic

$$\lambda(\mathbf{x}) = 2 \log \left[\frac{f(\mathbf{x} \mid \hat{\boldsymbol{\theta}})}{f(\mathbf{x} \mid \hat{\boldsymbol{\theta}}_{\mathbf{0}})} \right] = 2 \log \left[\frac{\sup_{\boldsymbol{\theta} \in \Theta} f(\mathbf{x} \mid \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta_0} f(\mathbf{x} \mid \boldsymbol{\theta})} \right] \tag{2.25}$$

The log and the multiplication with 2 is necessary, so that statistic is under the null hypothesis asymptotically chi-square distributed with degrees of freedom $\text{df} = \dim(\hat{\boldsymbol{\theta}}) - \dim(\hat{\boldsymbol{\theta}}_{\mathbf{0}})$ as shown by Wilks (1938).

Besides Wilks theorem, in the context of classification problems and latent class models, the asymptotic distribution of the test statistic λ can be assesed by bootstraping it. We will do this in the applied chapters below to find the optimal number of components $g \in \mathbb{N}$ for a mixture model:

(i) Therefore we start with fitting two models $f_0(\cdot \mid \hat{\boldsymbol{\theta}}_{\mathbf{0}})$ and $f_1(\cdot \mid \hat{\boldsymbol{\theta}}_{\mathbf{1}})$ under the null hypotheses $H_0 : g = g_0$ vs. the alternative $H_1 : g = g_1$ for $g_1 = g_0 + 1$ to the full observed data $\underset{\sim\mathbf{obs}}{\mathbf{x}}$.

(ii) We compute the LRT statistic λ_{obs} for the models based on the observed data $\underset{\sim\mathbf{obs}}{\mathbf{x}}$.

(iii) Then we assume H_0 holds and sample independently from $f_0(\cdot \mid \hat{\boldsymbol{\theta}}_{\mathbf{0}})$ datasets $\underset{\sim\mathbf{b}}{\mathbf{x}}$, $b = 1, ..., B$.

(iv) For each of these bootstrap-datasets $\underset{\sim\mathbf{b}}{\mathbf{x}}$ we fit one model under H_0 and one under H_1 and compute the LRT statistic λ_b for them.

(v) Finally the p-value of the test is the quantile of λ_{obs} of the ordered bootstrap replications $\lambda_b, b = 1, ..., B$. I.e. if there are j bootstrap replications $\lambda_b, b = 1, ..., B$ smaller than λ_{obs}, then it holds p-value$= \dfrac{j}{B+1}$.

Vuong Test Let $\mathbf{X}$ be a random vector following a true distribution with density h_0. Let $\mathbb{E}_0$ be the expected value regarding the true density h_0. Assume we have fitted two models $f(\cdot \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{f}})$ and $g(\cdot \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{g}})$ with the unique maximum likelihood estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{f}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$, which are interior points of Θ_f and Θ_g. Furthermore $\log f$ and $\log g$ must be in h_0-almost all x twice continuously differentiable on Θ_f and Θ_g and the respective derivatives are in h_0-almost all x dominated by h_0 integrable functions. Also in h_0-almost all x are $(\log f(x \mid \cdot))^2$ and $(\log g(x \mid \cdot))^2$ dominated by h_0-integrable functions. More details and explanations on these regularity conditions can be found in Vuong (1989).

Then the Vuong test tests the hypothesis

$$H_0 : \mathbb{E}_0 \left[\log \frac{f(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_f)}{g(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_g)} \right] = 0$$

meaning the models are equivalent against

$$H_1 : \mathbb{E}_0 \left[\log \frac{f(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_f)}{g(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_g)} \right] > 0$$

meaning the model f is better then model g , or

$$H_2 : \mathbb{E}_0 \left[\log \frac{f(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_f)}{g(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_g)} \right] < 0$$

meaning the model g is better then model f.

For the unadjusted Vuong test we go the following steps:

(i) For all observations x_i , $i = 1, .., n$ we compute $m_i = \log \left[\frac{f(\mathbf{X}_i \mid \hat{\boldsymbol{\theta}}_f)}{g(\mathbf{X}_i \mid \hat{\boldsymbol{\theta}}_g)} \right]$.

(ii) We compute

$$LR_n(\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_g) = \sum_{i=1}^{n} m_i \tag{2.26}$$

(iii) We compute $\hat{\omega}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(m_i - \frac{1}{n} LR_n(\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_g) \right)^2$

(iv) Compute the test statistic

$$\nu = \frac{LR_n(\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_g)}{\sqrt{n\hat{\omega}^2}} \tag{2.27}$$

(v) In Vuong (1989) it is shown, that under regularity conditions

$$\frac{1}{n} LR_n(\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_g) \xrightarrow{a.s.} \mathbb{E}_0 \left[\log \frac{f(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_f)}{g(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_g)} \right]$$

and

$$\hat{\omega}^2 \xrightarrow{a.s.} \operatorname{var}_0 \left[\log \frac{f(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_f)}{g(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_g)} \right].$$

Furthermore and most important it is shown for nonnested but also for overlapping models, that ν is asymptoticaly standard normal distributed under H_0. For that reason we can now reject H_0 in favor of H_1, if $\nu > \Phi^{-1}(1 - \frac{a}{2})$, or reject H_0 in favor of H_2, if $\nu < \Phi^{-1}(1 - \frac{a}{2})$.

For the adjusted Vuong test we change the log-likelihood ratio by adding a term penalizing the number of parameters and the number of observations:

$$L\tilde{R}_n(\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_g) = LR_n(\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_g) - K_n(f, g) \tag{2.28}$$

Let $k_1 = \dim(\hat{\boldsymbol{\theta}}_f)$ and $k_2 = \dim(\hat{\boldsymbol{\theta}}_g)$ be the numbers of parameters in f and g. Then two possible correction factors are the Akaike correction $K_n^A = k_1 - k_2$ and the Schwarz correction $K_n^S = \frac{log(n)}{2}(k_1 - k_2)$. These are inspired by the AIC and the BIC. Due to the division by n it is obvious, that

$$\tilde{\nu} = \frac{L\tilde{R}_n(\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_g)}{\sqrt{n\hat{\omega}^2}}$$

follows the same asymptotical distribution as the unadjusted ν under H_0. I.e. $\tilde{\nu} \sim \mathcal{N}(0, 1)$. Therefore the last step of adjusted Vuong test is similar to the unadjusted.

2.4 Modeling densities with directed acyclic graphs

In Chapter 2.2.3 we have already introduced graphs, pathes and cycles. In this chapter we introduce more basic graph terminology and explain how directed acyclic graphs can be used to model densities. The most of the following definitions and concepts can be found in Peters et al. (2017) and Koller and Friedman (2009).

Definition 2.44 (Directed edge and directed graph)**.** *Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $X, Y \in \mathcal{V}$. A pair of nodes can be connected by a directed or undirected edge. We say that there is an undirected edge between two nodes X and Y if $(X, Y) \in \mathcal{E}$ and $(Y, X) \in \mathcal{E}$. Analogous we say that there is an directed edge between X and Y if $(X, Y) \in \mathcal{E}$ and $(Y, X) \notin \mathcal{E}$, or $(Y, X) \in \mathcal{E}$ and $(X, Y) \notin \mathcal{E}$. We call $\mathcal{G}$ a directed graph, if all its edges are directed.*

Definition 2.45 (Parent and child nodes)**.** *Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $X, Y \in \mathcal{V}$. If $(X, Y) \in \mathcal{E}$ and $(Y, X) \notin \mathcal{E}$, then we call X a patent node of Y and Y a child node of X. We call $\pi(Y)$ the set of parent nodes of Y.*

Definition 2.46 (Directed Acyclic Graph (DAG))**.** *A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is called a Directed Acyclic Graph (DAG), if it is a directed graph and it contains no cycles.*

Given this background, we are now using DAGs for modeling densities: Let $\mathcal{V} = \{X_1, ..., X_d\}$ be a set of random variables with a joint density $f_{1,...,d}$. Now consider the decomposition

$$\begin{aligned} f_{1,\dots,d}(x_1, \dots, x_d) &= f_{d|1,...,d-1}(x_d|x_1, ..., x_{d-1})\, f_{1,...,d-1}(x_1, ..., x_{d-1}) \\ &= ... = \left[\prod_{i=2}^{d} f_{i|1,...,i-1}(x_i|x_1, ..., x_{i-1})\right] f_1(x_1) \end{aligned} \tag{2.29}$$

which has already been discussed by Czado (2019), among others. Now we find sets $P_i \subset \{1, ..., i-1\}$ such that

$$f(x_i|x_1, ..., x_{i-1}) = f\left(x_i|\{x_j : j \in P_i\}\right)$$

for all $i = 1, ..., d$ and define $\pi(X_i) = \{X_j : j \in P_i\}$. Due to the Markov assumptions it does not have to be $\pi(X_i) = \{X_1, ..., X_{i-1}\}$ for $i = 1, ..., d$.

Thus, we can construct a DAG on a set of random variables. The topological order of the random variables can be optimized mathematically or with prior knowledge. We

are doing both in the Chapters 4.3 and 6. In these chapters we are working with D-vine copula regression models as previously Czado and Scharl (2021) did, and whose concept was first developed for by Kraus and Czado (2017). There the conditional density of a random variable X_i, given its parent nodes $\pi(X_i) = \{X_1^{(i)}, ..., X_{d_i}^{(i)}\}$, is

$$f(x_i|\pi(X_i) = \pi(x_i)) = \left[\prod_{k=1}^{d_i-2} c_{i,k;1:k-1}\right] \cdot c_{i,d_i} \cdot f(x_i)$$

with $c_{i,d_i} = c_{i,d_i}\left(F(x_i), F(x_{d_i}^{(i)})\right)$ and

$$c_{i,k;1:k-1} = c_{i,k;1:k-1}\left(F_{i|1:k-1}\left(x_i|x_1^{(i)}, ..., x_{d_i}^{(i)}\right), F_{k|1:k-1}\left(x_k|x_1^{(i)}, ..., x_{d_i}^{(i)}\right)\right).$$

When we apply it later, the order of the D vine is not fixed and is estimated from the data. Still, we fix the corresponding parent sets.

Chapter 3

The Sachs dataset

3.1 Introduction and short biological background

In the following chapters, we are working with the "Sachs Protein Data" dataset. This was first analyzed by Sachs et al. (2005) and consists of measurements of multiple phosphorylated protein and phospholipid components in human immune system cells. The initial goal was to find signaling relationships and to model them with Bayesian networks.

To better evaluate the results of the following chapters, we need an understanding of the biological background of the data. The data are from 14 experiments in which human naive (i.e. inactive) CD4 T-lymphocytes were activated with anti-CD3 and anti-CD28 antibodies and infused with different activators or inhibitors. Only in experiments 8 and 9 no anti-CD3/CD28 was used, but the molecule phorbol 12-myristate 13-acetate (PMA) which activates protein kinase C (`PKC`) and the molecule b2 cyclic adenosine monophosphate β2 cAMP which activates `PKA`. All reagents act as activators or inhibitors on the various phosphoproteins and -lipids, which are available as variables in the dataset.

Table 3.1 shows which reagents were used as stimulants in the experiments, and which variables are influenced directly by the stimulants. In Table 3.2 is explicitely shown, which reagent influences which observed variable.

In the experiments, the cell reactions were stopped after 15 minutes and then the 11 phosphorylated proteins and phospholipids were measured simultaneously by multivariate flow cytometry in each cell. Thus, the individual observations are independent. The value of a variable is the quantitative amount of the respective molecules measured in the cell. The variables describe proteins and lipids that were phosphorylated at different sites (i.e. a phosphoryl group was attached to a specific amino acid). Serine (S), threonine (T) and tyrosine (Y) are considered here. However, which variable is phosphorylated at which position does not play an explicit role in the further master thesis. Therefore, reference is

Exp.	Stimulation	Directly influenced variables
1.	Anti-CD3/CD28	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+)
2.	Anti-CD3/CD28, ICAM-2	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+)
3.	Anti-CD3/CD28, akt-inhibitor	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), pakts473 (-)
4.	Anti-CD3/CD28, G0076	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+/-)
5.	Anti-CD3/CD28, Psitectorigenin	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), PIP2 (-)
6.	Anti-CD3/CD28, U0126	plcg (+), praf (+), pmek (+/-), p4442 (+/-), PKC (+)
7.	Anti-CD3/CD28, LY294002	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), pakts473 (+)
8.	PMA	PKC (+)
9.	β 2camp	PKA (+)
10.	Anti-CD3/CD28, ICAM-2, akt-inhib	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), pakts473 (-)
11.	Anti-CD3/CD28, ICAM-2, G0076	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+/-)
12.	Anti-CD3/CD28, ICAM-2, Psitectorigenin	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), PIP2 (-)
13.	Anti-CD3/CD28, ICAM-2, U0126	plcg (+), praf (+), pmek (+/-), p4442 (+/-), PKC (+)
14.	Anti-CD3/CD28, ICAM-2, LY294002	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), pakts473 (+)

Table 3.1: Stimulations used in the different experiments, as well as the variables, that are biochemically activated (+) or inhibited (-) by the stimulations. Variables directly influenced by the stimulants in opposite ways, are marked with (+/-).

Reagent	Influence on
Anti-CD3/CD28	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+)
ICAM-2	
β2 cAMP	PKA (+)
akt-inhibitor	pakts473 (-)
U0126	pmek (-), p4442 (-)
PMA	PKC (+)
G06976	PKC (-)
Psitectorigenin	PIP2 (-)
LY294002	pakts473 (+)

Table 3.2: Known biological effects of the reagents employed. Intercellular adhesion molecule–2 (ICAM-2) overall stimulated the cell, with no specific perturbation on the measured molecules. (Source: Sachs et al. 2005, p.525)

made to Sachs et al. (2005) and more detailed literature on biochemistry.

Since experiments 3-7 and 10-14 differ only by the additional intervention of the protein intercellular adhesion molecule 2 (ICAM-2), we might expect the results of these experiments to have similarities. Moreover, for the same reason, the observations of experiment 2 might have similarities with experiments 10-14. As described earlier, PMA activates `PKC` and G06976 inhibits `PKC`. This could cause experiment 8 to produce particularly different results compared to experiments 4 and 11. Also, experiment 9, which also did not use anti-CD3 and anti-CD28, may have unique features. We will investigate this question during the clustering in Chapter 5.

Not only do the activators and inhibitors used in the experiments influence the measured molecules (ie. variables), but they also influence each other. For example, `pmek` is the quantitative amount of mitogen-activated protein kinase kinase 1 and 2. As a kinase

kinase, this enzyme catalyzes the phosphorylation of proteins to mitogen-activated protein kinase 1 and 3, which was measured as variable `p4442`. Such relationships between variables exist partly directly, as just described, and partly indirectly via unmeasured variables. These relationships have been extensively explored and resulted in the consent graph shown in Figure 3.1.

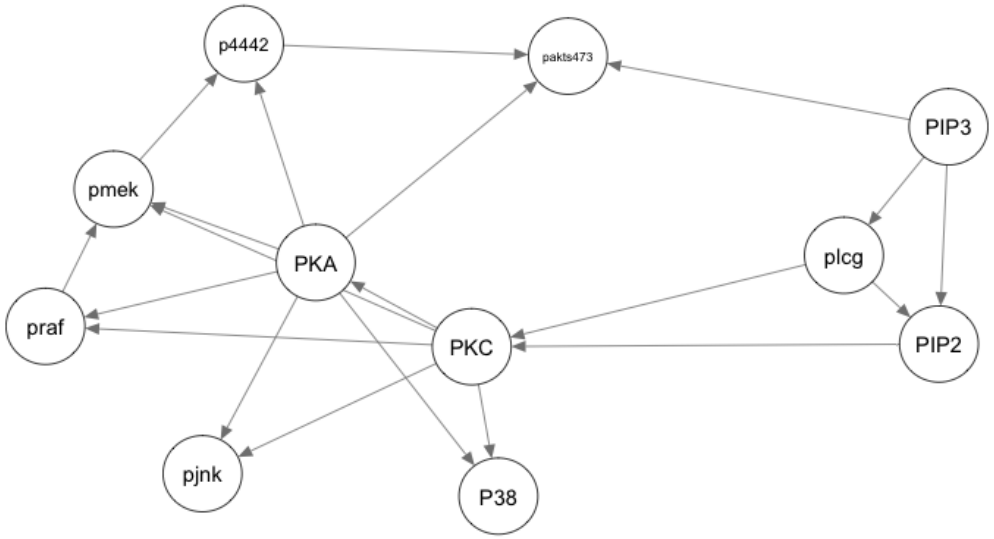


Figure 3.1: Consent DAG of the Sachs data. Arcs correspond to direct or indirect biochemical influences.

3.2 Data preprocessing

We are working with all 14 experiments and all 11 variables of the Sachs dataset. But instead of the original data, we apply the logarithm to all variables to receive the log-data. In the resulting 14-experiment-dataset we now have 1523 observations for which at least one variable equals zero. These would lead to zero-inflated margins as we can see especially for the variables `PKC` or `pjnk`, so we delete these observations. Therefore the final dataset (the log-data with removed zeros), with which we are working, has 14 experiments, 11 variables and a total of 10149 observations. In Figure 3.2 we can see pairwise scatter plots and empirical marginal densities of the resulting dataset.

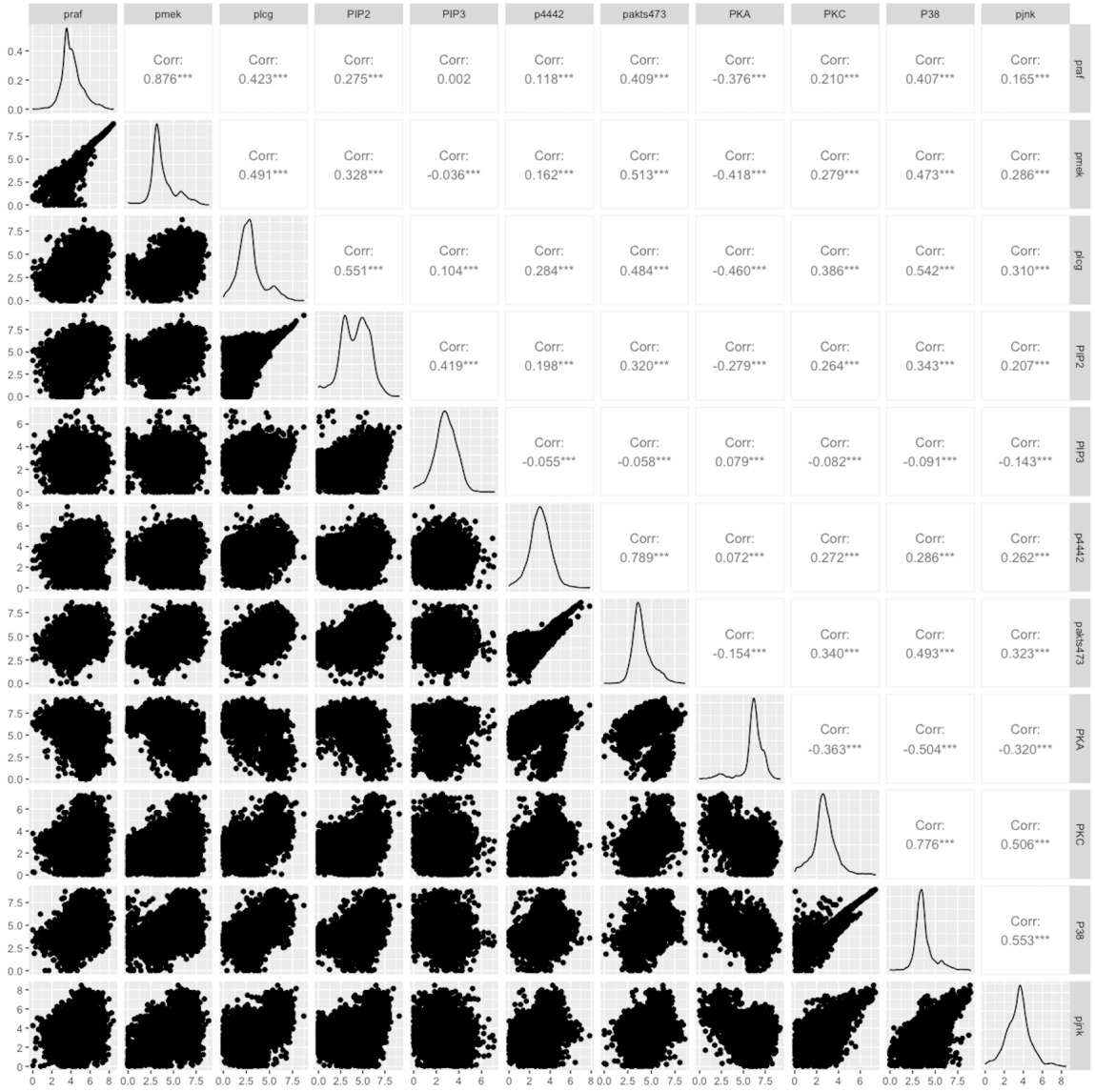


Figure 3.2: Pairwise scatter plots and empirical marginal densities of the log-data with zeros removed

Chapter 4

Analysis of full pooled data

In this chapter we would like to understand the dataset just discussed better. For simplicity, we do now not take into account that the data come from different experiments. This will be addressed in the next chapter with clustering. From Figure 3.1 one could assume that an approach with Gaussian distribution - apart from the clear multimodality of the variable `PIP2` - would be reasonable. For this reason we will start with a multivariate Gaussian analysis. Later in this chapter we will also use Vine copulas and marginals different from the Gaussian marginal to create a joint distribution.

4.1 Multivariate Gaussian analysis

First we would like to get an insight into how correlated the variables are with one another. To do this, we consider the coefficient Kendall's Tau. The results are shown in Figure 4.1. Green means the variables are more or less uncorrelated, red means they are very correlated (positively or negatively).

We can see that the variable pairs `praf` and `pmek`, as well as `p4442` and `pakts473` and `PKC` with `P38` are positively correlated with Kendall's Tau values between 0.587 - 0.66. Besides the pairs `PIP2` and `PIP3` and `PIP2` and `plcg` all other pairs of variables have absolute Kendall's Tau values below 0.3 what we interpret as weak correlation.

We now consider a multivariate Gaussian model for all $d = 11$ variables. Thus, we assume the data is $\mathcal{N}_{11}(\boldsymbol{\mu}, \Sigma)$ distributed. For the full probability density function see equation (2.2). For our data, $\boldsymbol{\mu}$ and Σ can be easily estimated and are given by

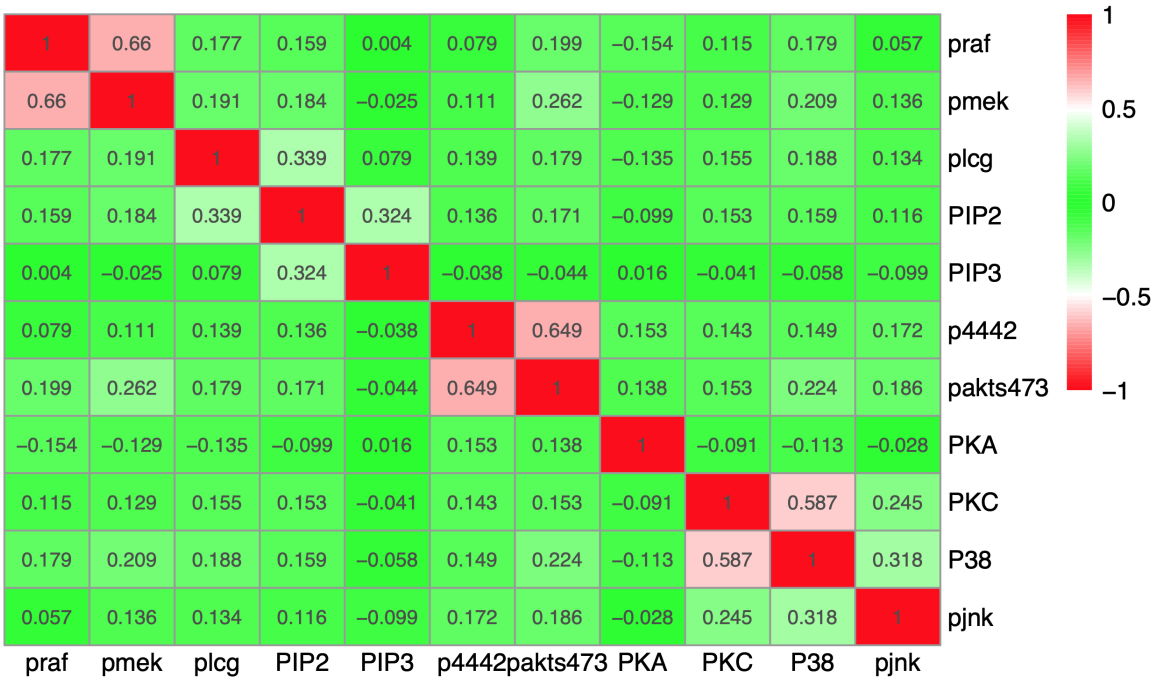


Figure 4.1: Empirical Kendall's τ of the variables for the pooled Sachs data.

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 4.096 \\ 3.794 \\ 2.887 \\ 4.126 \\ 2.820 \\ 3.035 \\ 4.022 \\ 6.106 \\ 2.694 \\ 3.765 \\ 3.457 \end{pmatrix}$$

and

$$\hat{\Sigma} = \begin{pmatrix} 1.091 & 1.213 & 0.597 & 0.447 & 0.002 & 0.122 & 0.444 & -0.510 & 0.225 & 0.470 & 0.208 \\ 1.213 & 1.755 & 0.879 & 0.676 & -0.045 & 0.214 & 0.706 & -0.718 & 0.379 & 0.694 & 0.457 \\ 0.597 & 0.879 & 1.826 & 1.159 & 0.136 & 0.383 & 0.680 & -0.806 & 0.535 & 0.810 & 0.506 \\ 0.447 & 0.676 & 1.159 & 2.418 & 0.627 & 0.308 & 0.516 & -0.561 & 0.420 & 0.591 & 0.389 \\ 0.002 & -0.045 & 0.136 & 0.627 & 0.923 & -0.053 & -0.058 & 0.099 & -0.081 & -0.097 & -0.165 \\ 0.122 & 0.214 & 0.383 & 0.308 & -0.053 & 0.993 & 0.817 & 0.093 & 0.278 & 0.315 & 0.315 \\ 0.444 & 0.706 & 0.680 & 0.516 & -0.058 & 0.817 & 1.080 & -0.208 & 0.362 & 0.567 & 0.405 \\ -0.510 & -0.718 & -0.806 & -0.561 & 0.099 & 0.093 & -0.208 & 1.680 & -0.483 & -0.723 & -0.501 \\ 0.225 & 0.379 & 0.535 & 0.420 & -0.081 & 0.278 & 0.362 & -0.483 & 1.050 & 0.879 & 0.626 \\ 0.470 & 0.694 & 0.810 & 0.591 & -0.097 & 0.315 & 0.567 & -0.723 & 0.879 & 1.223 & 0.738 \\ 0.208 & 0.457 & 0.506 & 0.389 & -0.165 & 0.315 & 0.405 & -0.501 & 0.626 & 0.738 & 1.458 \end{pmatrix}.$$

However, the estimated mean vector and the covariance matrix alone do not give us information on how well the multivariate Gaussian model fits the data. The question is to what extent specific properties of the data are captured by this model. To do this, we sample 10149 data points (as many as our observed dataset has) with this multivariate distribution and plot them against the measured data. For sampling the data we use the **mvtnorm** package of Hothorn (2014) in R. The determinant of the covariance matrix is 0.039. If the determinant was very close to zero, we would have numerical problems when sampling due to the necessary inversion of the covariance matrix. However, this is not the case here.

In Figure 4.2 we can see the pairs plots of the observed dataset in the lower triangle and the dataset sampled from the multivariate Gaussian Model in the upper triangle. If the distribution of the two datasets were identical, then the plots in the upper triangle would be the plots from the lower mirrored over the diagonal. There are some pairs of variables like `praf` and `plcg` or `PIP3` and `p4442`, for which the sampled data look like the original observed data. These bivariate relationships were well covered by the model. However, there are many pairs of variables for which the multivariate Gaussian model approach obviously does not fit. The specifics of the variable pair `praf` and `pmek` or almost all variables to `PKC` cannot be handled with this model. For this reason, we will work with a vine copula model in the following section. In this model, those relations such as `praf` and `pmek` will be better taken into account.

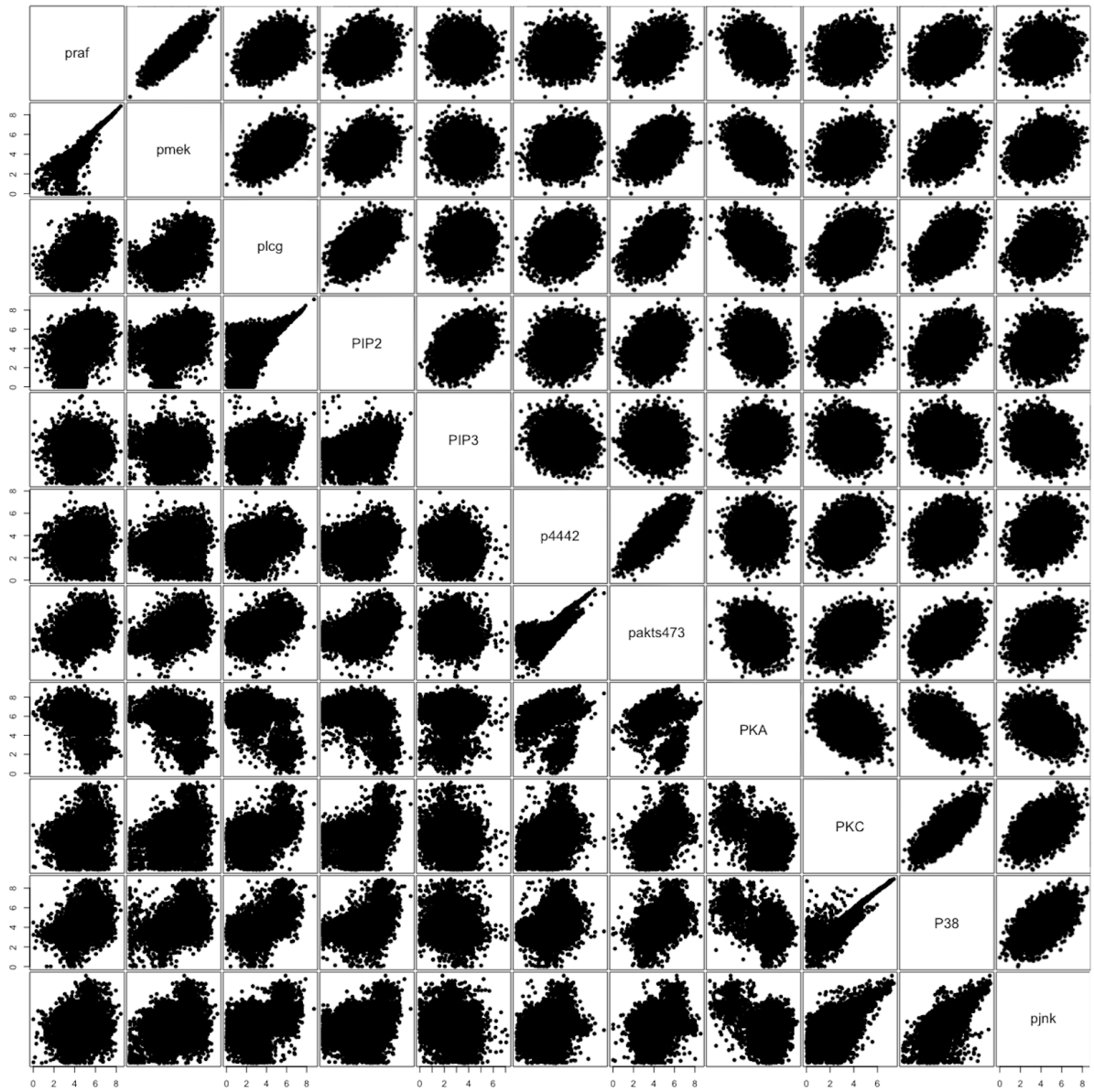


Figure 4.2: Pairs plots of observed dataset (lower triangle) and dataset sampled from the multivariate Gaussian Model (upper triangle).

4.2 Vine copula based analysis

We will now fit a vine copula model. For this we allow as candidates for the marginal distribution the normal, log-normal, logistic, log-logistic, gamma, t-fix, skew-normal and skew-t distribution. For the copula families we allow Gaussian, t, Clayton, Gumbel, Frank, Joe, BB1, BB6, BB8 copulas and their rotations as definined in Definition 2.18.

First, we estimate the marginal distributions to obtain the u-scale data. The results of this estimation are shown in Table 4.1. The marginal families and the parameters were fitted by optimizing the AIC. This means, we take the number of parameters into account. It is interesting that almost throughout skew distributions were chosen. Figure 4.3 shows the histograms and the curves of the estimated densities.

Variable	Family	Parameters	Marginal Loglikelihood
PIP3	skew-t	(2.821, 0.961, 23.909, 0.898)	-13951.23
plcg	skew-t	(2.887, 1.390, 5.062, 1.390)	-16872.77
PIP2	skew-normal	(4.145, 1.563, 0.756)	-18766.96
PKC	skew-t	(2.687, 1.067, 4.157, 1.016)	-14295.35
PKA	skew-t	(6.220, 9.896, 2.008, 0.886)	-14706.29
P38	skew-t	(3.790, 1.988, 2.250, 1.504)	-13612.65
pjnk	logistic	(3.460, 0.665)	-16179.19
praf	skew-t	(4.113, 1.087, 4.848, 1.407)	-14281.63
pmek	skew-t	(3.909, 3.797, 2.111, 1.670)	-15857.25
p4442	skew-t	(3.035, 0.998, 13.161, 0.964)	-14322.43
pakts473	skew-t	(4.035, 1.069, 5.321, 1.554)	-14014.56

Table 4.1: Marginal family and parameter estimates for the pooled Sachs data.

The results are mixed. We see a very good fit for some variables, but very poor results for others like `plcg`, `PIP2` or `PKA`. Problems here are often multimodality as it is well seen especially for `PIP2`. However, this is indicative of a mixture distribution with multiple components. Therefore, we will discuss this problem in the clustering chapter.

This gives us now a marginal distribution for each variable, so that for each observation $x_{i,j}, j = 1, ..., d$ we can compute the pseudo copula data $u_{i,j} = \hat{F}_j(x_{i,j})$ for $i = 1, ..., n$. With this, we transform the data to u-scale. The result is shown in Figure 4.4. If we look at the plots on the diagonal of Figure 4.4, the variables `plcg`, `PIP2` or `PKA` are particularly striking, as their plots look the least uniform for them. This is not surprising, as for these the marginal distributions did not fit ideally as we have seen just before.

With this data we can now fit a vine copula model. We use the R library **VineCopula** of Nagler et al. (2019). The structure is left open as R-Vine. As explained above, we use

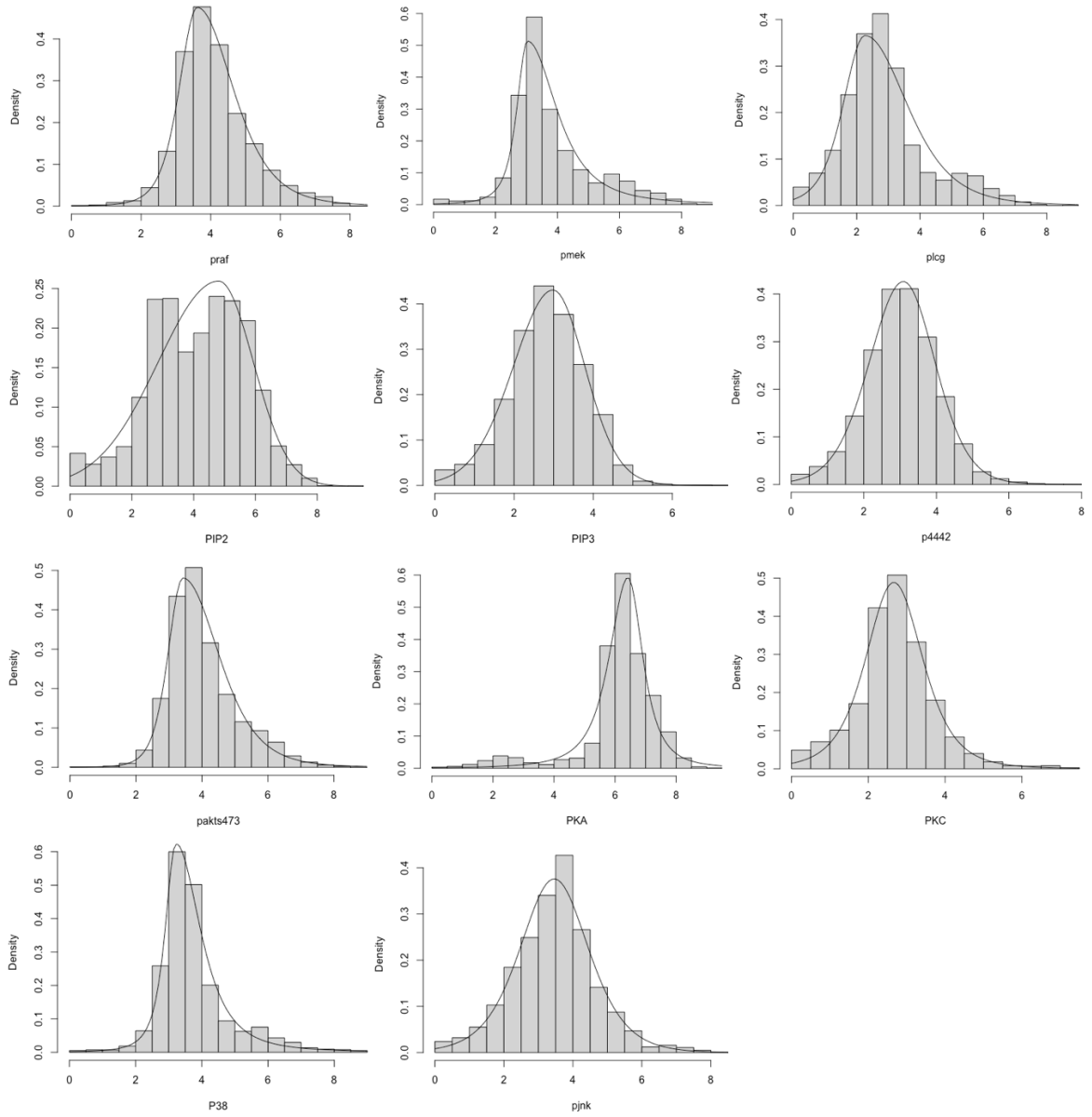


Figure 4.3: Histograms and estimated parametric densities according to Table 4.1 for all 11 variables of the pooled Sachs data.

the copula families Gaussian, t, Clayton, Gumbel, Frank, Joe, BB1,BB6,BB8 and their rotations. Due to the size, we do not include a table with the estimated parameters here. But in Figure 4.5 we can see the vine tree structure of the first two trees.

In the previous section, we finally sampled data from the discussed model. We then examined the results in a pairs plot. However, the results were mixed and but not satisfactory for some pairs of variables. Here we would like to proceed analogously and see if the vine copula model can deal with more specifics from the data. Therefore we sample again 10149 data points (as many as our observed dataset has) from this distribution and compare them to the measured data. However, it makes sense to compare the data directly on u-scale, because we have the contour data available in addition to the pairs plots and we would transform the data with the same marginal distributions to the x-scale. The observed data on u-scale are shown before in Figure 4.4; the sampled data in Figure 4.6.

Comparing Figure 4.4 and Figure 4.6, we immediately see that this model suits the data much better. Especially for e.g. the variable-pairs `praf` and `pmek` or `PKC` and `P38` the results are better. Nevertheless, the characteristics of e.g. `PKA` to almost all other variables have not yet been sufficiently reproduced. As already mentioned, problems already occurred when transforming the data from x-scale to u-scale, because, for example, the multimodality of some variables cannot be captured by our candidate marginal distributions when considering the entire pooled data without taking the individual experiments into account. We expect that these problems will no longer occur after clustering into multiple components.

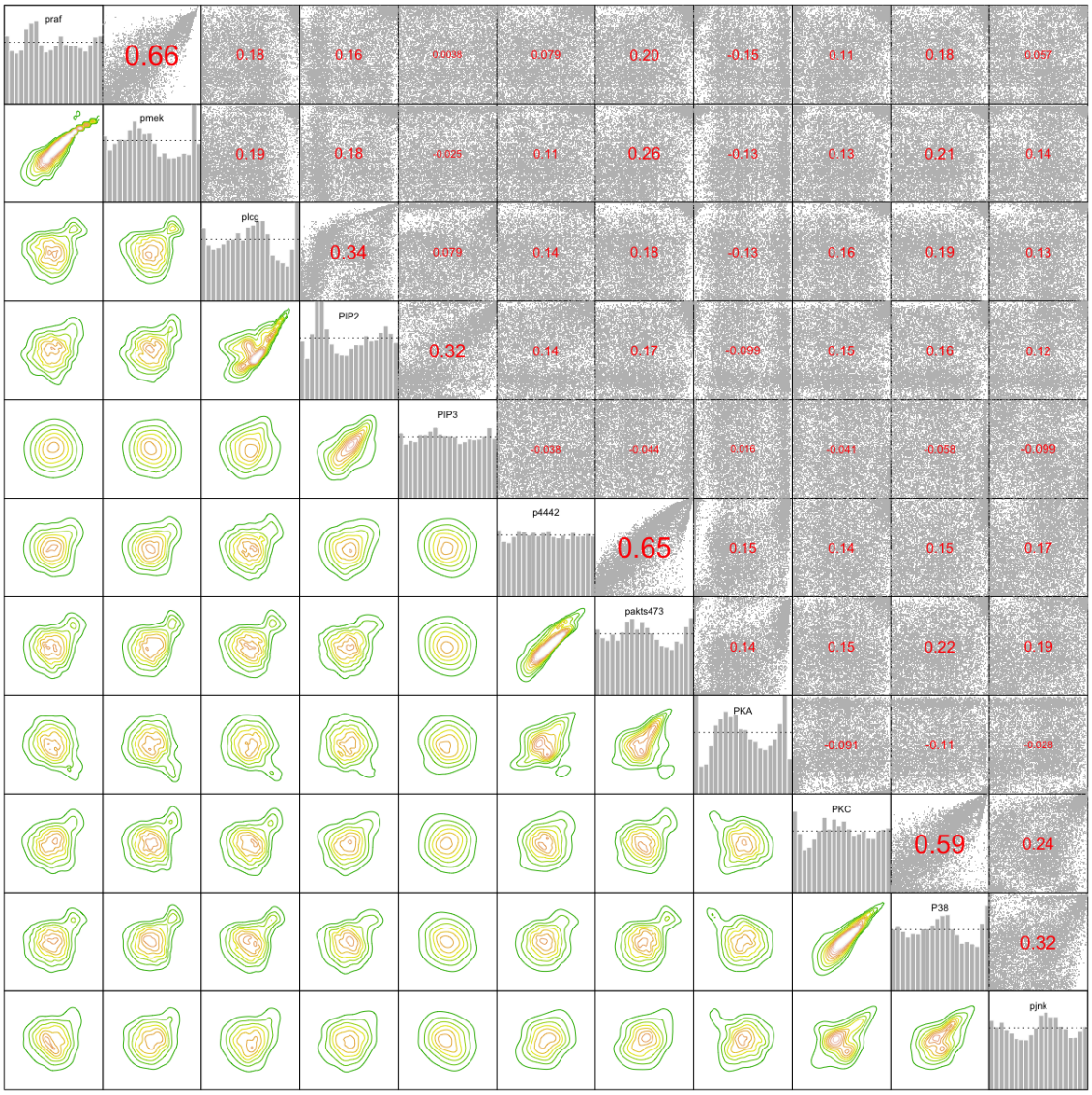
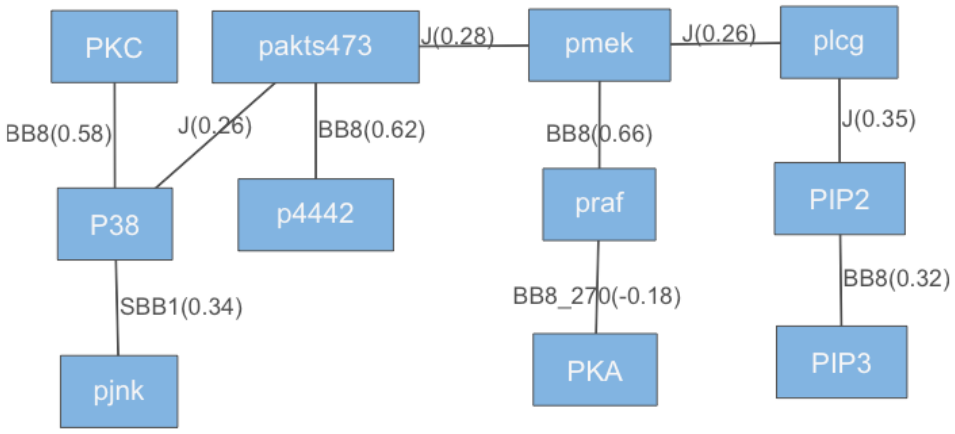
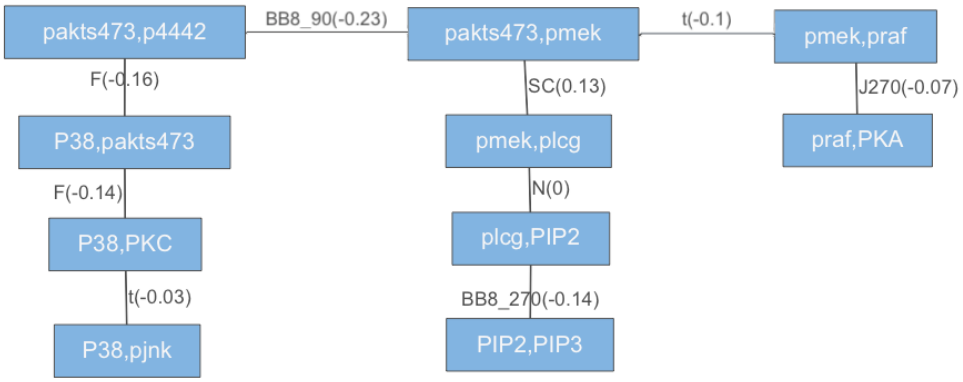


Figure 4.4: Normalized contours on the z-scale and pairs plots on the u-scale of the pooled data.



(a) Tree 1



(b) Tree 2

Figure 4.5: Estimated vine tree structures of the first two trees of the vine copula model of the pooled data.

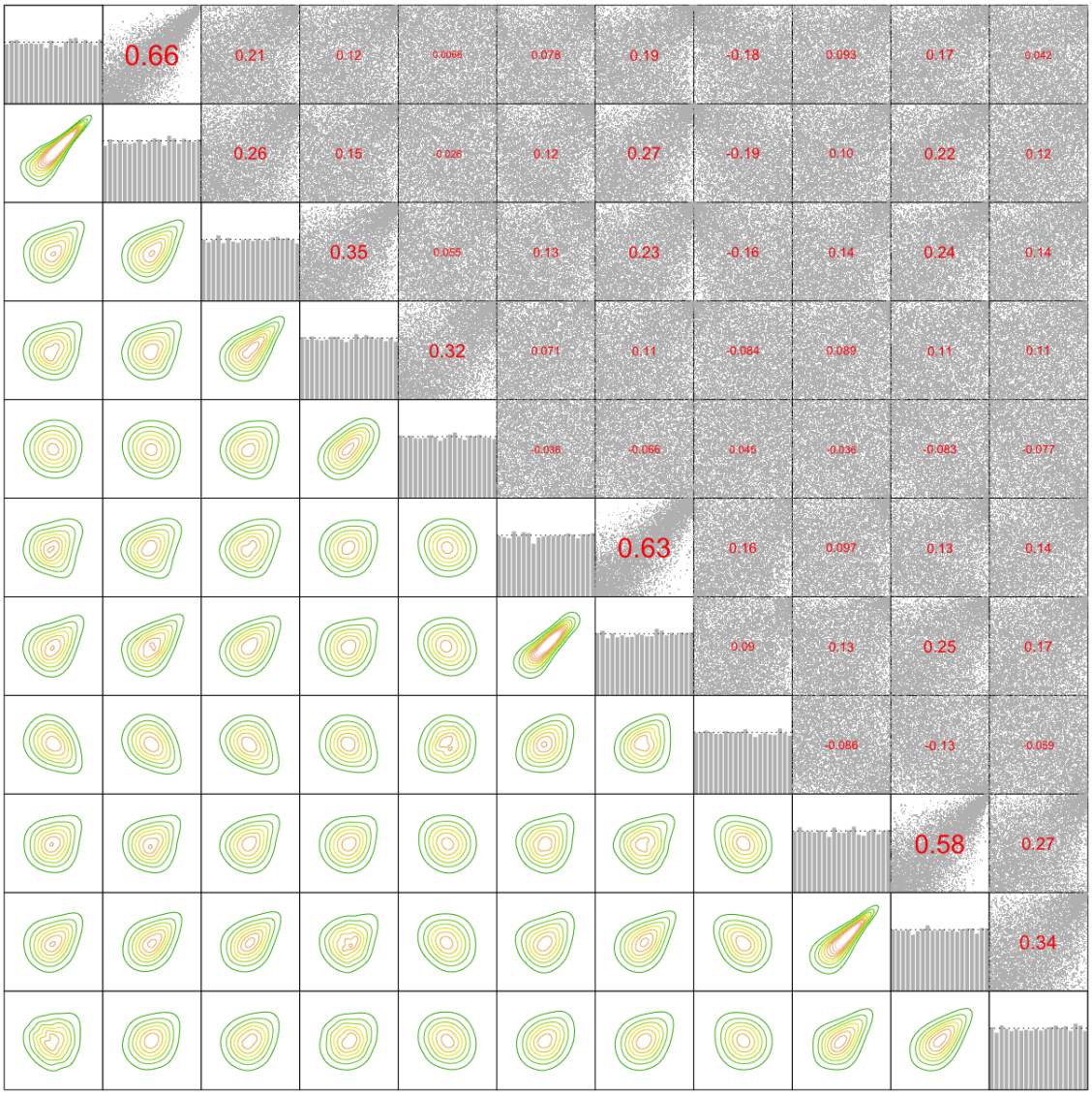


Figure 4.6: Normalized contours on the z-scale and pairs plots on the u-scale of the simulated data. Simulation is based on the estimated vine copula model of the pooled data.

4.3 Directed acyclic graphs

In the previous sections we used a multivariate Gaussian distribution and a vine copula model to fit the distribution of the data. Another approach to model the density are directed acyclic graphs (DAG). We want to fit three DAG model, by using the consent graph, which was introduced in Chapter 3.1. We use the **vinereg** library by Nagler and Kraus (2021). The corresponding D-vine regression models were introduced by Kraus and Czado (2017). For each variable except `PIP3` (which has no parent nodes) we estimate a regression D-vine considering the parent nodes. For the first model (given in Table 4.3), we transform the data to u-scale by fitting Gaussian margins, and then allow only the Gaussian copula family as well as the indepencence copula. We denote this model as $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$. In the second model (given in Table 4.4) we transfom the data to u-scale by the margins, we have already fitted in the previous section in Table 4.1. But in the second model we still use only the Gaussian copula family as well as the indepencence copula for the regressin D-vines ($\mathbf{M}_{Par}\mathbf{C}_{Gauss}$). In the third model (given in Table 4.5), we use again all the parametric margins fitted in Chapter 4.2 and we allow all parametric copula families ($\mathbf{M}_{Par}\mathbf{C}_{Par}$).

We start with fitting the Gaussian marginal distributions, which we will use for the first model. The results are given in Table 4.2.

Variable	Family	Parameters	Marginal Loglikelihood
PIP3	normal	2.82, 0.961	-13994.331
plcg	normal	2.887, 1.351	-17454.730
PIP2	normal	4.126, 1.555	-18880.995
PKC	normal	2.694, 1.025	-14648.882
PKA	normal	6.106, 1.296	-17032.228
P38	normal	3.765, 1.106	-15422.559
pjnk	normal	3.457, 1.207	-16312.928
praf	normal	4.096, 1.044	-14840.097
pmek	normal	3.794, 1.325	-17255.844
p4442	normal	3.035, 0.997	-14365.586
pakts473	normal	4.022, 1.039	-14791.794

Table 4.2: Parameter estimates for normal marginal distributions to the pooled Sachs data.

Besides the variables `PKC` and `pakts473`, the same arcs were modeled and except from the variable `p4442` the order of the variables in the D-vines similar. While comparing the second and the third model, we can see that of course, the values of the copula-loglikelihood, -AIC and -BIC are better in the third model with fewer restrictions regarding the copula families. On the other hand, it is very interesting that the values of the

copula-loglikelihood, -AIC and -BIC are better in the first model than in the second. The reason for this effect is that we fitted the first model only with the data at u-scale, which we had calculated exclusively with gaussian margins, while we dropped this restriction for the second model.

In the previous section, we finally sampled data from the discussed models. As we have fitted the D-vine regression models to the pooled data on u-scale, it would once again make sense to also compare the results from the sampling with the original data on u-scale. The sampling itself is not as trivial as from the previously discussed models. Here we apply the corresponding algorithm of Bevacqua et al. (2017) sequentially for all nodes respectively. The results are shown in Figure 4.7, Figure 4.8 and Figure 4.9.

We have to compare the results from the D-vine regression models shown in Figures 4.7, 4.8 and 4.9 to the original data on u-scale shown previously in Figure 4.4. It will also be interesting to compare them to the sampling results from the vine copula model of the last subchapter shown in Figure 4.6. It is interesting, that the three D-vine regression models have the same parent nodes for almost each node, but the order of the regression D-vines is different for the variable `p4442`. The D-vine regression models are optimized with respect to the AIC.

As one can see, the data sampled from the third model with all parametric copula families have on the one hand many similarities with the original data, which are not equally clear in the limited models with only the Gaussian copula family. Since one can not catch the effect of upper or lower tail dependence with copulas of the Gaussian family, these models have problems capturing the specifics of the data. The larger model from Table 4.5 has no Gaussian copulas at all. However, since we define the DAG through the consent graph, we also see properties of the original data that were not even captured by the D-vine regression model with all parametric copula families. For example, if we look at the three variables `praf`, `pmek` and `plcg`, we see that the relationship between `praf` and `pmek` is very well modeled. This is due to the fact that in the model the variable `pmek` depends on `praf`, `PKC` and`PKA`. In contrast, the relationship between `praf` and `plcg` or `pmek` and `plcg`, which is clearly visible in Figure 4.4, was not captured by our model, since `plcg` for example only indirectly influences `praf` via the node `PKC`. And through `PKC` and `praf` then `pmek`. This is also interesting when comparing the simulated data of the D-vine regression model to those of the vine copula model shown in Figure 4.6, where the correlations of `praf` and `plcg` or `pmek` and `plcg` are better visible. Again, the reason is that in the D-vine regression model we fixed the graph, whereas in the vine copula model it was estimated from the data.

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC
plcg	PIP3	55.626	-109.252	-102.026
PIP2	plcg, PIP3	2965.669	-5925.338	-5903.663
PKC	plcg, PIP2	2680.441	-5354.883	-5333.207
PKA	PKC	686.261	-1370.522	-1363.297
P38	PKC, PKA	6148.889	-12291.779	-12270.103
pjnk	PKC, PKA	2332.515	-4659.031	-4637.356
praf	PKA, PKC	1498.523	-2991.046	-2969.371
pmek	praf, PKC, PKA	9232.084	-18452.167	-18408.816
p4442	pmek, PKA	1200.909	-2395.818	-2374.143
pakts473	p4442, PKA	5334.243	-10662.486	-10640.81
Σ		32135.16	-64212.322	-64002.792

Variable	Pair copula	Family	Parameters	Tau
plcg	plcg, PIP3	gaussian	0.104	0.067
PIP2	PIP2, plcg	gaussian	0.551	0.372
	plcg, PIP3	gaussian	0.104	0.067
	PIP2, PIP3; plcg	gaussian	0.436	0.287
PKC	PKC, plcg	gaussian	0.386	0.252
	plcg, PIP2	gaussian	0.551	0.372
	PKC, PIP2; plcg	gaussian	0.066	0.042
PKA	PKA, PKC	gaussian	-0.295	-0.191
P38	P38, PKC	gaussian	0.776	0.566
	PKC, PKA	gaussian	-0.295	-0.191
	P38, PKA; PKC	gaussian	-0.387	-0.253
pjnk	pjnk, PKC	gaussian	0.506	0.338
	PKC, PKA	gaussian	-0.295	-0.191
	pjnk, PKA; PKC	gaussian	-0.173	-0.111
praf	praf, PKA	gaussian	-0.377	-0.246
	PKA, PKC	gaussian	-0.295	-0.191
	praf, PKC; PKA	gaussian	0.087	0.055
pmek	pmek, praf	gaussian	0.877	0.680
	praf, PKC	gaussian	0.210	0.135
	PKC, PKA	gaussian	-0.295	-0.191
	pmek, PKC; praf	gaussian	0.202	0.129
	praf, PKA; PKC	gaussian	-0.335	-0.217
	pmek, PKA; praf, PKC	gaussian	-0.148	-0.094
p4442	p4442, pmek	gaussian	0.162	0.104
	pmek, PKA	gaussian	-0.352	-0.229
	p4442, PKA; pmek	gaussian	0.16	0.102
pakts473	pakts473, p4442	gaussian	0.789	0.579
	p4442, PKA	gaussian	0.084	0.053
	pakts473, PKA; p4442	gaussian	-0.124	-0.079

Table 4.3: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ - pooled data: Summary of the D-vine regression model fitted to the pooled data with Gaussian margins, using only the Gaussian copula family and the indepencence copula.

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC
plcg	PIP3	65.103	-128.207	-120.982
PIP2	plcg, PIP3	2777.391	-5548.782	-5527.107
PKC	plcg, PIP2	2239.298	-4472.597	-4450.921
PKA	PKC	266.239	-530.477	-523.252
P38	PKC, PKA	4623.152	-9240.304	-9218.628
pjnk	PKC, PKA	1476.262	-2946.524	-2924.848
praf	PKA, PKC	766.946	-1527.891	-1506.216
pmek	praf, PKC, PKA	6604.347	-13196.695	-13153.344
p4442	PKA, pmek	805.02	-1604.04	-1582.365
pakts473	p4442, PKA, PIP3	5414.014	-10816.027	-10772.677
Σ		25037.772	-50011.544	-49780.34

Variable	Pair copula	Family	Parameters	Tau
plcg	plcg, PIP3	gaussian	0.113	0.072
PIP2	PIP2, plcg	gaussian	0.524	0.351
	plcg, PIP3	gaussian	0.113	0.072
	PIP2, PIP3; plcg	gaussian	0.439	0.289
PKC	PKC, plcg	gaussian	0.312	0.202
	plcg, PIP2	gaussian	0.524	0.351
	PKC, PIP2; plcg	gaussian	0.136	0.087
PKA	PKA, PKC	gaussian	-0.226	-0.145
P38	P38, PKC	gaussian	0.752	0.542
	PKC, PKA	gaussian	-0.226	-0.145
	P38, PKA; PKC	gaussian	-0.160	-0.102
pjnk	pjnk, PKC	gaussian	0.455	0.301
	PKC, PKA	gaussian	-0.226	-0.145
	pjnk, PKA; PKC	gaussian	-0.075	-0.048
praf	praf, PKA	gaussian	-0.280	-0.180
	PKA, PKC	gaussian	-0.226	-0.145
	praf, PKC; PKA	gaussian	0.128	0.082
pmek	pmek, praf	gaussian	0.821	0.613
	praf, PKC	gaussian	0.183	0.117
	PKC, PKA	gaussian	-0.226	-0.145
	pmek, PKC; praf	gaussian	0.146	0.093
	praf, PKA; PKC	gaussian	-0.249	-0.160
	pmek, PKA; praf, PKC	gaussian	-0.035	-0.023
p4442	p4442, PKA	gaussian	0.194	0.124
	PKA, pmek	gaussian	-0.263	-0.169
	p4442, pmek; PKA	gaussian	0.215	0.138
pakts473	pakts473, p4442	gaussian	0.795	0.585
	p4442, PKA	gaussian	0.194	0.124
	PKA, PIP3	gaussian	0.042	0.027
	pakts473, PKA; p4442	gaussian	-0.110	-0.070
	p4442, PIP3; PKA	gaussian	-0.064	-0.041
	pakts473, PIP3; p4442, PKA	gaussian	-0.021	-0.014

Table 4.4: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ - pooled data: Summary of the D-vine regression model fitted to the pooled data, using only the Gaussian copula family and the indepencence copula. The marginal distributions given in Table 4.1 were used.

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC
plcg	PIP3	81.774	-161.549	-154.324
PIP2	plcg, PIP3	4264.177	-8518.354	-8482.228
PKC	plcg	1141.21	-2280.42	-2273.195
PKA	PKC	865.729	-1729.458	-1722.232
P38	PKC, PKA	6794.476	-13578.951	-13542.826
pjnk	PKC, PKA	2744.762	-5479.523	-5443.397
praf	PKA, PKC	1722.171	-3436.341	-3407.441
pmek	praf, PKC, PKA	9782.278	-19544.555	-19472.304
p4442	PKA, pmek	1641.93	-3273.861	-3237.735
pakts473	p4442, PKA, PIP3	6729.569	-13437.137	-13357.661
Σ		35768.076	-71440.149	-71093.343

Variable	Pair copula	Family	Rotation	Parameters	Tau
plcg	plcg, PIP3	clayton	180	0.141	0.066
PIP2	PIP2, plcg	bb7	0	1.967, 0.071	0.362
	plcg, PIP3	clayton	180	0.141	0.066
	PIP2, PIP3; plcg	bb8	0	2.431, 0.882	0.327
PKC	PKC, plcg	joe	0	1.489	0.216
PKA	PKA, PKC	joe	90	1.337	-0.160
P38	P38, PKC	bb8	0	3.686, 0.991	0.581
	PKC, PKA	joe	270	1.337	-0.16
	P38, PKA; PKC	bb8	270	1.151, 0.981	-0.069
pjnk	pjnk, PKC	bb7	0	1.687, 0.165	0.316
	PKC, PKA	joe	270	1.337	-0.16
	pjnk, PKA; PKC	t	0	0.021, 13.489	0.013
praf	praf, PKA	bb8	270	1.432, 0.983	-0.178
	PKA, PKC	joe	90	1.337	-0.16
	praf, PKC; PKA	joe	0	1.148	0.078
pmek	pmek, praf	bb8	0	4.852, 0.989	0.662
	praf, PKC	bb8	0	1.364, 0.990	0.160
	PKC, PKA	joe	270	1.337	-0.160
	pmek, PKC; praf	bb7	0	1.145, 0.052	0.099
	praf, PKA; PKC	bb8	270	1.266, 0.985	-0.118
	pmek, PKA; praf, PKC	joe	270	1.072	-0.040
p4442	p4442, PKA	t	0	0.244, 3.360	0.157
	PKA, pmek	joe	90	1.375	-0.175
	p4442, pmek; PKA	bb8	0	1.233, 0.996	0.112
pakts473	pakts473, p4442	bb8	0	6.941, 0.786	0.633
	p4442, PKA	t	0	0.244, 3.360	0.157
	PKA, PIP3	clayton	0	0.066	0.032
	pakts473, PKA; p4442	t	0	-0.071, 6.862	-0.045
	p4442, PIP3; PKA	bb8	90	1.187, 0.724	-0.032
	pakts473, PIP3; p4442, PKA	bb8	90	1.108, 0.843	-0.028

Table 4.5: $\mathbf{M}_{Par}\mathbf{C}_{Par}$ - pooled data: Summary of the D-vine regression model fitted to the pooled data, using all parametric copula families. The marginal distributions given in Table 4.1 were used.

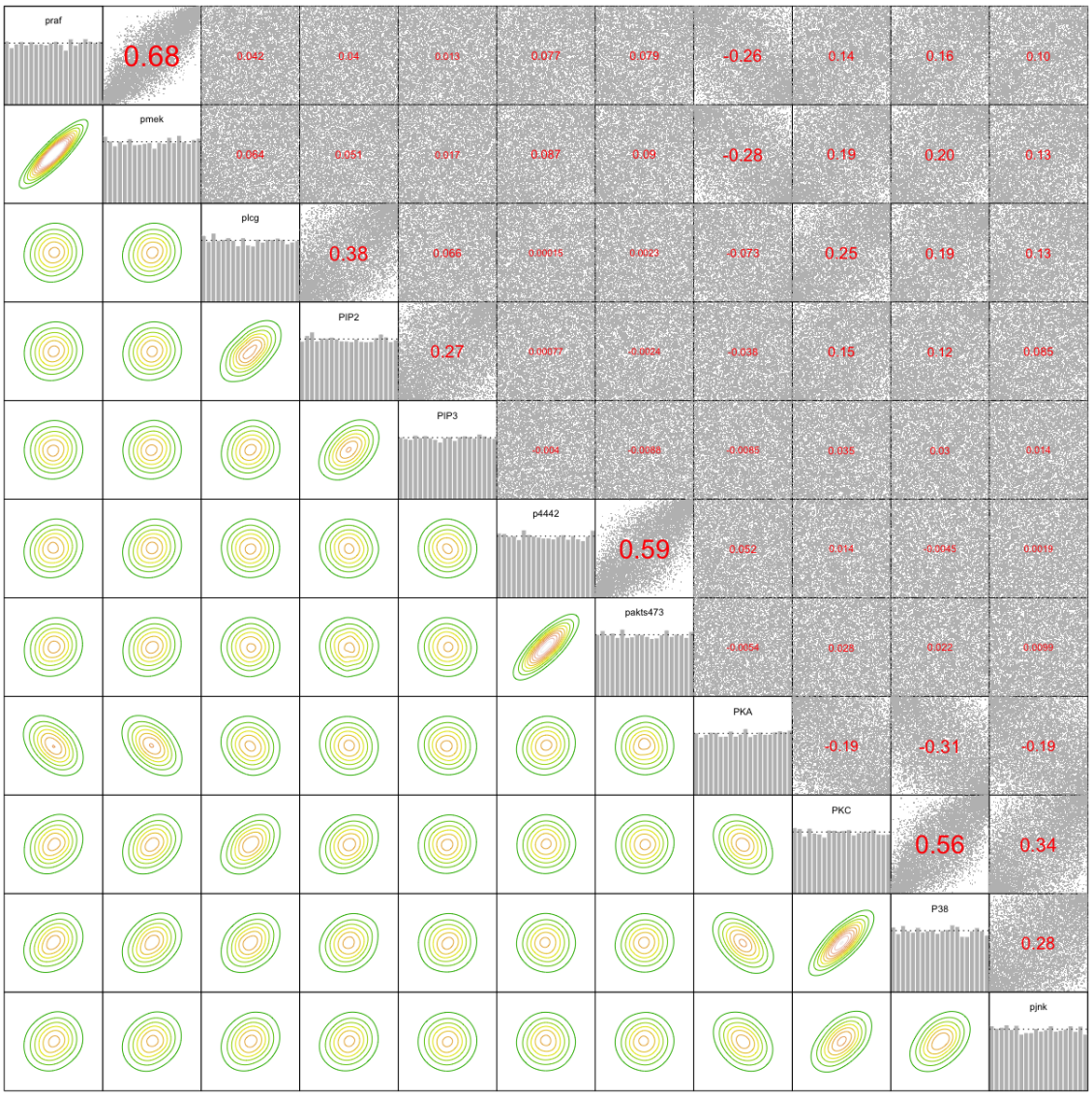


Figure 4.7: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ - pooled data: Normalized contours and pairs plots as well as Kendall's τ of the simulated data on u-scale. Simulation is based on the estimated D-vine regression model with only the Gaussian copula family and the indepencence copula allowed.

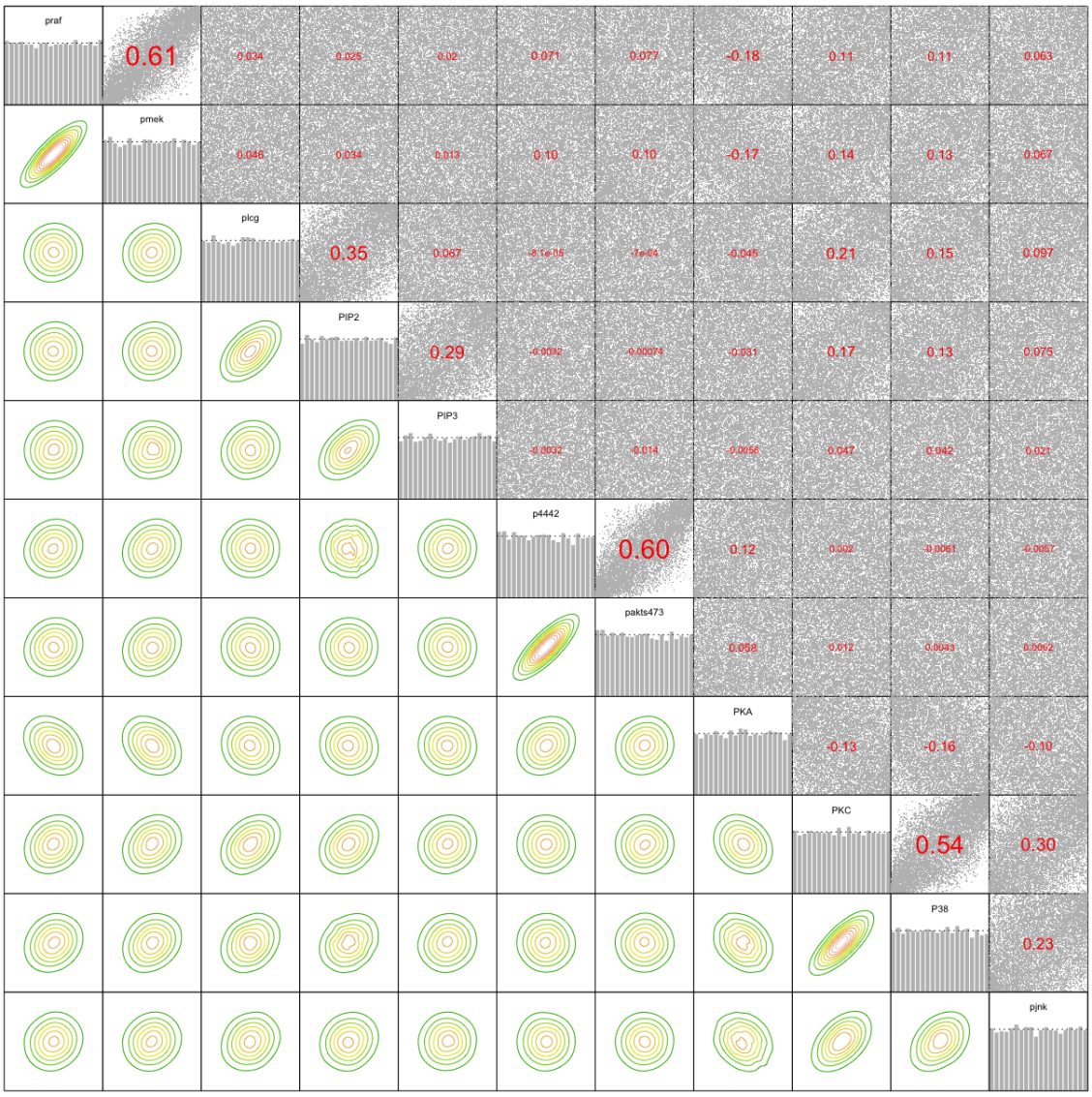


Figure 4.8: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ - pooled data: Normalized contours and pairs plots as well as Kendall's τ of the simulated data on u-scale. Simulation is based on the estimated D-vine regression model with only the Gaussian copula family and the indepencence copula allowed.

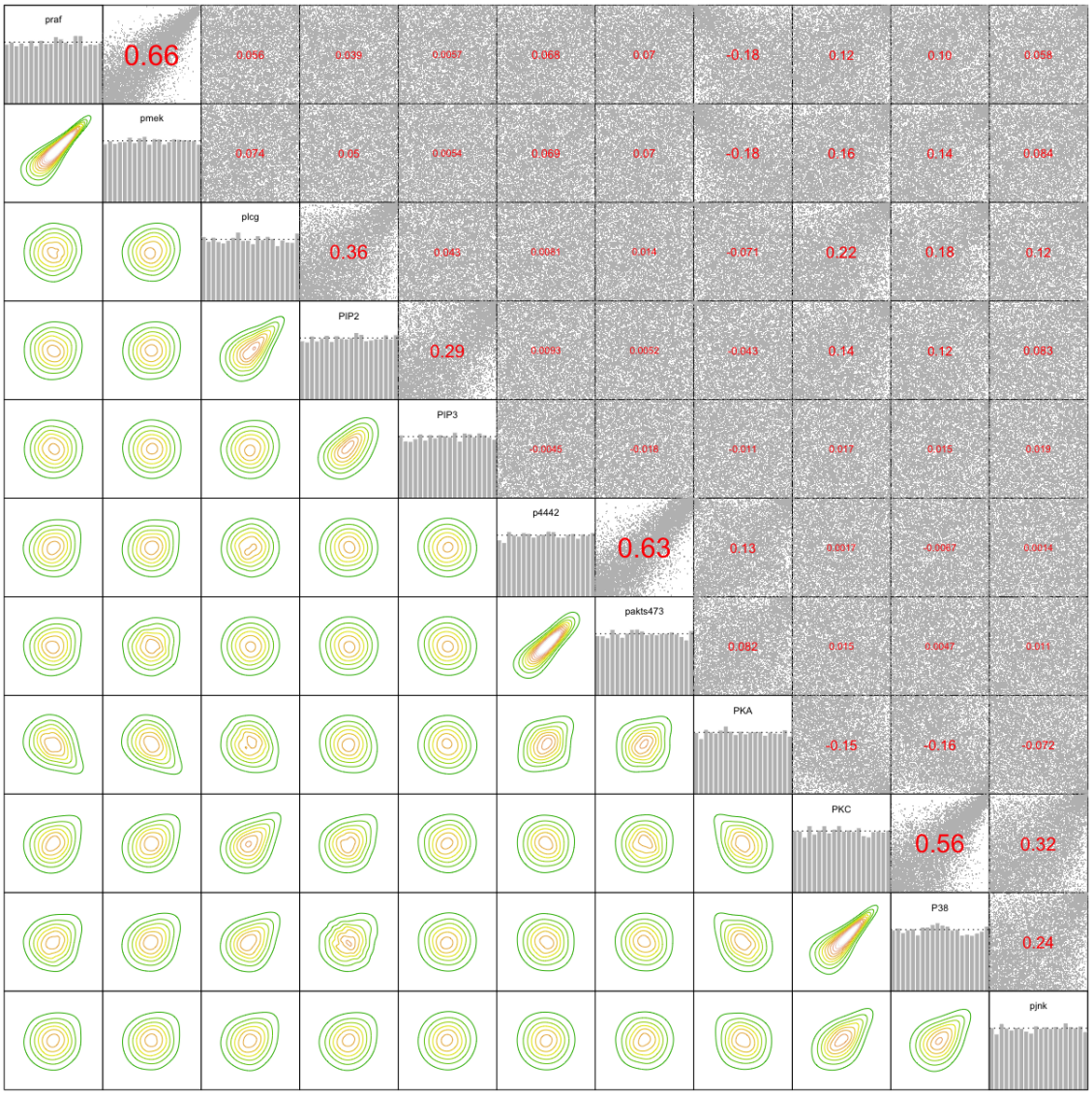


Figure 4.9: $\mathbf{M}_{Par}\mathbf{C}_{Par}$ - pooled data: Normalized contours and pairs plots as well as Kendall's τ of the simulated data on u-scale. Simulation is based on the estimated D-vine regression model of the pooled data with all parametric copula families.

Chapter 5

Clustering

After analyzing the pooled data in the previous chapter, we will focus on the clustering of the data in this chapter. Therefore we will first work with Gaussian mixture models and later with vine copula mixture models. Since the data come from 14 different experiments, a central question is obviously to what extent the data from the specific experiments are identified by a specific cluster.

5.1 Gaussian mixture models (GMM)

5.1.1 Optimal number of components selection

We will first focus on the optimal number of clusters. Based on this, we will then select a model and analyze it in more detail. We will focus on the differences between clusters i.e. for GMMs obviously the differences in mean vectors and differences between the covariance matrices.

We fit the GMMs using the **mclust** package (Scrucca et al. (2016)) in R. To ensure the replicability of the calculations, all calculations were performed with the seed key 111. In Figure 5.1 the BIC values of different Gaussian mixture models up to g=25 components are plotted depending on the parametrization of the covariance matrix and their number of components. The specifications of the models and the meaning of e.g. VVV have already been discussed in detail in the chapter of the theoretical background.

The top 3 models based on the BIC are VVV 16 with -228679, VVV 17 with -228757 and VVV 13 with -228893. But the differences are extrem small: When switching from the model with 13 clusters to the model with 16 clusters, the BIC improves by a factor of only 0.000935, i.e. not even a whole permille. So we need to investigate whether the better BIC value of the 16 cluster VVV model is reliable or random: the final GMM also

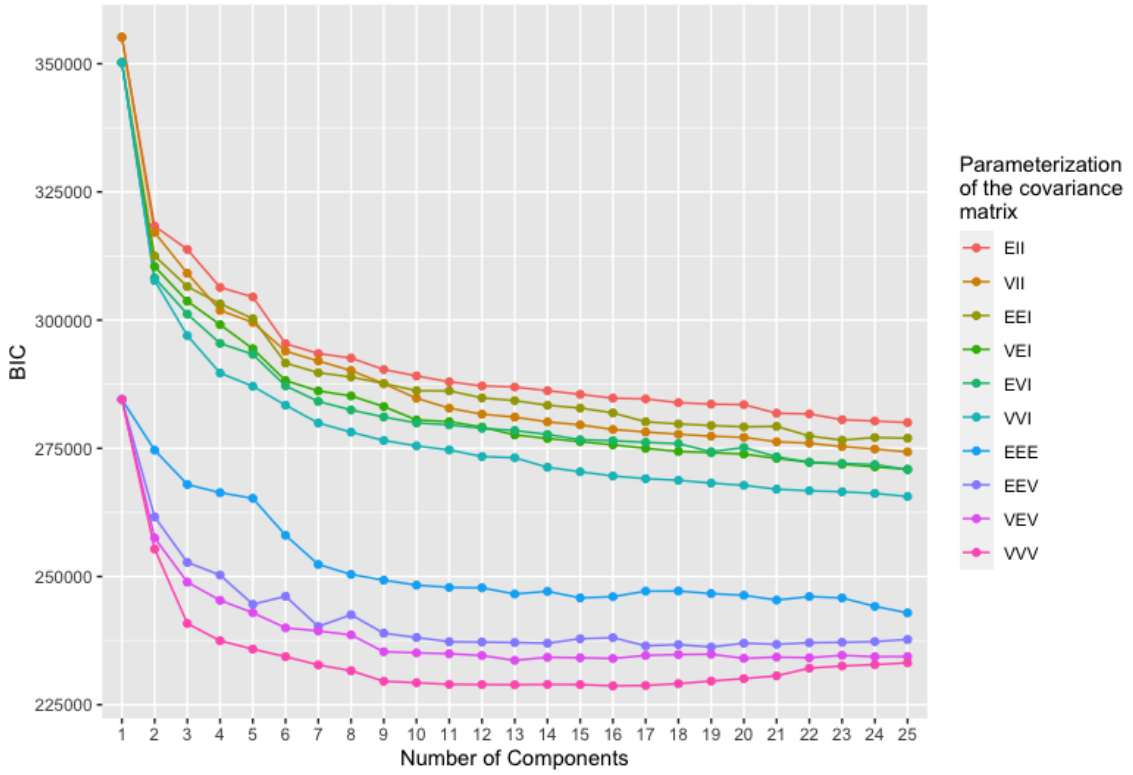


Figure 5.1: BIC values of Gaussian mixture models regarding their number of components and their parametrization of the covariance matrix.

depends on the random starting values, which we control with the seed key. If the BIC improvement of the 16 cluster model is reliable, then similar results should be obtained even if the starting conditions are changed, i.e. if the seed key is changed. If the seed 112 is used, then the 16 cluster VVV model is not even under the best three models. With a seed of 113 it is only the third-best model and with a seed of 9999 it is again only the third-best model.

For this reason, we do not want to rely on the BIC alone. There are different ways to find the optimal number of clusters. There are besides the usage of prior information or the optimization of the likelihood or of information criteria like the BIC, also approaches from test theory. We now use the likelihood ratio test, using the function **mclustBootstrapLRT** from the R package **mclust** (Scrucca et al.(2016)). Here we test step by step whether the likelihood advantage of a model with $g_1 = g_0 + 1$ components is significant compared to a model with only g_0 components. Since we already recognized in Figure 5.1 that the VVV models always perform better than the models with other parametrization of the covariance matrix, we only test VVV models. Here we do not work with the theoretical distribution, but we bootstrap the test statistic. Normally one chooses $B = n$, i.e. one generates as many bootstrap samples as observations in the original dataset. If we

want to test models with up to g=26 components, this means for our specific case that for each 10149 bootstrap samples 26 mixture models have to be fitted. This makes 263874 models. Since this could lead to major runtime issues for the VCMMs in the next chapter we would like to compare the results of several tests at this point: We run the parametric likelihood ratio test with up to 26 components for $B \in \{500, 1000, 2000, 10149\}$ and report the results in Table 5.1.

Number of Components	LRTS $B = 500$	P-Value 500	LRTS 1000	P-Value 1000	LRTS 2000	P-Value 2000	LRTS 10149	P-Value 10149
1 vs 2	26850.67	0.002	29889.08	0.001	29888.92	0.001	29889.63	0.001
2 vs 3	18289.98	0.002	15252.13	0.001	15252.54	0.001	15251.35	0.001
3 vs 4	4092.68	0.002	4092.62	0.001	4090.46	0.001	4091.41	0.001
4 vs 5	2500.52	0.002	2214.49	0.001	2504.83	0.001	2501.87	0.001
5 vs 6	3050.09	0.002	3012.15	0.001	2793.17	0.001	3048.81	0.001
6 vs 7	1639.36	0.002	1669.94	0.001	1955.78	0.001	821.41	0.001
7 vs 8	1139.62	0.002	1473.55	0.001	1205.66	0.001	2256.05	0.001
8 vs 9	2953.27	0.002	1060.19	0.001	821.95	0.001	1361.34	0.001
9 vs 10	1421.28	0.002	1051.06	0.001	1014.35	0.001	586.51	0.001
10 vs 11	131.27	**0.595**	2922.65	0.001	3374.15	0.001	709.17	0.001
11 vs 12	2460.91	0.002	1336.35	0.001	934.61	0.001	2800.57	0.001
12 vs 13	516.40	0.002	139.16	**0.244**	449.71	0.001	594.06	0.001
13 vs 14	752.99	0.002	643.83	0.001	567.74	0.001	819.86	0.001
14 vs 15	780.02	0.002	681.58	0.001	691.74	0.001	324.65	0.001
15 vs 16	888.59	0.002	427.27	0.001	410.52	0.001	899.73	0.001
16 vs 17	922.52	0.002	1122.22	0.001	2420.87	0.001	1996.90	0.001
17 vs 18	276.51	0.002	712.57	0.001	454.20	0.001	1236.69	0.001
18 vs 19	802.48	0.002	138.92	**0.536**	700.89	0.001	-961.03	0.025
19 vs 20	414.20	0.002	1075.14	0.001	119.12	**0.947**	1097.85	0.001
20 vs 21	759.19	0.002	324.72	0.001	553.81	0.001	1025.03	0.001
21 vs 22	803.12	0.002	1173.40	0.001	951.70	0.001	365.23	0.001
22 vs 23	95.06	**0.387**	565.27	0.001	287.33	0.001	125.92	**0.644**
23 vs 24	403.85	0.002	588.98	0.001	203.37	0.012	706.05	0.001
24 vs 25	158.96	**0.285**	522.47	0.001	463.56	0.001	269.51	0.001
25 vs 26	680.55	0.002	313.25	0.001	422.79	0.001	348.85	0.001

Table 5.1: Likelihood Ratio Test for GMM up to 26 components. All p-values are results from the bootstrap and not theoretical. Values below 0.001 are rounded up. P-values over 0.05 are printed bold. Note that 0.002 is the lowest possible p-value for $B = 500$.

The results of the bootstrapped likelihood ratio tests in Table 5.1 are not convincing. We first note that each test recommends a different number of components. If one test finds that the likelihood improvement between two models is not significant, there are always at least two other tests that find this likelihood improvement to be very significant and have p-values ≤ 0.001. Therefore our assumption that likelihood ratio tests with $B \leq 10149$ come to similar results and that we can reduce B in favor of runtime cannot be confirmed here. Second the likelihood improvements are only for models with high components not significant. In Figure 5.1 we see that the BIC values stabilize from about

9 to 10 components. Only the LRT with $B = 500$ finds that the likelihood increase between models 10 and 11 is not significant. However, in the test with $B = 10149$, which should be the most reliable, is the first non-significant likelihood improvement between models 22 and 23. To prevent overfitting we would like to work with as few components (and thus parameters) as possible. We also know that the data comes from 14 different experiments, so a model with 22 components has most likely too many components. Third, in the LRT with $B = 10149$ we see a negative likelihood increase when comparing models 18 and 19. Thus, the fit of the models decreases. Nevertheless though the bootstrapping a p-value < 0.05 is calculated. This raises clear doubts about the reliability of the results. Since no clear answer to the question of the optimal cluster can be given, and we also have the goal to find out to what extent the data from the specific experiments are split into specific clusters again, we will now continue to work with the 13 cluster VVV model, which has also always performed well with respect to the BIC. We have to consider that a GMM might be the wrong approach for our data. This would explain why, according to LRT, so many components are needed. In the next chapter we will adress this question by using a different approach.

Now that we have calculated the likelihood ratio test statistics for the 10149 bootstrap samples, we would like to visualize the results of Wilks (1938). It was shown that statistic is under the null hypothesis (here: $H_0 : g = 13$) asymptotically chi-square distributed with degrees of freedom df $= \dim(\hat{\boldsymbol{\theta}}) - \dim(\hat{\boldsymbol{\theta}}_{\mathbf{0}})$. Now in our specific case the 13 cluster model has 13 covariance matrices with 11^2 as well as 13 mean vectors with 11 parameters each. The 14 cluster model has 14 of each. So we come to df $= 132$. In Figure 5.2 the corresponding empirical distribution function is shown in black and the theoretical cdf in blue. We can see, that there is a not negligible difference between the curves.

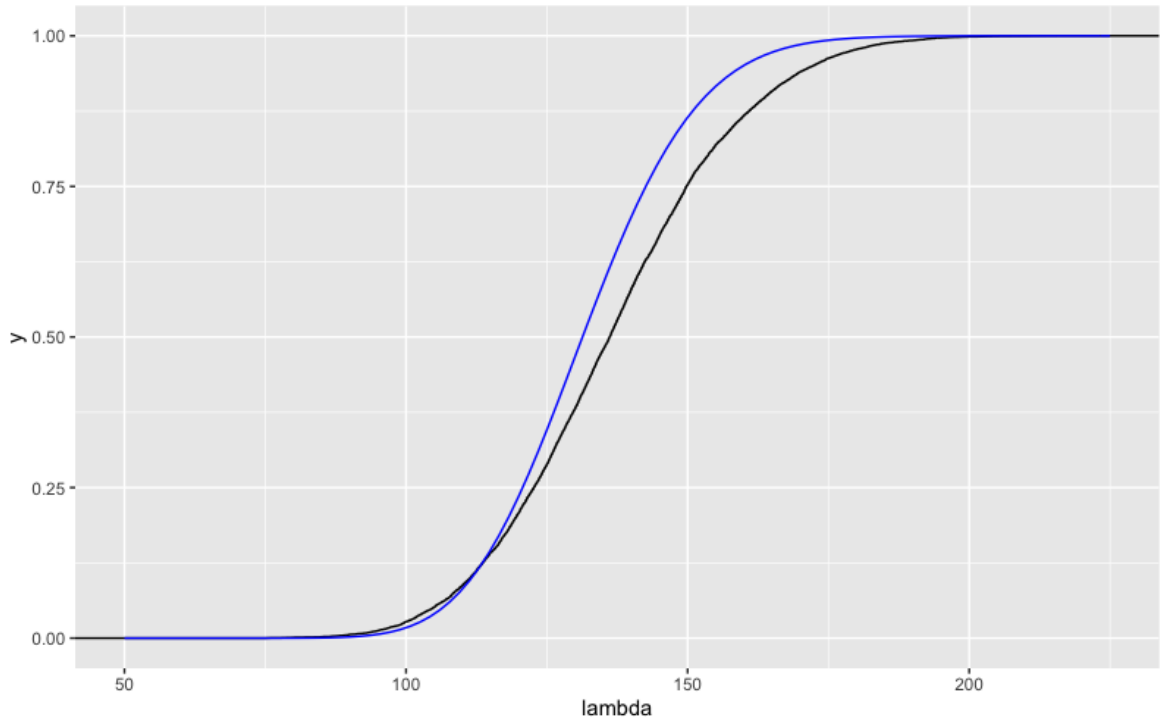


Figure 5.2: Theoretical cumulative distribution function of the χ^2-distribution with 132 degrees of freedom in blue and the ecdf of the test statistic of the 10149 bootstrap LRT in black.

5.1.2 Analysis of the clusters

One question we already asked at the beginning of the clustering chapter was how sharp the observations from the 14 experiments are split into specific clusters. We would like to investigate this question now. To do this, we calculate the percentage of the observations from the experiments, that were classified to a specific cluster. This means we create a table, with the experiments on the x-axis and the clusters on the y-axis. Then we look at the number of observations from one experiment assigned to a specific cluster and divide it by the total number of observations in this experiment. The sum of all entries in one column adds up to 1. The result is shown in Figure 5.3. For better readability it is shown as a heatmap.

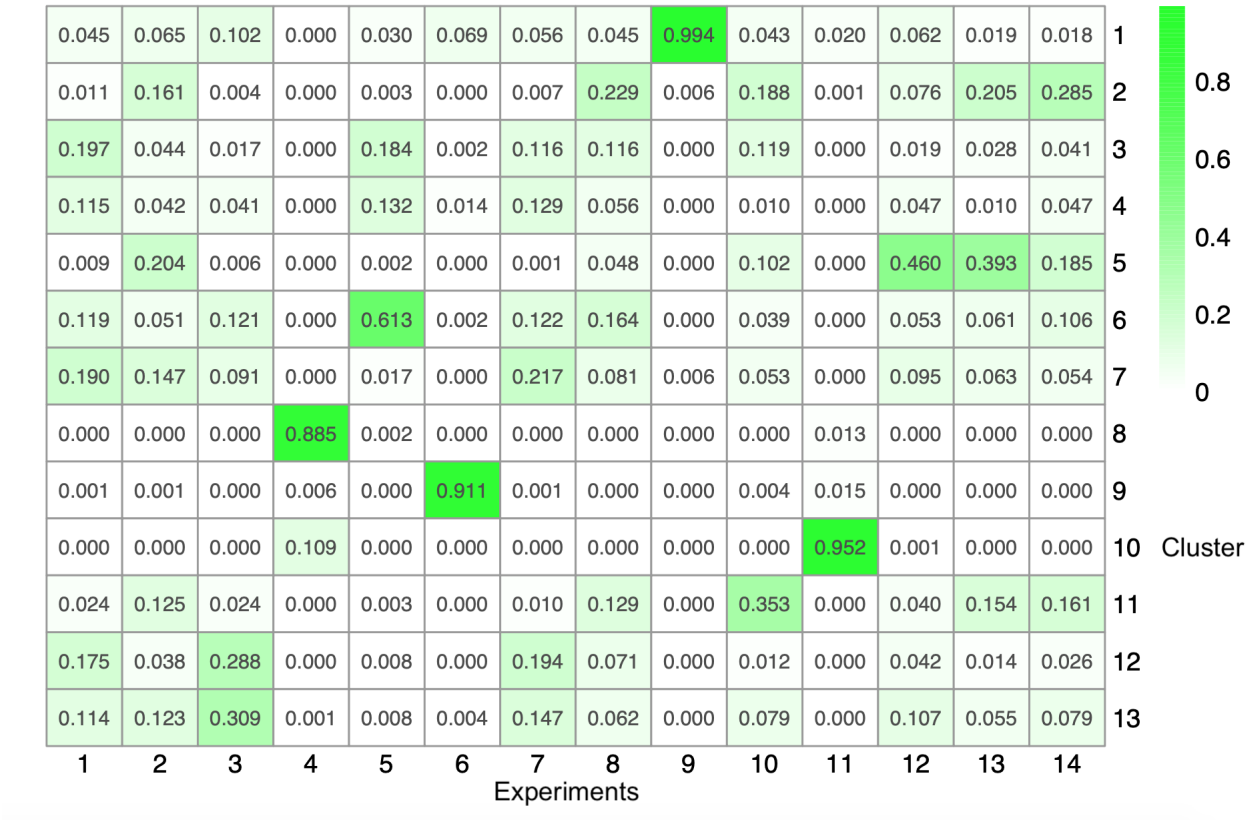


Figure 5.3: Percentage of observations in an experiment assigned to the clusters.

Some experiments are represented very well by exactly one cluster. For these, which include for example experiments E4, E6, E9 and E11, the clusters identify the experiments very well. It may be helpful to consider the biological background of the data to explain this pattern. Experiments E3-E7 and E10-E14 were perturbed with the same stimulations (see Table 3.1), but in experiments E10-E14 the stimulation ICAM-2 was

additionally applied. Here, for example, it is apparent that G0076, which was applied only in experiments E4 and E11, is clearly recognized in the data. Regarding the other well recognized experiments, we do not see such a pattern. For this reason, we would now like to better understand how similar the clusters are to each other. Since in a GMM each cluster is defined by its mean and a covariance matrix, we have to check the mean values. If an experiment breaks down into several specific clusters, but the mean values are far apart, then it may be that an experiment itself breaks down into two or more components. The reason for this could be unobserved variables or peculiarities within one or more experiments.

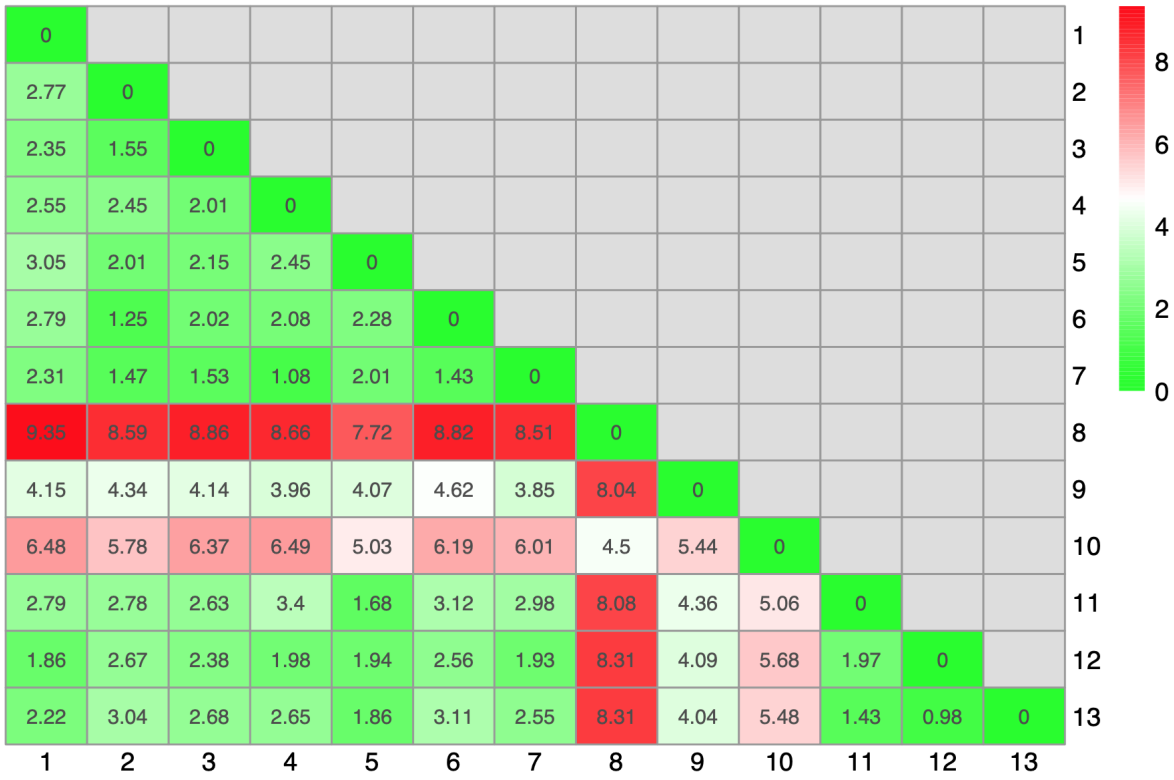


Figure 5.4: Euclidean distance between the clusters means of the GMM g=13

In Figure 5.4 we can see, that the smallest Euclidean distance is between the means of cluster C12 and C13 with a value of 0.98. The second smallest is between the means of cluster C4 and C7 with a value of 1.08. Also interesting is what we can see with cluster C8. It has by far the largest euclidean distance to the means of all other clusters. The reason for this is, that cluster C8 represents with 0.885 of the observations from experiment E4 very exactly this experiment.

Besides the clusters means we should also discuss the differences between the clusters

covariance matrices. We simply consider the difference of the matrices in the spectral norm, which is the matrix norm derived from the Euclidean norm. Let $A \in \mathbb{R}^{n \times m}$ be a matrix and $\lambda_{max}(\cdot)$ be the largest eigenvalue of a matrix. Then the spectral norm in defined as $||A||_2 = \max \frac{||A\mathbf{x}||_2}{||\mathbf{x}||_2} = \sqrt{\lambda_{max}(A^T A)}$ for $\mathbf{x} \in \mathbb{R}^m$. By the definition of a norm, we know that the spectral norm of the difference of two covariance matrices is greater than 0, but less than the sum of the spectral norms of the individual matrices. For this reason, we need to normalize by this factor. Therefore we do not consider $||\Sigma_i - \Sigma_j||_2$, but $\frac{||\Sigma_i - \Sigma_j||_2}{||\Sigma_i||_2 + ||\Sigma_j||_2}$ which takes values between 0 and 1. In Figure 5.5. we can see the results.

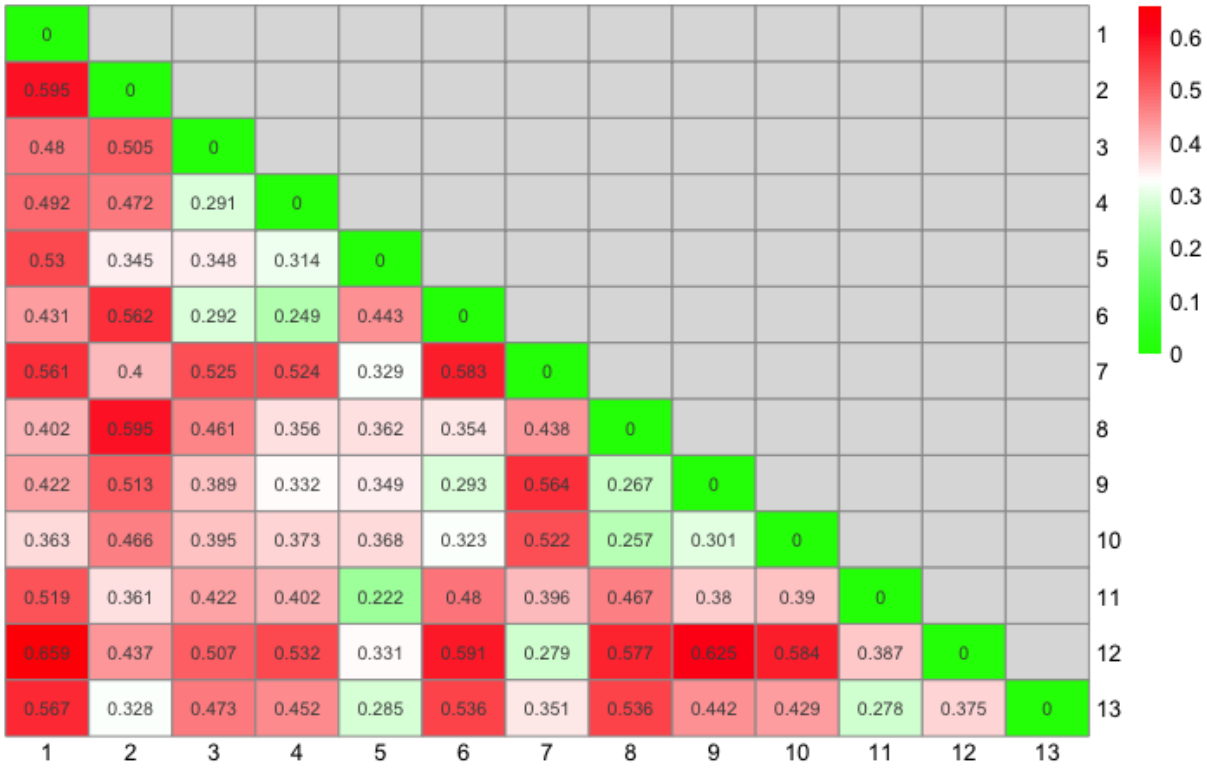


Figure 5.5: Normalized differences in covariance matrices of GMM g=13

In Figure 5.4 we saw, that especially the clusters C12 and C13 and the clusters C4 and C7 have mean values very close to each other. For this reason we are particularly interested in these clusters, whether they are also similar in the covariance matrix. This is clearly not the case for clusters C4 and C7, as they have one of the highest distances in the sense of the normalized spectral norm of the covariance matrices. This distance is indeed smaller for clusters C12 and C13, but in comparison to the others it is neither remarkably high nor remarkable low.

5.1.3 Data simulation

In the previous chapters we finally simulated data from the discussed models and compared the results in pairs plots. Here we want to proceed the same way and simulate 10149 datapoints from the distribution of the gmm with 13 components. The results are shown in Figure 5.6. If the two datasets were sampled from the identical same distribution they would -most likely- look like if they have been mirrored on the diagonal. We can see that the results of the mixture model are much better then the results from the multivariate Gaussian model, which we discussed in the previous chapter. Still we have seen in this section, that there are issues and the Gaussian mixture model fits not perfectly to the data. For that reason we will now continue with vine copula mixture models.

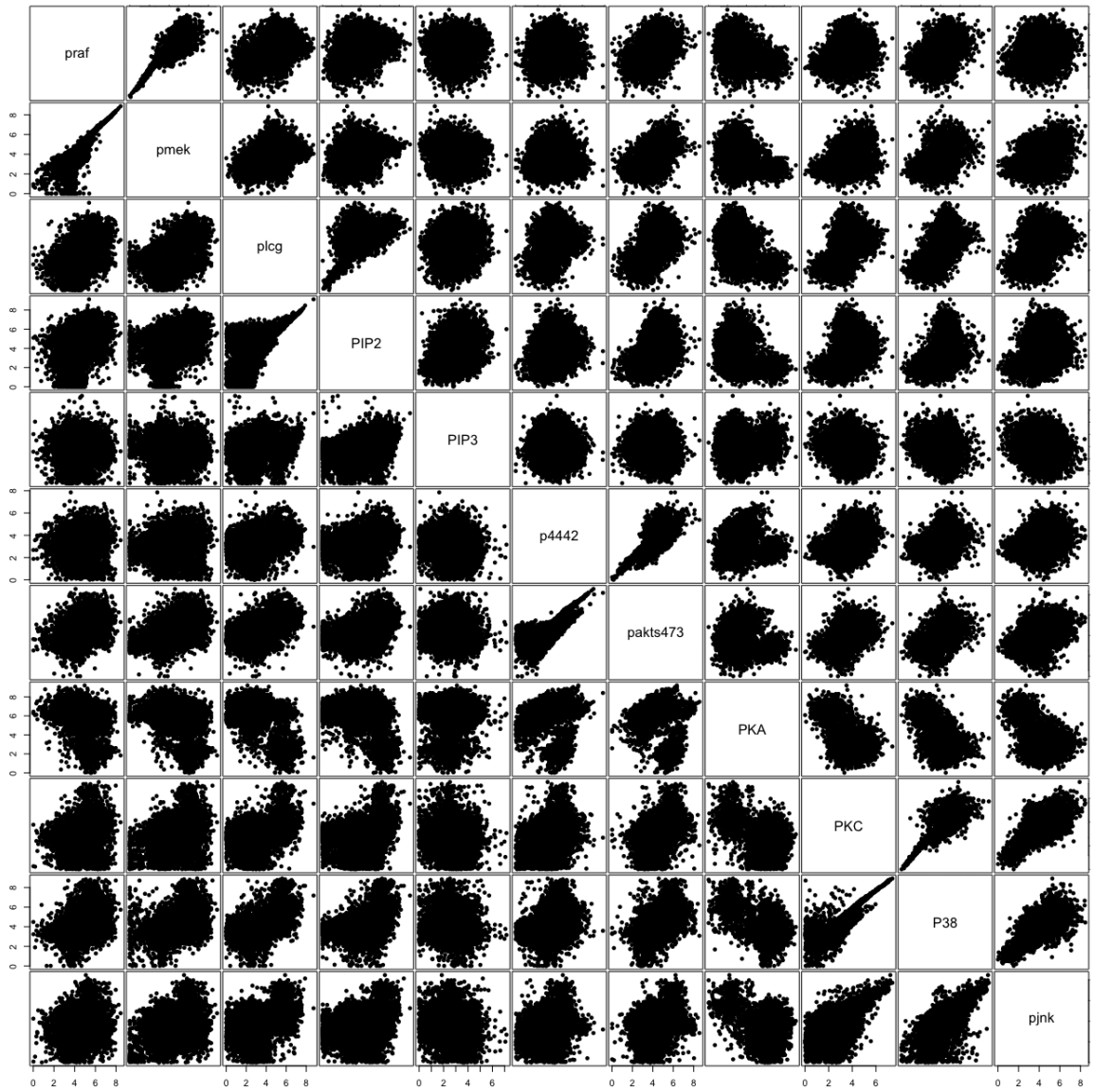


Figure 5.6: Pairs plots of observed dataset (lower triangle) and dataset sampled from the 13 component Gaussian mixture model (upper triangle).

5.2 Vine copula mixture models (VCMM)

5.2.1 VCMMs for the Sachs data

Now we are considering vine copula mixture models according to Sahin and Czado (2021). We describe them with VCMM(g,C,M), where g stands for the number of clusters, C for the allowed copula families and M for the allowed marginal distributions. For that reason we work with different settings regarding the allowed copula families and marginal distributions. If a copula family is included, it means that all its rotations (which are 0, 90, 180 and 270 degrees) are also allowed.

$C \in \{\text{NC}, \text{LC}\}$ with NC = {Gaussian}
and LC = {Gaussian, t, Clayton, Gumbel, Frank, Joe, BB1, BB6, BB8}

$M \in \{\text{NM}, \text{SM}, \text{LM}\}$ with NM = {normal},
SM = {normal, log normal, logistic, log logistic, gamma, t fix}
where t-fix is the Student t distribution with fixed df=3.
LM = SM $\cup$ {skew normal, skew t} =
= {normal, log normal, logistic, log logistic, gamma, t fix, skew normal, skew t}

In Figure 5.7 the BIC values of different VCMMs are plotted, depending on their number of components as well as their allowed copula families and marginal distributions.

We can see, that the models with the large set of copula families (LC) and the large set of marginal distributions (LM) have the best BIC and log-likelihood values. We would like to understand these models better and examine in Table 5.2 how often the different marginals are used. The two skew distributions were by far the most frequently used in the LM-models, but the normal distribution and the logistic distribution were also used regularly. In the two SM-models the logistic, the normal and the gamma distribution were used the most often.

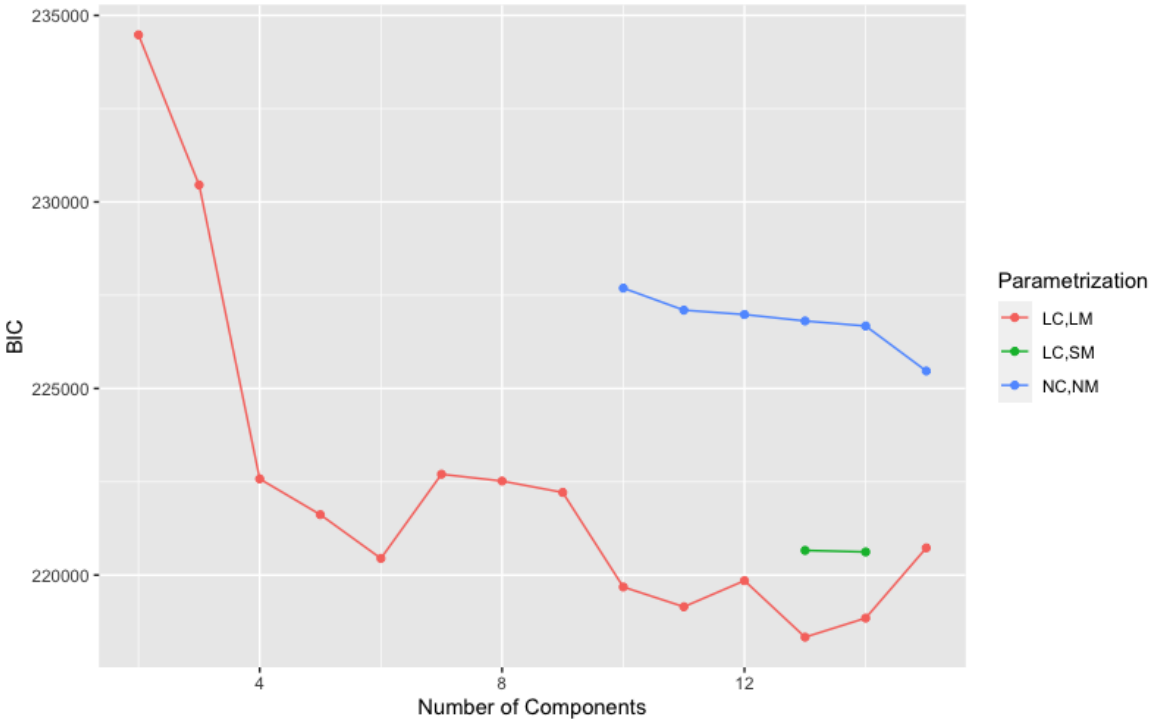


Figure 5.7: BIC values of vine copula mixture models with different parametrizations.

	norm	log-norm	logis	log-logis	gamma	t-fix	skew-norm	skew-t
VCMM(13,LC,SM)	0.30	0.03	0.37	0.07	0.08	0.14	0.00	0.00
VCMM(14,LC,SM)	0.30	0.03	0.35	0.08	0.08	0.16	0.00	0.00
VCMM(2,LC,LM)	0.00	0.00	0.18	0.00	0.00	0.00	0.09	0.73
VCMM(3,LC,LM)	0.00	0.00	0.18	0.00	0.00	0.00	0.21	0.61
VCMM(4,LC,LM)	0.05	0.00	0.18	0.00	0.00	0.02	0.20	0.55
VCMM(5,LC,LM)	0.02	0.00	0.16	0.00	0.00	0.02	0.22	0.58
VCMM(6,LC,LM)	0.05	0.00	0.11	0.00	0.00	0.03	0.27	0.55
VCMM(7,LC,LM)	0.05	0.00	0.18	0.00	0.00	0.01	0.23	0.52
VCMM(8,LC,LM)	0.02	0.00	0.12	0.02	0.01	0.01	0.26	0.55
VCMM(9,LC,LM)	0.06	0.00	0.13	0.02	0.00	0.04	0.28	0.46
VCMM(10,LC,LM)	0.05	0.00	0.11	0.00	0.02	0.04	0.36	0.43
VCMM(11,LC,LM)	0.05	0.01	0.09	0.02	0.00	0.02	0.37	0.44
VCMM(12,LC,LM)	0.08	0.00	0.11	0.02	0.02	0.05	0.28	0.46
VCMM(13,LC,LM)	0.10	0.01	0.13	0.00	0.03	0.06	0.28	0.38
VCMM(14,LC,LM)	0.08	0.02	0.14	0.01	0.03	0.05	0.30	0.39
VCMM(15,LC,LM)	0.07	0.00	0.15	0.01	0.02	0.04	0.30	0.41

Table 5.2: percentages of the marginal distribution families utilized in different VCMMs.

5.2.2 Simulation setup for number of components selection

A first and important question in the choice of the number of components for the VCMM is, how reliable the metrics we use are. That is whether the BIC or the also computed integrated completed likelihood (ICL, see Biernacki et al. (2000)) recommend the model that is closest to the true model. To check this we use the following setup: Due to the running times, we reduce the dataset to the four variables `PIP3`, `PKA`, `p4442` and `pakts473` instead of the 11 we are normally working with. Then we fit a 2 and a 6 component VCMM with large set of copula families (LC) and the large set of marginal distributions (LM) to this reduced, four dimensional dataset. With different seeds we sample 100 datasets with 200 observations each from this 2-component VCMM. For each of these datasets we fit VCMMs with up to 8 components. The VCMM algorithm optimizes the marginal parameters and the pair copula parameters by maximizing the log-likelihood. As the log-likelihood is basically a sum with as many terms as we have observations, we can reduce the running time perceptible by working with the smaller datasets. For the resulting models we compute the mean value of the loglik, BIC and ICL. We do this, because this way we don't have the problems that the models depends so much on the random initial clustering. We do the same for the 6-component VCMM where we sample 100 datasets with 600 observations each and fit models with up to 12 components.

The results are shown in Figure 5.8 and Figure 5.9. We can see that for both, the 2- and the 6-component VCMM, the loglik gets much better for models with more components, but it grows slower with higher number of components. The likelihood ratio test works with this fact. For both the 2- and the 6-component VCMM, the BIC and the ICL clearly find the right number of components.

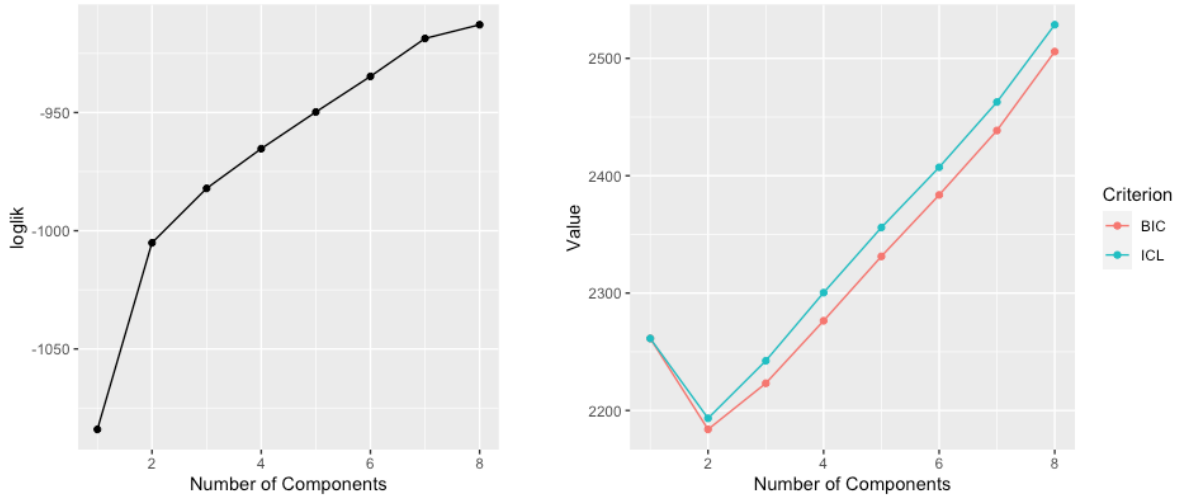


Figure 5.8: Log-likelihood (left), BIC and ICL (right) mean values for VCMM models fitted to datasets simulated from a 2-component VCMM. Only the the four variables PIP3, PKA, p4442 and pakts473 of the Sachs dataset were used to fit the VCMM sampled from.

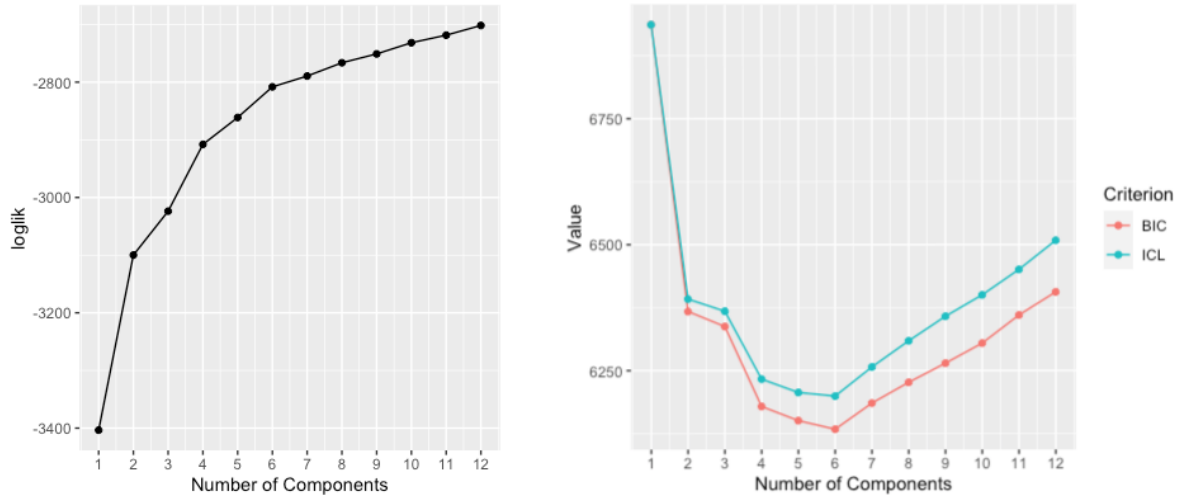


Figure 5.9: Log-likelihood (left), BIC and ICL (right) mean values for VCMM models fitted to datasets simulated from a 6-component VCMM. Only the the four variables PIP3, PKA, p4442 and pakts473 of the Sachs dataset were used to fit the VCMM sampled from.

5.2.3 Likelihood ratio test and Vuong test

Now we want to specify a number of components. Since the variation of the BIC values between the models, which differ only in the number of clusters, is small, we use selected tests again. We again use the likelihood ratio test and the Vuong test, which have been explained in Chapter 2.3.2. Due to the runing time we do not bootstrap the LRT this time, but use Wilks theorem (Wilks, 1938), which we have already visualized in the last subchapter. That means we calculate the p-values theoretically with the χ^2-distribution. The results are shown in Table 5.3.

Number of components	LRTS	df	marginal df	copula df	p-value	5% quantile of certain χ^2-distribution
2 vs. 3	5009.84	107	35	72	<1.11e-16	132.14
3 vs. 4	8578.30	76	32	44	<1.11e-16	97.35
4 vs. 5	1853.34	97	41	56	<1.11e-16	120.99
5 vs. 6	2040.85	94	36	58	<1.11e-16	117.63
6 vs. 7	-1408.32	92	30	62	> 1 - 1.11e-16	115.39
7 vs. 8	1207.18	111	43	68	<1.11e-16	136.60
8 vs. 9	830.93	57	23	34	<1.11e-16	75.62
9 vs. 10	3493.39	104	36	68	<1.11e-16	128.80
10 vs. 11	1416.43	96	39	57	<1.11e-16	119.87
11 vs. 12	-45.02	71	30	41	> 1 - 1.11e-16	91.67
12 vs. 13	1796.29	46	11	35	<1.11e-16	62.83
13 vs. 14	387.46	82	40	42	<1.11e-16	104.14
14 vs. 15	-792.38	118	40	78	> 1 - 1.11e-16	144.35

Table 5.3: Likelihood ratio test for VCMM(g,LC,LM) up to 15 components. All p-values are theoretical results from the asymptotic χ^2-distribution. To provide comparability with the results from the mclust package p-value $= 1 - \text{cdf}_{\chi^2}$ is shown. The differences of the exact p-values to 0 or 1 are smaller then the machine precision. The fifth column shows the theoretical threshold over which the LRTS must lie, so that the increase in likelihood is significant at level $\alpha = 0.05$.

First we need to make clear what the likelihood ratio test is here testing for: We assume two models VCMM(g,LC,LM) and VCMM(g+1,LC,LM) to be fitting equally good in the sense of the likelihood. In this case we would choose the models with less components. Now the LRT tests, if the likelihood of the one model is perceptible better than the likelihood of the other model. Up to VCMM(6,LC,LM), there are always large likelihood impovements of the models with more clusters over the models with less clusters. But for the likelihood ratio test of VCMM(6,LC,LM) against VCMM(7,LC,LM) this does not hold anymore. Based on the results of the likelihood ratio test, we should choose the 6 cluster model instead of the 7 cluster model.

In addition, we would like to apply the unadjusted and adjusted Vuong test here.

Again, we calculate the p-values theoretically. The test statistic of the Vuong test ν is standard normal distributed. The results are shown in Table 5.4.

Number of components	Unadjusted Vuong ν	unadjusted p-value	Akaike adjusted ν	Akaike adjusted p	Schwarz adjusted ν	Schwarz adjusted p
2 vs. 3	-23.12	0	-22.13	0	-18.56	0
3 vs. 4	-35.14	0	-34.52	0	-32.27	0
4 vs. 5	-10.36	1.89e-25	-9.28	8.86e-21	-5.36	8.864e-21
5 vs. 6	-14.86	2.92e-50	-13.49	8.65e-42	-8.55	8.651e-42
6 vs. 7	8.56	1	9.68	1	13.72	1
7 vs. 8	0.94	0.825	1.97	9.76e-01	5.70	0.976
8 vs. 9	-4.32	7.87e-06	-3.73	9.74e-05	-1.59	9.741e-05
9 vs. 10	-14.83	4.82e-50	-13.95	1.68e-44	-10.76	1.682e-44
10 vs. 11	-7.70	6.77e-15	-6.66	1.40e-11	-2.89	1.399e-11
11 vs. 12	0.22	0.588	0.93	0.824	3.49	0.824
12 vs. 13	-9.79	6.28e-23	-9.29	7.89e-21	-7.48	7.89e-21
13 vs. 14	-2.79	0.003	-1.61	0.054	2.66	0.054
14 vs. 15	4.73	0.999	6.14	1	11.23	1

Table 5.4: Vuong Test for VCMM(g,LC,LM) up to 15 components. All p-values are theoretical results from the standard normal distribution. P-values below 1e-100 are rounded to 0.

We have to remember what the p-value tells us here: p-values < 0.05 suggest to choose the model with more components, p-values > 0.95 suggest to choose the model with less components, otherwise we assume that the models are equally good. In this case the VCMM(6,LC,LM) is strongly recommended over the VCMM(7,LC,LM). Also seems the VCMM(7,LC,LM) to have a significantly better likelihood than the VCMM(8,LC,LM). Up to model VCMM(6,LC,LM), always the models with more components seem to be significantly better fitting to the data then the models with less components. The models with 11 and 12 clusters seem to be equally good fitting.

We can therefore summarize, that in addition to the model VCMM(13,LC,LM), which has the best BIC, the model VCMM(6,LC,LM) is also interesting for closer examination. Therefore, we now want to check with a Vuong test whether the improvement of the likelihood is significant when using the 13 cluster model instead of the 6 cluster model: We compute the statistics: unadjusted $\nu = -28.41$ with p-value $=$ 7.75e$-$178, and Schwarz adjusted $\nu = -7.67$ with p-value $= 8.71\text{e} - 15$. The Schwarz adjusted Vuong test is more interesting in this case, as the different number of parameters has more impact here. For both the unadjusted and the Schwarz adjusted Vuong test, the VCMM(13,LC,LM) is significantly better, then the VCMM(6,LC,LM).

5.2.4 Analysis of the mixture weights

We recall Equation (2.9), which defines the density of a mixture model with $g \in \mathbb{N}$ components. The $\pi_j^{(g)}$-values for $j = 1, ..., g$ play an important role. If the $\pi_j^{(g)}$-value for a component is close to zero, the density of this component hardly matters for the model. Since it holds $\sum_{j=1}^{g} \pi_j^{(g)} = 1$, we know that on average $\pi_j^{(g)} \approx 1/g$ for $j = 1, ..., g$. Now consider this in a context where we want to specify the number of components for a model. If several models have very similar values with respect to other criteria like log-likelihood or BIC, it could make sense to set a threshold $\alpha \in (0, 1/g)$ under which we omit component j as soon as $\pi_j^{(g)} < \alpha$.

In Figure 5.7 we saw two local minima with respect to the BIC values. For this reason, the VCMMs with 6 and 13 clusters and their neighborhoods are particularly interesting. For that reason Table 5.5 shows (only for these interesting cases $g \in \{6, 7, 13, 14\}$) the $\pi_j^{(g)}$-values from the models VCMM(g,LC,LM) and their means $1/g$.

Component	6	7	13	14
1	0.281	0.191	0.170	**0.027**
2	0.140	0.139	0.070	0.060
3	**0.056**	**0.053**	**0.030**	0.038
4	0.181	0.202	**0.030**	0.120
5	0.164	0.174	0.038	0.060
6	0.178	0.121	0.080	0.151
7		0.120	**0.015**	**0.015**
8			0.061	0.058
9			0.171	0.142
10			0.142	0.086
11			0.069	0.069
12			0.050	0.048
13			0.073	0.073
14				0.053
1/g	0.167	0.143	0.077	0.071

Table 5.5: Estimated $\pi_j^{(g)}$-Values for the 6,7,13 and 14 component VCMMs. Values with $\pi_j^{(g)} < 0.5 * 1/g$ are printed bold.

It is interesting that although the BIC difference between the 12 component VCMM and the 13 component VCMM is large, also for the 13 component VCMM exceptionally small $\pi_j^{(13)}$-values occur. The same applies for the 6 component VCMM, which has a remarkably small $\pi_3^{(13)}$ value for cluster 3. We want to investigate these peculiarities by comparing the results with Figure 5.10 and Figure 5.11. In which we can see how the different experiments were divided into the clusters. In Figure 5.10 is shown, how many

percent of observations in an experiment is assigned to the specific VCMM(6,LC,LM)-clusters. The sum of all entries in one column adds up to 1. For better readability it is shown as a heatmap. The same holds for Figure 5.11 with the VCMM(13,LC,LM).

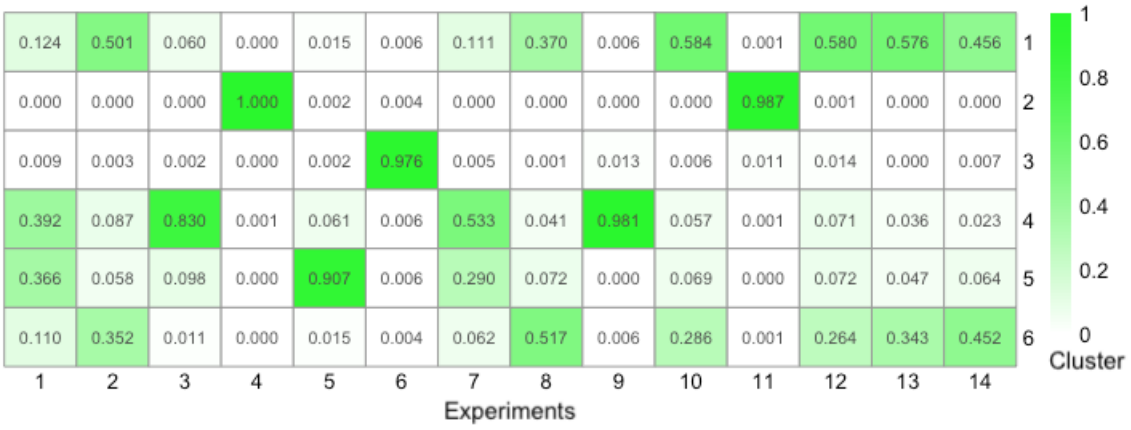


Figure 5.10: Percentage of observations in an experiment assigned to the VCMM(6,LC,LM)- clusters.

In VCMM(6,LC,LM), cluster C3 matches experiment E6 very closely. Thus for this model, we dismiss the threshold approach described above for our specific case, since the small $\pi_j^{(6)}$-values are due to the specifics of individual experiments. While cluster C7 of the VCMM(13,LC,LM) matches experiment E9 very closely, clusters C3 and C4 are very heterogeneous. Clearly it would not make sense to leave out all components with $\pi_j^{(13)} < 0.5 * 1/g$ due to cluster C7. However, we should check what Figure 5.11 would look like, if we create from VCMM(13,LC,LM) a new model with fixed $\pi_3^{(13*)} = \pi_4^{(13*)} = 0$. This would effectively be an 11 component VCMM, but since it is not created and optimized by the VCMM algorithm we dont want to call it VCMM(11,LC,LM). After setting $\pi_3^{(13*)}$ and $\pi_4^{(13*)}$ to zero, we need to adjust all other $\pi_j^{(13*)}$-values proportionally, so that $\sum_{j=1}^{13} \pi_j^{(13*)} = 1$ still holds. Therefore we set $\pi_j^{(13*)} := \frac{\pi_j^{(13)}}{1-\pi_3^{(13)}-\pi_4^{(13)}}$. We can now use the resulting 11 component VCMM to re-cluster the data. The result can be seen in Figure 5.12 analogous to Figure 5.11.

Obviously clusters C3 and C4 are empty now. The observations in these clusters have been split to the other clusters without significantly changing the results shown in Figure 5.11. I.e. the experiments that were previously very specifically assigned to one cluster are still so. This is good and shows that the threshold approach described above can be helpful when working with vine copula mixture models.

Before we move on to the next topic, we would like to address one last question regarding the mixture weights: In CM-step 1 of the VCMM algorithm, the π_j are calculated

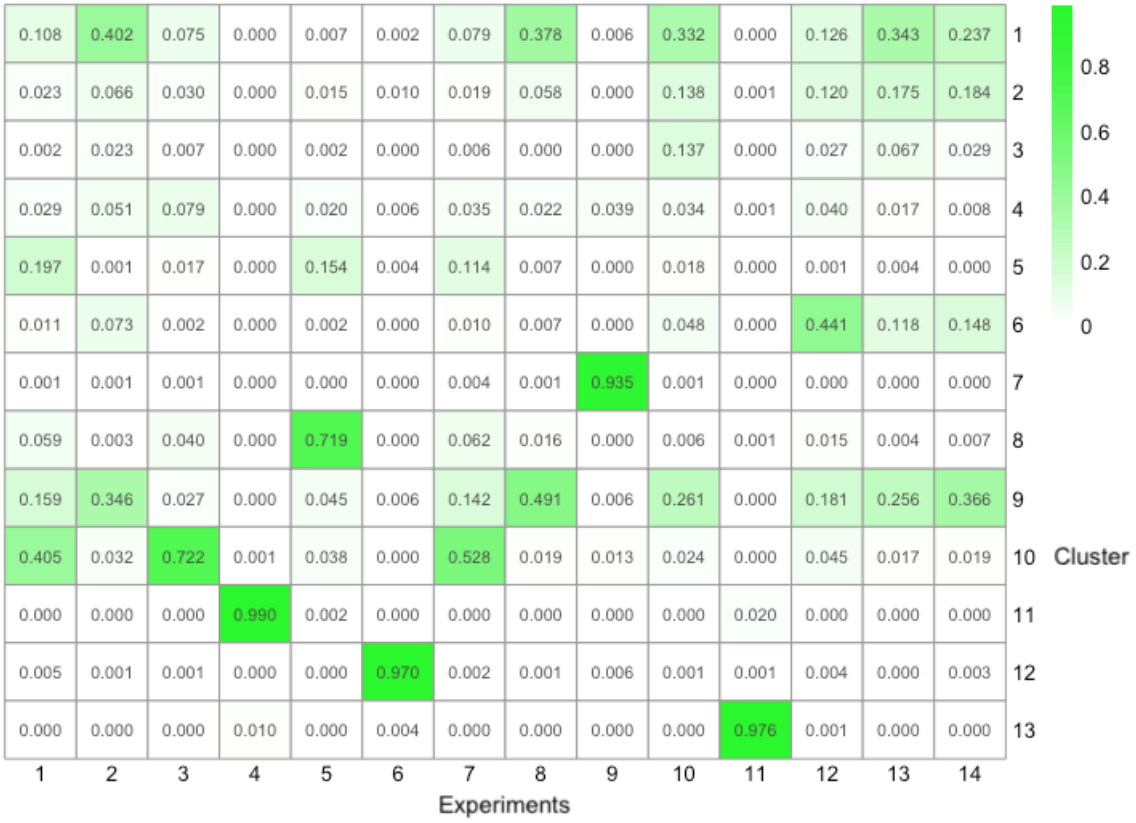


Figure 5.11: Percentage of observations in an experiment assigned to the VCMM(13,LC,LM)- clusters.

via the maximum likelihood method given the updated posterior probabilities of the observations. Thus, it is non-trivial, but likely in view of the previous results of our analysis, that the size of π_j is directly related to the number of observations assigned to this cluster. In this case, one could possibly interpret the mixture weights as the theoretical size of the cluster. For completeness, we should check this assumption for the full interpretability of the VCMM results. In Figure 5.13 and 5.14 we see the relative size of the VCMM(6,LC,LM)- and VCMM(13,LC,LM)-clusters (ie. the number of observations in the respective cluster divided by the total number of observations 10149), as well as the π_j for the respective cluster. The lines are obviously very similar, confirming our assumption.

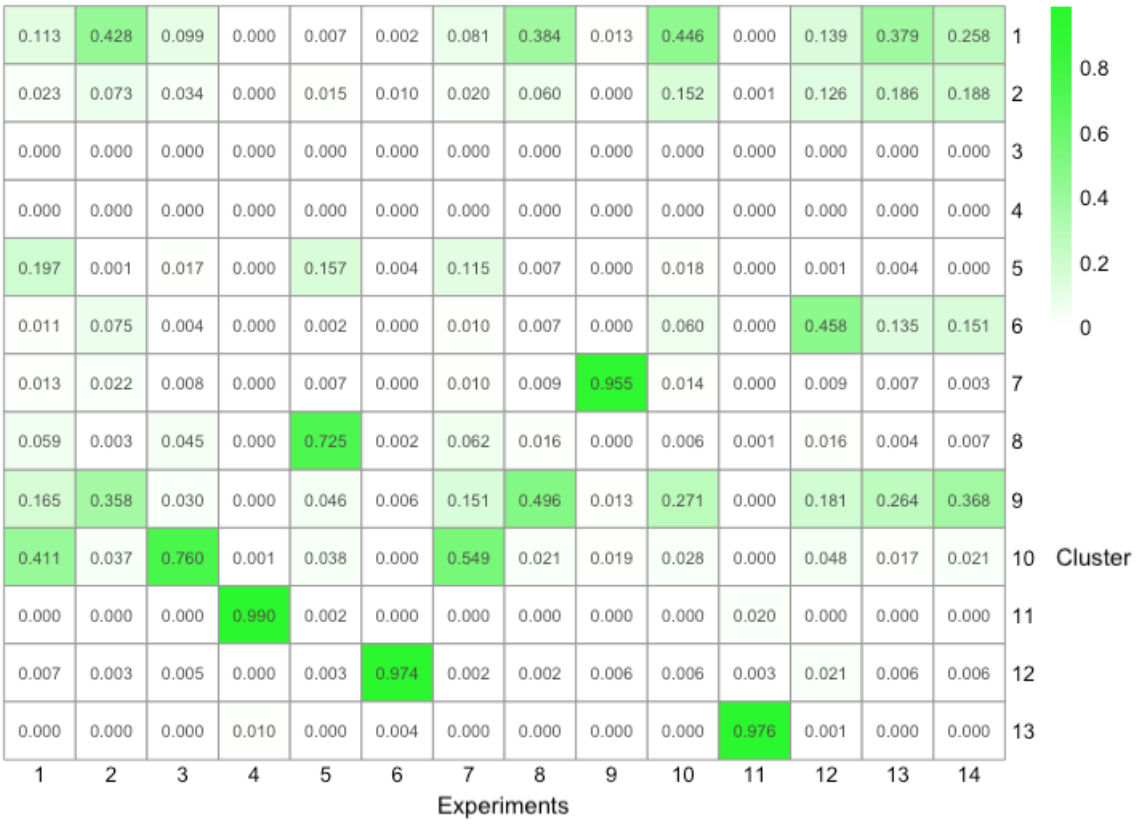


Figure 5.12: Percentage of observations in an experiment assigned to clusters of the the adjusted $\pi_j^{(13*)}$ model derived from the VCMM(13,LC,LM). It is basically the VCMM(13,LC,LM), but $\pi_3^{(13*)} = \pi_4^{(13*)} = 0$ were set to zero manually and all other $\pi_j^{(13*)}$-values have been adjusted proportionally.

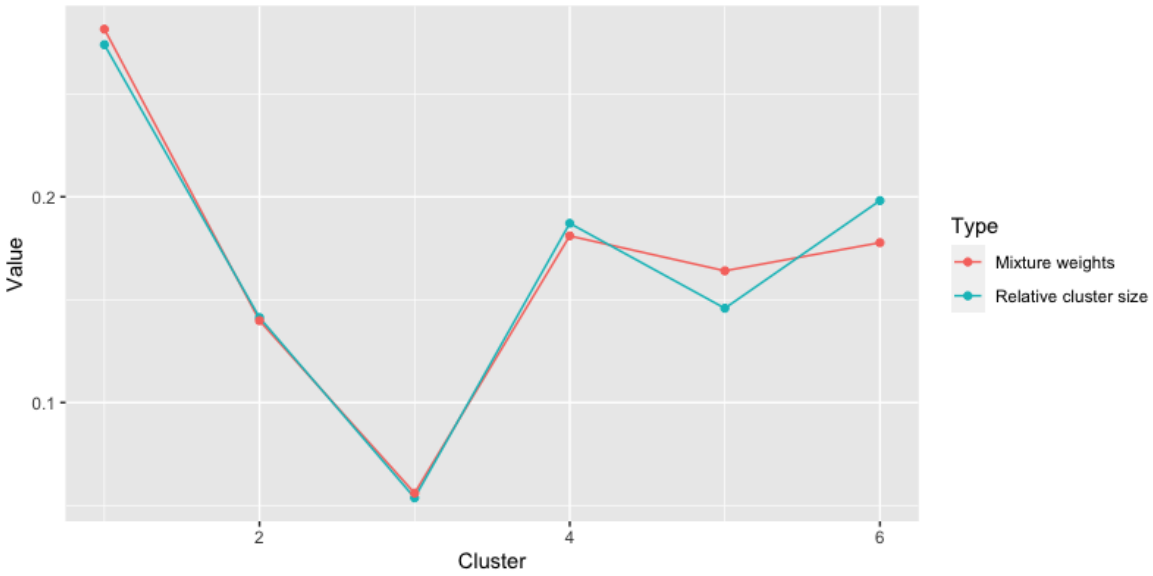


Figure 5.13: Percentage of observations in assigned to the VCMM(6,LC,LM)-clusters and mixture weights $\pi_j^{(6)}$ of the respective clusters.

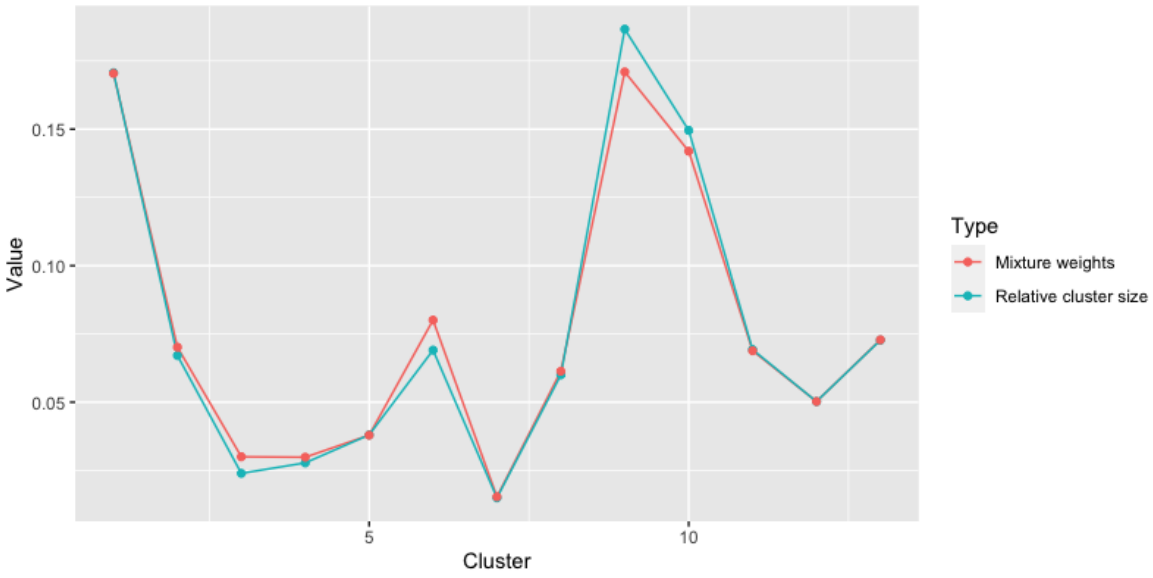


Figure 5.14: Percentage of observations in assigned to the VCMM(13,LC,LM)-clusters and mixture weights $\pi_j^{(13)}$ of the respective clusters.

5.2.5 Results of the VCMMs and their biological context

In Figures 5.10 and 5.11 we have already seen the percentage of observations in an experiment assigned to the specific VCMM clusters. We would like to discuss these figures in more detail: Both models recognize the same experiments well as distinct clusters. These experiments which are assigned as a whole to a cluster are experiments E3, E4, E5, E6, E9 and E11. The clusters C11 and C13 from VCMM(13,LC,LM) were merged to cluster C2 from VCMM(6,LC,LM). Clusters C7 and C10 from VCMM(13,LC,LM) were largely merged to cluster C4 from VCMM(6,LC,LM). Experiment E6 always stands separately and does not share a cluster with any other experiment in both models.

Exp.	Stimulation	Directly influenced variables
1.	Anti-CD3/CD28	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+)
2.	Anti-CD3/CD28, ICAM-2	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+)
3.	Anti-CD3/CD28, akt-inhibitor	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), pakts473 (-)
4.	Anti-CD3/CD28, G0076	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+/-)
5.	Anti-CD3/CD28, Psitectorigenin	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), PIP2 (-)
6.	Anti-CD3/CD28, U0126	plcg (+), praf (+), pmek (+/-), p4442 (+/-), PKC (+)
7.	Anti-CD3/CD28, LY294002	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), pakts473 (+)
8.	PMA	PKC (+)
9.	β 2camp	PKA (+)
10.	Anti-CD3/CD28, ICAM-2, akt-inhib	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), pakts473 (-)
11.	Anti-CD3/CD28, ICAM-2, G0076	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+/-)
12.	Anti-CD3/CD28, ICAM-2, Psitectorigenin	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), PIP2 (-)
13.	Anti-CD3/CD28, ICAM-2, U0126	plcg (+), praf (+), pmek (+/-), p4442 (+/-), PKC (+)
14.	Anti-CD3/CD28, ICAM-2, LY294002	plcg (+), praf (+), pmek (+), p4442 (+), PKC (+), pakts473 (+)

Table 5.6: Stimulations used in the different experiments, as well as the variables, that are biochemically activated (+) or inhibited (-) by the stimulations. Variables directly influenced by the stimulants in opposite ways, are marked with (+/-).

In Table 5.6 we can see 4 peculiarities, that could possibly help to explain the results:

(i) In the experiments E3-E7 and E10-E14, there is one pair each in which another common stimulation was used in addition to anti-CD3/CD28.

(ii) Experiments E2 and E10-E14 have the similarity that always anti-CD3/CD28 and ICAM-2 were used. However, we know that ICAM-2 has no specific direct but only indirect effect on the measured variables.

(iii) Experiments E8 and E9 are the only ones where anti-CD3/CD28 was not applied.

(iv) Some reagents affect the same variables. These are anti-CD3/CD28, PMA and G06976 with effect on `PKC`, anti-CD3/CD28 and U0126 on `pmek` and `p4442`, and akt-inhibitor and LY294002 on `pakts473`.

Regarding these points, we would now like to discuss our results:

(i) Reagent G0076, which was applied in experiment E4 and E11, has an exceptional effect. Experiment E4 and E11 are together in one cluster in the 6 component model. And in the 13 component model they are not, but both are assigned to an individual cluster. We cannot see this in the other pairs: The experiments E10, E12, E13 and E14 are all split across several clusters, but all across the same. These were not identified by the VCMM.

(ii) Experiments E2 and E10-E14 were not assigned to a specific cluster, but they were all distributed to the same multiple clusters. The observations of these experiments are in cluster C1 and C6 in the 6 component model, and cluster C1, C2, C6 and C9 in the 13 component model. This suggests that the data of these experiments have some common features. However, we know from Chapter 3 that the drug ICAM-2 was used here as a "general perturbation" (Sachs et al. (2005), p.525) and the molecules directly affected by it were not measured for this data set. Thus, we detect the effect of ICAM-2 in the data only indirectly.This would be a likely explanation for the fact that the affected experiments, while all split similar, were not split into one or more experiment-specific clusters in either model discussed here.

(iii) The effect of the β 2camp used in experiment E9 seems to be very individual for the variables considered, because in both models the observations of E9 have been assigned very uniquely to exactly one cluster. This is different for PMA used in experiment E8. It is interesting that the observations of experiment E8 were partitioned very similarly to the experiments in which ICAM-2 was also used (ie. the experiments discussed in (ii)). Further biochemical analysis would be necessary to determine whether PMA produces similar effects on hidden variables.

(iv) Anti-CD3/CD28 and PMA both activate `PKC`, whereas G06976 inhibits `PKC`. The observations of experiments E4 and E11, in which G06976 was applied, were very precisely assigned to one cluster. This suggests that the data points in which `PKC` was inhibited have strong features that were clearly found by the VCMM algorithm. In contrast, experiment E8 in which PMA was used was not identified as one cluster. U0126 was applied in experiments E6 and E13. While experiment E6 matches with a single cluster in each of the models discussed here, we see no particularities in experiment E13 - or similarities to experiment E6. The data from these experiments seem not to behave any more or less differently to experiments E8 and E9 (the only in which anti-CD3/CD28 was not applied) than other experiments. When we compare experiment E3, where akt-inhibitor was applied, to experiment E7, where LY294002 was applied, we have the following setting: Both have influence on the variable `pakts473`, but in exactly different ways: LY294002 activates `pakts473` and as the name suggests akt-inhibitor inhibits it. Nevertheless these observations from these experiments were split very similarly in both models, as we can

see for VCMM(6,LC,LM) in cluster 4 and for VCMM(13,LC,LM) in cluster C10. Thus, the data have notable similarities.

5.2.6 VCMM performance at optimal initial conditions

In the previous subchapter, we have discussed in detail how the observations of the different experiments are split among the VCMM clusters. At this point, it would be interesting to know what result would be obtained, if the VCMM algorithm had the optimal (i.e. the true) partitioning of observations to the experiments at some iteration. This could help us to understand what is feasible. That means we fit a 14 component VCMM, skipping the initial clustering and using the true partitioning to the experiments as initial clustering instead. The results are shown in Figure 5.15.

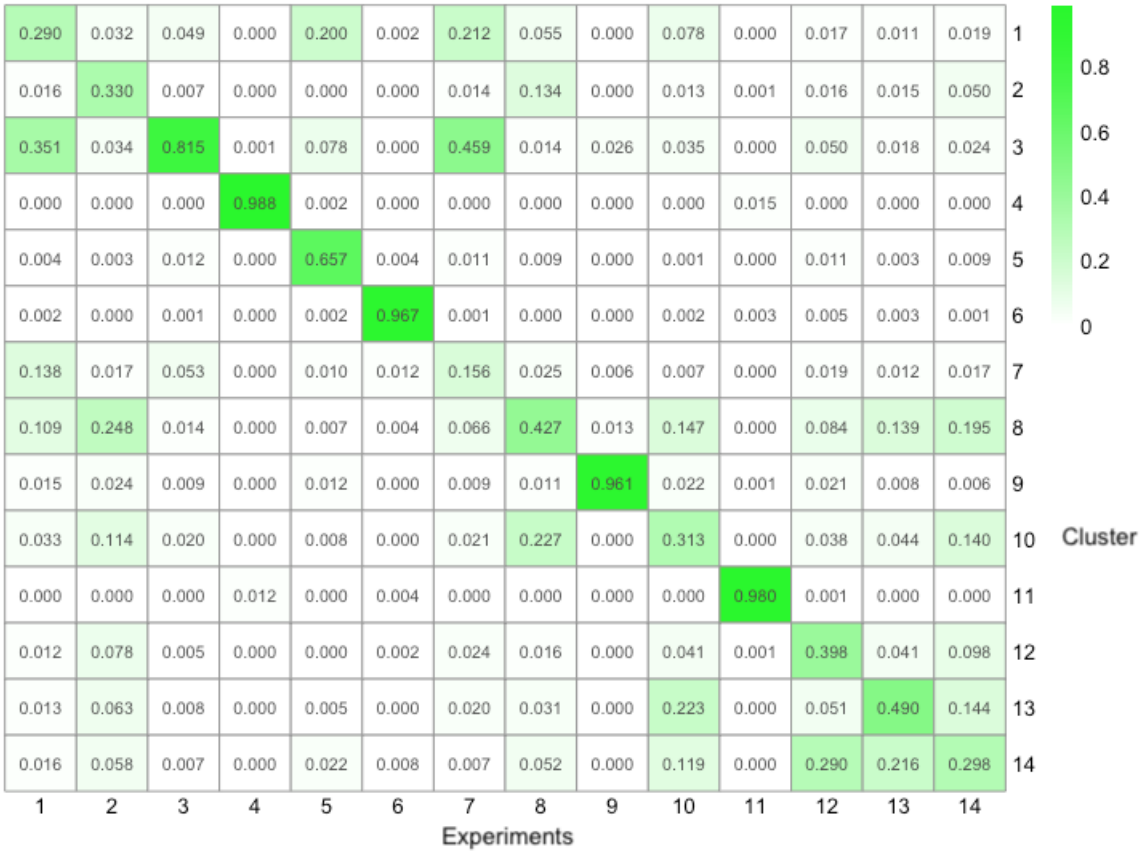
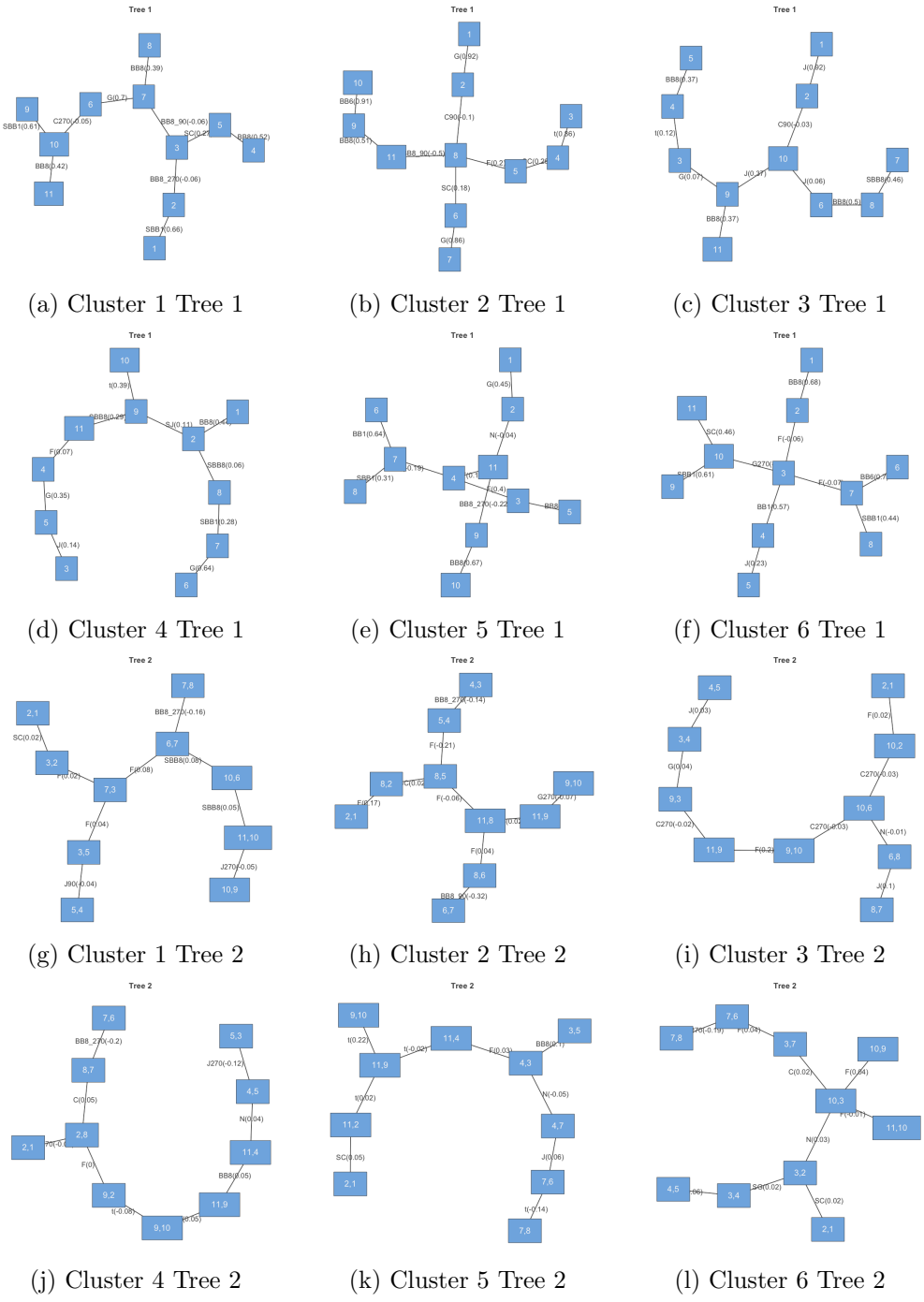


Figure 5.15: Percentage of observations in an experiment assigned to the clusters of a 14 component VCMM, fitted to the true experiment partitioning as initial clustering.

As expected, high values are noted on the diagonal. Apart from this the similarity to the results in Figure 5.10 and 5.11 is remarkable: Again, E3-E6 and E9 and E11 are exactly the experiments that were very accurately identified by this model. In contrast, experiments E10 and E12-E14 have again all been split into different clusters, but together into the same clusters. In addition, experiments E1 andE 7 were once again clustered on the same clusters. This result supports our assumption that the VCMM(6,LC,LM) and

VCMM(13,LC,LM) models, which we discussed in more detail before, already produced very good results.

5.2.7 Vine tree structures



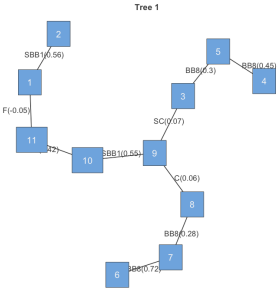
(a) Cluster 1 Tree 1 (b) Cluster 2 Tree 1 (c) Cluster 3 Tree 1

(d) Cluster 4 Tree 1 (e) Cluster 5 Tree 1 (f) Cluster 6 Tree 1

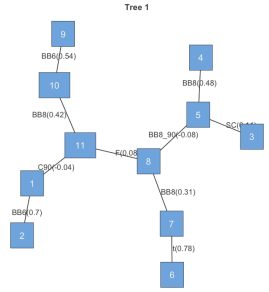
(g) Cluster 1 Tree 2 (h) Cluster 2 Tree 2 (i) Cluster 3 Tree 2

(j) Cluster 4 Tree 2 (k) Cluster 5 Tree 2 (l) Cluster 6 Tree 2

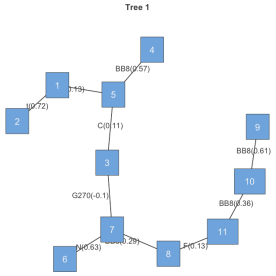
Figure 5.16: Vine tree structures of the clusters of VCMM(6,LC,LM). The variables correspond to: 1=praf, 2=pmek, 3=plcg, 4=PIP2, 5=PIP3, 6=p4442, 7=pakts473, 8=PKA, 9=PKC,10=P38, 11= pjnk



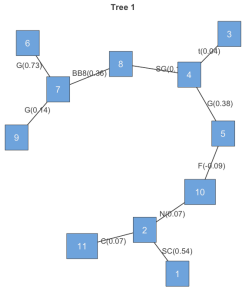
(a) Cluster 1 Tree 1



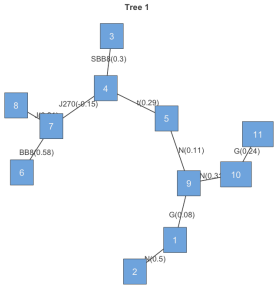
(b) Cluster 2 Tree 1



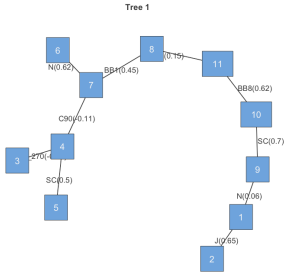
(c) Cluster 3 Tree 1



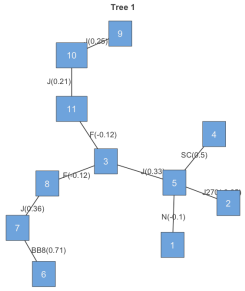
(d) Cluster 4 Tree 1



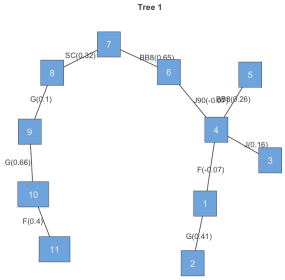
(e) Cluster 5 Tree 1



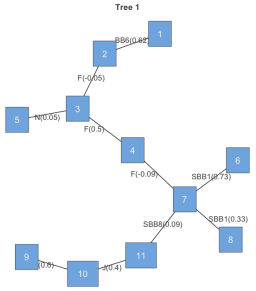
(f) Cluster 6 Tree 1



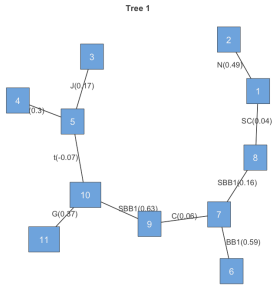
(g) Cluster 7 Tree 1



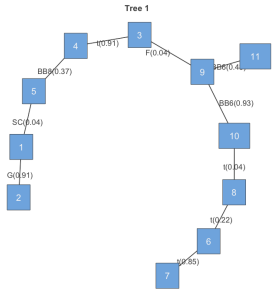
(h) Cluster 8 Tree 1



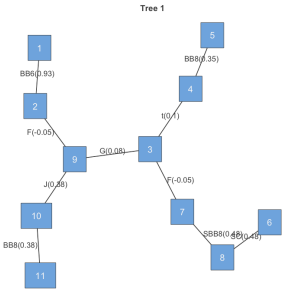
(i) Cluster 9 Tree 1



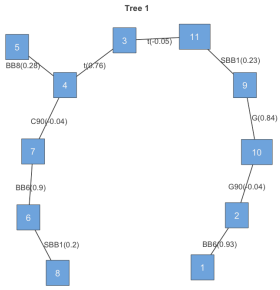
(j) Cluster 10 Tree 1



(k) Cluster 11 Tree 1

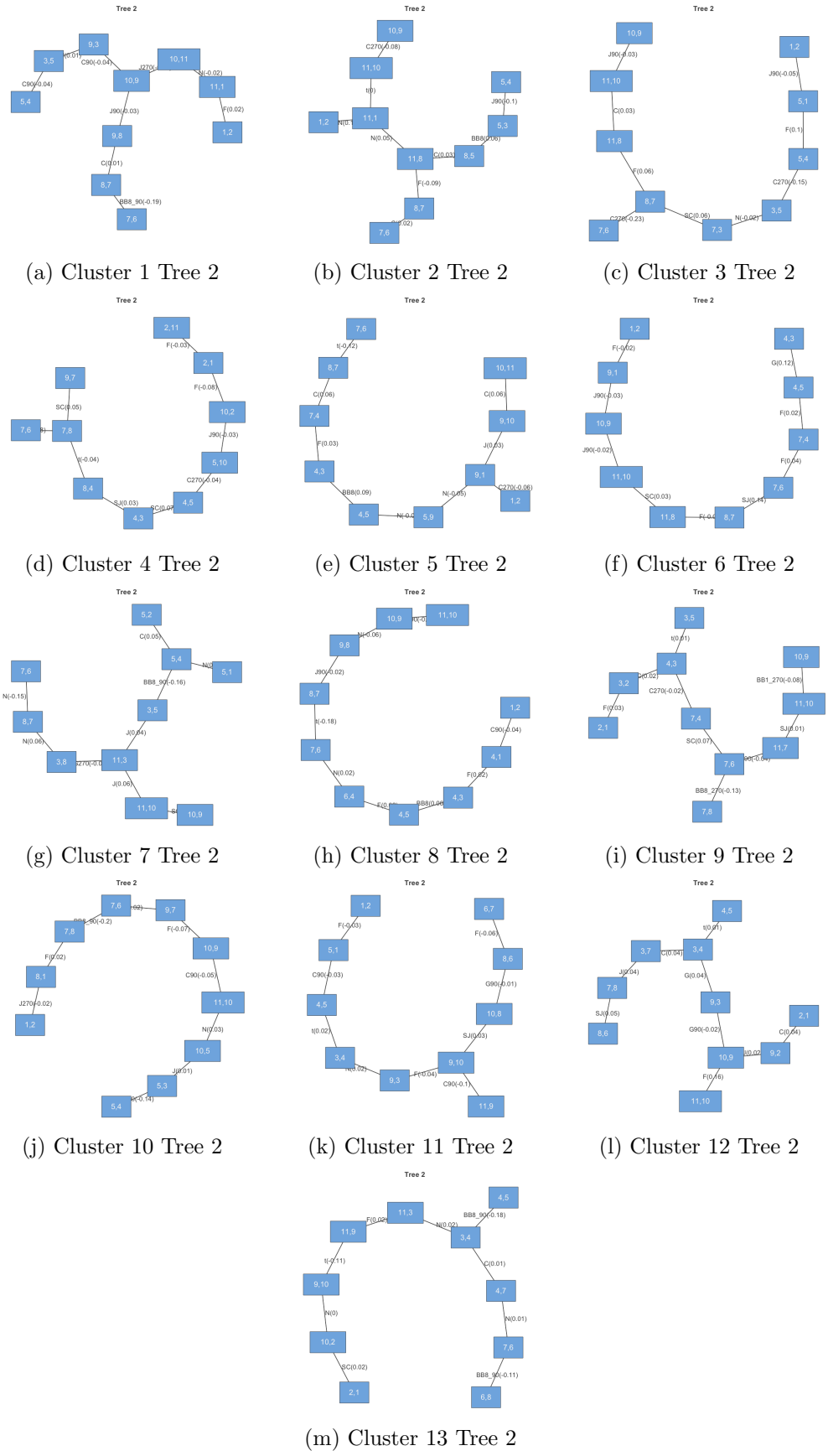


(l) Cluster 12 Tree 1



(m) Cluster 13 Tree 1

Figure 5.17: Vine tree structures of the first trees of the clusters of VCMM(13,LC,LM). The variables correspond to: 1=praf, 2=pmek, 3=plcg, 4=PIP2, 5=PIP3, 6=p4442, 7=pakts473, 8=PKA, 9=PKC,10=P38, 11= pjnk



(a) Cluster 1 Tree 2 (b) Cluster 2 Tree 2 (c) Cluster 3 Tree 2

(d) Cluster 4 Tree 2 (e) Cluster 5 Tree 2 (f) Cluster 6 Tree 2

(g) Cluster 7 Tree 2 (h) Cluster 8 Tree 2 (i) Cluster 9 Tree 2

(j) Cluster 10 Tree 2 (k) Cluster 11 Tree 2 (l) Cluster 12 Tree 2

(m) Cluster 13 Tree 2

Figure 5.18: Vine tree structures of the second trees of the clusters of VCMM(13,LC,LM). The variables correspond to: 1=praf, 2=pmek, 3=plcg, 4=PIP2, 5=PIP3, 6=p4442, 7=pakts473, 8=PKA, 9=PKC,10=P38, 11= pjnk

5.2.8 Data simulation

In the previous chapters, we finally simulated data from the discussed best fitting models, and compared the simulated data to our observations. We want to do the same here, by again simulating 10149 data points from the VCMM(6,LC,LM) and VCMM(13,LC,LM).

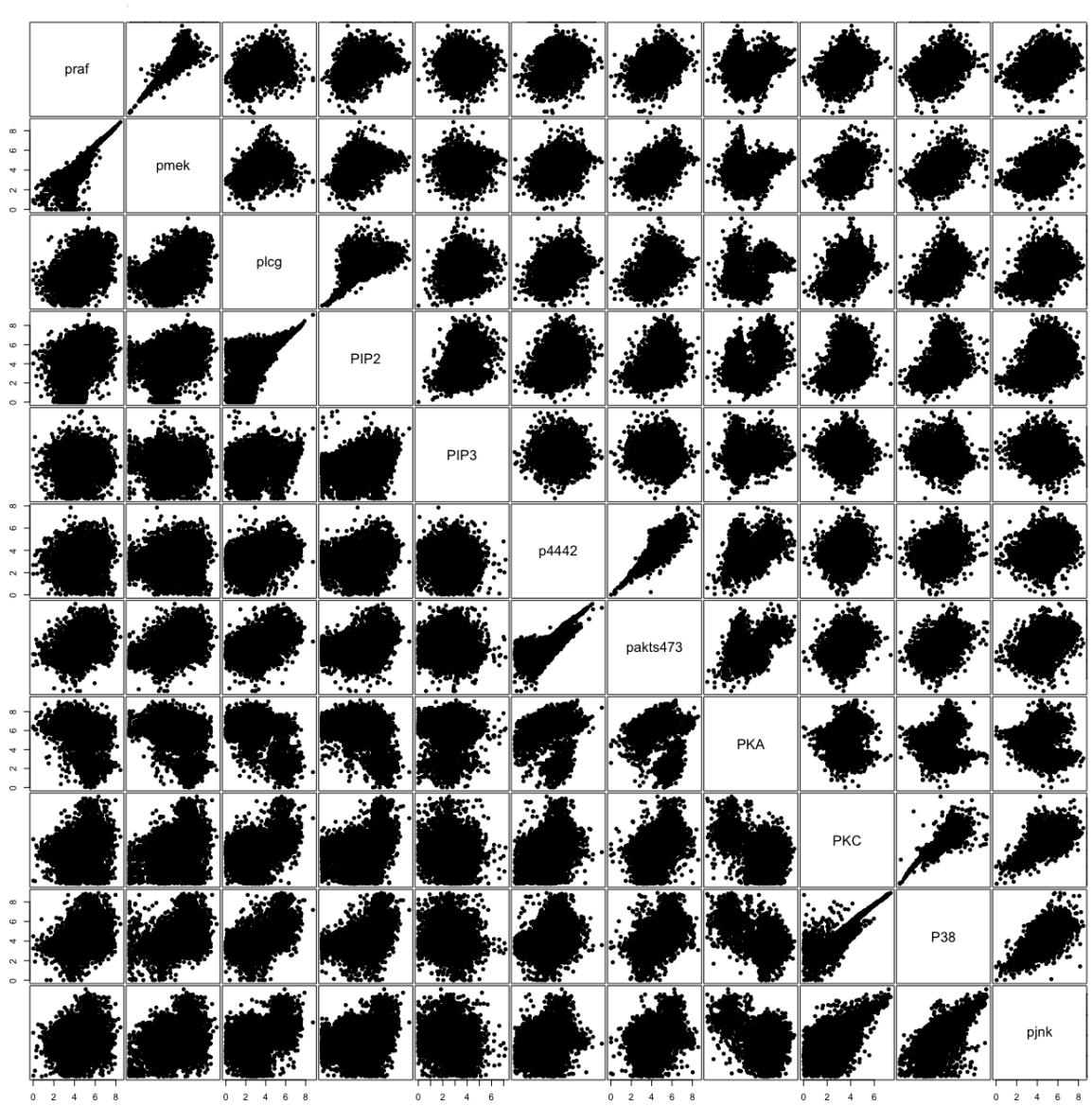


Figure 5.19: Pairs plots of observed dataset (lower triangle) and dataset sampled from the VCMM(6,LC,LM) (upper triangle).

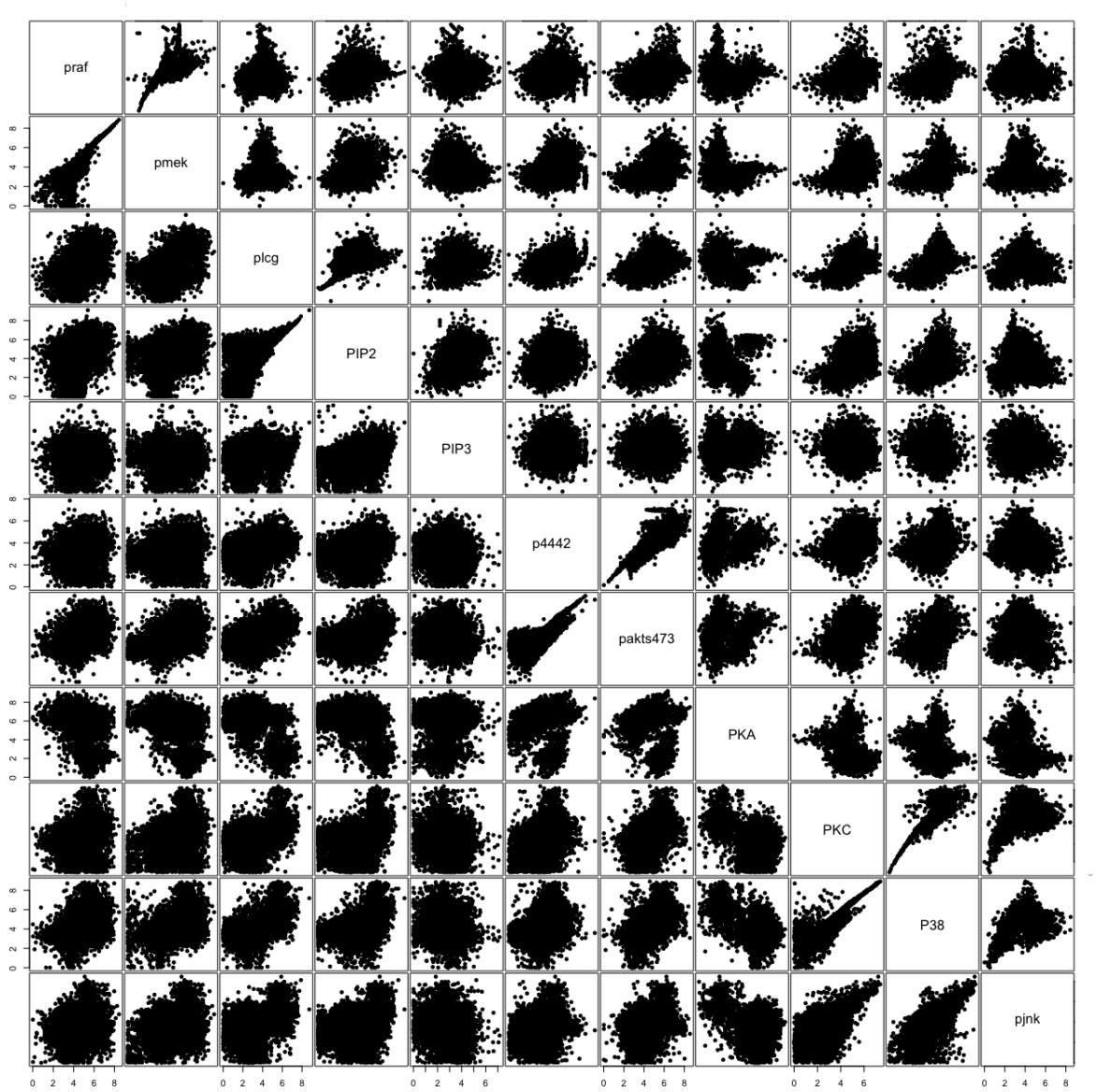


Figure 5.20: Pairs plots of observed dataset (lower triangle) and dataset sampled from the VCMM(13,LC,LM) (upper triangle).

The results are shown in Figure 5.19 and Figure 5.20. If the datasets were sampled from the identical same distribution they would -most likely- look like if they have been mirrored on the diagonal. We can see that the results of the VCMMs are better then the results from the GMM. Still we can see, that there are for both models differences between the original and the simulated data. Overall, Figure 5.19 and 5.20 alone do not clearly indicate which of the models is better, because for some variable pairs the VCMM(6,LC,LM) seems to fit better, and for some variables the VCMM(13,LC,LM) seems to recover more characteristics.

Chapter 6

Causal Analysis

In this chapter we are at first fitting D-vine regression models (see Kraus and Czado (2017)) to different groups of the Sachs dataset. In the second step, we then combine these to obtain causal D-vine regression models and sample data from them afterwards. In the following subchapter we will compare the causal models with the D-vine regression models fitted simply to the entire pooled data set in Chapter 4.3. Subsequent to this, we analyze how all the discussed D-vine regression models handle the tails of the distributions, which shows, that the choice of the used copula families - or restrictions of these lead to different results. Finally and as an outlook we apply vine copula mixture models to selected causal datasets to see if substructures exist within them.

6.1 D-Vine regression model fitting

In Table 3.1 we have seen, that the data comes from 14 different experiments, in which different stimulations were applied. These stimulations have direct influence on specific variables. Still the variables also influence each other, as illustrated in Figure 3.1. We call the DAG from Figure 3.1 the consent graph. These influences are resulting from biochemical processes within the observed T-cells, and are well researched. The biological background was discussed in short manner in the previous chapters, but it can be found in more depth in Sachs et al. (2005).

In this chapter we will apply causal analysis, by dividing the dataset in groups with only these observations, which have not been influenced directly for a specific variable. It should be noted, that there are groups of variables like `praf`, `plcg`, `pmek` or `p4442`, which are all influenced in the same experiments. Another such group are the variables `PIP3`, `P38` and `pjnk`, which are influenced in no experiment directly.

At first we fit marginal distributions to the variables of interest of the specific group.

Here we allow for some models only the univariate normal distribution, and for others all the parametric marginal distributions we have already worked with in the previous chapters. With these we then transform the data (of the specific group) to u-scale. To these observations on u-scale we are fitting D-vine regression models with the **vinereg** library by Nagler and Kraus (2021). In a first step, we work with data transformed by Gaussian margins, and then allow only the Gaussian copula family as well as the indepencence copula for the D-vine regression. Inviews of the Gaussian margins and the Gaussian copula families, we call these models $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$. For a second model we use all the parametric margins, we have already worked with in the previous chapters, but still restrict the copula families to the Gaussian copula family as well as the indepencence copula, therefore we describe it as $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$. In a third model class, we then use again all the parametric margins and we also allow all parametric copula families, which we denote as $\mathbf{M}_{Par}\mathbf{C}_{Par}$. In all the models we use the topological order induced by the consent graph, which is given by `PIP3`, `plcg`, `PIP2`, `PKC`, `PKA`, `P38`, `pjnk`, `praf`, `pmek`, `p4442`, `pakts473`.

6.1.1 Variable plcg

Since `PIP3` has no parent, we do not fit a D-vine regression model for it, but start with the variable `plcg` which is a child node of `PIP3`.

(a) Gaussian marginal parameter estimates:

Variable	Family	Parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PIP3	normal	2.641, 1.001	8, 9	1028	-1459.97
plcg	normal	3.034, 0.7	8, 9	1028	-1091.108

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with Gaussian margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
plcg	PIP3	105.355	-208.711	-203.776	8, 9	1028

(c) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with Gaussian margins and Gaussian copulas :

Variable	Pair copula	Family	Parameters	Tau
plcg	plcg, PIP3	gaussian	0.431	0.284

Table 6.1: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b) and the copula parameters (panel c). Only the observations of the Sachs data, which have not been directly perturbed in plcg have been used.

(a) Parametric marginal families and parameter estimates:

Variable	Family	Parameters	Number of parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PIP3	normal	(2.641, 1.001)	2	8, 9	1028	-1459.97
plcg	t-fix	(3.050, 0.462)	2	8, 9	1028	-1031.63

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
plcg	PIP3	99.275	-196.55	-191.615	8, 9	1028

(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with parametric margins and Gaussian copulas:

Variable	Pair copula	Family	Parameters	Tau
plcg	plcg, PIP3	gaussian	0.424	0.279

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and parametric copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
plcg	PIP3	219.967	-435.934	-426.063	8, 9	1028

(e) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copulas fitted for the regression D-vine with parametric margins and parametric copulas:

Variable	Pair copula	Family	Rotation	Parameters	Tau
plcg	plcg, PIP3	bb8	0	1.914, 0.999	0.334

Table 6.2: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal) + $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b and d) and the copula parameters (panel c and e). Only the observations of the Sachs data, which have not been directly perturbed in plcg have been used.

6.1.2 Variable PIP2

(a) Gaussian marginal parameter estimates:

Variable	Family	Parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PIP3	normal	2.875, 0.971	1-4, 6-11, 13-14	8738	-12141.029
plcg	normal	3.045, 1.351	1-4, 6-11, 13-14	8738	-15029.257
PIP2	normal	4.346, 1.387	1-4, 6-11, 13-14	8738	-15259.27

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with Gaussian margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
PIP2	plcg, PIP3	2660.696	-5315.392	-5294.166	1-4, 6-11, 13-14	8738

(c) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with Gaussian margins and Gaussian copulas :

Variable	Pair copula	Family	Parameters	Tau
PIP2	PIP2, plcg	gaussian	0.546	0.368
	plcg, PIP3	gaussian	0.053	0.034
	PIP2, PIP3; plcg	gaussian	0.472	0.313

Table 6.3: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b) and the copula parameters (panel c). Only the observations of the Sachs data, which have not been directly perturbed in PIP2 have been used.

(a) Parametric marginal families and parameter estimates:

Variable	Family	Parameters	Number of parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PIP3	skew-t	(2.876, 0.971, 23.888, 0.881)	4	1-4, 6-11, 13-14	8738	-12095.24
plcg	skew-t	(3.053, 1.413, 4.632, 1.421)	4	1-4, 6-11, 13-14	8738	-14500.60
PIP2	normal	(4.346, 1.387)	2	1-4, 6-11, 13-14	8738	-15259.27

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
PIP2	plcg, PIP3	2194.22	-4382.44	-4361.214	1-4, 6-11, 13-14	8738

(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with parametric margins and Gaussian copulas:

Variable	Pair copula	Family	Parameters	Tau
PIP2	PIP2, plcg	gaussian	0.478	0.318
	plcg, PIP3	gaussian	0.056	0.036
	PIP2, PIP3; plcg	gaussian	0.461	0.305

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and parametric copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
PIP2	plcg, PIP3	3758.478	-7508.957	-7480.655	1-4, 6-11, 13-14	8738

(e) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copulas fitted for the regression D-vine with parametric margins and parametric copulas:

Variable	Pair copula	Family	Rotation	Parameters	Tau
PIP2	PIP2, plcg	joe	0	2.002	0.356
	plcg, PIP3	clayton	180	0.091	0.043
	PIP2, PIP3; plcg	bb8	0	3.826, 0.688	0.354

Table 6.4: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal) + $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b and d) and the copula parameters (panel c and e). Only the observations of the Sachs data, which have not been directly perturbed in PIP2 have been used.

6.1.3 Variable PKC

(a) Gaussian marginal parameter estimates:

Variable	Family	Parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
plcg	normal	3.024, 0.661	9	155	-155.177
PIP2	normal	4.213, 1.138	9	155	-239.523
PKC	normal	1.388, 0.984	9	155	-216.887

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with Gaussian margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
PKC		0	0	0	9	155

Table 6.5: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Estimation results with regard to the margins (panel a) and the copula-loglikelihood (panel b). Only the observations of the Sachs data, which have not been directly perturbed in PKC have been used.

(a) Parametric marginal families and parameter estimates:

Variable	Family	Parameters	Number of parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
plcg	t-fix	(3.039, 0.460)	2	9	155	-149.296
PIP2	skew-normal	(4.247, 1.152, 0.718)	3	9	155	-238.33
PKC	skew-normal	(1.364, 1.010, 8.669)	3	9	155	-193.89

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
PKC		0	0	0	9	155

(c) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and parametric copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
PKC	PIP2	1.311	-0.622	2.421	9	155

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copulas fitted for the regression D-vine with parametric margins and parametric copulas:

Variable	Pair copula	Family	Rotation	Parameters	Tau
PKC	PKC, PIP2	joe	270	1.157	-0.082

Table 6.6: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal) + $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b and c) and the copula parameters (panel d). Only the observations of the Sachs data, which have not been directly perturbed in PKC have been used.

6.1.4 Variable PKA

(a) Gaussian marginal parameter estimates:

Variable	Family	Parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKC	normal	2.714, 1.012	1-8, 10-14	9994	-14302.386
PKA	normal	6.107, 1.304	1-8, 10-14	9994	-16833.374

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with Gaussian margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
PKA	PKC	713.791	-1425.582	-1418.373	1-8, 10-14	9994

(c) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with Gaussian margins and Gaussian copulas :

Variable	Pair copula	Family	Parameters	Tau
PKA	PKA, PKC	gaussian	-0.306	-0.198

Table 6.7: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b) and the copula parameters (panel c). Only the observations of the Sachs data, which have not been directly perturbed in PKA have been used.

(a) Parametric marginal families and parameter estimates:

Variable	Family	Parameters	Number of parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKC	skew-t	(2.708, 1.055, 4.104, 1.035)	4	1-8, 10-14	9994	-13929.14
PKA	skew-t	(6.221, 9.842, 2.008, 0.880)	4	1-8, 10-14	9994	-14526.86

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
PKA	PKC	282.082	-562.164	-554.955	1-8, 10-14	9994

(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with parametric margins and Gaussian copulas:

Variable	Pair copula	Family	Parameters	Tau
PKA	PKA, PKC	gaussian	-0.234	-0.15

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and parametric copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
PKA	PKC	870.333	-1738.666	-1731.456	1-8, 10-14	9994

(e) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copulas fitted for the regression D-vine with parametric margins and parametric copulas:

Variable	Pair copula	Family	Rotation	Parameters	Tau
PKA	PKA, PKC	joe	90	1.345	-0.163

Table 6.8: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal) + $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b and d) and the copula parameters (panel c and e). Only the observations of the Sachs data, which have not been directly perturbed in PKA have been used.

6.1.5 Variable P38

(a) Gaussian marginal parameter estimates:

Variable	Family	Parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKC	normal	2.694, 1.025	1-14	10149	-14648.882
PKA	normal	6.106, 1.296	1-14	10149	-17032.228
P38	normal	3.765, 1.106	1-14	10149	-15422.559

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with Gaussian margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
P38	PKC, PKA	6148.889	-12291.779	-12270.103	1-14	10149

(c) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with Gaussian margins and Gaussian copulas :

Variable	Pair copula	Family	Parameters	Tau
P38	P38, PKC	gaussian	0.776	0.566
	PKC, PKA	gaussian	-0.295	-0.191
	P38, PKA; PKC	gaussian	-0.387	-0.253

Table 6.9: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b) and the copula parameters (panel c). Only the observations of the Sachs data, which have not been directly perturbed in P38 have been used.

(a) Parametric marginal families and parameter estimates:

Variable	Family	Parameters	Number of parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKC	skew-t	(2.687, 1.067, 4.157, 1.016)	4	1-14	10149	-14295.35
PKA	skew-t	(6.220, 9.896, 2.008, 0.886)	4	1-14	10149	-14706.29
P38	skew-t	(3.790, 1.988, 2.250, 1.504)	4	1-14	10149	-13612.65

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
P38	PKC, PKA	4623.152	-9240.304	-9218.628	1-14	10149

(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with parametric margins and Gaussian copulas:

Variable	Pair copula	Family	Parameters	Tau
P38	P38, PKC	gaussian	0.752	0.542
	PKC, PKA	gaussian	-0.226	-0.145
	P38, PKA; PKC	gaussian	-0.16	-0.102

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and parametric copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
P38	PKC, PKA	6794.476	-13578.951	-13542.826	1-14	10149

(e) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copulas fitted for the regression D-vine with parametric margins and parametric copulas:

Variable	Pair copula	Family	Rotation	Parameters	Tau
P38	P38, PKC	bb8	0	3.686, 0.991	0.581
	PKC, PKA	joe	270	1.337	-0.16
	P38, PKA; PKC	bb8	270	1.151, 0.981	-0.069

Table 6.10: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal) + $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b and d) and the copula parameters (panel c and e). Only the observations of the Sachs data, which have not been directly perturbed in P38 have been used.

6.1.6 Variable pjnk

(a) Gaussian marginal parameter estimates:

Variable	Family	Parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKC	normal	2.694, 1.025	1-14	10149	-14648.882
PKA	normal	6.106, 1.296	1-14	10149	-17032.228
pjnk	normal	3.457, 1.207	1-14	10149	-16312.928

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with Gaussian margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
pjnk	PKC, PKA	2332.515	-4659.031	-4637.356	1-14	10149

(c) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with Gaussian margins and Gaussian copulas :

Variable	Pair copula	Family	Parameters	Tau
pjnk	pjnk, PKC	gaussian	0.506	0.338
	PKC, PKA	gaussian	-0.295	-0.191
	pjnk, PKA; PKC	gaussian	-0.173	-0.111

Table 6.11: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b) and the copula parameters (panel c). Only the observations of the Sachs data, which have not been directly perturbed in pjnk have been used.

(a) Parametric marginal families and parameter estimates:

Variable	Family	Parameters	Number of parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKC	skew-t	(2.687, 1.067, 4.157, 1.016)	4	1-14	10149	-14295.35
PKA	skew-t	(6.220, 9.896, 2.008, 0.886)	4	1-14	10149	-14706.29
pjnk	logistic	(3.460, 0.665)	2	1-14	10149	-16179.19

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
pjnk	PKC, PKA	1476.262	-2946.524	-2924.848	1-14	10149

(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with parametric margins and Gaussian copulas:

Variable	Pair copula	Family	Parameters	Tau
pjnk	pjnk, PKC	gaussian	0.455	0.301
	PKC, PKA	gaussian	-0.226	-0.145
	pjnk, PKA; PKC	gaussian	-0.075	-0.048

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and parametric copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
pjnk	PKC, PKA	2744.762	-5479.523	-5443.397	1-14	10149

(e) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copulas fitted for the regression D-vine with parametric margins and parametric copulas:

Variable	Pair copula	Family	Rotation	Parameters	Tau
pjnk	pjnk, PKC	bb7	0	1.687, 0.165	0.316
	PKC, PKA	joe	270	1.337	-0.16
	pjnk, PKA; PKC	t	0	0.021, 13.489	0.013

Table 6.12: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal) + $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b and d) and the copula parameters (panel c and e). Only the observations of the Sachs data, which have not been directly perturbed in pjnk have been used.

6.1.7 Variable praf

(a) Gaussian marginal parameter estimates:

Variable	Family	Parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKC	normal	2.433, 0.915	8, 9	1028	-1366.871
PKA	normal	6.461, 0.655	8, 9	1028	-1023.593
praf	normal	3.242, 0.704	8, 9	1028	-1097.672

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with Gaussian margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
praf	PKC	1.72	-1.44	3.495	8, 9	1028

(c) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with Gaussian margins and Gaussian copulas :

Variable	Pair copula	Family	Parameters	Tau
praf	praf, PKC	gaussian	-0.058	-0.037

Table 6.13: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b) and the copula parameters (panel c). Only the observations of the Sachs data, which have not been directly perturbed in praf have been used.

(a) Parametric marginal families and parameter estimates:

Variable	Family	Parameters	Number of parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKC	skew-t	(2.432, 0.953, 4.628, 0.862)	4	8, 9	1028	-1340.88
PKA	skew-t	(6.473, 0.669, 7.566, 1.555)	4	8, 9	1028	-970.96
praf	skew-t	(3.255, 0.717, 6.211, 1.202)	4	8, 9	1028	-1075.40

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
praf	PKC, PKA	13.554	-21.108	-6.302	8, 9	1028

(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with parametric margins and Gaussian copulas:

Variable	Pair copula	Family	Parameters	Tau
praf	praf, PKC	gaussian	-0.061	-0.039
	PKC, PKA	gaussian	0.142	0.091
	praf, PKA; PKC	gaussian	-0.046	-0.029

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and parametric copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
praf	PKC, PKA	25.073	-42.146	-22.404	8, 9	1028

(e) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copulas fitted for the regression D-vine with parametric margins and parametric copulas:

Variable	Pair copula	Family	Rotation	Parameters	Tau
praf	praf, PKC	clayton	90	0.096	-0.046
	PKC, PKA	bb8	180	1.272, 0.957	0.104
	praf, PKA; PKC	frank	0	-0.31	-0.034

Table 6.14: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal) + $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b and d) and the copula parameters (panel c and e). Only the observations of the Sachs data, which have not been directly perturbed in praf have been used.

6.1.8 Variable pmek

(a) Gaussian marginal parameter estimates:

Variable	Family	Parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKC	normal	2.433, 0.915	8, 9	1028	-1366.871
PKA	normal	6.461, 0.655	8, 9	1028	-1023.593
praf	normal	3.242, 0.704	8, 9	1028	-1097.672
pmek	normal	2.586, 1.043	8, 9	1028	-1501.829

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with Gaussian margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
pmek	PKC, praf, PKA	144.491	-278.982	-254.305	8, 9	1028

(c) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with Gaussian margins and Gaussian copulas :

Variable	Pair copula	Family	Parameters	Tau
pmek	pmek, PKC	gaussian	0.388	0.254
	PKC, praf	gaussian	-0.058	-0.037
	praf, PKA	indep		0
	pmek, praf; PKC	gaussian	0.267	0.172
	PKC, PKA; praf	gaussian	0.132	0.084
	pmek, PKA; PKC, praf	gaussian	0.151	0.097

Table 6.15: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b) and the copula parameters (panel c). Only the observations of the Sachs data, which have not been directly perturbed in pmek have been used.

(a) Parametric marginal families and parameter estimates:

Variable	Family	Parameters	Number of parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKC	skew-t	(2.432, 0.953, 4.628, 0.862)	4	8, 9	1028	-1340.88
PKA	skew-t	(6.473, 0.669, 7.566, 1.555)	4	8, 9	1028	-970.96
praf	skew-t	(3.255, 0.717, 6.211, 1.202)	4	8, 9	1028	-1075.40
pmek	skew-t	(2.593, 1.111, 4.026, 0.799)	4	8, 9	1028	-1458.15

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
pmek	PKC, praf, PKA	143.342	-274.683	-245.071	8, 9	1028

(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with parametric margins and Gaussian copulas:

Variable	Pair copula	Family	Parameters	Tau
pmek	pmek, PKC	gaussian	0.331	0.215
	PKC, praf	gaussian	-0.061	-0.039
	praf, PKA	gaussian	-0.054	-0.035
	pmek, praf; PKC	gaussian	0.318	0.206
	PKC, PKA; praf	gaussian	0.139	0.089
	pmek, PKA; PKC, praf	gaussian	0.171	0.109

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and parametric copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
pmek	praf, PKC, PKA	313.688	-609.376	-564.958	8, 9	1028

(e) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copulas fitted for the regression D-vine with parametric margins and parametric copulas:

Variable	Pair copula	Family	Rotation	Parameters	Tau
pmek	pmek, praf	joe	0	1.752	0.295
	praf, PKC	clayton	90	0.096	-0.046
	PKC, PKA	bb8	180	1.272, 0.957	0.104
	pmek, PKC; praf	bb8	180	1.505, 0.98	0.199
	praf, PKA; PKC	frank	0	-0.31	-0.034
	pmek, PKA; praf, PKC	bb8	180	1.207, 0.965	0.084

Table 6.16: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal) + $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b and d) and the copula parameters (panel c and e). Only the observations of the Sachs data, which have not been directly perturbed in pmek have been used.

6.1.9 Variable p4442

(a) Gaussian marginal parameter estimates:

Variable	Family	Parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKA	normal	6.461, 0.655	8, 9	1028	-1023.593
pmek	normal	2.586, 1.043	8, 9	1028	-1501.829
p4442	normal	3.131, 0.976	8, 9	1028	-1433.533

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with Gaussian margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
p4442	PKA, pmek	155.235	-304.47	-289.664	8, 9	1028

(c) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with Gaussian margins and Gaussian copulas :

Variable	Pair copula	Family	Parameters	Tau
p4442	p4442, PKA	gaussian	0.482	0.32
	PKA, pmek	gaussian	0.179	0.115
	p4442, pmek; PKA	gaussian	0.072	0.046

Table 6.17: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b) and the copula parameters (panel c). Only the observations of the Sachs data, which have not been directly perturbed in p4442 have been used.

(a) Parametric marginal families and parameter estimates:

Variable	Family	Parameters	Number of parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PKA	skew-t	(6.473, 0.669, 7.566, 1.555)	4	8, 9	1028	-970.96
pmek	skew-t	(2.593, 1.111, 4.026, 0.799)	4	8, 9	1028	-1458.15
p4442	skew-normal	(3.130, 0.976, 0.870)	3	8, 9	1028	-1428.85

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
p4442	PKA, pmek	140.876	-275.751	-260.945	8, 9	1028

(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with parametric margins and Gaussian copulas:

Variable	Pair copula	Family	Parameters	Tau
p4442	p4442, PKA	gaussian	0.457	0.302
	PKA, pmek	gaussian	0.184	0.118
	p4442, pmek; PKA	gaussian	0.063	0.04

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and parametric copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
p4442	PKA, pmek	225.232	-440.464	-415.787	8, 9	1028

(e) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copulas fitted for the regression D-vine with parametric margins and parametric copulas:

Variable	Pair copula	Family	Rotation	Parameters	Tau
p4442	p4442, PKA	bb7	0	1.729, 0.109	0.314
	PKA, pmek	bb8	180	1.313, 0.978	0.131
	p4442, pmek; PKA	frank	0	0.627	0.069

Table 6.18: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal) + $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b and d) and the copula parameters (panel c and e). Only the observations of the Sachs data, which have not been directly perturbed in p4442 have been used.

6.1.10 Variable pakts473

(a) Gaussian marginal parameter estimates:

Variable	Family	Parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PIP3	normal	4.176, 1.257	1-2, 4-6, 8-9, 11-13	6795	-10500.41
PKA	normal	5.931, 1.462	1-2, 4-6, 8-9, 11-13	6795	-12223.191
p4442	normal	3.06, 1.054	1-2, 4-6, 8-9, 11-13	6795	-10001.389
pakts473	normal	4.176, 1.121	1-2, 4-6, 8-9, 11-13	6795	-10417.916

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with Gaussian margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
pakts473	p4442, PKA, PIP3	3781.127	-7552.253	-7518.133	1-2, 4-6, 8-9, 11-13	6795

(c) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with Gaussian margins and Gaussian copulas :

Variable	Pair copula	Family	Parameters	Tau
pakts473	pakts473, p4442	gaussian	0.781	0.571
	p4442, PKA	indep		
	PKA, PIP3	gaussian	0.08	0.051
	pakts473, PKA; p4442	gaussian	-0.308	-0.2
	p4442, PIP3; PKA	gaussian	-0.058	-0.037
	pakts473, PIP3; p4442, PKA	gaussian	0.062	0.039

Table 6.19: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b) and the copula parameters (panel c). Only the observations of the Sachs data, which have not been directly perturbed in pakts473 have been used.

(a) Parametric marginal families and parameter estimates:

Variable	Family	Parameters	Number of parameters	Estimated with data from experiments	Sample size	Marginal loglikelihood
PIP3	skew-t	(2.727, 0.944, 19.930, 0.918)	4	1-2, 4-6, 8-9, 11-13	6795	-9807.9
PKA	t-fix	(6.282, 0.760)	2	1-2, 4-6, 8-9, 11-13	6795	-11021.34
p4442	skew-t	(3.060, 1.056, 12.358, 0.954)	4	1-2, 4-6, 8-9, 11-13	6795	-9969.96
pakts473	skew-t	(4.200, 1.137, 7.140, 1.643)	4	1-2, 4-6, 8-9, 11-13	6795	-9918.73

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and Gaussian copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
pakts473	p4442, PKA, PIP3	3498.433	-6984.866	-6943.922	1-2, 4-6, 8-9, 11-13	6795

(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Copulas fitted for the regression D-vine with parametric margins and Gaussian copulas:

Variable	Pair copula	Family	Parameters	Tau
pakts473	pakts473, p4442	gaussian	0.781	0.57
	p4442, PKA	gaussian	0.079	0.05
	PKA, PIP3	gaussian	0.05	0.032
	pakts473, PKA; p4442	gaussian	-0.24	-0.154
	p4442, PIP3; PKA	gaussian	-0.062	-0.04
	pakts473, PIP3; p4442, PKA	gaussian	0.028	0.018

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copula-loglikelihood, -AIC and -BIC for the regression D-vine with parametric margins and parametric copulas:

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
pakts473	p4442, PKA, PIP3	4686.65	-9355.299	-9293.884	1-2, 4-6, 8-9, 11-13	6795

(e) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Copulas fitted for the regression D-vine with parametric margins and parametric copulas:

Variable	Pair copula	Family	Rotation	Parameters	Tau
pakts473	pakts473, p4442	bb6	0	1.414, 2.009	0.596
	p4442, PKA	t	0	0.154, 5.188	0.098
	PKA, PIP3	clayton	0	0.056	0.027
	pakts473, PKA; p4442	bb8	270	1.55, 0.942	-0.183
	p4442, PIP3; PKA	clayton	270	0.069	-0.034
	pakts473, PIP3; p4442, PKA	clayton	180	0.03	0.015

Table 6.20: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal) + $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Estimation results with regard to the margins (panel a), the copula-loglikelihood (panel b and d) and the copula parameters (panel c and e). Only the observations of the Sachs data, which have not been directly perturbed in pakts473 have been used.

6.2 Model summary

After fitting these different (sub-)models on the corresponding data sets, we now want to combine them into causal models for all 11 variables. These causal models are given in the following Tables.

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
plcg	PIP3	105.355	-208.711	-203.776	8, 9	1028
PIP2	plcg, PIP3	2660.696	-5315.392	-5294.166	1-4, 6-11, 13-14	8738
PKC		0	0	0	9	155
PKA	PKC	713.791	-1425.582	-1418.373	1-8, 10-14	9994
P38	PKC, PKA	6148.889	-12291.779	-12270.103	1-14	10149
pjnk	PKC, PKA	2332.515	-4659.031	-4637.356	1-14	10149
praf	PKC	1.720	-1.440	3.495	8, 9	1028
pmek	PKC, praf, PKA	144.491	-278.982	-254.305	8, 9	1028
p4442	PKA, pmek	155.235	-304.470	-289.664	8, 9	1028
pakts473	p4442, PKA, PIP3	3781.127	-7552.253	-7518.133	1-2, 4-6, 8-9, 11-13	6795

Variable	Pair copula	Family	Parameters	Tau
plcg	plcg, PIP3	gaussian	0.431	0.284
PIP2	PIP2, plcg	gaussian	0.546	0.368
	plcg, PIP3	gaussian	0.053	0.034
	PIP2, PIP3; plcg	gaussian	0.472	0.313
PKA	PKA, PKC	gaussian	-0.306	-0.198
P38	P38, PKC	gaussian	0.776	0.566
	PKC, PKA	gaussian	-0.295	-0.191
	P38, PKA; PKC	gaussian	-0.387	-0.253
pjnk	pjnk, PKC	gaussian	0.506	0.338
	PKC, PKA	gaussian	-0.295	-0.191
	pjnk, PKA; PKC	gaussian	-0.173	-0.111
praf	praf, PKC	gaussian	-0.058	-0.037
pmek	pmek, PKC	gaussian	0.388	0.254
	PKC, praf	gaussian	-0.058	-0.037
	praf, PKA	indep		0
	pmek, praf; PKC	gaussian	0.267	0.172
	PKC, PKA; praf	gaussian	0.132	0.084
	pmek, PKA; PKC, praf	gaussian	0.151	0.097
p4442	p4442, PKA	gaussian	0.482	0.320
	PKA, pmek	gaussian	0.179	0.115
	p4442, pmek; PKA	gaussian	0.072	0.046
pakts473	pakts473, p4442	gaussian	0.781	0.571
	p4442, PKA	indep		0
	PKA, PIP3	gaussian	0.080	0.051
	pakts473, PKA; p4442	gaussian	-0.308	-0.200
	p4442, PIP3; PKA	gaussian	-0.058	-0.037
	pakts473, PIP3; p4442, PKA	gaussian	0.062	0.039

Table 6.21: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Node-wise summary of all copulas fitted for the causal D-vine regression model. The set of feasible copula families was restricted to Gaussian and independence copulas only. For the transformation on u-scale only Gaussian margins were used.

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
plcg	PIP3	102.407	-202.814	-197.879	8, 9	1028
PIP2	plcg, PIP3	2194.220	-4382.44	-4361.214	1-4, 6-11, 13-14	8738
PKC		0	0	0	9	155
PKA	PKC	282.082	-562.164	-554.955	1-8, 10-14	9994
P38	PKC, PKA	4623.152	-9240.304	-9218.628	1-14	10149
pjnk	PKC, PKA	1476.262	-2946.524	-2924.848	1-14	10149
praf	PKC, PKA	13.554	-21.108	-6.302	8, 9	1028
pmek	PKC, praf, PKA	143.342	-274.683	-245.071	8, 9	1028
p4442	PKA, pmek	140.876	-275.751	-260.945	8, 9	1028
pakts473	p4442, PKA, PIP3	3498.433	-6984.866	-6943.922	1-2, 4-6, 8-9, 11-13	6795

Variable	Pair copula	Family	Parameters	Tau
plcg	plcg, PIP3	gaussian	0.424	0.279
PIP2	PIP2, plcg	gaussian	0.478	0.318
	plcg, PIP3	gaussian	0.056	0.036
	PIP2, PIP3; plcg	gaussian	0.461	0.305
PKA	PKA, PKC	gaussian	-0.234	-0.150
P38	P38, PKC	gaussian	0.752	0.542
	PKC, PKA	gaussian	-0.226	-0.145
	P38, PKA; PKC	gaussian	-0.160	-0.102
pjnk	pjnk, PKC	gaussian	0.455	0.301
	PKC, PKA	gaussian	-0.226	-0.145
	pjnk, PKA; PKC	gaussian	-0.075	-0.048
praf	praf, PKC	gaussian	-0.061	-0.039
	PKC, PKA	gaussian	0.142	0.091
	praf, PKA; PKC	gaussian	-0.046	-0.029
pmek	pmek, PKC	gaussian	0.331	0.215
	PKC, praf	gaussian	-0.061	-0.039
	praf, PKA	gaussian	-0.054	-0.035
	pmek, praf; PKC	gaussian	0.318	0.206
	PKC, PKA; praf	gaussian	0.139	0.089
	pmek, PKA; PKC, praf	gaussian	0.171	0.109
p4442	p4442, PKA	gaussian	0.457	0.302
	PKA, pmek	gaussian	0.184	0.118
	p4442, pmek; PKA	gaussian	0.063	0.040
pakts473	pakts473, p4442	gaussian	0.781	0.570
	p4442, PKA	gaussian	0.079	0.050
	PKA, PIP3	gaussian	0.050	0.032
	pakts473, PKA; p4442	gaussian	-0.240	-0.154
	p4442, PIP3; PKA	gaussian	-0.062	-0.040
	pakts473, PIP3; p4442, PKA	gaussian	0.028	0.018

Table 6.22: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Node-wise summary of all copulas fitted for the causal D-vine regression model. The set of feasible copula families was restricted to Gaussian and independence copulas only.

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
plcg	PIP3	219.967	-435.934	-426.063	8, 9	1028
PIP2	plcg, PIP3	3758.478	-7508.957	-7480.655	1-4, 6-11, 13-14	8738
PKC	PIP2	1.311	-0.622	2.421	9	155
PKA	PKC	870.333	-1738.666	-1731.456	1-8, 10-14	9994
P38	PKC, PKA	6794.476	-13578.951	-13542.826	1-14	10149
pjnk	PKC, PKA	2744.762	-5479.523	-5443.397	1-14	10149
praf	PKC, PKA	25.073	-42.146	-22.404	8, 9	1028
pmek	praf, PKC, PKA	313.688	-609.376	-564.958	8, 9	1028
p4442	PKA, pmek	225.232	-440.464	-415.787	8, 9	1028
pakts473	p4442, PKA, PIP3	4686.650	-9355.299	-9293.884	1-2, 4-6, 8-9, 11-13	6795

Variable	Pair copula	Family	Rotation	Parameters	Tau
plcg	plcg, PIP3	bb8	0	1.914, 0.999	0.334
PIP2	PIP2, plcg	joe	0	2.002	0.356
	plcg, PIP3	clayton	180	0.091	0.043
	PIP2, PIP3; plcg	bb8	0	3.826, 0.688	0.354
PKC	PKC, PIP2	joe	270	1.157	-0.082
PKA	PKA, PKC	joe	90	1.345	-0.163
P38	P38, PKC	bb8	0	3.686, 0.991	0.581
	PKC, PKA	joe	270	1.337	-0.160
	P38, PKA; PKC	bb8	270	1.151, 0.981	-0.069
pjnk	pjnk, PKC	bb7	0	1.687, 0.165	0.316
	PKC, PKA	joe	270	1.337	-0.160
	pjnk, PKA; PKC	t	0	0.021, 13.489	0.013
praf	praf, PKC	clayton	90	0.096	-0.046
	PKC, PKA	bb8	180	1.272, 0.957	0.104
	praf, PKA; PKC	frank	0	-0.310	-0.034
pmek	pmek, praf	joe	0	1.752	0.295
	praf, PKC	clayton	90	0.096	-0.046
	PKC, PKA	bb8	180	1.272, 0.957	0.104
	pmek, PKC; praf	bb8	180	1.505, 0.98	0.199
	praf, PKA; PKC	frank	0	-0.310	-0.034
	pmek, PKA; praf, PKC	bb8	180	1.207, 0.965	0.084
p4442	p4442, PKA	bb7	0	1.729, 0.109	0.314
	PKA, pmek	bb8	180	1.313, 0.978	0.131
	p4442, pmek; PKA	frank	0	0.627	0.069
pakts473	pakts473, p4442	bb6	0	1.414, 2.009	0.596
	p4442, PKA	t	0	0.154, 5.188	0.098
	PKA, PIP3	clayton	0	0.056	0.027
	pakts473, PKA; p4442	bb8	270	1.55, 0.942	-0.183
	p4442, PIP3; PKA	clayton	270	0.069	-0.034
	pakts473, PIP3; p4442, PKA	clayton	180	0.030	0.015

Table 6.23: $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Node-wise summary of all copulas fitted for the causal D-vine regression model. The set of feasible copula families was all parametric copula families.

6.3 Sampling

Again we want to sample $n = 10149$ (which is the sample size of the pooled data) data from the models, which is shown in Figures 6.2, 6.3 and 6.4. Since we have fitted the causal D-vine regression models to the data on u-scale, it makes sense to compare the results from the sampling to the original data und u-scale. The original (pooled) data shown in Figure 6.1 is transformed to u-scale by the marginal distributions given in Table 4.1. Analogous to Chapter 4.3, the sampled data here are generated by applying sequentially the corresponding algorithm of Bevacqua et al. (2017) to all nodes from Tables 6.21, 6.22 and 6.23.

Here, the differences between the original and the simulated data are particularly interesting, as they provide information about the result of our causal analysis. We pay particular attention to two criteria, the strength of Kendall's τ (i) and the tail dependence (ii):

(i)a The first variables that we notice are those that are very strongly correlated in the pooled data: These are `praf-pmek` ($\tau_{orig}^{praf-pmek} = 0.66$), `p4442-pakts473` ($\tau_{orig}^{p4442-pakts473} = 0.65$) and `PKC-P38` ($\tau_{orig}^{PKC-P38} = 0.59$). For the latter two, the correlation in the causal models are similarly strong with $\tau_{M_{Gauss}C_{Gauss}}^{p4442-pakts473} = 0.53$, $\tau_{M_{Par}C_{Gauss}}^{p4442-pakts473} = 0.55$ and $\tau_{M_{Par}C_{Par}}^{p4442-pakts473} = 0.58$, as well as $\tau_{M_{Gauss}C_{Gauss}}^{PKC-P38} = 0.57$, $\tau_{M_{Par}C_{Gauss}}^{PKC-P38} = 0.54$ and $\tau_{M_{Par}C_{Par}}^{PKC-P38} = 0.59$, respectively. However, it is special that the correlation between `praf` and `pmek` is much lower with $\tau_{M_{Gauss}C_{Gauss}}^{praf-pmek} = 0.14$, $\tau_{M_{Par}C_{Gauss}}^{praf-pmek} = 0.18$ or $\tau_{M_{Par}C_{Par}}^{praf-pmek} = 0.21$. This cannot be due to the fact that the variables in the graph we have given only indirectly affect each other. Instead, according to Table 6.23, the dependence of `pmek` and `praf` was explicitly measured in a bivariate copula. This bivariate copula is an unrotated Joe copula with only one parameter, i.e. the parameter is already defined by the correlation (and could be estimated through inversion of Kendalls tau).

(i)b Similarly interesting is the high correlation ($\tau_{M_{Par}C_{Par}}^{plcg-PIP2} = \tau_{M_{Gauss}C_{Gauss}}^{plcg-PIP2} = 0.45$ and $\tau_{M_{Par}C_{Gauss}}^{plcg-PIP2} = 0.4$) of `plcg` and `PIP2` in our model, which is much higher than the correlation of these variables ($\tau_{orig}^{plcg-PIP2} = 0.34$) in the pooled data. Also, this must come explicitly from the data of our causal analysis, since in our model with all parametric copula families `PIP2` depends directly on `plcg`. For the all parametric Model, the bivariate copula of `PIP2` and `plcg` is a Joe copula with one parameter as shown in Table 6.23. For the two other models it is of course a Gaussian copula.

(i)c Another large difference in the strength of correlation is between the variables `plcg`

and PKC: the simulated data have correlations of $\tau^{plcg-PKC}_{M_{Par}C_{Par}} = -0.038$, $\tau^{plcg-PKC}_{M_{Par}C_{Gauss}} = 0.01$ and $\tau^{plcg-PKC}_{M_{Gauss}C_{Gauss}} = -0.027$, while the correlation in the original data is only $\tau^{plcg-PKC}_{orig} = 0.16$. In our model with all parametric copula families PKC only depends on PIP2, which itself depends on plcg, so these variables are connected indirectly. In both Gaussian models PKC has no parent nodes (ie. its independet of the other variables), which leeds to the τ-value of almost zero.

(i)d The same holds for the pair plcg and PIP3: plcg was simulated via PIP3, but the correlation in the simulated data is $\tau^{plcg-PIP3}_{M_{Par}C_{Par}} = 0.34$, $\tau^{plcg-PIP3}_{M_{Par}C_{Gauss}} = \tau^{plcg-PIP3}_{M_{Gauss}C_{Gauss}} = 0.28$ respectively, whereas in the original data it is only $\tau^{plcg-PIP3}_{orig} = 0.079$. This is very interesting and seems to come from the fitting with the causal data, in which the ignored external influences on plcg lead to a stronger correlation.

(ii)a The comparison of tail dependence is also interesting, but only possible for the D-vine regression model with all parametric copulas and not for the Gaussian models: while in many cases the tail dependence of the simulated data is similar to that of the original data, there are mostly no tail dependencies identifiable for the variable praf or pakts473.

(ii)a A particularly interesting pair is pmek - PKC: in the original data, these are upper tail dependent, while in the simulated data they are lower tail dependent. The variable pmek was generated (among other) by PKC.

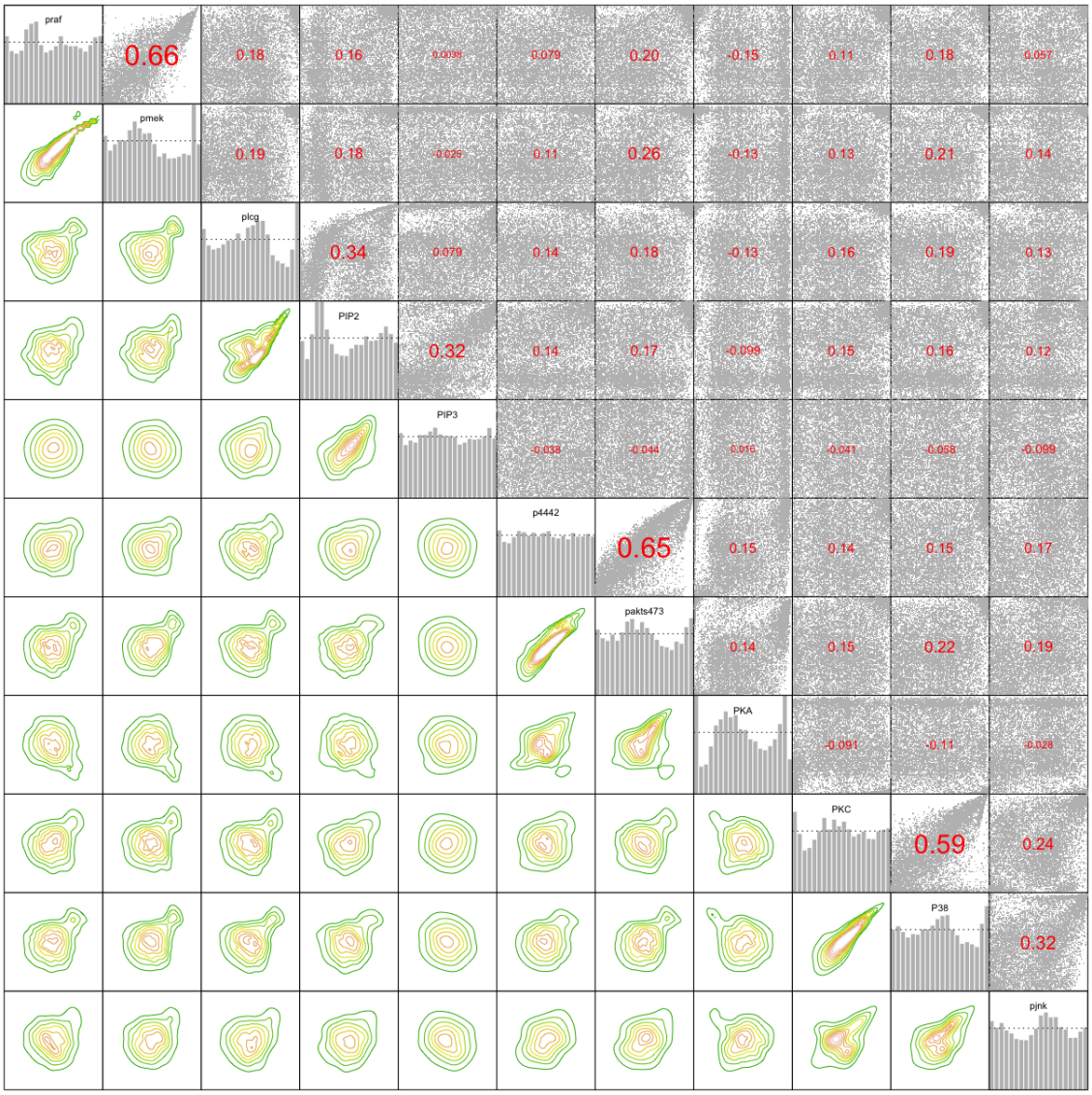


Figure 6.1: Normalized contour and pairs plots as well as Kendall's τ_{orig} of the pooled data on u-scale. The data is transformed by the marginal distributions given in Table 4.1.

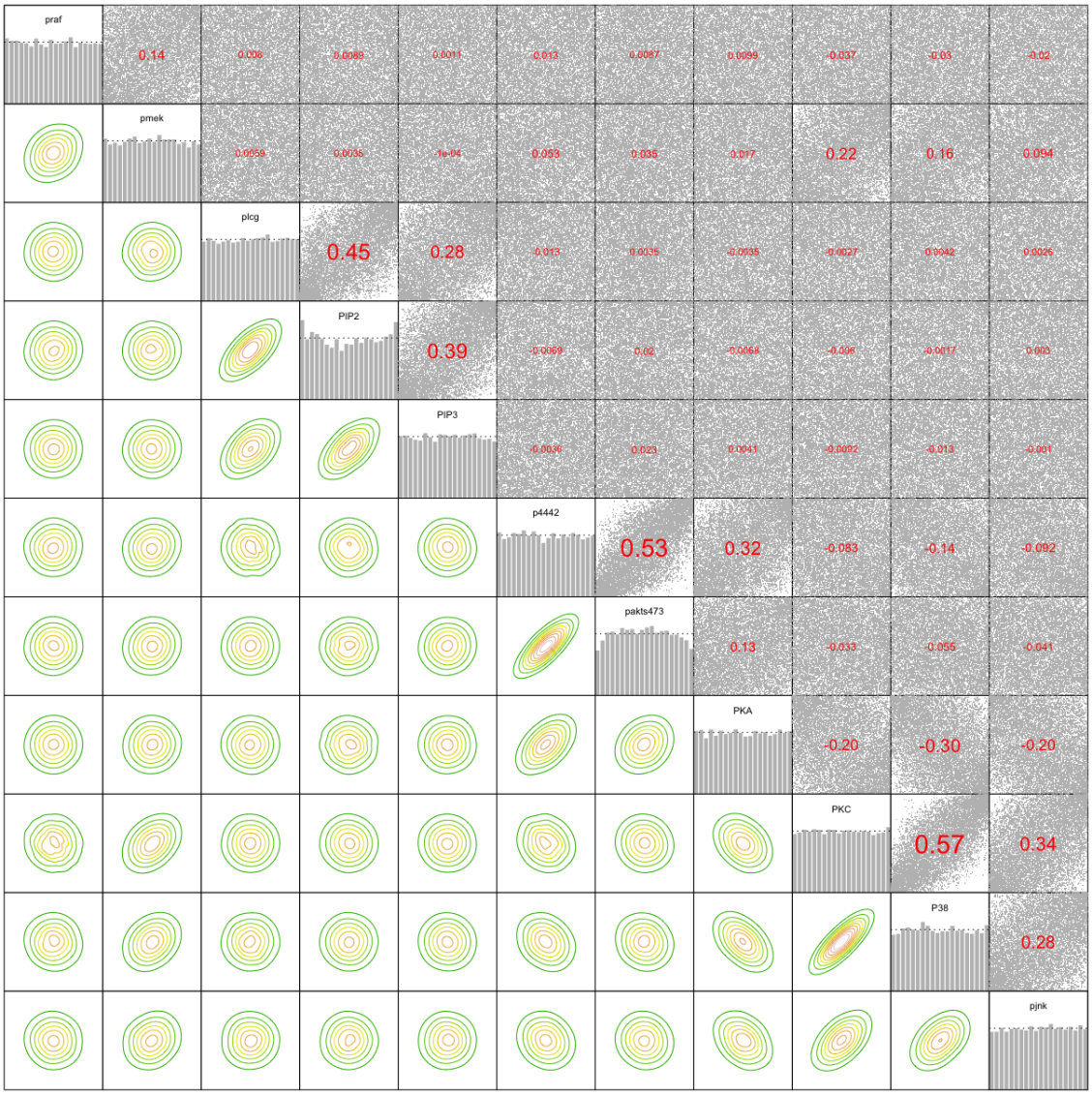


Figure 6.2: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ (causal): Normalized contours and pairs plots as well as Kendall's τ of the simulated data on u-scale. The data is simulated from the causal D-vine regression model given in Table 6.21.

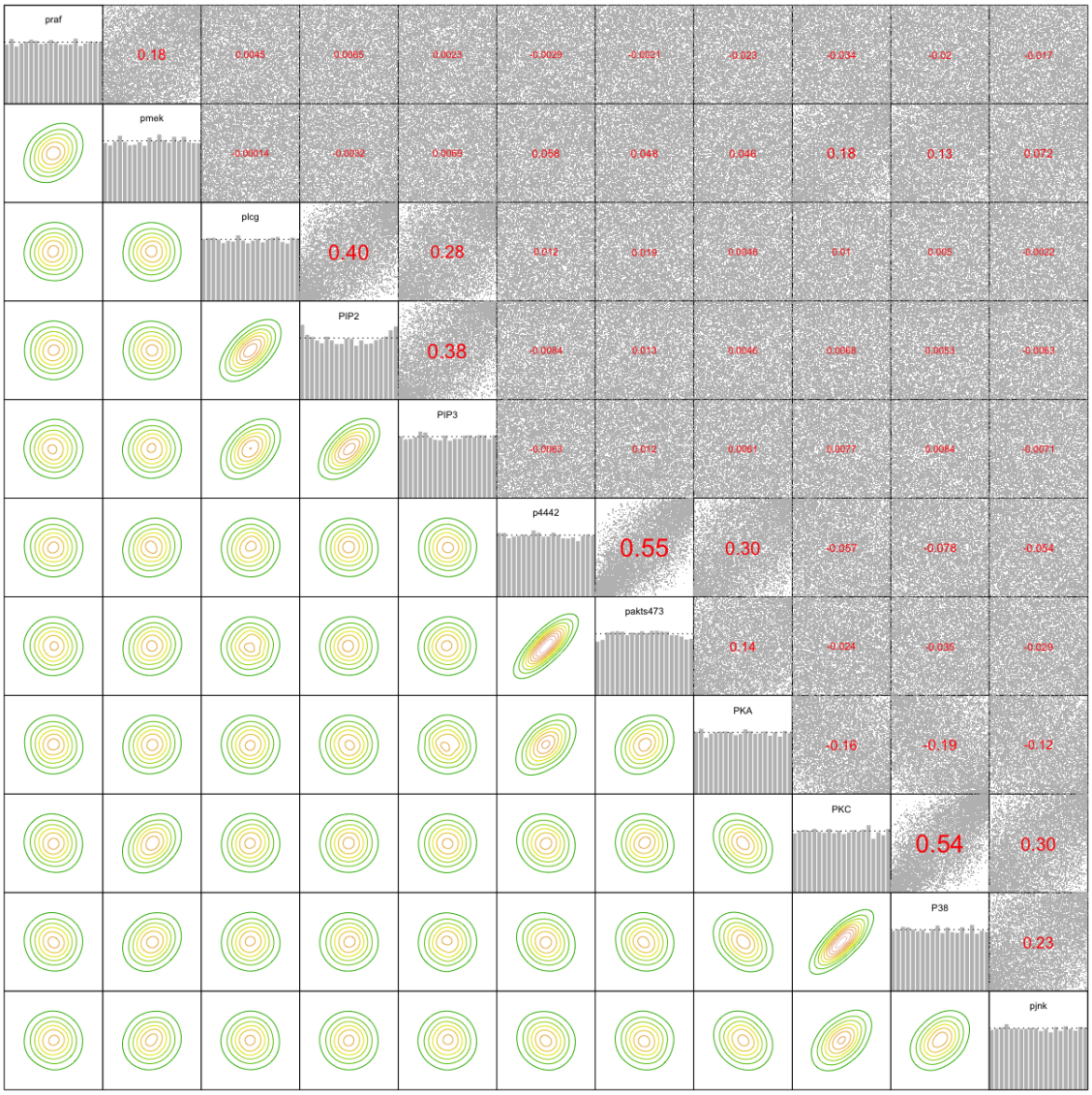


Figure 6.3: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ (causal): Normalized contours and pairs plots as well as Kendall's τ of the simulated data on u-scale. The data is simulated from the causal D-vine regression model given in Table 6.22.

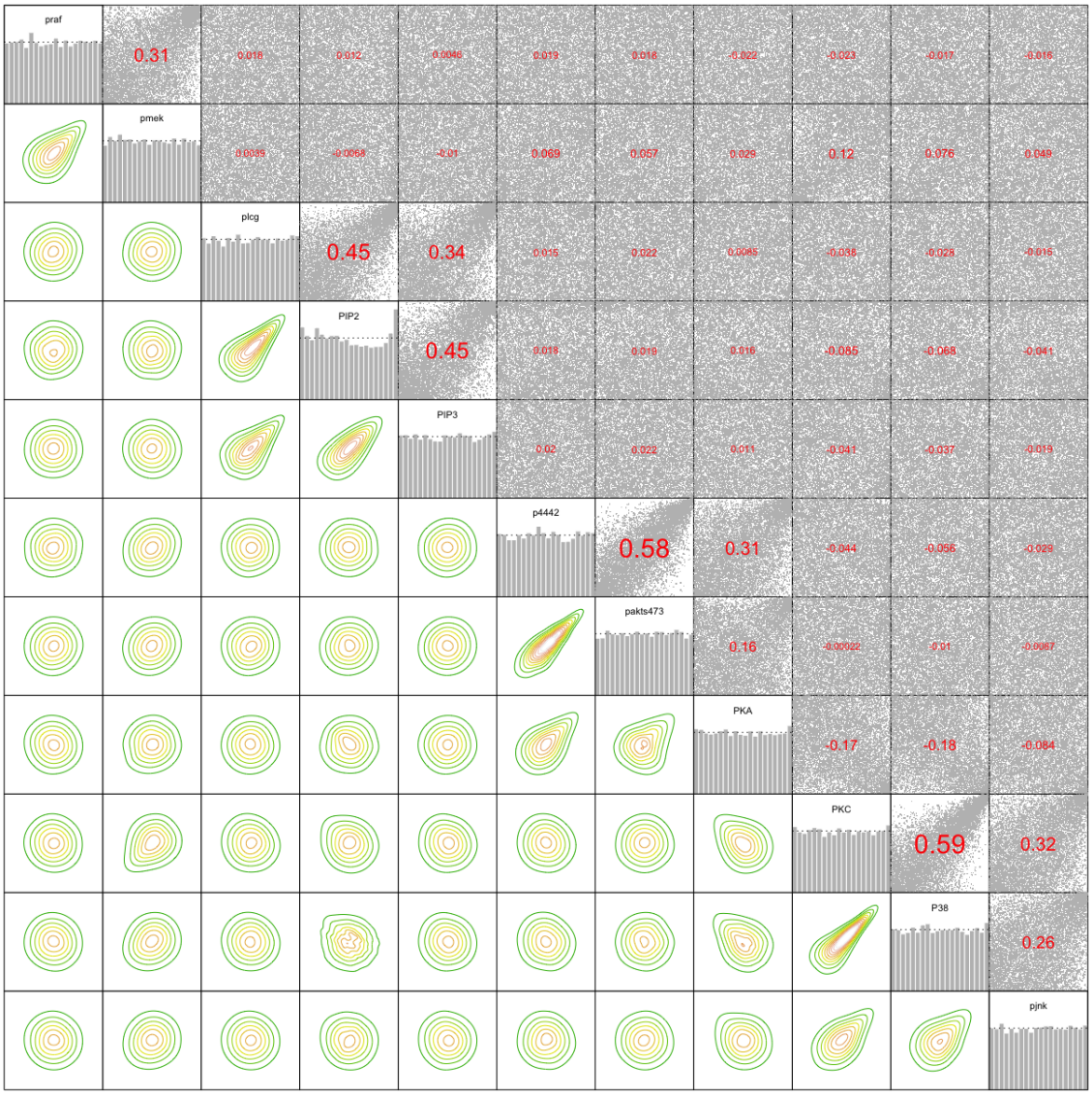


Figure 6.4: $\mathbf{M}_{Par}\mathbf{C}_{Par}$ (causal): Normalized contours and pairs plots as well as Kendall's τ of the simulated data on u-scale. The data is simulated from the causal D-vine regression model given in Table 6.23.

6.4 Comparison of pooled and causal models

In this chapter we want to compare the three models fitted to the causal data: One with Gaussian margins and Gaussian copula familes only ($\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$); one with more parametric margins but Gaussian copula families only ($\mathbf{M}_{Par}\mathbf{C}_{Gauss}$) and finally one with more parametric margins and copula families ($\mathbf{M}_{Par}\mathbf{C}_{Par}$). We did the same in chapter 4.3 with the pooled data. Therefore we want to compare now the three pairs of models with the same specifications.

The $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$-models differ in the nodes `PKC` and `praf` regarding their parent nodes. In the model for the causal data, `PKC` has no parent nodes anymore, while it has two in the model for the pooled data. On the other hand, in all the following nodes, where `PKC` is a parent node according to the consent graph, it is directly connected to the respective regression variable in the regrssion D-vine. Besides that, the models also differ in the nodes `pmek` and `p4442` in the order of the respective regression D-vines.

In the $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$-models, we see similar results: Again `PKC` has no parent nodes in the model for the causal data and two parent nodes in the model for the pooled data. Beides that, again `PKC` seems to have strong influence on the variables, whose parent node it is itself.

For the $\mathbf{M}_{Par}\mathbf{C}_{Par}$-models, `PKC` has `plcg` as parent node in the model fitted to the pooled data and `PIP2` as parent node in the model fitted to the causal data. Besides that, the only change in the regression D-vines is the different order of `PKC` and `PKA` in the node `praf`. Regarding the DAG, this makes the models very similar. Another interesting fact regarding the $\mathbf{M}_{Par}\mathbf{C}_{Par}$-models is for example the large difference in correlation between `plcg-PIP3` in the node `plcg`, and between `pmek-praf` in the node of `pmek`.

In Table 6.24 we have the AIC on the original scale, i.e. on x-scale. This means we are takting the marginal likelihoods into account. One can clearly see, that the AIC is getting lower with less restrictions on the margins and the copula families. The AIC values from the D-vine regression models fitted to the pooled data is not comparible to them of the causal D-vine regression models, since they are mainly (besides the variables P38 and pjnk) fitted to different datasets with different sample sizes.

Node	$\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ pooled	$\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ pooled	$\mathbf{M}_{Par}\mathbf{C}_{Par}$ pooled	$\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ causal	$\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ causal	$\mathbf{M}_{Par}\mathbf{C}_{Par}$ causal
plcg	62792.87	61535.79	61502.45	4901.45	4788.39	4555.27
PIP2	94740.77	93655.14	90685.57	79555.72	79347.78	76221.26
PKC	96620.33	95419.56	94611.74	1231.17	1175.03	1174.41
PKA	61995.70	57488.80	56289.82	60853.94	56365.84	55189.33
P38	81921.56	76012.28	71673.63	81921.56	76012.28	71673.63
pjnk	91335.04	87435.14	84902.14	91335.04	87435.14	84902.14
praf	90057.37	85062.65	83154.20	6986.83	6777.37	6756.33
pmek	109109.90	105116.30	98768.48	9716.95	9448.10	9113.40
p4442	94917.50	88191.90	86522.08	7625.44	7462.17	7297.46
pakts473	109715.40	103205.00	100583.90	78749.56	74478.99	72108.56

Table 6.24: Nodewise AIC on original scale.

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Sample size
plcg	PIP3	55.626	-109.252	-102.026	10149
PIP2	plcg, PIP3	2965.669	-5925.338	-5903.663	10149
PKC	plcg, PIP2	2680.441	-5354.883	-5333.207	10149
PKA	PKC	686.261	-1370.522	-1363.297	10149
P38	PKC, PKA	6148.889	-12291.779	-12270.103	10149
pjnk	PKC, PKA	2332.515	-4659.031	-4637.356	10149
praf	PKA, PKC	1498.523	-2991.046	-2969.371	10149
pmek	praf, PKC, PKA	9232.084	-18452.167	-18408.816	10149
p4442	pmek, PKA	1200.909	-2395.818	-2374.143	10149
pakts473	p4442, PKA	5334.243	-10662.486	-10640.81	10149

(a) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ - pooled

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
plcg	PIP3	105.355	-208.711	-203.776	8, 9	1028
PIP2	plcg, PIP3	2660.696	-5315.392	-5294.166	1-4, 6-11, 13-14	8738
PKC		0	0	0	9	155
PKA	PKC	713.791	-1425.582	-1418.373	1-8, 10-14	9994
P38	PKC, PKA	6148.889	-12291.779	-12270.103	1-14	10149
pjnk	PKC, PKA	2332.515	-4659.031	-4637.356	1-14	10149
praf	PKC	1.720	-1.440	3.495	8, 9	1028
pmek	PKC, praf, PKA	144.491	-278.982	-254.305	8, 9	1028
p4442	PKA, pmek	155.235	-304.470	-289.664	8, 9	1028
pakts473	p4442, PKA, PIP3	3781.127	-7552.253	-7518.133	1-2, 4-6, 8-9, 11-13	6795

(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ - causal

Variable	Pair copula	Tau
plcg	plcg, PIP3	0.067
PIP2	PIP2, plcg	0.372
	plcg, PIP3	0.067
	PIP2, PIP3; plcg	0.287
PKC	PKC, plcg	0.252
	plcg, PIP2	0.372
	PKC, PIP2; plcg	0.042
PKA	PKA, PKC	-0.191
P38	P38, PKC	0.566
	PKC, PKA	-0.191
	P38, PKA; PKC	-0.253
pjnk	pjnk, PKC	0.338
	PKC, PKA	-0.191
	pjnk, PKA; PKC	-0.111
praf	praf, PKA	-0.246
	PKA, PKC	-0.191
	praf, PKC; PKA	0.055
pmek	pmek, praf	0.680
	praf, PKC	0.135
	PKC, PKA	-0.191
	pmek, PKC; praf	0.129
	praf, PKA; PKC	-0.217
	pmek, PKA; praf, PKC	-0.094
p4442	p4442, pmek	0.104
	pmek, PKA	-0.229
	p4442, PKA; pmek	0.102
pakts473	pakts473, p4442	0.579
	p4442, PKA	0.053
	pakts473, PKA; p4442	-0.079

(c) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ - pooled

Variable	Pair copula	Tau
plcg	plcg, PIP3	0.284
PIP2	PIP2, plcg	0.368
	plcg, PIP3	0.034
	PIP2, PIP3; plcg	0.313
PKC		
PKA	PKA, PKC	-0.198
P38	P38, PKC	0.566
	PKC, PKA	-0.191
	P38, PKA; PKC	-0.253
pjnk	pjnk, PKC	0.338
	PKC, PKA	-0.191
	pjnk, PKA; PKC	-0.111
praf	praf, PKC	-0.037
pmek	pmek, PKC	0.254
	PKC, praf	-0.037
	praf, PKA	0
	pmek, praf; PKC	0.172
	PKC, PKA; praf	0.084
	pmek, PKA; PKC, praf	0.097
p4442	p4442, PKA	0.320
	PKA, pmek	0.115
	p4442, pmek; PKA	0.046
pakts473	pakts473, p4442	0.571
	p4442, PKA	0
	PKA, PIP3	0.051
	pakts473, PKA; p4442	-0.200
	p4442, PIP3; PKA	-0.037
	pakts473, PIP3; p4442, PKA	0.039

(d) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ - causal

Table 6.25: Overview of the pooled and causal $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$-models.

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Sample size
plcg	PIP3	65.103	-128.207	-120.982	10149
PIP2	plcg, PIP3	2777.391	-5548.782	-5527.107	10149
PKC	plcg, PIP2	2239.298	-4472.597	-4450.921	10149
PKA	PKC	266.239	-530.477	-523.252	10149
P38	PKC, PKA	4623.152	-9240.304	-9218.628	10149
pjnk	PKC, PKA	1476.262	-2946.524	-2924.848	10149
praf	PKA, PKC	766.946	-1527.891	-1506.216	10149
pmek	praf, PKC, PKA	6604.347	-13196.695	-13153.344	10149
p4442	PKA, pmek	805.020	-1604.040	-1582.365	10149
pakts473	p4442, PKA, PIP3	5414.014	-10816.027	-10772.677	10149

(a) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ - pooled

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
plcg	PIP3	102.407	-202.814	-197.879	8, 9	1028
PIP2	plcg, PIP3	2194.22	-4382.44	-4361.214	1-4, 6-11, 13-14	8738
PKC		0	0	0	9	155
PKA	PKC	282.082	-562.164	-554.955	1-8, 10-14	9994
P38	PKC, PKA	4623.152	-9240.304	-9218.628	1-14	10149
pjnk	PKC, PKA	1476.262	-2946.524	-2924.848	1-14	10149
praf	PKC, PKA	13.554	-21.108	-6.302	8, 9	1028
pmek	PKC, praf, PKA	143.342	-274.683	-245.071	8, 9	1028
p4442	PKA, pmek	140.876	-275.751	-260.945	8, 9	1028
pakts473	p4442, PKA, PIP3	3498.433	-6984.866	-6943.922	1-2, 4-6, 8-9, 11-13	6795

(b) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ - causal

Variable	Pair copula	Tau
plcg	plcg, PIP3	0.072
PIP2	PIP2, plcg	0.351
	plcg, PIP3	0.072
	PIP2, PIP3; plcg	0.289
PKC	PKC, plcg	0.202
	plcg, PIP2	0.351
	PKC, PIP2; plcg	0.087
PKA	PKA, PKC	-0.145
P38	P38, PKC	0.542
	PKC, PKA	-0.145
	P38, PKA; PKC	-0.102
pjnk	pjnk, PKC	0.301
	PKC, PKA	-0.145
	pjnk, PKA; PKC	-0.048
praf	praf, PKA	-0.180
	PKA, PKC	-0.145
	praf, PKC; PKA	0.082
pmek	pmek, praf	0.613
	praf, PKC	0.117
	PKC, PKA	-0.145
	pmek, PKC; praf	0.093
	praf, PKA; PKC	-0.160
	pmek, PKA; praf, PKC	-0.023
p4442	p4442, PKA	0.124
	PKA, pmek	-0.169
	p4442, pmek; PKA	0.138
pakts473	pakts473, p4442	0.585
	p4442, PKA	0.124
	PKA, PIP3	0.027
	pakts473, PKA; p4442	-0.070
	p4442, PIP3; PKA	-0.041
	pakts473, PIP3; p4442, PKA	-0.014

(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ - pooled

Variable	Pair copula	Tau
plcg	plcg, PIP3	0.279
PIP2	PIP2, plcg	0.318
	plcg, PIP3	0.036
	PIP2, PIP3; plcg	0.305
PKC		
PKA	PKA, PKC	-0.150
P38	P38, PKC	0.542
	PKC, PKA	-0.145
	P38, PKA; PKC	-0.102
pjnk	pjnk, PKC	0.301
	PKC, PKA	-0.145
	pjnk, PKA; PKC	-0.048
praf	praf, PKC	-0.039
	PKC, PKA	0.091
	praf, PKA; PKC	-0.029
pmek	pmek, PKC	0.215
	PKC, praf	-0.039
	praf, PKA	-0.035
	pmek, praf; PKC	0.206
	PKC, PKA; praf	0.089
	pmek, PKA; PKC, praf	0.109
p4442	p4442, PKA	0.302
	PKA, pmek	0.118
	p4442, pmek; PKA	0.040
pakts473	pakts473, p4442	0.570
	p4442, PKA	0.050
	PKA, PIP3	0.032
	pakts473, PKA; p4442	-0.154
	p4442, PIP3; PKA	-0.040
	pakts473, PIP3; p4442, PKA	0.018

(d) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ - causal

Table 6.26: Overview of the pooled and causal $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$-models.

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Sample size
plcg	PIP3	81.774	-161.549	-154.324	10149
PIP2	plcg, PIP3	4264.177	-8518.354	-8482.228	10149
PKC	plcg	1141.210	-2280.420	-2273.195	10149
PKA	PKC	865.729	-1729.458	-1722.232	10149
P38	PKC, PKA	6794.476	-13578.951	-13542.826	10149
pjnk	PKC, PKA	2744.762	-5479.523	-5443.397	10149
praf	PKA, PKC	1722.171	-3436.341	-3407.441	10149
pmek	praf, PKC, PKA	9782.278	-19544.555	-19472.304	10149
p4442	PKA, pmek	1641.930	-3273.861	-3237.735	10149
pakts473	p4442, PKA, PIP3	6729.569	-13437.137	-13357.661	10149

(a) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ - pooled

Variable	D-vine order	Copula loglik	Copula AIC	Copula BIC	Estimated with data from experiments	Sample size
plcg	PIP3	219.967	-435.934	-426.063	8, 9	1028
PIP2	plcg, PIP3	3758.478	-7508.957	-7480.655	1-4, 6-11, 13-14	8738
PKC	PIP2	1.311	-0.622	2.421	9	155
PKA	PKC	870.333	-1738.666	-1731.456	1-8, 10-14	9994
P38	PKC, PKA	6794.476	-13578.951	-13542.826	1-14	10149
pjnk	PKC, PKA	2744.762	-5479.523	-5443.397	1-14	10149
praf	PKC, PKA	25.073	-42.146	-22.404	8, 9	1028
pmek	praf, PKC, PKA	313.688	-609.376	-564.958	8, 9	1028
p4442	PKA, pmek	225.232	-440.464	-415.787	8, 9	1028
pakts473	p4442, PKA, PIP3	4686.650	-9355.299	-9293.884	1-2, 4-6, 8-9, 11-13	6795

(b) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ - causal

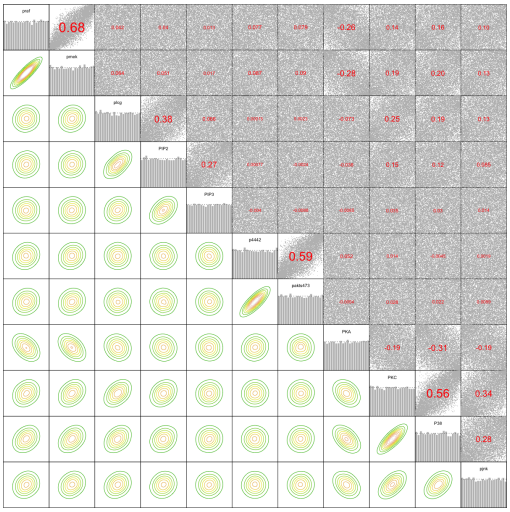
Variable	Pair copula	Family	Tau
plcg	plcg, PIP3	clayton	0.066
PIP2	PIP2, plcg	bb7	0.362
	plcg, PIP3	clayton	0.066
	PIP2, PIP3; plcg	bb8	0.327
PKC	PKC, plcg	joe	0.216
PKA	PKA, PKC	joe	-0.160
P38	P38, PKC	bb8	0.581
	PKC, PKA	joe	-0.160
	P38, PKA; PKC	bb8	-0.069
pjnk	pjnk, PKC	bb7	0.316
	PKC, PKA	joe	-0.160
	pjnk, PKA; PKC	t	0.013
praf	praf, PKA	bb8	-0.178
	PKA, PKC	joe	-0.160
	praf, PKC; PKA	joe	0.078
pmek	pmek, praf	bb8	0.662
	praf, PKC	bb8	0.160
	PKC, PKA	joe	-0.160
	pmek, PKC; praf	bb7	0.099
	praf, PKA; PKC	bb8	-0.118
	pmek, PKA; praf, PKC	joe	-0.040
p4442	p4442, PKA	t	0.157
	PKA, pmek	joe	-0.175
	p4442, pmek; PKA	bb8	0.112
pakts473	pakts473, p4442	bb8	0.633
	p4442, PKA	t	0.157
	PKA, PIP3	clayton	0.032
	pakts473, PKA; p4442	t	-0.045
	p4442, PIP3; PKA	bb8	-0.032
	pakts473, PIP3; p4442, PKA	bb8	-0.028

(c) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ - pooled

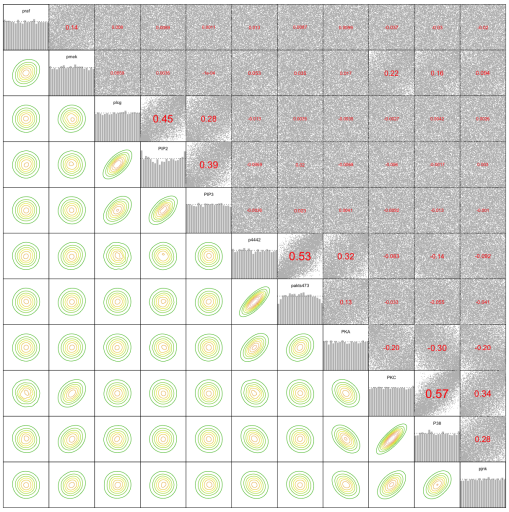
Variable	Pair copula	Family	Tau
plcg	plcg, PIP3	bb8	0.334
PIP2	PIP2, plcg	joe	0.356
	plcg, PIP3	clayton	0.043
	PIP2, PIP3; plcg	bb8	0.354
PKC	PKC, PIP2	joe	-0.082
PKA	PKA, PKC	joe	-0.163
P38	P38, PKC	bb8	0.581
	PKC, PKA	joe	-0.160
	P38, PKA; PKC	bb8	-0.069
pjnk	pjnk, PKC	bb7	0.316
	PKC, PKA	joe	-0.160
	pjnk, PKA; PKC	t	0.013
praf	praf, PKC	clayton	-0.046
	PKC, PKA	bb8	0.104
	praf, PKA; PKC	frank	-0.034
pmek	pmek, praf	joe	0.295
	praf, PKC	clayton	-0.046
	PKC, PKA	bb8	0.104
	pmek, PKC; praf	bb8	0.199
	praf, PKA; PKC	frank	-0.034
	pmek, PKA; praf, PKC	bb8	0.084
p4442	p4442, PKA	bb7	0.314
	PKA, pmek	bb8	0.131
	p4442, pmek; PKA	frank	0.069
pakts473	pakts473, p4442	bb6	0.596
	p4442, PKA	t	0.098
	PKA, PIP3	clayton	0.027
	pakts473, PKA; p4442	bb8	-0.183
	p4442, PIP3; PKA	clayton	-0.034
	pakts473, PIP3; p4442, PKA	clayton	0.015

(d) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ - causal

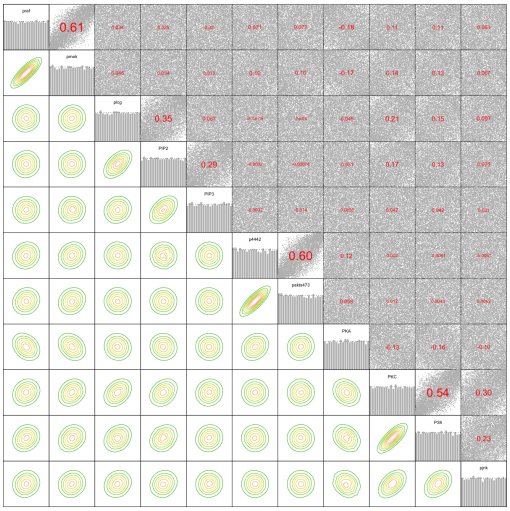
Table 6.27: Overview of the pooled and causal $\mathbf{M}_{Par}\mathbf{C}_{Par}$-models.



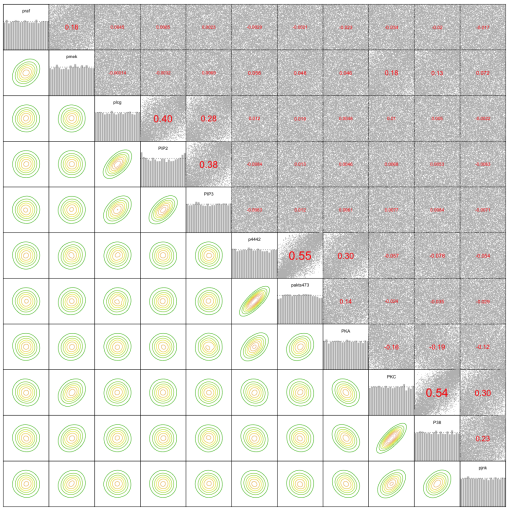
(a) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ - pooled



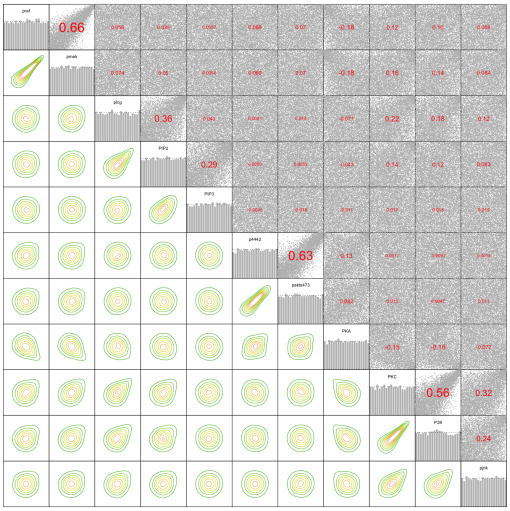
(b) $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ - causal



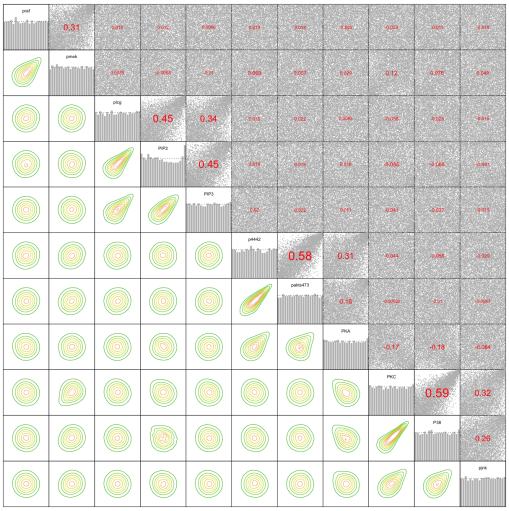
(c) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ - pooled



(d) $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$ - causal



(e) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ - pooled



(f) $\mathbf{M}_{Par}\mathbf{C}_{Par}$ - causal

Figure 6.5: Normalized contours and pairs plots as well as Kendall's τ for the simulated data of the different models on u-scale.

6.5 Quantile sampling

In this section, we want to find out, which impact the choice of the margins and the feasible copula families has on the tails. Therefore, we apply the following steps: First we generate data for a variable $U^{(1)}$ without parent nodes by sampling it uniform (i). Then we fix first the values of the 0.05- and later the 0.95-quantile (ii). Then we sample the child nodes of $U^{(1)}$ given the 0.05- or 0.95-quantiles (iii). For this child node then fix again the value of the 0.05- and later the 0.95-quantile (iv), and sample its child node again only given the corresponding quantiles. We repeat this procedure (v) until we sample the last node of the DAG (vi)(which has no childnodes).

(i) Sample the first node $U^{(1)} \sim uniform(0, 1)$

(ii) Fix quantiles $\hat{u}^{(1)}_{0.05} := q_{0.05}(U^{(1)})$ and $\hat{u}^{(1)}_{0.95} := q_{0.95}(U^{(1)})$

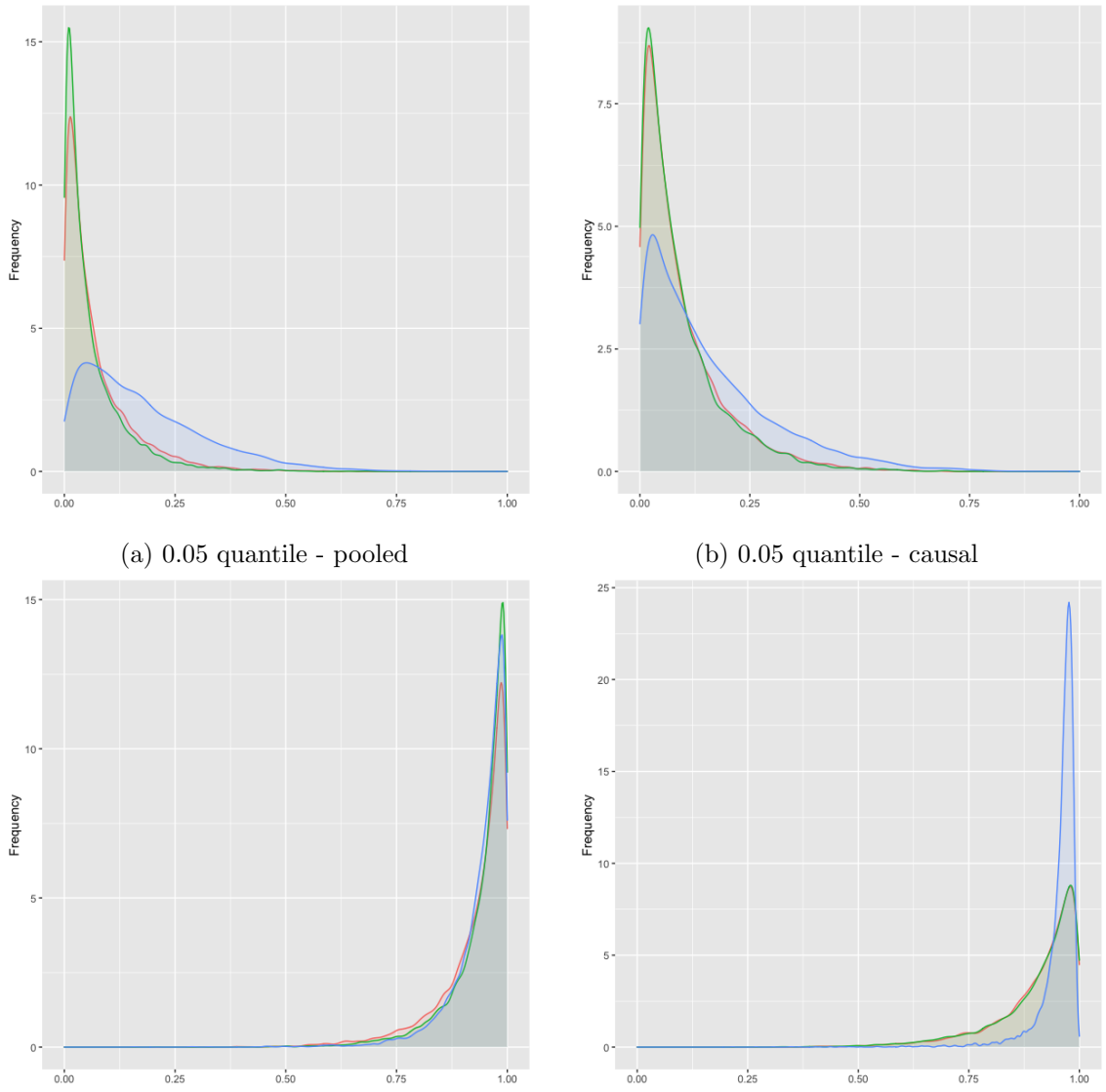
(iii) Sample child node $U^{(2)}_{0.05}|U^{(1)} = \hat{u}^{(1)}_{0.05}$ and $U^{(2)}_{0.95}|U^{(1)} = \hat{u}^{(1)}_{0.95}$

(iv) Fix quantiles $\hat{u}^{(2)}_{0.05} := q_{0.05}(U^{(2)}_{0.05})$ and $\hat{u}^{(2)}_{0.95} := q_{0.95}(U^{(2)}_{0.95})$

(v) ....

(vi) Sample child node $U^{(d)}_{0.05}|U^{(1)} = \hat{u}^{(1)}_{0.05}, ..., U^{(d-1)} = \hat{u}^{(d-1)}_{0.05}$
and $U^{(d)}_{0.95}|U^{(1)} = \hat{u}^{(1)}_{0.95}, ..., U^{(d-1)} = \hat{u}^{(d-1)}_{0.95}$

By appling this method to every node of a DAG model, we generate data that can be helpfull to understand how the tails are modeled. In Figures 6.6 to 6.9, we see histograms of the variable `pakts473` (which is the last node in the topological order of the consent graph). The results shown in Figures 6.6 to 6.9 are generated by applying the method described above to the D-vine regression models discussed in the previous chapters. The most important and obvious result we see in the figures is that the choice of pair copulas has a large impact on the tails. It does matter for the tails with which pair copulas the D-vine regression models are fitted.



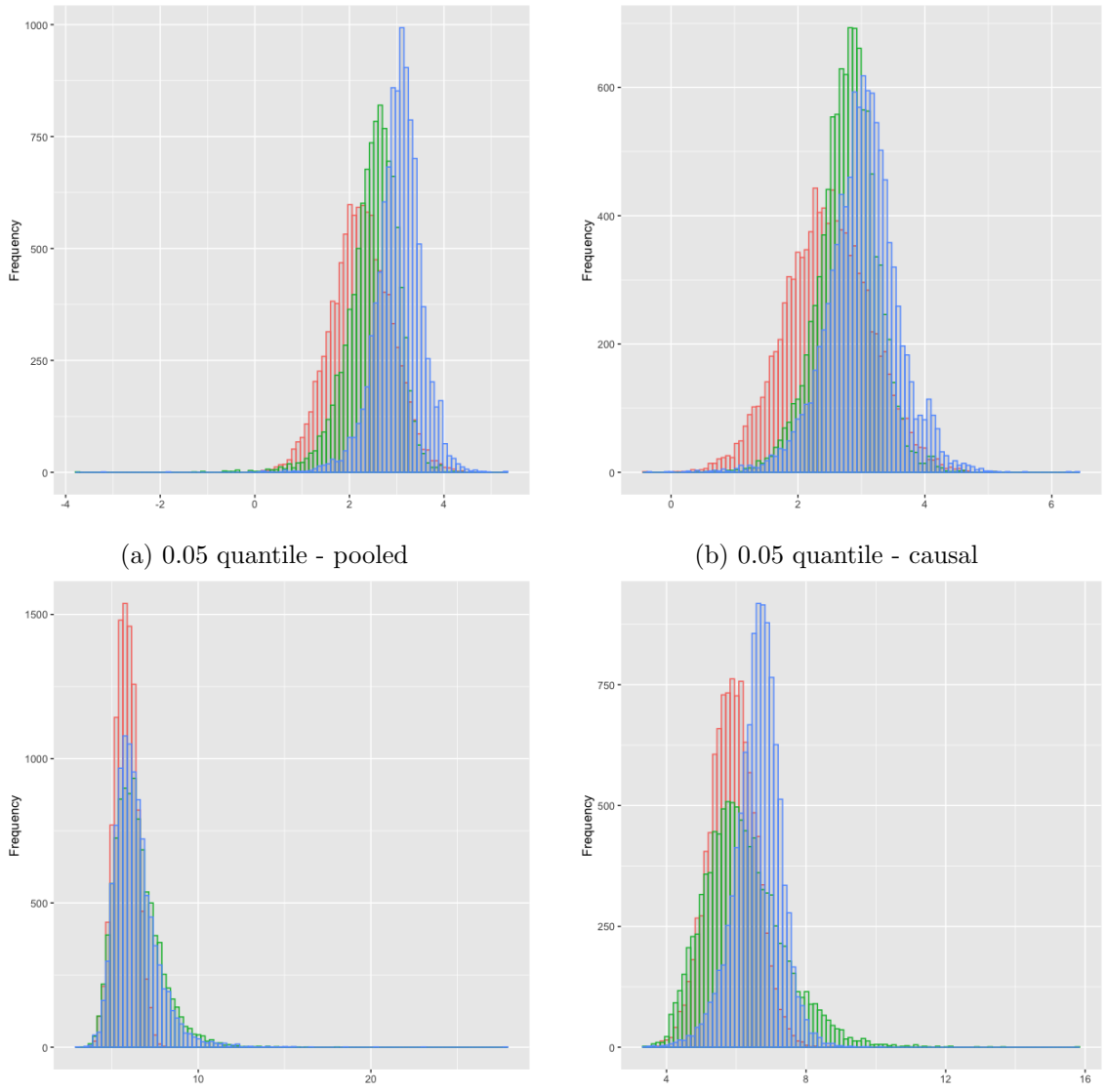
(a) 0.05 quantile - pooled

(b) 0.05 quantile - causal

(c) 0.95 quantile - pooled

(d) 0.95 quantile - causal

Figure 6.6: Empirical density plots for the sampling of 10149 points each of variable pakts473 on u-scale. For the sampling different models were used: Red: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$; Green: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$; Blue: $\mathbf{M}_{Par}\mathbf{C}_{Par}$.



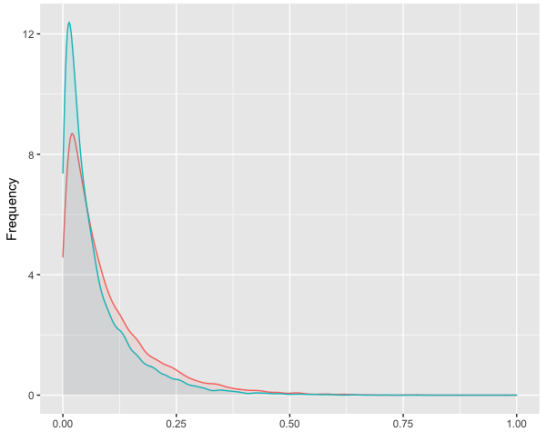
(a) 0.05 quantile - pooled

(b) 0.05 quantile - causal

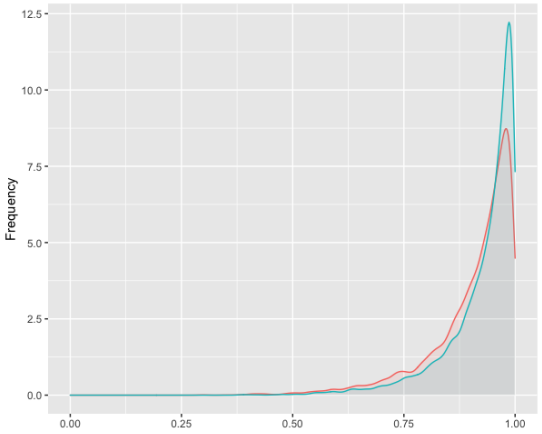
(c) 0.95 quantile - pooled

(d) 0.95 quantile - causal

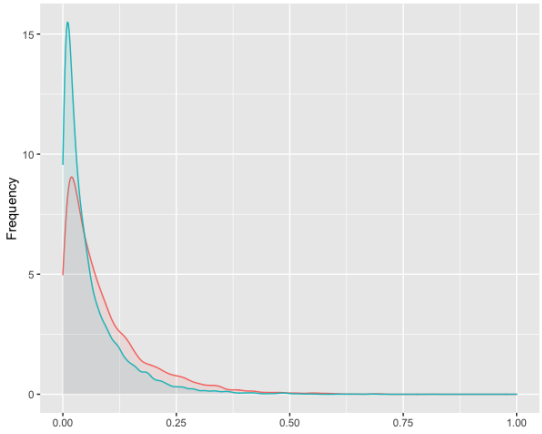
Figure 6.7: Histograms for the sampling of 10149 points each of variable pakts473 on x-scale. For the transformation the respective distributions from Tables 4.1 and 4.2 as well as 6.19 (a) and 6.20 (a) were used. For the sampling different models were used: Red: $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$; Green: $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$; Blue: $\mathbf{M}_{Par}\mathbf{C}_{Par}$.



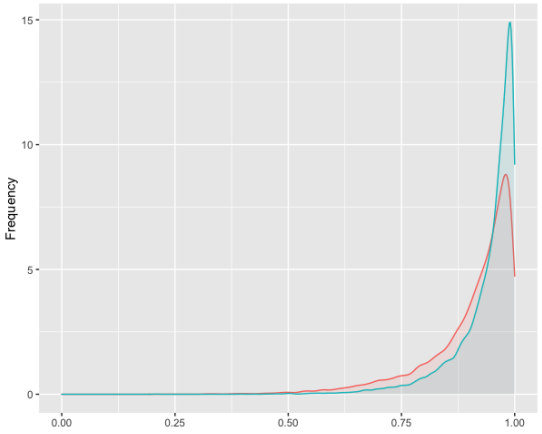
(a) 0.05 quantile - $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$



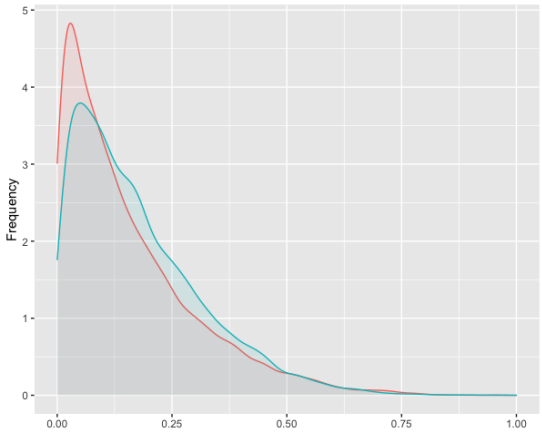
(b) 0.95 quantile - $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$



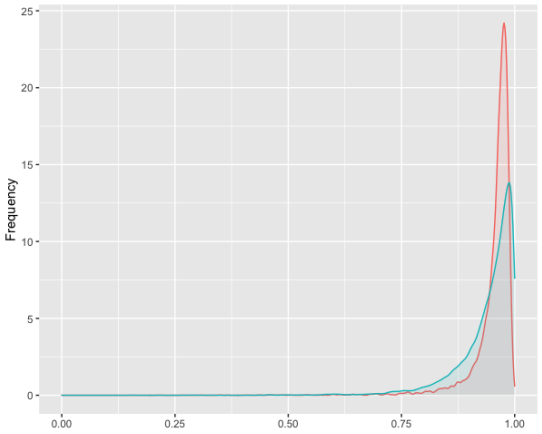
(c) 0.05 quantile - $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$



(d) 0.95 quantile - $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$

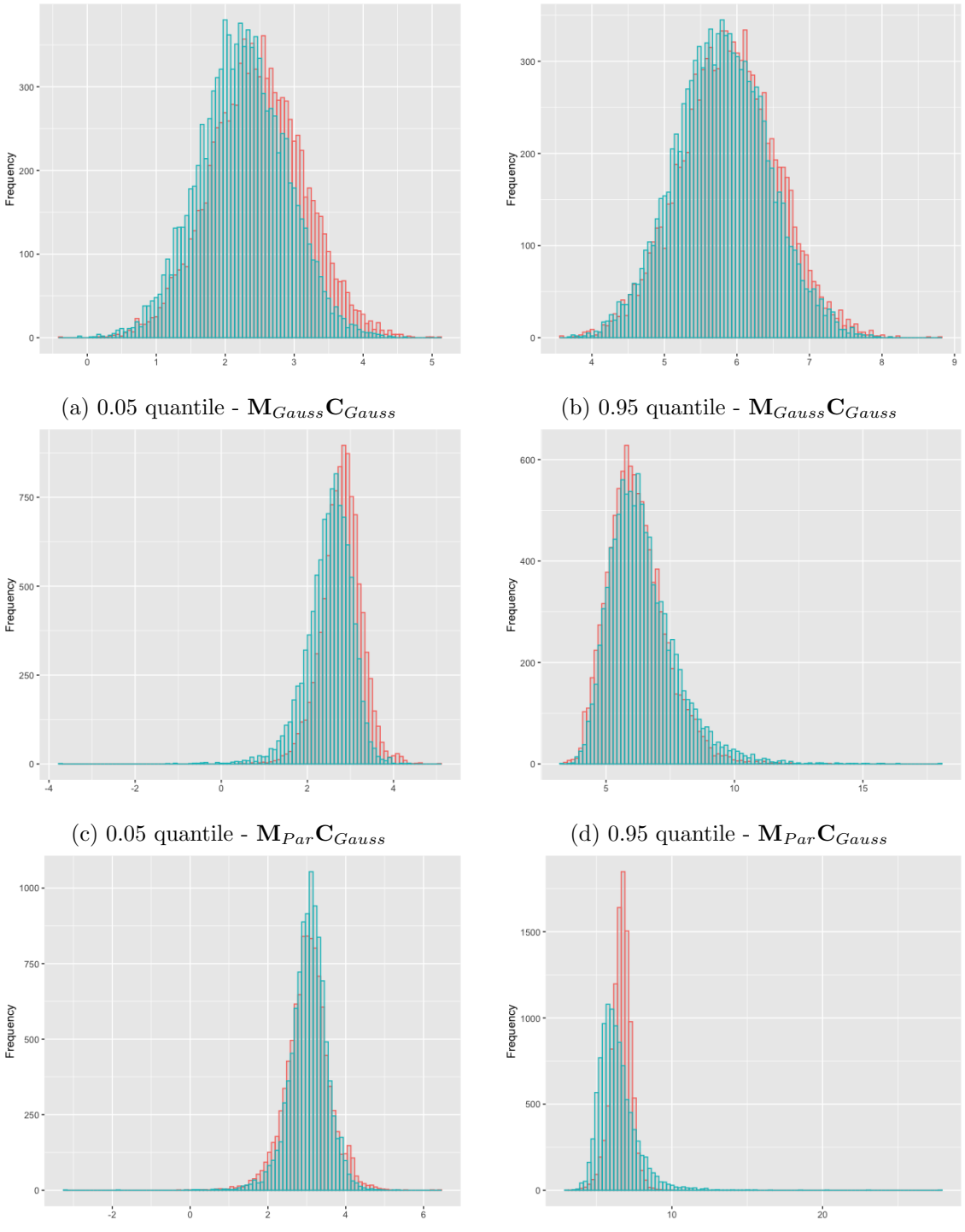


(e) 0.05 quantile - $\mathbf{M}_{Par}\mathbf{C}_{Par}$



(f) 0.95 quantile - $\mathbf{M}_{Par}\mathbf{C}_{Par}$

Figure 6.8: Empirical density plots for the sampling of 10149 points each of variable pakts473 on u-scale. The models were fitted on different data sets: Red: causal; Green: pooled.



(a) 0.05 quantile - $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$

(b) 0.95 quantile - $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$

(c) 0.05 quantile - $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$

(d) 0.95 quantile - $\mathbf{M}_{Par}\mathbf{C}_{Gauss}$

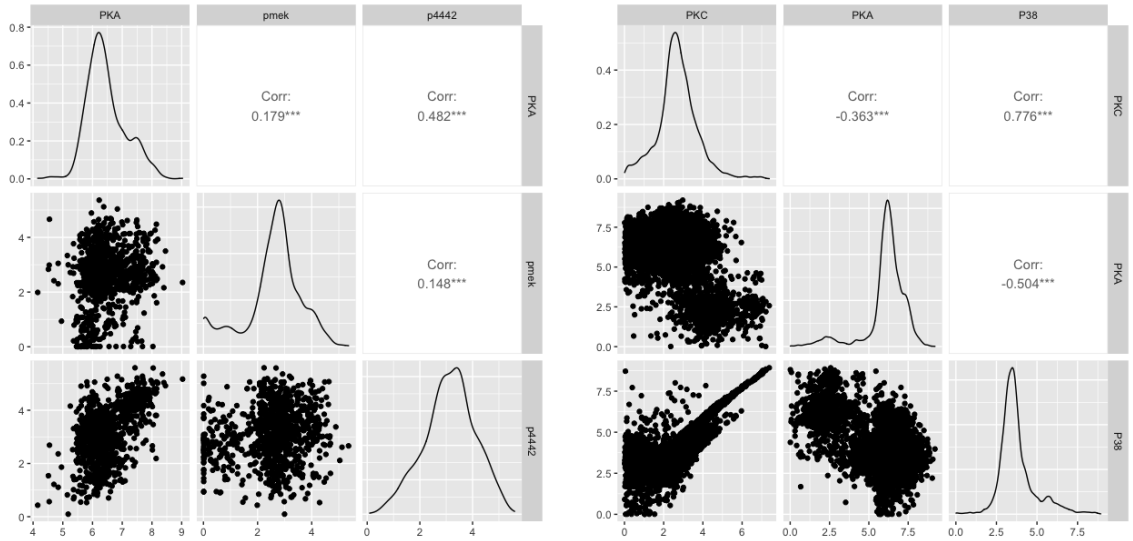
(e) 0.05 quantile - $\mathbf{M}_{Par}\mathbf{C}_{Par}$

(f) 0.95 quantile - $\mathbf{M}_{Par}\mathbf{C}_{Par}$

Figure 6.9: Histograms for the sampling of 10149 points each of variable pakts473 on x-scale. For the transformation the respective distributions from Tables 4.1 and 4.2 as well as 6.19 (a) and 6.20 (a) were used. The models were fitted on different data sets: Red: causal; Green: pooled.

6.6 VCMM clustering on causal data

In the previous sections, we used the causal data to fit D-vine regression models. Now we want to understand the causal datasets better and analyze if there are substructures within these. Examplary we can see the causal datasets for the variables `p4442` and `P38` on x-scale in Figure 6.10. While the causal dataset of `p4442` consists of the experiments 8 and 9 and has a sample size of 1028, the variable `P38` has not been perturbed in any experiment, i.e. the causal dataset for `P38` has 10149 observations from experiments 1-14. Figure 6.10 supports the assumption that there may be substructures in the causal data. Therefore we are now working with the vine copula mixture models again. The algorithm from the library **vineclust** by Sahin (2021) allows us to specify a vine structure to be used in all clusters.



(a) p4442 (b) P38

Figure 6.10: Causal datasets of p4442 and P38 on x-scale. The sample sizes are 1028 and 10149.

From the consent graph we get 1-3 parent nodes for each variable (except `PIP3`), therefore the causal datasets contain 2-4 variables each. For a two variable dataset there is only one possible vine structure. For a three variable dataset, there are three possible vine structures, which are all D-vines, but of which one is problematic: As shown in Kraus and Czado (2017) the corresponding conditional distribution function for a D-vine structure can be expressed in a closed form if and only if the response Y is a leaf node. This means, if the child node is connected to the other two variables by more than one arc, then numeric integration is necessary. We have the same problem for the four variable datasets, as there are 12 D-vines and 12 C-vines possible, of which for only 6 D-vines and

6 C-vines the corresponding conditional distribution function has a closed form solution as explained by Tepegjozova et al. (2021).

For that reason we are focussing here on the two examples of the variables `p4442` and `P38`. We fit VCMMs to the datasets, while we fix the vine structures to the respective two structures that do not require numeric integration. In Figures 6.11 and 6.12 the BIC values of different VCMMs are plotted, depending on their number of components as well as their D-vine structures, which have fixed fixed.

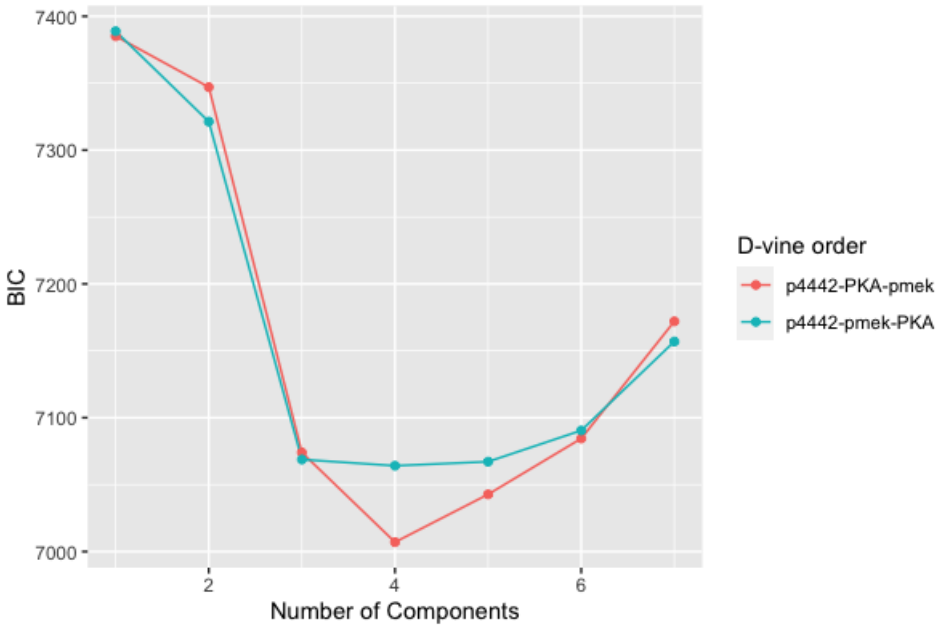


Figure 6.11: BIC values of vine copula mixture models for variable p4442 with different D-vine structures and numbers of components.

It is clear, that for the variable `p4442` the VCMM with four components and the D-vine structure `p4442-PKA-pmek` has the best BIC and for the variable `P38` the VCMM with six components and the D-vine structure `P38-PKC-PKA` has the best BIC.Since our simulation setup in Chapter 5.2.2 and e.g. Figure 5.6, which looks rather similar to Figure 6.11, we assume that this estimation of the number of components based on the BIC is reliable.

It should be mentioned here, that besides the D-vine regression model $\mathbf{M}_{Gauss}\mathbf{C}_{Gauss}$ - pooled, in every of the previously fitted D-vine models the structure `p4442-PKA-pmek` was chosen in the node `p4442`. The same holds for the structure `P38-PKC-PKA`, which was chosen in every previously fitted D-vine models for the node `P38`.

Since the D-vine structures are fixed, it makes no sense to plot the vine structures. Instead, we give an overview of the copulas used in the model with `p4442-PKA-pmek` and

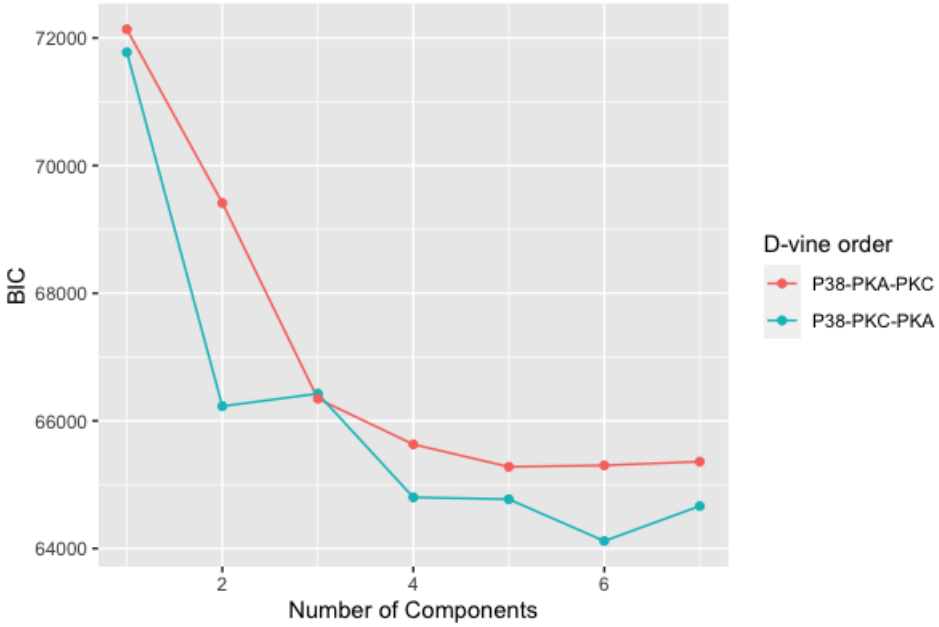


Figure 6.12: BIC values of vine copula mixture models for variable P38 with different D-vine structures and numbers of components.

four clusters in Table 6.28, and analogously for the model with `P38-PKC-PKA` and six clusters in Table 6.29.

Cluster (Weight)	Variables	Family	Rotation	Parameters	Tau
1	p4442, PKA	Joe	180	1.6	0.251
(0.218)	PKA, pmek	Frank	0	-0.792	**-0.087**
	p4442, pmek; PKA	Frank	0	-0.437	-0.048
2	p4442, PKA	Joe	0	1.688	**0.277**
(0.165)	PKA, pmek	Frank	0	0.903	0.099
	p4442, pmek; PKA	Clayton	90	-0.143	-0.067
3	p4442, PKA	Frank	0	-1.814	**-0.195**
(0.404)	PKA, pmek	Gaussian	0	0.109	0.069
	p4442, pmek; PKA	Clayton	180	0.03	0.015
4	p4442, PKA	BB8	0	1.608, 0.965	0.217
(0.213)	PKA, pmek	Gumbel	180	1.082	0.076
	p4442, pmek; PKA	Joe	270	-1.057	-0.032

Table 6.28: Structure of the VCMM to the dataset in which p4442 has not been perturbed. The vine structure of p4442-PKA-pmek was fixed.

Cluster (Weight)	Variables	Family	Rotation	Parameters	Tau
1	P38, PKC	BB1	0	0.467, 1.207	0.328
(0.161)	PKC, PKA	Clayton	180	0.086	0.041
	P38, PKA; PKC	Clayton	180	0.03	0.015
2	P38, PKC	Gumbel	0	14.59	0.931
(0.069)	PKC, PKA	Joe	180	1.048	0.027
	P38, PKA; PKC	Joe	180	1.025	0.014
3	P38, PKC	BB8	0	1.34, 0.999	0.16
(0.114)	PKC, PKA	Gaussian	0	0.019	0.012
	P38, PKA; PKC	Gaussian	0	-0.019	-0.012
4	P38, PKC	BB8	0	3.098, 0.991	0.521
(0.259)	PKC, PKA	Joe	180	1.109	0.059
	P38, PKA; PKC	Frank	0	0.569	0.063
5	P38, PKC	BB8	180	2.444, 0.794	0.266
(0.338)	PKC, PKA	Gaussian	0	-0.068	-0.043
	P38, PKA; PKC	Clayton	270	-0.084	-0.04
6	P38, PKC	BB6	0	2.48, 2.862	0.806
(0.058)	PKC, PKA	Joe	180	1.024	0.014
	P38, PKA; PKC	Joe	90	-1.025	-0.014

Table 6.29: Structure of the VCMM to the dataset in which P38 has not been perturbed. The vine structure of P38-PKC-PKA was fixed.

Chapter 7

Conclusion

In this master thesis, we approached the investigation of the Sachs dataset from different directions. First of all, we started with an analysis of the full pooled data. After the multivariate Gaussian approach, which did not sufficiently fit the features of the data, we started with a vine copula model and D-vine regression models. For them we allowed Gaussian and non-Gaussian marginal distributions and different parametric copula family sets. When sampling data, the vine copula model approach seemed to perform best. But this was also due to greater flexibility, as we already specified the graph for all D-vine regression models.

Pooled data: In the clustering chapter, we always used different tools, such as information criteria in Chapters 5.1.1, 5.2.1 and 5.2.2 and methods from test theory in Chapters 5.1.1 and 5.2.3, to find the best number of components. The analysis showed that the Gaussian mixture model approach did not fit the data properly, while the vine copula mixture models performed much better due to their flexibility. With the VCMMs, we were able to separate the observations of certain experiments very precisely from those of other experiments. With the VCMMs it was possible to find known substructures in the data. Since the VCMMs were the special focus of this thesis, we analyzed not only the Sachs data with the VCMMs, but also certain properties of the VCMMs themselves: For example, we investigate the VCMMs with respect to the interpretability of mixture weights. Also we studied in the simulation setup how reliable the known information criteria are for VCMMs.

Causal data: In the last chapter we fitted causal models. Again we allowed different marginal distributions and copulas, while we were espicially working with D-vine regression models. This allowed us to build models, of which we can expect, that all external influences are removed. In this chapter we also showed that the choice of copula families plays a major role. The $\mathbf{M}_{Par}\mathbf{C}_{Par}$-models had lower nodewise AIC values then the

$\mathbf{M}_{Par}\mathbf{C}_{Gauss}$-models, which had lower AIC values then the only Gaussian models. Already the fitted copula families in the $\mathbf{M}_{Par}\mathbf{C}_{Par}$-models, where we set the fewest restrictions beforehand, show that tail dependence is important for this data set. Therefore, we finally sampled data specifically in the 0.05 and 0.95 quantiles, where it became apparent that the tails were modeled different by the models with only Gaussian copula families. One question that could not be definitely clarified in this master's thesis is how homogeneous the causal datasets themselves are. For example, in the last chapter, applying VCMMs to those showed that substructures may exist here as well. For example, the analysis with VCMMs reliably suggested a division of the causal dataset of variable p4442 into four components.

Bibliography

[1] Akaike, H.: *Information Theory and an Extension of the Likelihood Ratio Principle*. Proceedings of the Second International Symposium of Information Theory,1973.

[2] Azzalini, A., Capitanio, A.: *The Skew-Normal and Related Families*. Cambridge University Press, 2013.

[3] Bedford, T., Cooke, R. M.: *Vines: A new graphical model for dependent random variables*. Annals of Statistics, 30(4). 2002.

[4] Bevacqua, E., Maraun, D., Haff, I., Widmann, M., Vra, M.: *Multivariate statistical modelling of compound events via pair-copula constructions: analysis of floods in Ravenna (Italy)*. Hydrology and Earth System Sciences, 21, 2701–2723, 2017.

[5] Biernacki, C., Celeux, G., Govaert, G.:*Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7), 2000.

[6] Brechmann, E. C.: *Truncated and simplified regular vines and their applications*. Technische Universität München, Department of Mathematics, 2010.

[7] Celeaux, G., Govaert, G.: *Gaussian parsimonious clustering models*. Pattern Recognition, Vol. 28, No. 5, 1995.

[8] Chang, B., Joe, H.: *Prediction based on conditional distributions of vine copulas*. Computational Statistics and Data Analysis, Vol. 139, pp. 45–63, 2019.

[9] Czado, C., Schmidt, T.: *Mathematische Statistik*. Springer, 2011.

[10] Czado, C.: *Analyzing Dependent Data with Vine Copulas*. Springer, 2019.

[11] Czado, C., Scharl, S.: *Analysis of an Intervential Protein Experiment Using a Vine Copula Based Structural Equation Models*. 2021.

[12] Delignette-Muller, M. L., Dutang, C.: *Fitdistrplus. An R Package for Fitting Distributions*. Journal of Statistical Software, 64(4), 2015.

[13] Dose, K.: *Biochemie: Eine Einführung*. Springer, 1996.

[14] Fernández, C., Steel, M. F. J.: *On Bayesian Modelling of Fat Tails and Skewness*. Econometrics, CentER Discussion Paper, Vol. 1996-58, pp. 3-8, 1996.

[15] Garcia C.: *A Simple Procedure for the Comparison of Covariance Matrices*. BMC Evolutionary Biology, 2012.

[16] Hothorn, T.: *mvtnorm: Multivariate Normal and t Distributions*. 2014.

[17] Kendall, M.G.: *A new measure of rank correlation*. Biometrika, Vol. 30, Issue 1-2, pp. 81–93, 1938.

[18] Koller, D., Friedman, N.: *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[19] Kraus, D., Czado C.: *D-vine copula based quantile regression* . Computational Statistics and Data Analysis, 110, 2017.

[20] Kullback, S., Leibler, R. A..: *On Information and Sufficiency*. Annals of Mathematical Statistics, 22, 79 - 86, 1951.

[21] McLachlan G.T.: *On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture*. Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 36, No. 3, 1987.

[22] Nagarajan, R., Scutari, M., Lebre, S.: *Bayesian Networks in R with Applications in Systems Biology*. Springer, 2013.

[23] Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Erhardt, T.: *VineCopula. Statistical Inference of Vine Copulas*. 2019.

[24] Nagler, T., Kraus, D.: *vinereg: D-Vine Quantile Regression*. 2021.

[25] Peters, J., Janzing, D., Schölkopf, B.: *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

[26] R Core Team: *R. A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2021.

[27] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., Nolan, G.: *Causal Protein- Signaling Networks Derived from Multiparameter Single-Cell Data*. Sciencemag 308, 2005.

[28] Sahin, Ö., Czado C.: *Vine copula mixture models and clustering for non-Gaussian data*. Econometrics and Statistics, 2021.

[29] Sahin, Ö.: *vineclust*. 2021.

[30] Sahin, Ö.: *Statistical Analysis of the Number of Sales of a Pharmacy Product*. Technische Universität München, Department of Mathematics, 2019.

[31] Scharl S.: *D-Vine Regression Based Bayesian Networks Applied to the Sachs Dataset*. Technische Universität München, Department of Mathematics, 2021.

[32] Schwarz, G.: *Estimating the Dimension of a Model*. The Annals of Statistics, 6, 461–464, 1978.

[33] Scrucca, L., Fop, M., Murphy, T. B., Raftery, A. E.: *Mclust 5. Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models.*. R Journal, 8, 2016.

[34] Sklar, A.: *Fonctions de Repartition a n Dimensions et Leurs Marges*. Publications de l'Institut Statistique de l'Universite de Paris, 8, pp. 229–23, 1959.

[35] Tepegjozova, M., Zhou, J., Claeskens, G., Czado, C.: *Nonparametric C- and D-vine based quantile regression*. Technische Universität München, Department of Mathematics, 2021.

[36] Vuong, Q.: *Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses*. Econometrica Vol. 57, No. 2 (Mar., 1989), pp. 307-333, 1989.

[37] Wilks, S.: *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*. Annals of Mathematical Statistics 9 (1) 60 - 62, 1938.

[38] Zhang, Q., Shi, X.: *A Mixture Copula Bayesian Network Model for Multimodal Genomic Data*. Cancer Informatics 16, 2017.