Technische Universität München

Zentrum Mathematik

Modeling dependence among meteorological measurements and tree ring data

Diplomarbeit

von

Michael Pachali

Themenstellerin:	Prof. Claudia Czado, Ph.D.
Betreuer:	Eike Brechmann
	Dr. Christian Zang
Abgabetermin:	27. September 2012

Hiermit erkläre ich, dass ich die Diplomarbeit selbstständig angefertigt und nur die angegebenen Quellen verwendet habe.

Garching, den 27. September 2012

Acknowledgments

The author gratefully acknowledges help, support and motivation by numerous people throughout this thesis project.

First and foremost my deepest thanks go to Prof. Claudia Czado, Ph.D., for her excellent supervision, inspiring input, motivation and valuable support in the last months.

In the same way I would like to express my sincere gratitude to Eike Brechmann for his excellent supervision, profound knowledge, invaluable support and his availability at any time.

Furthermore, my deep thanks go to Dr. Christian Zang for his supervision, motivation and excellent cooperation, especially in questions of climate and dendrological issues.

Without our frequent team meetings with fruitful discussions that helped me to develop a deep understanding of the subject as well as without all the guidance and help that I received from you three, this thesis would not have been possible.

In addition, I want to thank Prof. Dr. Annette Menzel for her cooperation and valuable suggestions as well as Dr. Christoph Dittmar, environmental research and education in Mistelbach, for providing the tree ring data. My thanks are also due to the Deutsche Wetterdienst (DWD) for providing the meteorological data from Hohenpeissenberg.

Finally, I want to thank my parents for their continuous and loving support during my years of study, and all of my friends around me who made those years wonderful.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

This thesis is divided into three parts, in which we

1. model dependencies among meteorological variables by *regular vines* (R-vines) and
2. model tree ring data by *linear mixed models* and
3. summarize the main points of Part I and II and give an outlook for future research.

Thus our work applies to the field of climate research and dendrology from a statistical point of view with a clear focus on the statistical methodology. Each part is considered separately.

Our first part is based on R-vines which present a recent field of research. Vines denote a flexible class of modeling high-dimensional dependencies, introduced by Bedford and Cooke [2001, 2002] and discussed in Kurowicka and Cooke [2006], which use only bivariate copulas as building blocks. The special cases of canonical vines (C-vines) and D-vines were considered by Aas et al. [2009] to derive multivariate copulas using pair copula decomposition. However, the more general class of R-vines is less restrictive in the dependency structure on the data and present a suitable graphical model to describe pair copula constructions. Recently, Dißmann et al. [2011] developed a sequential, heuristic method based on graph theory to fit an appropriate R-vine copula specification to a given dataset. We will use this method to model the dependence structure among meteorological variables in our first part.

While Dißmann et al. [2011] or Brechmann and Czado [2012] apply R-vines to financial data, modeling the multivariate distribution of meteorological quantities is also in the focus of interest. Whether for forecasting, pricing of weather derivatives or other studies in climate science or hydrology, statistical information about dependencies between different weather variables are needed. For example, Möller et al. [2012] propose a method for post-processing an ensemble of multivariate forecasts in order to obtain a joint predictive distribution of weather by using Bayesian model averaging and copulas. Copulas serve also as a base for the study of Bokusheva [2010] modeling the dependence structure between weather-based crop insurance respectively weather derivatives yields and weather variables. Further works with application in the field of hydrology, e.g. by Genest and Favre [2007] or Kao and Govindaraju [2010], are also using copula approaches to model multivariate dependencies among droughts and meteorological quantities. Genest and Favre [2007] even indicated that application of copulas in hydrology is still in its nascent stages and their full potential for analyzing hydrologic problems is yet to be realized [Kao and Govindaraju, 2010, p. 122]. Therefore our flexible and less restrictive R-vine approach based on pair copula constructions to model high-dimensional dependencies among meteorological measurements seems consequential.

Our model will be built on data coming from the meteorological observatory in Hohenpeissenberg which is located about 80km southwest of Munich, southern Germany, in the foothills of the Alps at an altitude of 1000m above sea level. In detail we consider observations of six variables, namely daily mean, minimum and maximum temperature as well as daily mean humdity, daily mean air pressure and daily total amount of precipitation from a time span of 1950-2009. We divide these sixty years into 12 subperiods of five years each and fit appropriate R-vine models to data from three of them (i.e. from periods 1955-1959, 1980-1984 and 2005-2009). In order to do this, based on the Theorem of Sklar [1959] and the theory of pair copula constructions, we need to model the marginal distributions of our variables first. In case of the temperature variables and daily mean air pressure we were inspired by Campbell and Diebold [2005] to capture autoregression and seasonal effects by linear regressions. The detection of slightly skewed distributed residuals is modeled by skew normal and skew t distributions respectively (according to Azzalini [1985] and Azzalini and Capitanio [2003]) and the results are compared over time. Further, we assume that daily mean humidity is beta distributed and hence we will use a beta regression model (introduced by Ferrari and Cribari-Neto [2004]) in order to capture autoregression and seasonal effects too. Modeling the behavior of daily total precipitation is a bit more demanding since the variable often takes values equal to zero. Therefore we will concentrate on the two stage chain-dependent-process model of Stern and Coe [1984], using binomial and gamma regressions to model the daily rainfall amount based on our data.

The distribution of daily precipitation is not continuous, since it has an additional point mass at zero. So in this case our classical six-dimensional R-vine approach is not straightforward anymore. We will handle this scenario by extending the model of Erhardt and Czado [2012] from insurance data to our application. The idea in doing so is simple, namely modeling the dependence structure of the first five variables without precipitation by a five-dimensional R-vine in the classical way and we connect the variable of positive precipitation amount (modeled in the course of the above mentioned two step regression) to the established vine to get a six-dimensional R-vine copula specification on rain days. Simulations from the whole model are then straightforward using the modeled rain success probability "day d has rain", also fitted in the course of the marginal distribution model of precipitation.

We compare the resulting R-vine structures for the three periods, i.e which dependencies are modeled among the variables and especially whether any changes in the common dependence structures are detectable over these time spans. We end this part with simulations from our models and compare several simulated parameters such as, e.g., the dependence measure Kendall's tau for different pairs of variables with the empirical ones.

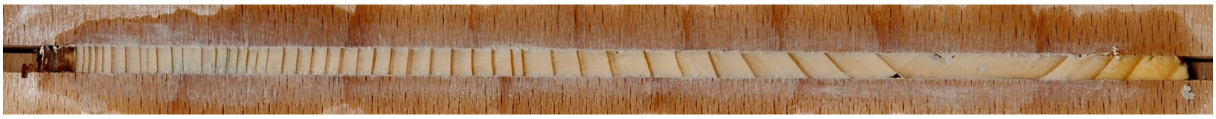


Figure 1.1: Drill core from a tree containing several year rings. © Christoph Dittmar

In the second part of the thesis, we consider tree ring width data of two tree species, namely Norway spruce (*Picea abies* (L.) Karst.) and silver fir (*Abies alba* Mill.), also measured in the region of Hohenpeissenberg. An example of a drill core from a tree containing a number of year rings is illustrated in Figure 1.1. In our seasonal climate, both species produce one tree ring per year, but however, the tree ring growth cannot proceed faster than permitted by the most limiting factor, according to Liebig's law of the minimum (Fritts [1976]). Thus, the tree-growth is limited by environmental factors and series of tree rings can be seen as archives of a tree's reaction to these factors [Zang, 2010, p. 7]. While lower treeline and forest border sites are most sensitive to precipitation, latitudinal and altitudinal treeline sites are most sensitive to temperature variations (see, e.g., in Zang [2010]). However, our observed trees are located on neither treeline or border. Thus, the limiting factors are not uniform and easy to identify across the investigated populations and correlations between growth and climate may be complicated to interpret (Friedrichs et al. [2008] and Leal et al. [2008]). Zang [2010] mentions that the greatest methodological challenge of his work is the extraction of quantititative information about limiting factors of tree growth. He presents several approaches such as dendroclimatic calibration (multiple linear relationships between climate and tree growth) or multivariate benchmarking using dendroclimatic archetypes. We present a further statistical methodology by linear mixed models to examine the relationships between climate and tree growth.

In our analysis we consider several seasonal means of meteorological quantities which we already used in the first part. In detail we calculate seasonal means´(the seasons are selected based on previous dendrological studies) of daily mean temperature, air pressure, humidity and total precipitation. In addition, we calculate the highest numbers of consecutive days without rain (longest dry periods/droughts) of the selected seasons to model the (multiple) linear relationships between climate occurrences and tree ring widths. Linear mixed models (LMMs) provide a flexible analytic tool to model these kinds of clustered longitudinal data in case of tree rings, in which the residuals are normally distributed but may not be independent or not have a constant variance in contrast to linear models. A LMM may include fixed-effect parameters associated with one or more covariates, such as the meteorological quantities in our case, *and* random effects associated with one or more random factors (like e.g. a randomly sampled specific tree effects). In order to do not distort the pure influence of different meteorological variables on our dependent year ring variable we will remove any non-climatic variance out of our series (detrending by splines) first. Appropriate random effects are then selected based on likelihood ratio tests while the fixed effects are chosen by the smallest AIC (due to Akaike [1973]) of the different considered models. We compare the results for both tree species, i.e. the modeled relationships between the climate variables and tree growth as well as the behavior of the modeled variances of standardized tree ring widths compared to the empirical counterparts and

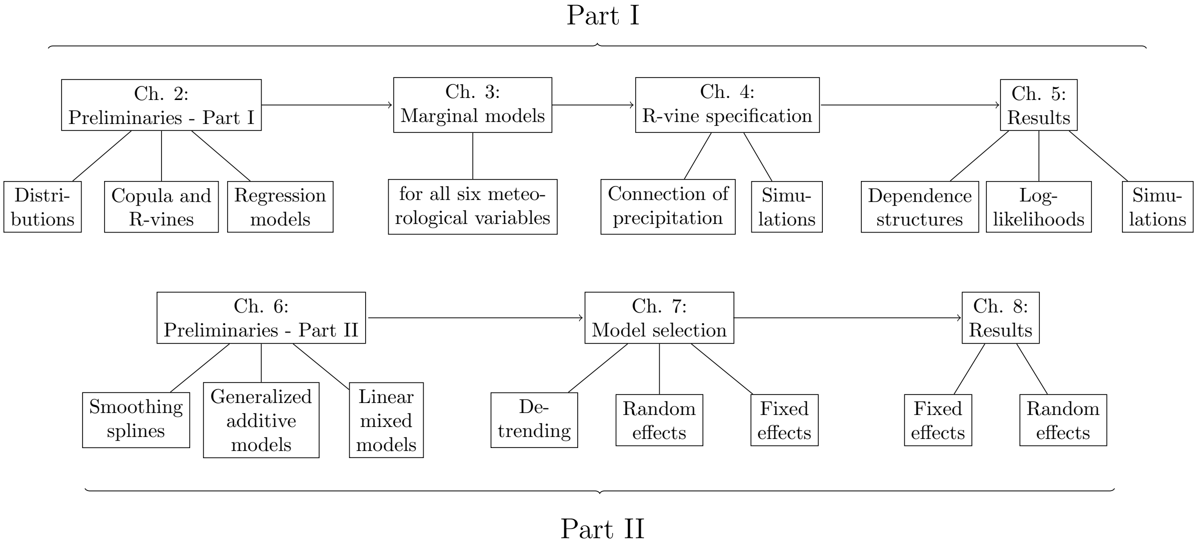


Figure 1.2: Path through the thesis.

whether it is influenced by climate occurrences.

Thus, the thesis is organized as follows. Part I starts with chapter 2 in which we present basic concepts that are needed throughout the first part of this work, namely several distributions with their properties as well as results from the theory of copulas, dependence measures and an introduction into R-vines based on graph theoretical concepts. A method how to select an appropriate R-vine copula specification is also given as well as several regression models for the marginal behavior modeling of the considered meteorological variables. Subsequently, these (six) marginal models are set up in Chapter 3 case by case with corresponding results, compared over the periods. In chapter 4, we specify our R-vine model with an emphasis on how to include the variable of precipitation into our modeling. It is presented by an algorithm and afterwards we outline how to simulate from our model. Chapter 5 summerizes the results of the three fitted R-vine models for each period and compares the modeled dependence structures and their corresponding log-likelihood values. Simulations from the models complete the first part.

Chapter 6 introduces Part II with basic concepts that are needed throughout the second part of the thesis such as smoothing spline interpolation or generalized additive models which will be needed to detrend the year ring data in an initial step. In addition, linear mixed models are subsequently defined together with methods for parameter estimation and model selection. This will be used in Chapter 7, because after detrending the raw ring width series, appropriate random and fixed effects will be selected to yield the final models for both considered tree species. The results of the modeled relationships between tree growth and climate variables as well as the behavior of the modeled variances of the detrended ring series are offered in Chapter 8, which complete the second part.

The thesis closes with a summary of the main points and an outlook for future research for each Part I and II respectively in Chapter 9 and 10. It completes the final third part of this work. The described structures of Part I and II are still summerized in Figure 1.2.

Part I

Modeling dependencies among meteorological variables

Chapter 2

Preliminaries - Part I

In this chapter we give a brief introduction to the basic concepts that are needed throughout this part and moreover throughout this thesis. In detail we start to introduce several distributions with their properties, which we will need for our work. In a second step, we introduce copulas as basis for regular vines, since the joint distribution function of a random vector can be represented by a copula (theorem of Sklar, 1959). In this context we present the important classes of elliptical and Archimedean copulas, give an overview of bivariate copula families which we will meet again during this work and show their relationships to the common dependence measures Kendall's τ and tail dependence. Since elliptical and Archimedean copula approaches for modeling multivariate dependence are quite restrictive, bivariate copula resp. pair copula constructions offer a more flexible and intuitive way of modeling multivariate distributions [Aas et al., 2009, p. 183]. However, the number of possible pair copula constructions is large, therefore they need to be classfied using regular vines which are based on graph theoretical concepts (Dißmann et al. [2011]). Nevertheless, beforehand, we need to model the marginal distributions of our wheather variables. Therefore we introduce several regression models to capture seasonality and autoregression for our variables later. In the end, we would like to test the residuals out of regression models to be independent in form of a Ljung-Box test.

2.1 Distributions

To model the multivariate depence among wheather variables, we will meet a number of different distributions to fit the marginal behavior of each variable. Additionally, we will be confronted with some basic distributions in the section of copulas. Therefore, we will introduce the range of them, their properties and behavior. We mainly follow Czado and Schmidt [2011] and Georgii [2007] for the most familiar distributions.

2.1.1 Continuous uniform distribution

The *continuous uniform distribution* is a continuous distribution defined on an interval $[a, b]$, $a < b$.

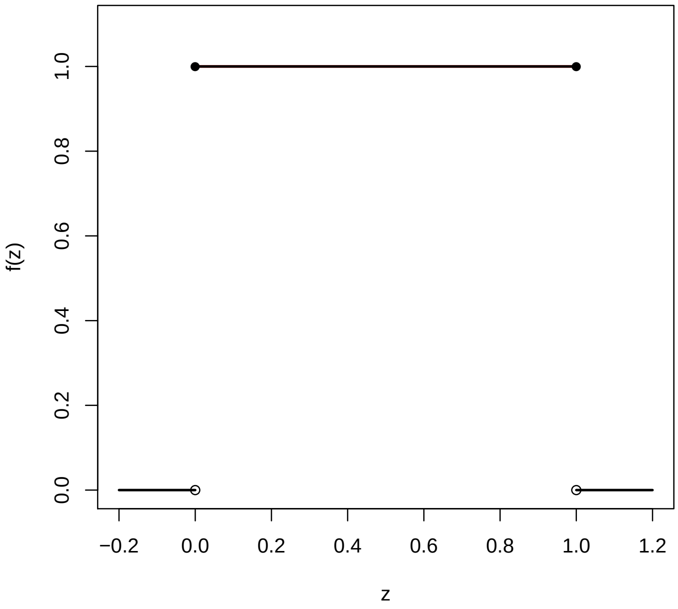


Figure 2.1: Density function of the standard uniform distribution.

Definition 2.1 (Uniform distribution.) *We denote $Z \sim U(a, b)$ to be uniform distributed, if it has the following density function*

$$f(z) = \frac{1}{b-a} 1_{[a,b]}(z) = \begin{cases} \frac{1}{b-a}, & \textit{for } a \leq z \leq b, \\ 0, & \textit{for } z < a \textit{ and } z > b. \end{cases} \tag{2.1}$$

The corresponding cumulative distribution function (cdf) is given by

$$P(Z \leq z) = F(z) = \begin{cases} 0, & \text{for } z < a, \\ \frac{x-a}{b-a}, & \text{for } a \leq z < b, \\ 1, & \text{for } z \geq b. \end{cases} \tag{2.2}$$

It holds, that

$$E[Z] = \frac{a+b}{2} \text{ and } Var(Z) = \frac{(b-a)^2}{12}.$$

In the special case $Z \sim U(0, 1)$, we say Z is *standard uniform distributed* with corresponding density function $f(z) = 1_{[0,1]}(z)$ and $P(Z \leq z) = z$.

2.1.2 Normal distribution

The most important distribution in statistics is of course the *normal (or Gaussian) distribution.* It is defined as follows:

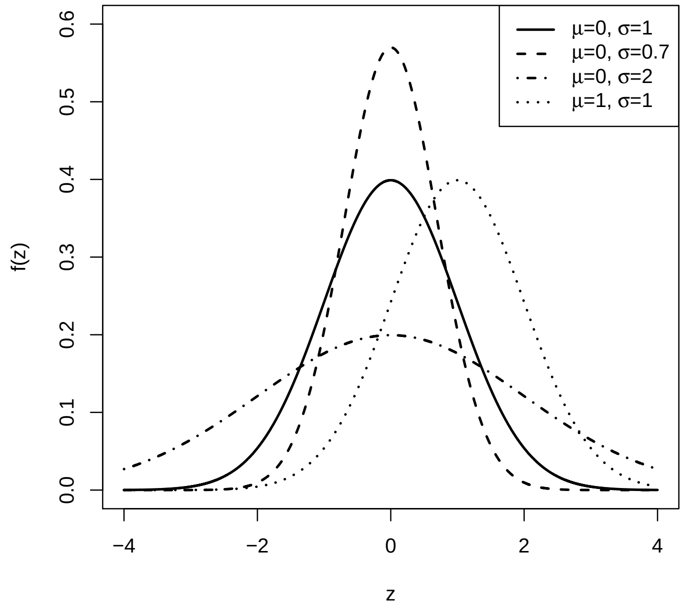


Figure 2.2: Density function of the normal distribution for different parameters. The solid line corresponds to the standard normal distribution

Definition 2.2 (Normal distribution.) *We say $Z \sim \mathcal{N}(\mu, \sigma^2)$ is normal (or Gaussian) distributed, if $\mu \in \mathbb{R}$, $\sigma > 0$ and the density function of Z is given by*

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}. \tag{2.3}$$

The normal cdf follows

$$P(Z \leq z) = F(z) = \int\limits_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \, dt, \tag{2.4}$$

respectively and it holds, that

$$E[Z] = \mu \text{ and } Var(Z) = \sigma^2. \tag{2.5}$$

The distribution with $\mu = 0$ and $\sigma = 1$ is called the *standard normal distribution* and one denotes $\phi(z) := \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ for the density and

$$\Phi(z) := \int_{-\infty}^{z} \phi(t) \, dt \tag{2.6}$$

for the cdf of a standard normal distribution.

In addition, the normal distribution is symmetric around μ, i.e. $P(Z - \mu \geq z) = P(Z - \mu \leq -z)$. Its skewness, defined as $\gamma_1 := E\left[\left(\frac{Z-\mu}{\mu}\right)^3\right]$, is equal to 0.

2.1.3 Multivariate normal distribution

The classical approach to model multivariate dependence among variables is to assume that they are multivariate normal (or Gaussian) distributed. Therefore we introduce this kind of distribution in the following:

Definition 2.3 (Multivariate normal distribution.) *An n-dimensional random vector $\boldsymbol{Z} = (Z_1, ..., Z_n)^{'}$ is called multivariate (or Gaussian) normal distributed of dimension n, if there exist $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^{'} \in \mathbb{R}^n$ and $L \in \mathbb{R}^{n \times m}$ with rank$(L) = m$, so that*

$$\boldsymbol{Z} = L\boldsymbol{X} + \boldsymbol{\mu}, \tag{2.7}$$

where $\boldsymbol{X} = (X_1, ..., X_m)^{'}$ and X_i are iid with $X_i \sim \mathcal{N}(0, 1)$.

In this case, we say $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$, where $\Sigma = LL^{'}$. If $n = m$, we say, that $\boldsymbol{Z}$ follows a *non-singular* normal distribution. In contrast, $\boldsymbol{Z}$ follows a *singular* normal distribution in case of $n > m$.

The matrix $\Sigma \in \mathbb{R}^{n \times n}$ denotes the corresponding *variance-covariance matrix* of $\boldsymbol{Z}$, that defines the dependence among the compenents of the random vector. It contains the entries

$$\Sigma_{ij} = Cov(Z_i, Z_j), \ 1 \leq i, j \leq n.$$

It holds, that

$$E[\boldsymbol{Z}] = \boldsymbol{\mu} \text{ and } Var(\boldsymbol{Z}) = \Sigma, \tag{2.8}$$

and the density function of $\boldsymbol{Z}$ is defined by

$$f(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu})^{'}\Sigma^{-1}(\boldsymbol{z} - \boldsymbol{\mu})\right), \tag{2.9}$$

where $\boldsymbol{z} = (z_1, ..., z_n)^{'} \in \mathbb{R}^n$ and $|\Sigma|$ is the determinant of Σ. The corresponding distribution function will be denoted by Φ_Σ.

Further properties are:

(i) The multivariate normal distribution is stable under linear transformations, i.e. if $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{b} \in \mathbb{R}^n$ and $C \in \mathbb{R}^{k \times n}$, it follows that

$$C\boldsymbol{Z} + \boldsymbol{b} \sim \mathcal{N}_k\left(C\boldsymbol{\mu} + \boldsymbol{b}, C\Sigma C^{'}\right).$$

(ii) The components of a multivariate normal distribution are normal distributed, i.e. if $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$, it holds that $Z_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$, for $i = 1, ..., n$.

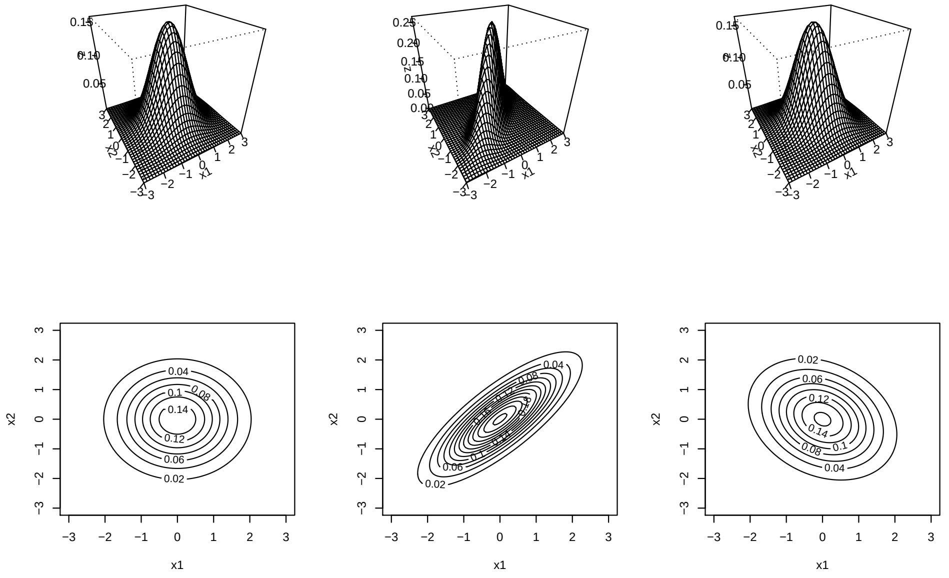


Figure 2.3: Joint density of a bivariate standard normal distribution with different correlation parameters and their corresponding contour plots. The left panel corresponds to a standard bivariate normal distribution with correlation $\rho = 0$ among both variables, the middle panel refers to a correlation $\rho = 0.8$ and the right panel based on $\rho = -0.25$.

(iii) If $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ and $\Sigma_{ij} = 0$, for $i, j \in \{1, ..., n\}$ and $i \neq j$, it follows that Z_i and Z_j are independent. Hence, if Σ is diagonal then $Z_1, ..., Z_n$ are independent.

(iv) The multivariate normal distribution belongs to the class of elliptical distributions. These will be defined in Section 2.18.

2.1.4 Skew normal distribution

In Chapter 3 we will see that some of the wheather variables are skewed in contrast to the assumption of the normal distribution. The *skewed normal distribution* offers an appropriate extension to fit the data well. Thereby, the normal distribution will be a special case of it. The following definition and listed properties are based on Azzalini [1985], Azzalini and Dalla Valle [1996] and Azzalini and Capitanio [1999].

Definition 2.4 (Skew normal distribution.) *A random variable Z is said to be skew normal distributed, if it has the density function*

$$f(z) = \frac{2}{\omega}\phi\left(\frac{z-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{z-\xi}{\omega}\right)\right) = \frac{1}{\omega\pi}e^{-\frac{(z-\xi^2)}{2\omega^2}}\int_{-\infty}^{\alpha\left(\frac{z-\xi}{\omega}\right)} e^{\frac{t^2}{2}}\,dt, \qquad (2.10)$$

where $\phi(z)$ is the density function and $\Phi(z)$ is the cdf of the standard normal distribution (cp. (2.6)), respectively. $\xi \in \mathbb{R}$ is called location parameter, $\omega > 0$ stands for the scale

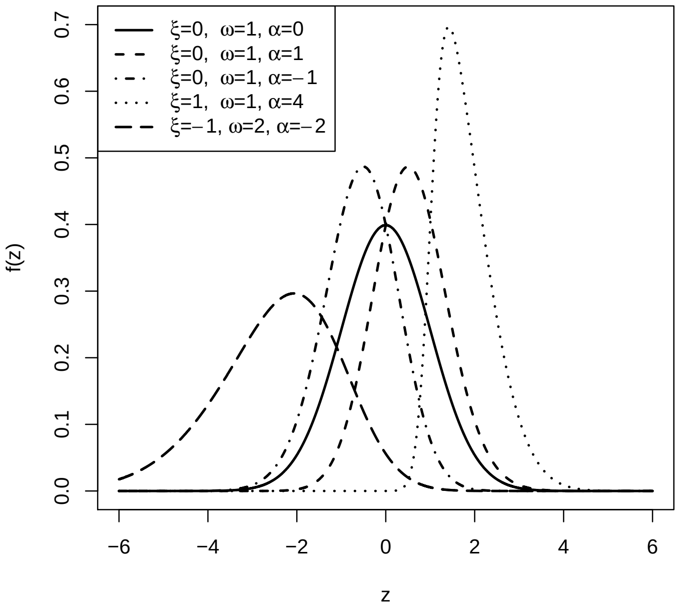


Figure 2.4: Density function of the skew normal distribution for different parameters. The solid line corresponds to the standard normal distribution.

parameter and $\alpha \in \mathbb{R}$ corresponds to the shape paramter of a skew normal distribution. Then we say,

$$Z \sim \mathcal{SN}(\xi, \omega, \alpha).$$

The corresponding cdf results in

$$P(Z \leq z) = F(z) = \Phi\left(\frac{z-\xi}{\omega}\right) - 2T\left(\frac{z-\xi}{\omega}, \alpha\right), \tag{2.11}$$

where $T(h, a)$ is *Owen's T function*[1]. It is defined as follows

$$T(h, a) := \frac{1}{2\pi}\int_0^a \frac{e^{-\frac{1}{2}h^2(1+x^2)}}{1+x^2}\,dx, \ -\infty < h, a < \infty.$$

Then, mean and variance are given by

$$E[Z] \ = \ \xi + \omega\delta\sqrt{\frac{2}{\pi}}, \text{ where } \delta := \frac{\alpha}{\sqrt{1+\alpha^2}}, \tag{2.12}$$

$$Var(Z) \ = \ \omega^2\left(1 - \frac{2\delta^2}{\pi}\right). \tag{2.13}$$

[1] It is named after Donald Bruce Owen, who introduced the formula in 1956. The function $T(h, a)$ gives the probability of the event ($X > h$ and $0 < Y < aX$), where X and Y are independent standard normal random variables [Azzalini, 1985, p. 173].

In case of a skew normal distribution, the special interest is focused on the skewness, which can be calculated as

$$\gamma_1 := E\left[\left(\frac{Z - E[Z]}{\sqrt{Var(Z)}}\right)^3\right] = \frac{4-\pi}{2}\frac{(\delta\sqrt{2/\pi})^3}{(1-2\delta^2/\pi)^{3/2}} \in (-1, 1). \tag{2.14}$$

It can be shown, that the skewness γ_1 is driven by the shape parameter α, in a way, that the absolute value of the skewness γ_1 increases as the absolute value of α increases. Moreover, if $\alpha > 0$, then the distribution is right skewed (positively skewed) and it is left skewed (negatively skewed) if $\alpha < 0$.

As mentioned above, if $\alpha = 0$, it follows that $T\left(\frac{z-\xi}{\omega}, \alpha\right) = 0$ as well as $\delta = 0$ and the distribution of Z results in a normal distribution, i.e $Z \sim \mathcal{N}(\xi, \omega^2)$.
Some further properties are:

(i) The density of $\mathcal{SN}(0, 1, 0)$ is the density of the standard normal distribution $\mathcal{N}(0, 1)$.

(ii) If $Z \sim \mathcal{SN}(0, 1, \alpha)$, then $-Z \sim \mathcal{SN}(0, 1, -\alpha)$.

(iii) If $Z \sim \mathcal{SN}(0, 1, \alpha)$, it follows that $1 - P(Z \leq -z) = P((-Z) \leq z)$.

(iv) The cdf of $Z \sim \mathcal{SN}(0, 1, 1)$ is $P(Z \leq z) = (\Phi(z))^2$.

(v) The skew normal distribution can be extended to the multivariate case. For further details we refer to Azzalini and Dalla Valle [1996].

2.1.5 t-distribution

In contrast to the normal distribution, some data may be better described by using a heavy-tailed distribution such as the *(Student's) t-distribution* (e.g. when outliers are expected). However, it is related to the normal distribution, i.e. it is symmetric and belongs to the class of elliptical distributions as well. Its definition is given by:

Definition 2.5 (t-distribution.) *A random variable Z follows a t-distribution with $\nu > 0$ degrees of freedom, if its density function corresponds to*

$$f(z) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)\Gamma(1/2)\sqrt{\nu}}\left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{2.15}$$

for all $z \in \mathbb{R}$, where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$ is the Gamma function. We say

$$Z \sim t_\nu.$$

It results that

$$\begin{aligned} E[Z] &= 0 \text{ for } \nu > 1, & (2.16)\\ Var(Z) &= \frac{\nu}{\nu-2} \text{ for } \nu > 2. & (2.17)\end{aligned}$$

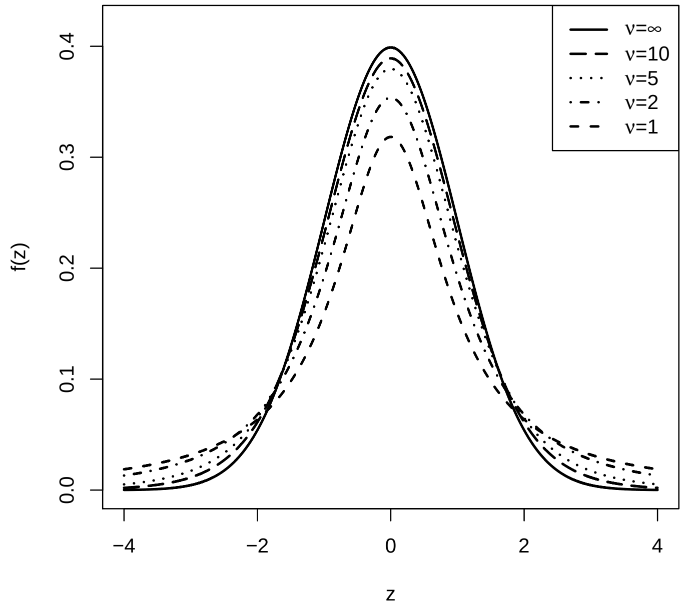


Figure 2.5: Density function of the t-distribution for different degrees of freedom ν. The solid line corresponds to the standard normal distribution.

Note, that $E[Z]$ and $Var(Z)$ do not exist for $\nu \leq 1$ and $Var(Z) = \infty$ for $1 < \nu \leq 2$. Similiar to the normal distribution, the skewness of the t-distribution is also equal to 0 (for $\nu > 3$, otherwise it is undefined).

The t-distribution becomes closer to the normal distribution as ν increases. As $\nu \to \infty$, Z approaches a standard normal distribution. Two further important properties are:

(i) The t-distribution with ν degrees of freedom can be defined as the distribution of the random variable Z with

$$Z = \frac{X}{\sqrt{V/\nu}}, \tag{2.18}$$

where X is standard normal distributed, V has a chi-squared distribution with ν degrees of freedom and X and V are independent.

(ii) For $\mu \in \mathbb{R}$, $(X + \mu)\sqrt{\frac{\nu}{V}}$, with X and V defined as in (i), follows a noncentral t-distribution with noncentrality parameter μ.

2.1.6 Multivariate t-distribution

To make things complete, we briefly introduce the *multivariate t-distribution*, since it can also model multivariate dependence with symmetric tail dependence, as we will see in Section 2.2.3. The definition is based on Demarta and McNeil [2005].

Definition 2.6 (Multivariate t-distribution.) *The n-dimensional random vector $\boldsymbol{Z} = (Z_1, ..., Z_n)'$ is said to have a (non-singular) multivariate t-distribution with $\nu > 0$ degrees of freedom, mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and positive-definite dispersion or scatter matrix $\Sigma \in \mathbb{R}^{n \times n}$, denoted $\boldsymbol{Z} \sim t_n(\nu, \boldsymbol{\mu}, \Sigma)$, if its density is given by*

$$f(\boldsymbol{z}) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma(\nu/2)\sqrt{(\pi\nu)^n \, |\Sigma|}} \left(1 + \frac{(\boldsymbol{z}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+n}{2}}, \tag{2.19}$$

where $\boldsymbol{z} = (z_1, ..., z_n)' \in \mathbb{R}^n$ and $|\Sigma|$ is the determinant of Σ.

Note, that the covariance matrix is not equal to Σ, but $Cov(\boldsymbol{Z}) = \frac{\nu}{\nu-2}\Sigma$ and it is only defined if $\nu > 2$. We refer to the *standard multivariate t-distribution*, if $\boldsymbol{\mu} = \boldsymbol{0}$.

2.1.7 Skew t-distribution

Similar to the skew normal distribution, since some wheather variables have corresponding characteristics, we are interested in a distribution that is related to the t-distribution but provide some skwewness in its density. The *skew t-distribution* offers this. Azzalini and Capitanio [2003] focused on this special case of skew elliptical densities.

If one wants to introduce an asymmetric variant of the t-distribution, the intuitive way is to replace the normal variate in (2.18) by a skew normal one. Hence, we denote a random variable Z to be skew t-distributed, if

$$Z = \xi + \frac{X}{\sqrt{V/\nu}}, \tag{2.20}$$

where $X \sim \mathcal{SN}(0, \omega, \alpha)$, i.e. X is skew normal distributed, thus has density (2.10) with $\xi = 0$. V has a chi-squared distribution with $\nu > 0$ degrees of freedom and X and V are independent. This yields to the following definition:

Definition 2.7 (Skew t-distribution.) *A random variable Z follows a skew t-distribution with $\nu > 0$ degrees of freedom, if it has the following density*

$$f(z) = 2t_\nu(z)T_{\nu+1}\left(\alpha\left(\frac{z-\xi}{\omega}\right)\left(\frac{\nu+1}{\left(\frac{z-\xi}{\omega}\right)^2 + \nu}\right)^{\frac{1}{2}}\right), \tag{2.21}$$

where t_ν is the density function of t-distribution with ν degrees of freedom (cp. (2.15)) and $T_{\nu+1}$ corresponds to the cdf of a t-distribution with $\nu + 1$ degrees of freedom. Similar to the skew normal distribution, $\xi \in \mathbb{R}$ denotes the location and $\omega > 0$ the scale parameter. Shape parameter $\alpha \in \mathbb{R}$ drives the skewness. Then, we say

$$Z \sim skewt(\xi, \omega, \alpha, \nu). \tag{2.22}$$

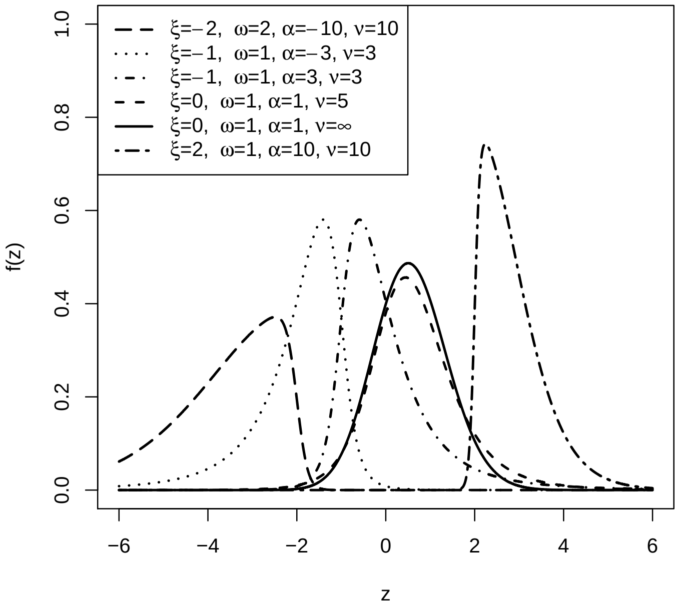


Figure 2.6: Density function of the skew t-distribution for different parameters. The solid line corresponds to a skew normal distribution with location parameter $\xi = 0$, scale parameter $\omega = 1$ and shape parameter $\alpha = 1$.

It can be shown, that the corresponding distribution function is given by

$$P(Z \leq z) = F(z) = E_{\frac{V}{\nu}}[F_X(z\sqrt{\nu})|V = \nu^2],$$

where X and V are the random variables from (2.20) and $F_X(z)$ denotes the corresponding cdf of the skew normal distribution of X.

Mean and variance are obtained by

$$E[Z] = \xi + \omega\delta(\nu/\pi)^{1/2}\frac{\Gamma\left(\frac{1}{2}(\nu-1)\right)}{\Gamma(\frac{1}{2}\nu)}, \text{ for } \nu > 1, \tag{2.23}$$

$$Var(Z) = \omega^2\frac{\nu}{\nu-2} - (E[Z])^2, \text{ for } \nu > 2, \tag{2.24}$$

where $\delta := \frac{\alpha}{\sqrt{1+\alpha^2}}$.

After some algebra, the index of skewness turns out to be

$$\gamma_1 = \mu\left[\frac{\nu(3-\delta^2)}{\nu-3} - \frac{3\nu}{\nu-2} + 2\mu^2\right]\left[\frac{\nu}{\nu-2} - \mu^2\right]^{-3/2}, \text{ for } \nu > 3, \tag{2.25}$$

where $\mu := \frac{E[Z]-\xi}{\omega}$.

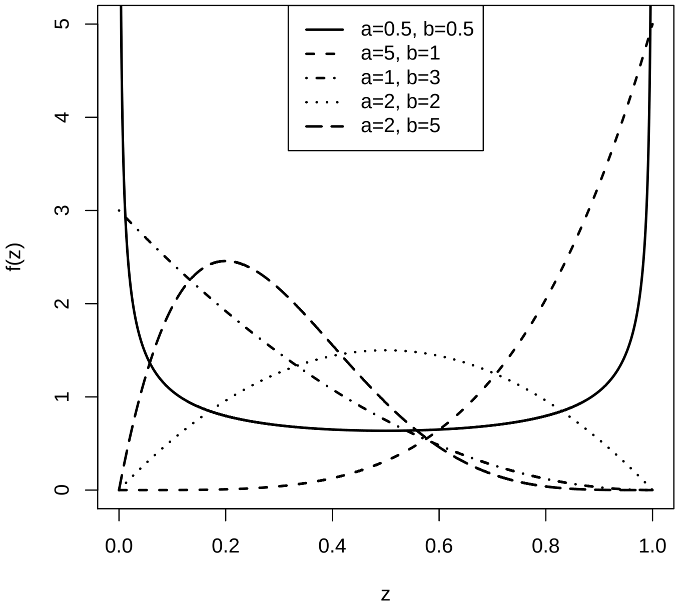


Figure 2.7: Density function of the beta distribution for different parameters.

The properties of the skew t-distribution are quite similar to the ones of the skew normal distribution, i.e. the skewness (2.25) is driven by shape parameter α. Hence, the absolute value of γ_1 increases as the absolute value of α increases. If $\alpha > 0$, then the distribution is right skewed (positively skewed) and it is left skewed (negatively skewed) if $\alpha < 0$. Further, if $\alpha = 0$, then Z follows a t-distribution with ν degrees of freedom, i.e. $Z \sim t_\nu$. Additionally, on can easily see in (2.21), that if $\nu \to \infty$, Z turns out to be skew normal distributed with the corresponding parameters.

Note, however, there are many ways to develop and define a skew t-distribution, so there exists no unique one. But Azzalini and Capitanio [2003] clarify that all ways lead to Equation (2.21) to be the density of a skew t-distribution.

2.1.8 Beta distribution

The *beta distribution* is a generalization of the uniform distribution, describing continuous random numbers with values in $[0, 1]$. Therefore it is suited to model the distribution of a variable of, e.g., daily relative humidity, since humidity is measured between 0% and 100% (i.e. one examines $\frac{humidity}{100\%}$ instead).

Definition 2.8 (Beta distribution.) *A random number Z is called to be beta distributed with parameters $a, b > 0$, if it has the density*

$$f(z) = \frac{1}{B(a,b)} z^{a-1}(1-z)^{b-1} 1_{[0,1]}(z), \tag{2.26}$$

where $B(a,b) := \int_0^1 t^{a-1}(1-t)^{b-1}\,dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ *is the beta-function. We say*

$$Z \sim Beta(a,b).$$

Mean and variance are given by

$$E[Z] = \frac{a}{(a+b)} \text{ and } Var(Z) = \frac{ab}{(1+a+b)(a+b)^2}. \tag{2.27}$$

Note, for the special case $a = b = 1$, one gets the uniform distribution on $[0,1]$. One further remark: If random numbers X, Y are independent and $X \sim Gamma(a,b)$ respectively $Y \sim Gamma(a,c)$, then $\frac{X}{X+Y} \sim Beta(b,c)$. The *gamma distribution* will be presented in the next subsection.

However we will use a different parameterization according to Ferrari and Cribari-Neto [2004], i.e. $\mu := a/(a+b)$ and $\phi := a+b$ and so (2.26) changes to

$$f(z;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} z^{\mu\phi-1}(1-z)^{(1-\mu)\phi-1}, 0 < z < 1 \tag{2.28}$$

with $0 < \mu < 1$ and $\phi > 0$, where ϕ is known as the precision parameter since, for fixed μ, the larger ϕ the smaller the variance of Z; ϕ^{-1} is called the dispersion parameter. It holds that $E[Z] = \mu$ and $Var(Z) = \mu(1-\mu)/(1+\phi)$.

2.1.9 Gamma distribution

The gamma distribution describes continuous and positive random variables, like, e.g. the daily amount of rainfall when it rains. Therefore we introduce and define this kind of distribution as follows:

Definition 2.9 (Gamma distribution.) *A random variable Z is said to be gamma distributed with parameters* $a, \lambda > 0$, *if it has the following density function*

$$f(z) = 1_{]0,\infty[}(z)\frac{\lambda^a}{\Gamma(a)} z^{a-1} e^{-\lambda z}. \tag{2.29}$$

If Z is gamma distributed, we write

$$Z \sim Gamma(a,\lambda).$$

It holds, that $cZ \sim Gamma(a, \frac{\lambda}{c})$. Therefore one defines λ^{-1} as scale parameter, while a determines the form of the distribution. It follows, that

$$E[Z] = \frac{a}{\lambda} \text{ and } Var(Z) = \frac{a}{\lambda^2}. \tag{2.30}$$

Three further interesting properties are:

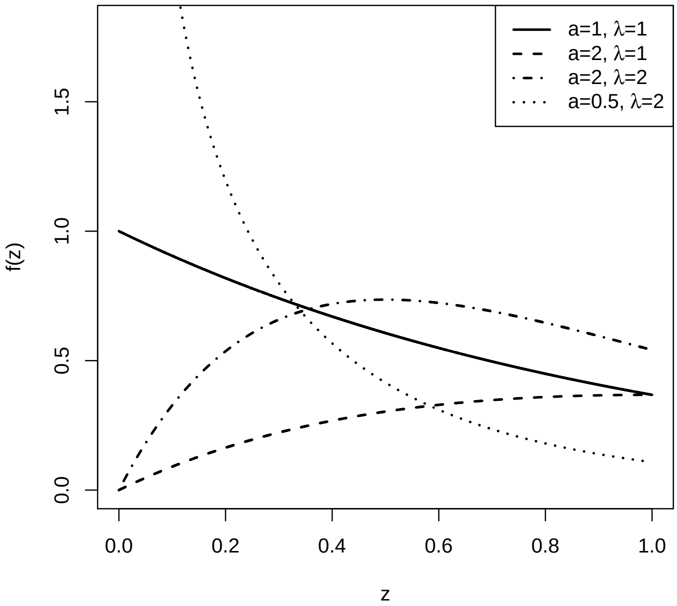


Figure 2.8: Density function of the gamma distribution for different parameters.

(i) The sum of independent $Z_i \sim Gamma(a_i, \lambda)$ distributed variables is again gamma distributed, i.e.

$$\sum_{i=1}^{n} Z_i \sim Gamma\left(\sum_{i=1}^{n} a_i, \lambda\right).$$

(ii) If $Z \sim \chi^2_n$, then $Z \sim Gamma(\frac{n}{2}, \frac{1}{2})$.

(iii) For $Z \sim Gamma(1, \lambda)$, one gets an exponential distributed random variable Z with parameter λ.

As for the beta distribution, we will use a alternative representation of density (2.29) to model the positive daily rain amount according to Stern and Coe [1984]. It is called *mean parametrization* and given by

$$f(z) = \frac{1}{\Gamma(\kappa)} \left(\frac{\kappa}{\mu}\right)^{\kappa} z^{\kappa-1} \exp\left(-\frac{\kappa}{\mu} z\right), \tag{2.31}$$

where $\kappa := a$ and $\mu := \frac{a}{\lambda}$. One can easily verify, that

$$E[Z] = \mu \text{ and } Var(Z) = \frac{\mu^2}{\kappa}. \tag{2.32}$$

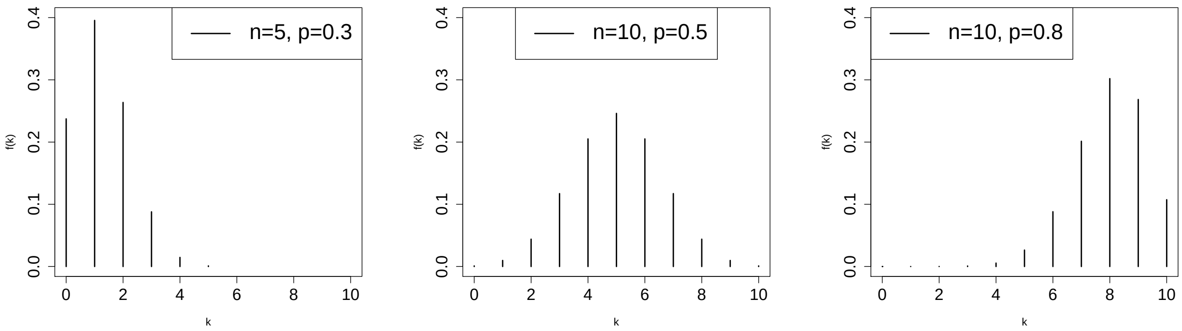


Figure 2.9: Probability mass function of the binomial distribution for different parameters.

2.1.10 Binomial distribution

As last step in this section we introduce the discrete *binomial distribution*. It is the distribution of the *number of successes* in a sequence of n independent yes/no experiments, each with success probability p.

Definition 2.10 *We say $Z \sim Bin(n,p)$ is binomial distributed, if $p \in (0,1)$ and for every $k \in \{0,...,n\}$ holds, that*

$$f(k) = P(Z = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.33}$$

For $n = 1$, we get a success/failure experiment, i.e. a random number, that only assumes the values 0 or 1. It is called *Bernoulli distribution*. Every binomial distributed random variable equals the sum of n Bernoulli random numbers.
The cdf of the binomial distribution for $k \in \mathbb{N}$ is given by

$$F(k) = P(Z \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}.$$

Mean and variance follow as

$$E[Z] = np \text{ and } Var(Z) = np(1-p). \tag{2.34}$$

2.2 Copulas

What are Copulas? Nelsen [2006] answers in his book "[...]copulas are functions that join or "couple" multivariate distribution functions to their one-dimensional marginal distribution functions. Alternatively, copulas are multivariate distribution functions whose one-dimensional margins are uniform on the interval (0,1)" [Nelsen, 2006, p. 1]. This property is an advantage over classical approaches like e.g. multivariate normal distributions, since copulas allow modeling multivariate dependence among random vectors seperatly of their margins. So the variables do not have to be characterized by the same parametric family of univariate distributions anymore [Genest and Favre, 2007, p. 347]. In this connection the most important theorem was introduced by Abe Sklar [1959] that describes the above mentioned properties and explains the name "copula".

We will now formally present the basic concepts and their properties, following mainly Nelsen [2006] and Brechmann [2010][2]. For illustration, we will start with the bivariate case and generalize afterwards.

Definition 2.11 (Bivariate Copula.) *A bivariate copula is a function $C : [0,1]^2 \mapsto [0,1]$ having the following proporties:*

1. *For every u_1, u_2 in $[0,1]$*
$$C(u_1, 0) = 0 = C(0, u_2)$$
and
$$C(u_1, 1) = u_1 \text{ and } C(1, u_2) = u_2$$

2. *For every $u_{11}, u_{21}, u_{12}, u_{22} \in [0,1]$ such that $u_{11} \leq u_{21}$ and $u_{12} \leq u_{22}$,*
$$C(u_{21}, u_{22}) - C(u_{21}, u_{12}) - C(u_{11}, u_{22}) + C(u_{11}, u_{12}) \geq 0$$

A simple example for a copula is the bivariate independence copula $\Pi^2(u_1, u_2) = u_1 u_2$. One can easily verify, that the properties of Definition 2.11 are fulfilled.

The definition of a copula can be generalized to the multivariate case. While the first property of Definition 2.11 can directly be transferred to the n-dimensional case, one needs more work for the second property (in detail see [Nelsen, 2006, p. 45]).

Definition 2.12 (Copula.) *A n-dimensional copula is a function $C : [0,1]^n \mapsto [0,1]$ with the following proporties:*

1. *For every $\boldsymbol{u} = (u_1, ..., u_n)'$ in $[0,1]^n$,*
$$C(u) = 0 \text{ if at least one coordinate of } \boldsymbol{u} \text{ is } 0,$$
and if all coordinates of $\boldsymbol{u}$ are 1 except u_i, then
$$C(1, ..., 1, u_i, 1, ..., 1) = u_i.$$

2. *C is n-increasing.*

[2] You can find further information e.g. in Joe [1997] and Genest and Favre [2007]

For the next theorem we need two of the following functions, for $\boldsymbol{u} = (u_1, ..., u_n)' \in [0,1]^n$:

$$\begin{aligned} M^n(\boldsymbol{u}) &:= min(u_1, u_2, ..., u_n); && (2.35)\\ \Pi^n(\boldsymbol{u}) &:= u_1 u_2 \cdots u_n; && (2.36)\\ W^n(\boldsymbol{u}) &:= max(u_1 + u_2 + ... + u_n - n + 1, 0). && (2.37) \end{aligned}$$

It can be shown, that M^n and Π^n are n-dimensional copulas for all $n \geq 2$, whereas W^n fails to be an n-dimensional copula for any $n > 2$ (i.e. W^n is a copula only for $n = 2$) [Nelsen, 2006, p. 47].

With these results, one gets an important property of copulas, stated in the following theorem:

Theorem 2.13 (Fréchet-Hoeffding bounds.) *Let C be a n-dimensional copula. Then for every $\boldsymbol{u} \in [0,1]^n$,*

$$W^n(\boldsymbol{u}) \leq C(\boldsymbol{u}) \leq M^n(\boldsymbol{u}). \tag{2.38}$$

You can regard a copula as an n-dimensional distribution function with uniform margins, i.e. $C(u_1, ..., u_n) = P(U_1 \leq u_1, ..., U_n \leq u_n)$ for uniform random variables $U_1, ..., U_n$. Often you are also interested in the *joint survival function*

$$\bar{C}(u_1, ..., u_n) = P(U_1 > u_1, ..., U_n > u_n).$$

For the bivariate case, $n = 2$, one gets

$$\bar{C}(u_1, u_2) = P(U_1 > u_1, U_2 > u_2) = 1 - u_1 - u_2 + C(u_1, u_2) \tag{2.39}$$

As mentioned above, the most important theorem in this connection is the theorem of Sklar [1959]. It shows the role of a copula that "couples" a joint distribution function to its univariate margins and therefore describes multivariate dependence. Here we only present it for the continuous case, since it is just relevant for our work.[3]

Theorem 2.14 (Sklar.) *Let F be a n-dimensional distribution function with continuous margins $F_1, ..., F_n$. Then there exists a unique copula C such that for all $\boldsymbol{x} = (x_1, ..., x_n)'$ $\in (\mathbb{R} \cup \{-\infty, \infty\})^n$,*

$$F(\boldsymbol{x}) = C(F_1(x_1), ..., F_n(x_n)). \tag{2.40}$$

Conversely, if C is a copula and $F_1, ..., F_n$ are distribution functions, then the function F defined by (2.40) is a joint distribution function with margins $F_1, ..., F_n$.

So if $X_i \sim F_i$, $i = 1, ..., n$, where $F_1, ..., F_n$ are continuous and invertible, and $\boldsymbol{X} = (X_1, ..., X_n)' \sim F$, then Sklar's Theorem 2.14 states that there exists a *unique* copula such that (2.40) is fulfilled. Thus, C describes the dependence between $X_1, ..., X_n$.
Furthermore one easily gets the appropriate copula, according to (2.40), by

$$C(\boldsymbol{u}) = F(F_1^{-1}(u_1), ..., F_n^{-1}(u_n)). \tag{2.41}$$

[3] A proof can be found, e.g., in [Nelsen, 2006, pp. 46-47].

This method is called *inversion method*. The Gaussian and t copulas are examples of copula families, which are constructed like this (see (2.44) and (2.46)). Further methods are presented in [Nelsen, 2006, pp. 51].

Besides the joint distribution function, simultaneously one is also interested in calculating the joint density function resp. the *copula density* c . It can be derived by partially differentiating and applying the chain rule as [Brechmann, 2010, p. 18]

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{\partial^n C(F_1(x_1), ..., F_n(x_n))}{\partial x_1 ... \partial x_n} = \frac{\partial^n C(F_1(x_1), ..., F_n(x_n))}{\partial F_1(x_1) ... \partial F_n(x_n)} f_1(x_1) \cdots f_n(x_n) \qquad (2.42) \\ \Leftrightarrow c(F_1(x_1), ..., F_n(x_n)) &:= \frac{\partial^n C(F_1(x_1), ..., F_n(x_n))}{\partial F_1(x_1) ... \partial F_n(x_n)} = \frac{f(\boldsymbol{x})}{f_1(x_1) \cdots f_n(x_n)}, \end{aligned}$$

where $f_1, ..., f_n$ and f are the corresponding density functions to $F_1, ..., F_n$ and F.

In case of the n-dimensional independence copula $\Pi^n(\boldsymbol{u}) := u_1 u_2 \cdots u_n$ from (2.36), we get

$$\begin{aligned} \pi^n(F_1(x_1), ..., F_n(x_n)) &:= \frac{\partial^n \Pi^n(F_1(x_1), ..., F_n(x_n))}{\partial F_1(x_1) ... \partial F_n(x_n)} = \frac{\partial^n [F_1(x_1) \cdots F_n(x_n)]}{\partial F_1(x_1) ... \partial F_n(x_n)} = 1, \\ &\overset{(2.42)}{\Leftrightarrow} f(\boldsymbol{x}) = f_1(x_1) \cdots f_n(x_n). \qquad (2.43) \end{aligned}$$

Equation (2.43) represents the well-known density if random variables $X_1, ..., X_n$ are independent.

This yields to the following result.

Theorem 2.15 *For $n \geq 2$, let $X_1, ..., X_n$ be continuous random variables. Then*

$X_1, ..., X_n$ are independent, iff the n-dimensional Copula C of $X_1, ..., X_n$ is Π^n.

Furthermore, perfect negative (for $n = 2$) and positive dependence can be detected by the Fréchet-Hoeffding bounds.

Theorem 2.16 *For $n \geq 2$, let $X_1, ..., X_n$ be continuous random variables with copula C. Then*

1. *each of the random variables $X_1, ..., X_n$ is almost surely a strictly increasing function of any of the others iff $C = M^n$, and*
2. *X_1 and X_2 are almost surely strictly decreasing functions of each other iff $C = W^2$.*

In addition, another important property of copulas is given by the following theorem:

Theorem 2.17 *Let $X_1, ..., X_n$ be continuous random variables with copula C. Then C is invariant under strictly increasing transformations of $X_1, ..., X_n$.*

Since we will meet a lot of them during the later analysis, we want to introduce two important classes of copulas known as elliptical and Archimedean copulas in the following. These copulas find a wide range of applications, since they can easily be constructed and have many nice properties, such as symmetry and associativity, within these classes. In addition, in case of Archimedean copulas, the great variety of families of copulas which belong to this class, shows their importance [Nelsen, 2006, p. 109].

2.2.1 Elliptical copulas

Elliptical copulas are generated by elliptical distributions using the inversion method, stated in (2.41), see Hult and Lindskog [2002] or Owen and Rabinovitch [1983]. Therefore we need the following definition:

Definition 2.18 (Elliptical distribution.) *Let $\boldsymbol{\mu}$ be a fixed n-component vector, i.e. $\boldsymbol{\mu} \in \mathbb{R}^n$, and Σ a $(n \times n)$-positive definite symmetric matrix, i.e. $\Sigma \in \mathbb{R}^{n \times n}$. Then, a n-dimensional random vector $\boldsymbol{X} = (X_1, ..., X_n)'$ is said to be elliptical distributed, if the density function of $\boldsymbol{X}$, $f_{\boldsymbol{X}}(\boldsymbol{x})$, has the following representation*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = c_n \left|\Sigma\right|^{-1/2} \xi\left((\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

with normalizing constant $c_n \in \mathbb{R}$ and some function ξ which is independent of n.

We have two famous examples for elliptical distributions:

1. The multivariate normal distribution is elliptical with $c_n = (2\pi)^{-n/2}$ and $\xi(k) = \exp(-\frac{1}{2}k)$ $\forall k > 0$.
2. The multivariate t-distribution with $c_n = (\pi n)^{-n/2}\Gamma\left(\frac{\nu+n}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)$ and $\xi(k) = \left(1+\frac{k}{\nu}\right)^{-(\nu+n)/2}$, $\forall k > 0$, and $\nu > 0$ are the degrees of freedom.

Using the inversion method from (2.41), based on Sklar's Theorem 2.14, we can construct the correspoding copulas in both cases, i.e.

1. The multivariate Gaussian copula with $\boldsymbol{u} = (u_1, ..., u_n)'$:

$$C_{Gaus,K}(\boldsymbol{u}) = \Phi_K\left(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)\right) \tag{2.44}$$

$$= \frac{1}{2\pi^{n/2}}\left|K\right|^{-\frac{1}{2}} \int_{-\infty}^{\Phi^{-1}(u_1)} \cdots \int_{-\infty}^{\Phi^{-1}(u_n)} \exp\left(-\frac{1}{2}\boldsymbol{x}'K^{-1}\boldsymbol{x}\right) du_1 \cdots du_n,$$

where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function (cdf) Φ, Φ_K the multivariate standard normal cdf with symmetric positive defnite correlation matrix $K \in [-1,1]^{n \times n}$ and $\boldsymbol{x} = (\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))' \in \mathbb{R}^n$.

Accordingly, the multivariate Gaussian copula density:

$$c_{Gaus,K}(\boldsymbol{u}) = \frac{\frac{1}{(2\pi)^{n/2}\sqrt{|K|}} \exp\left(-\frac{1}{2}\boldsymbol{x}'K^{-1}\boldsymbol{x}\right)}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_i^2\right)} = \left|K\right|^{-\frac{1}{2}} \exp\left(\frac{1}{2}\boldsymbol{x}'(I_n - K^{-1})\boldsymbol{x}\right). \tag{2.45}$$

2. The multivariate t copula, with $\boldsymbol{u} = (u_1, ..., u_n)'$ is given by (see Demarta and McNeil [2005] or Kurowicka and Joe [2010]):

$$C_{t,\nu,K}(\boldsymbol{u}) = t_{K,\nu}\left(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_n)\right) \tag{2.46}$$

$$= \int_{-\infty}^{t_\nu^{-1}(u_1)} \cdots \int_{-\infty}^{t_\nu^{-1}(u_n)} \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)^n\left|K\right|}} \left(1 + \frac{\boldsymbol{x}'K^{-1}\boldsymbol{x}}{\nu}\right)^{-\frac{\nu+n}{2}} du_1 \cdots du_n,$$

where t_ν^{-1} is the inverse of the cdf of the univariate standard t distribution with $\nu > 0$ degrees of freedom, $t_{K,\nu}$ the corresponding cdf of the multivariate standard t distribution with correlation matrix $K \in [-1,1]^{n\times n}$ and degrees of freedom ν and $\boldsymbol{x} = (t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_n))' \in \mathbb{R}^n$.

The t copula density is then given by:

$$c_{t,\nu,K}(\boldsymbol{u}) = \frac{1}{\sqrt{|K|}} \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}\right)^n \frac{\prod_{i=1}^n \left(1 + \frac{t_\nu^{-1}(u_i)^2}{\nu}\right)^{\frac{\nu+1}{2}}}{\left(1 + \frac{\boldsymbol{x}'K^{-1}\boldsymbol{x}}{\nu}\right)^{\frac{\nu+n}{2}}} \tag{2.47}$$

where $K \in [0,1]^n$ is the correlation matrix of the joint density of the standard t distribution and ν the degrees of freedom.

2.2.2 Archimedean copulas

To explain Archimedean copulas, one needs the following definition beforehand [Nelsen, 2006, p. 151-152]:

Definition 2.19 (Completely monotonicity.) *A function $g(t)$ is completely monotonic on an interval J if it is continuous there and has derivatives of all orders that alternate in sign, i.e, if it satisfies*

$$(-1)^k \frac{d^k}{dt^k} g(t) \geq 0,$$

for all t in the interior of J and $k = 0, 1, 2, ...$

And the following theorem explains Archimedean copulas:

Theorem 2.20 (Archimedean copula.) *Let $\varphi : [0,1] \mapsto [0,\infty]$ be a continuous strictly decreasing function, such that $\varphi(0) = \infty$ and $\varphi(1) = 0$, and let φ^{-1} denote the inverse of φ. If $C : [0,1]^n \mapsto [0,1]$ is given by*

$$C(\boldsymbol{u}) = \varphi^{-1}\left(\varphi(u_1) + \varphi(u_2) + ... + \varphi(u_n)\right),$$

where $\boldsymbol{u} = (u_1, ..., u_n)' \in [0,1]^n$, then C is a n-dimensional copula, so called n-dimensional Archimedean copula with generator φ, for all $n \geq 2$ iff φ^{-1} is completely monotonic on $[0,\infty)$.

In the bivariate case, the assumptions of complete monotonicity and $\varphi(0) = \infty$ are not necessary. Here, it is sufficient to assume $\varphi(0) \leq \infty$ and the *pseudo-inverse* $\varphi^{[-1]}$ of a convex generator function φ is considered instead of φ^{-1} [Nelsen, 2006, p. 110]. The pseudo-inverse $\varphi^{[-1]}$ of a continuous, strictly decreasing function $\varphi : [0,1] \mapsto [0,\infty]$ is defined by:

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

So if $\varphi(0) = \infty$, then it follows that $\varphi^{[-1]} = \varphi^{-1}$ and φ is called *strict* .

For example, $W^2(u_1, u_2) = \max(u_1 + u_2 - 1, 0)$, from (2.37), is an Archimedean copula with $\varphi(t) = 1 - t$, for $t \in [0, 1]$. $\Pi^2(u_1, u_2) = u_1 u_2$ with $\varphi(t) = -\log(t)$ is strictly Archimedean, but $M^2(u_1, u_2) = \min(u_1, u_2)$ from (2.35) is not an Archimedean copula.

Since most commonly used generators depend only on one or at most two parameters (see the following section), modeling dependency of n-dimensional random vectors through Archimedean copulas is quite restrictive. Elliptical copulas have correlation parameters for each pair of variables, therefore one concentrates more on elliptical copulas for dimensions $n > 2$, while Archimedean copulas are rather used in the bivariate case.

2.2.3 Bivariate copula families

Bivariate copulas, also called *pair-copulas*, constitute the building blocks of the joint vine distribution that we will use for our model later. So we here introduce several bivariate copula families, belonging to elliptical and Archimedean copulas, which we will meet again throughout the thesis. Extended descriptions as well as further copula families can be found, e.g., in Nelsen [2006] or Joe [1997]. We start with the most familiar bivariate copula families, such as Gaussian-, t-, Clayton-, Gumbel-, Frank- and Joe- Copulas, then describe the rotated counterparts and give an example of two-parameters Archimedean copulas in the end. The corresponding scatter and contour plots for standard normal margins can be found in Appendix A. In addition, the connection of these families to dependence measures like Kendall's τ or tail dependence parameters will be shown in the next section.

We summarize the properties of the different bivariate copula families in tables to make it comparable. Notice that the Gaussian- and the t-copulas differ from the rest in such a way that they are elliptical copulas which are also *reflection symmetric*, i.e. if (U_1, U_2) follows one of the two copulas, then $(1 - U_1, 1 - U_2)$ is distributed as the same copula. In contrast, the other described families are all Archimedean copulas which are non-symmetric with respect to reflection, except Frank copulas which are also reflection symmetric. But all copulas studied here, except the family of t-copulas, since they depend also on the degrees of freedom ν, depend only on one parameter (in the bivariate case). An example of a two-parameters Archimedean copula is given afterwards. Furthermore we are interested in which parameter values lead the corresponding copula to be equal to the Fréchet-Hoeffding bounds M^2 (2.35) and W^2 (2.37) as well as to be equal to the bivariate independent copula Π^2 from (2.36).

Note, in case of Gaussian and t-copulas, Φ_θ represents the bivariate standard normal distribution and $t_{\theta,\nu}$ the bivariate standard t-distribution with $\nu > 0$ degrees of freedom, each with the corresponding correlation parameter $\theta \in (-1, 1)$. Furthermore it holds $x_i = \Phi^{-1}(u_i)$ and $x_i = t_\nu^{-1}(u_i)$ respectively, for $i = 1, 2$, where Φ is the univariate standard normal distribution and t_ν the univariate standard t-distribution with $\nu > 0$ degrees of freedom.

Copula	$C_\theta(u_1, u_2)$	Generator $\varphi(t)$	Parameter
Gaussian	$\Phi_\theta\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right)$	/	$\theta \in (-1, 1)$
t	$t_{\theta,\nu}\left(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2)\right)$	/	$\theta \in (-1,1)$, $\nu > 0$
Clayton	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$	$\frac{1}{\theta}(t^\theta - 1)$	$\theta > 0$
Gumbel	$\exp\left[-\left((-\log u_1)^\theta + (-\log u_2)^\theta\right)^{\frac{1}{\theta}}\right]$	$(-\log t)^\theta$	$\theta \geq 1$
Frank	$-\frac{1}{\theta}\log\left[1 + \frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right]$	$-\log\left[\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right]$	$\theta \in \mathbb{R}\backslash\{0\}$
Joe	$1 - \left[(1-u_1)^\theta + (1-u_2)^\theta - (1-u_1)^\theta(1-u_2)^\theta\right]^{\frac{1}{\theta}}$	$-\log\left[1-(1-t)^\theta\right]$	$\theta > 1$

Table 2.1: Part I: Properties of the most familiar bivariate copula families.

Copula	$C_\theta = W^2$	$C_\theta = \Pi^2$	$C_\theta = M^2$	Further properties
Gaussian	for $\theta \to -1$	for $\theta = 0$	for $\theta \to 1$	elliptical & symmetric
t	for $\theta \to -1$	for $\theta = 0$ and $\nu \to \infty$	for $\theta \to 1$	elliptical & symmetric
Clayton	/	for $\theta \to 0$	for $\theta \to \infty$	Archimedean & non-symmetric
Gumbel	/	for $\theta = 1$	for $\theta \to \infty$	Archimedean & non-symmetric
Frank	for $\theta \to -\infty$	for $\theta \to 0$	for $\theta \to \infty$	Archimedean & symmetric
Joe	/	for $\theta \to 1$	for $\theta \to \infty$	Archimedean & non-symmetric

Table 2.2: Part II: Properties of the most familiar bivariate copula families.

Copula	$c_\theta(u_1, u_2)$
Gaussian	$\frac{1}{\sqrt{1-\theta^2}}\exp\left(-\frac{\theta^2(x_1^2+x_2^2))-2\theta x_1 x_2}{2(1-\theta^2)}\right)$
t	$\frac{\Gamma\left(\frac{\nu+2}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)}{\nu\pi dt_\nu(x_1)dt_\nu(x_2)\sqrt{1-\theta^2}} \times \left(1 + \frac{x_1^2+x_2^2-2\theta x_1 x_2}{\nu(1-\theta^2)}\right)^{-\frac{\nu+2}{2}}$
Clayton	$(1+\theta)(u_1 u_2)^{-1-\theta} \times (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}-2}$
Gumbel	$\frac{C_\theta(u_1,u_2)}{u_1 u_2}\frac{(\log u_1 \cdot \log u_2)^{\theta-1}}{\left((-\log u_1)^\theta+(-\log u_2)^\theta\right)^{2-\frac{1}{\theta}}} \times \left[\left((-\log u_1)^\theta + (-\log u_2)^\theta\right)^{\frac{1}{\theta}} + \theta - 1\right]$
Frank	$\theta(e^{-\theta} - 1)\frac{e^{\theta(u_1+u_2)}}{\left[e^{-\theta}-1+(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)\right]^2}$
Joe	$\left[(1-u_1)^\theta + (1-u_2)^\theta - (1-u_1)^\theta(1-u_2)^\theta\right]^{\frac{1}{\theta}-2}(1-u_1)^{\theta-1}(1-u_2)^{\theta-1}$ $\times\left[\theta - 1 + (1-u_1)^\theta + (1-u_2)^\theta - (1-u_1)^\theta(1-u_2)^\theta\right]$

Table 2.3: Part III: Properties of the most familiar bivariate copula families.

Rotated copulas

Instead of considering the distribution of $(U_1, U_2) \in [0,1]^2$ one can also consider the copulas of $(U_1, 1-U_2)$, or $(1-U_1, 1-U_2)$, or $(1-U_1, U_2)$, which are known as the rotated versions of the copula of (U_1, U_2). Their densities are their original densities rotated by 90, 180 and 270 degrees respectively and thus the idea only makes sense for reflection-asymmetric copulas, i.e. for Clayton, Gumbel and Joe copulas.

According to Lemma 2.4.4 from Nelsen [2006], one gets the following copulas and densities for the three cases:

90°: $(U_1, U_2) \in [0,1]^2$ follows a rotated copula by 90 degrees with parameter θ iff $(1-U_1, U_2)$ is distributed as one of the above mentioned copulas with parameter $-\theta$,

$$C_\theta^{90^\circ}(u_1, u_2) = u_2 - C_{(-\theta)}(1-u_1, u_2),$$

where C is one of the Clayton, Gumbel or Joe copulas. Its density is given by

$$c_\theta^{90^\circ}(u_1, u_2) = c_{(-\theta)}(1-u_1, u_2),$$

where c is the corresponding density of C.

180°: $(U_1, U_2) \in [0,1]^2$ follows a rotated copula by 180 degrees with parameter θ iff $(1-U_1, 1-U_2)$ is distributed as one of the above mentioned copulas with parameter θ. A rotated copula by 180 degrees is also called *survival copula.*

$$C_\theta^{180^\circ}(u_1, u_2) = u_1 + u_2 - 1 + C_\theta(1-u_1, 1-u_2),$$

where C is one of the Clayton, Gumbel or Joe copulas. Its density is given by

$$c_\theta^{180^\circ}(u_1, u_2) = c_\theta(1-u_1, 1-u_2),$$

where c is the corresponding density of C.

270°: $(U_1, U_2) \in [0,1]^2$ follows a rotated copula by 270 degrees with parameter θ iff $(U_1, 1-U_2)$ is distributed as one of the above mentioned copulas with parameter $-\theta$,

$$C_\theta^{270^\circ}(u_1, u_2) = u_1 - C_{(-\theta)}(u_1, 1-u_2),$$

where C is one of the Clayton, Gumbel or Joe copulas. Its density is given by

$$c_\theta^{270^\circ}(u_1, u_2) = c_{(-\theta)}(u_1, 1-u_2),$$

where c is the corresponding density of C.

Two-parametric Archimedean copula

At last we give an example for a two-parametric Archimedean copula, called *Clayton-Gumbel copula/BB1* (cp. Joe [1997]). It is a generalization of the one-parametric Clayton and Gumbel families. It has the generator $\varphi(t) = (t^{-\theta} - 1)^{\delta}$ and so

$$C^{C-G}_{\theta,\delta}(u_1, u_2) = \left[1 + \left[(u_1^{-\theta} - 1)^{\delta} + (u_2^{-\theta} - 1)^{\delta}\right]^{\frac{1}{\delta}}\right]^{-\frac{1}{\theta}},$$

where $\theta > 0$ and $\delta \geq 1$. It obtains the independence copula Π^2 for $\theta \to 0$ and $\delta = 1$ as well as W^2 for $\theta \to \infty$ and $\delta \to \infty$. In addition, it becomes a Clayton copula for $\delta = 1$ and a Gumbel copula if $\theta \to 0$.
Further properties and further two-parametric Archimedean copulas, such as the Joe-Clayton copula, can be found in the above mentioned references and in Schepsmeier [2010].

2.2.4 Pair copula constructions of general multivariate distributions

Multivariate copulas are often limited in modeling the range of various dependence structures. The Gaussian copula can model the whole correlation structure but does not allow for tail dependence. In contrast, the t-copula allows for tail dependence, but it cannot model asymmetric tail dependence, i.e. when upper and lower tail dependence are not the same. Therefore one uses Archimedean copulas, which are stated above. Nevertheless, elliptical and Archimedean copulas do not allow for different dependency patterns between pairs of variables (Czado [2012] and Kurowicka and Joe [2010]). *Pair copula constructions* as building blocks for our later model can overcome these shortcomings. Based on Aas et al. [2009], pair copula constructions are described as a simple and flexible way to specify multivariate dependence.

We start with a 3-dimensional example for illustration and generalize afterwards.

Example 1 (Pair copula construction in 3 dimensions.) *Let $\boldsymbol{X} = (X_1, X_2, X_3)'$ a 3-dimensional random vector with joint density function f univariate densities f_1, f_2 and f_3. F_1, F_2 and F_3 denote the corresponding marginal distribution functions. According to the definition of conditional densities [Czado and Schmidt, 2011, pp. 20-21], we get*

$$f(x_1, x_2, x_3) = f_3(x_3) f(x_2|x_3) f(x_1|x_2 x_3). \tag{2.48}$$

Sklar's Theorem 2.14 and Equation (2.42) indicate

$$f(x_1, x_2, x_3) = c_{123}\left(F_1(x_1), F_2(x_2), F_3(x_3)\right) f_1(x_1) f_2(x_2) f_3(x_3), \tag{2.49}$$

where c_{123} is the density of a three-dimensional copula. In the bivariate case, it follows that

$$f(x_2, x_3) = c_{23}\left(F_2(x_2), F_3(x_3)\right) f_2(x_2) f_3(x_3)$$

for a bivariate copula density c_{23}. Thus, using again the definition of conditional densities,

$$f(x_2|x_3) = \frac{f(x_2, x_3)}{f_3(x_3)} = c_{23}\left(F_2(x_2), F_3(x_3)\right) f_2(x_2). \tag{2.50}$$

Similarly, one can decompose

$$f(x_1|x_2,x_3) = \frac{f(x_1,x_3|x_2)}{f(x_3|x_2)} = c_{13|2}\left(F(x_1|x_2), F(x_3|x_2)\right) f(x_1|x_2), \tag{2.51}$$

where $c_{13|2}$ is an appropriate pair copula for $F(x_1|x_2)$ and $F(x_3|x_2)$. Decompose $f(x_1|x_2)$ as in (2.50) and one gets

$$f(x_1|x_2,x_3) = c_{13|2}\left(F(x_1|x_2), F(x_3|x_2)\right) c_{12}\left(F_1(x_1), F_2(x_2)\right) f_1(x_1).$$

Combining all into Equation (2.48):

$$\begin{aligned} f(x_1,x_2,x_3) &= c_{12}\left(F_1(x_1), F_2(x_2)\right) c_{23}\left(F_2(x_2), F_3(x_3)\right) c_{13|2}\left(F(x_1|x_2), F(x_3|x_2)\right) \\ &\quad \times f_1(x_1) f_2(x_2) f_3(x_3). \end{aligned}$$

If we look at Equation (2.49), one can see that the trivariate copula density is constructed by bivariate copulas as building blocks:

$$\begin{aligned} c_{123}\left(F_1(x_1), F_2(x_2), F_3(x_3)\right) &= c_{12}\left(F_1(x_1), F_2(x_2)\right) c_{23}\left(F_2(x_2), F_3(x_3)\right) \\ &\quad \times c_{13|2}\left(F(x_1|x_2), F(x_3|x_2)\right). \end{aligned}$$

Pair copulas are well studied, understood and applied (see Kurowicka and Joe [2010] and Section 2.2.3). Note that since in Equation (2.48) the variables can be permuted in $3! = 6$ *ways, this decomposition is not unique.*

In addition, we assume that the pair copula $c_{13|2}$ in Equation (2.51) is independent of the conditioning variable X_2, i.e.,

$$c_{13|2}\left(F(x_1|x_2), F(x_3|x_2); x_2\right) \equiv c_{123}\left(F(x_1|x_2), F(x_3|x_2)\right).$$

According to Aas et al. [2009], this assumption is necessary in order to facilitate statistical inference. Hobæk Haff et al. [2010] call this the simplified pair copula construction and show that it typically is a good approximiation to the correct decomposition.

Now we want to decompose a n-dimensional random vector $\boldsymbol{X} = (X_1, X_2, ..., X_n)'$ with joint density (using again the definition of conditional densities):

$$f(\boldsymbol{x}) = f_n(x_n) f(x_{n-1}|x_n) f(x_{n-2}|x_{n-1}, x_n) \cdots f(x_1|x_2, ..., x_{n-1}, x_n). \tag{2.52}$$

Like in Example 1, we can decompose each term in (2.52) into marginal densities and appropriate bivariate copulas using the general formula

$$f(x|\boldsymbol{v}) = c_{x,v_j|\boldsymbol{v}_{-j}}\left(F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j})\right) f(x|\boldsymbol{v}_{-j}),$$

where $\boldsymbol{v}$ is an m-dimensional vector, v_j an arbitrary component of $\boldsymbol{v}$ and $\boldsymbol{v}_{-j}$ denotes the $(m-1)$-dimensional vector $\boldsymbol{v}$ without v_j.

Hint: The above stated pair copulas are applied to marginal conditional distributions of the form $F(x|\boldsymbol{v})$. These can be obtained for every j (Joe [1996]) by

$$F(x|\boldsymbol{v}) = \frac{\partial C_{x,v_j|\boldsymbol{v}_{-j}}\left(F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j})\right)}{\partial F(v_j|\boldsymbol{v}_{-j})}, \tag{2.53}$$

where $C_{x,v_j|\boldsymbol{v}_{-j}}$ is the bivariate copula distribution function. For the special case, where v is univariate, we have

$$F(x|v) = \frac{\partial C_{x,v}\left(F(x), F(v)\right)}{\partial F(v)}.$$

Now we have seen how a multivariate density can be decomposed into the product of pair copulas and marginal densities. But this decomposition is not unique and therefore for high-dimensional distributions, there is a significant number of possible pair copulas constructions. Aas et al. [2009] show for example that there are 240 different constructions for a five-dimensional density. For that reason to help organising them, Bedford and Cooke [2001, 2002] have introduced a graphical model denoted as the *regular vine*. We will take care of it in detail later. Beforehand, we introduce the dependence measures Kendall's τ and tail dependence and their connections to our presented bivariate copula families.

2.3 Dependence measures

In the previous section we have seen the pair copula construction to model multivariate dependence. Thus, modeling the dependence between two random variables becomes an important feature to explain dependence among large numbers of variables.

2.3.1 Pearson's product moment correlation

The *Pearson's product moment correlation* is very popular to measure bivariate dependence because it is often straightforward to calculate. It is defined as follows:

Definition 2.21 (Pearson's product moment correlation.) *The product moment correlation of random variables X and Y with finite expectations $E[X], E[Y]$ and finite variances is*

$$corr(X,Y) = \frac{E[XY] - E[X]E[Y]}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}. \tag{2.54}$$

If we have given N pairs of samples $(x_i, y_i), i = 1, ..., N$, from the random vector (X, Y), then we calculate the corresponding sample or empirical product moment correlation as follows:

$$\widehat{corr}(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{2.55}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

Nevertheless, for uncertainty analysis it has several disadvantages [Kurowicka and Cooke, 2006, p. 28].

1. The product moment correlation is not defined if the expectations and variances of X and Y are not finite (e.g. for the Cauchy distribution).

2. *corr* is only a measure of linear dependence and not invariant under non-linear strictly increasing transformations.

3. The value of *corr* depends on marginal distributions.

The *association measures* Kendall's τ and Spearman's ρ can overcome these issues.

2.3.2 Kendall's τ

Kendall's τ belongs to the so-called "association measures" to model any type of dependence between two random variables. In contrast, the term "correlation coefficient" is reserved for a measure of the linear dependence (e.g. the Pearson's product moment correlation of Definition 2.21). Proofs of the following theorems can be found in Nelsen [2006].

Before we can define Kendall's τ, we have to introduce the definition of *concordance.*

Definition 2.22 (Concordance.) *Two pairs of observations (x_i, y_i) and (x_j, y_j) from the continuous random vector (X, Y) are called concordant, if*

$$x_i < x_j \textit{ and } y_i < y_j, \textit{ or if } x_i > x_j \textit{ and } y_i > y_j, \textit{ respectively if } (x_i - x_j)(y_i - y_j) > 0.$$

Similarly, the pairs are discordant if

$$(x_i - x_j)(y_i - y_j) < 0.$$

The case $(x_i - x_j)(y_i - y_j) = 0$ cannot occur, when X and Y are continuous.

The idea of associated measures is to investigate, whether "large" values of one variable are "associated" with "large" values of the other and similarly for "small" values.

Kendall's τ is now defined as probability of concordance minus the probability of discordance of two random variables X and Y:

Definition 2.23 (Kendall's τ.) *Let X and Y two random variables and (X_1, Y_1), (X_2, Y_2) two independent and identically distributed copies of (X, Y), then Kendall's τ is defined as follows:*

$$\tau(X, Y) = P\left((X_1 - X_2)(Y_1 - Y_2) > 0\right) - P\left((X_1 - X_2)(Y_1 - Y_2) < 0\right) \tag{2.56}$$

The sample or empirical version of Kendall's τ is calculated in the following way: Let $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ denote a random sample of N observations from the continuous random vector (X, Y). There are $\binom{N}{2}$ distinct pairs of (x_i, y_i) and (x_j, y_j) of observations in the sample. Each pair is either concordant oder discordant, so let c denote the number of concordant pairs and d the number of discordant pairs. Then the sample or empirical Kendall's τ is defined as

$$\hat{\tau}(X, Y) = \frac{c - d}{c + d} = \frac{c - d}{\binom{N}{2}}.$$

Equivalently, $\hat{\tau}$ is the probability of concordance minus the probability of discordance for a pair of observations (x_i, y_i) and (x_j, y_j) that is chosen randomly from the sample.

An important relationship between Kendall's τ and copulas is stated in the following theorem.

Theorem 2.24 *Let X_1 and X_2 be continuous random variables with copula C. Then*

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

In case of Archimedean copulas, one gets the expression of Kendall's τ in terms of the generator φ:

$$\tau(X_1, X_2) = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)}\, dt$$

2.3.3 Spearman's ρ

The population version of Spearman's ρ, as further measure of association, is also defined in terms of concordance.

Definition 2.25 (Spearman's ρ.) *Let (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) be independent and identically distributed copies of the continuous random vector (X, Y). Then Spearman's ρ is defined as*

$$\rho(X, Y) = 3\left[P\left((X_1 - X_2)(Y_1 - Y_3) > 0\right) - P\left((X_1 - X_2)(Y_1 - Y_3) < 0\right)\right] \qquad (2.57)$$

So Spearman's ρ is defined to be proportional to the probability of concordance minus the probability of discordance of the two vectors (X_1, Y_1), (X_2, Y_3). The copula of (X_1, Y_1) is C, but since (X_2, Y_3) are independent, their copula is Π^2. The empirical version of Spearman's ρ is defined as the correlation of the pairs of ranks and can be found, e.g., in Nelsen [2006]. Corresponding to Theorem 2.24, we have

Theorem 2.26 *Let X_1 and X_2 be continuous random variables with copula C. Then*

$$\rho(X_1, X_2) = 12 \int_{[0,1]^2} C(u_1, u_2)\, du_1 du_2 - 3 = 12 \int_{[0,1]^2} u_1 u_2\, dC(u_1, u_2) - 3$$

It can be shown that $\tau(X, Y), \rho(X, Y) = 1$ if Y is almost surely an increasing function of X. Accordingly, $\tau(X, Y), \rho(X, Y) = -1$ if Y is almost surely an decreasing function of X. Further, since Kendall's τ and Spearman's ρ can be expressed in terms of copulas of two random variables, both measures are invariant under strictly increasing transformations (follows from Theorem 2.17) and independent of the marginal distributions of X and Y.

2.3.4 Tail dependence

While the above introduced measures describe the dependency between two random variables on the whole space $[0,1]^2$, tail dependence measures the dependence between the variables in the upper-right quadrant and in the lower-left quadrant of $[0,1]^2$. Hence,

Definition 2.27 (Tail dependence.) *Let X_1 and X_2 random variables with corresponding marginal distribution functions F_1, F_2. The lower tail dependence parameter λ^{lower} of X_1 and X_2 is given by:*

$$\lambda^{lower} = \lim_{t \to 0_+} P(X_2 \le F_2^{-1}(t) | X_1 \le F_1^{-1}(t)). \tag{2.58}$$

The upper tail dependence parameter λ^{upper} is defined as:

$$\lambda^{upper} = \lim_{t \to 1_-} P(X_2 > F_2^{-1}(t) | X_1 > F_1^{-1}(t)), \tag{2.59}$$

if the limits exist.

The connection to copulas is stated in the following theorem:

Theorem 2.28 *Let X_1 and X_2 be continuous random variables with copula C. If the limits (2.58) and (2.59) exist, then*

$$\lambda^{lower} = \lim_{t \to 0_+} \frac{C(t,t)}{t}, \tag{2.60}$$

and

$$\lambda^{upper} = \lim_{t \to 1_-} \frac{\bar{C}(t,t)}{1-t} = \lim_{t \to 1_-} \frac{1 - 2t + C(t,t)}{1-t}, \tag{2.61}$$

where $\bar{C}$ is the joint survival function from (2.39).

For example, the lower tail dependence parameter for the idependent copula Π^2 is given by

$$\lambda^{lower} = \lim_{t \to 0_+} \frac{\Pi^2(t,t)}{t} = \lim_{t \to 0_+} \frac{t^2}{t} = \lim_{t \to 0_+} t = 0.$$

The upper Fréchet-Hoeffding bound M^2, from (2.37), shows perfect lower tail dependence:

$$\lambda^{lower} = \lim_{t \to 0_+} \frac{M^2(t,t)}{t} = \lim_{t \to 0_+} \frac{t}{t} = \lim_{t \to 0_+} 1 = 1.$$

The tail dependence parameters of the bivariate copula families discussed in Section 2.2.3, except the rotated ones, can be found in the literature, e.g. in Nelsen [2006] or in Demarta and McNeil [2005]. The tail dependence parameters for the rotated copulas can be found in parts in Brechmann [2010].

Copula parameters can be expressed in terms of Kendall's τ and/or tail dependence parameters. This is shown in the following Table 2.4 for the most familiar bivariate copula families from Section 2.2.3.

Copula	Kendall's τ	λ^{lower}	λ^{upper}
Gaussian	$\frac{2}{\pi}\arcsin(\theta)$	0	0
t	$\frac{2}{\pi}\arcsin(\theta)$	$2t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{\frac{1-\theta}{1+\theta}}\right)$	$=\lambda^{lower}$
Clayton	$\frac{\theta}{\theta+1}$	$2^{-1/\theta}$	0
Gumbel	$1-\frac{1}{\theta}$	0	$2-2^{1/\theta}$
Frank[4]	$1-\frac{4}{\theta}+4\frac{D_1(\theta)}{\theta}$	0	0
Joe	$1+\frac{4}{\theta^2}\int_0^1 t\log(t)(1-t)^{2(1-\theta)/\theta}\,dt$	0	$2-2^{1/\theta}$

Table 2.4: Kendall's τ and tail dependence parameters of the most familiar bivariate copula families.

2.4 Regular vines

Bedford and Cooke [2001, 2002] introduced a graphical model to classify different pair copula constructions, called *regular vines*. In order to define them, we need a few basics in graph theory, which are given in the following subsection.

2.4.1 Graph theory

The following definitions and theorems can be found in Diestel [2010]. We start with the definition of a *graph*.

Definition 2.29 (Graph, node, edge, degree, path, cycle.) *A graph is a pair $G = (N, E)$ of sets such that $E \subseteq \{\{x, y\} : x, y \in N\}$. The elements of E are called edges of the graph G, the elements of N are its nodes. Further,*

- *$d(v)$ denotes the degree of v, i.e. the number of neighbors of a node $v \in N$.*
- *A path is a graph $G = (N, E)$ with $N = \{v_0, v_1, ..., v_k\}$ and $E = \{\{v_0, v_1\}, \{v_1, v_2\}, ..., \{v_{k-1}, v_k\}\}$.*
- *A cycle is a path with $v_0 = v_k$.*

If there is a function $w : E \to \mathbb{R}$, then G is called *weighted* and denoted by $G = (N, E, w)$, i.e., weights are assigned to each edge. G is called *complete*, if $E = \{\{x, y\} : \text{ for all } x, y \in N\}$.

[4] $D_1(\theta) = \int_0^\theta \frac{c/\theta}{\exp(x)-1}\,dx$ (Debye function)

A graph G is called *connected* if any two of its nodes are linked by a path in G. *Trees* are graphs, which are connected and do not contain cycles. They are characterized by the following theorem:

Theorem 2.30 (Characterization of trees.) *The following statements are equivalent for a graph $T = (N, E)$:*

(i) T is a tree.

(ii) Any two nodes of T are linked by a unique path in T.

(iii) T is minimally connected, i.e., T is connected but $T - e$ is disconnected for every edge $e \in E$. $T - e$ denotes the graph with removed edge e.

(iv) T is maximally acyclic, i.e., T contains no cycle but $T + \{x, y\}$ does contain a cycle for any two non-adjacent nodes $x, y \in N$. $T + \{x, y\}$ denotes the graph with additional edge $\{x, y\}$.

A tree with a *root node* v_0, i.e. a tree that has a node v_0 with $d(v_0) = |N| - 1$, is called *star*. It holds, that $d(v) = 1 \; \forall v \in N \backslash \{v_0\}$.

A *subgraph* of a graph $G = (N, E)$ is a graph $G^{'} = (N^{'}, E^{'})$ with $N^{'} \subseteq N$ and $E^{'} \subseteq E$. A subgraph $T = (N_T, E_T)$, which is a tree with $N_T = N$, is called *spanning tree* of a graph $G = (N, E)$.

2.4.2 Definition regular vines

A regular vine can be described as a nested set of trees, where the edges of tree i are the nodes of tree $i+1$, and where two edges in tree i are joined by an edge in tree $i+1$ only if they share a common node. It is based on pair copula constructions, seen in Section 2.2.3, i.e. edges will correspond to pair copulas which then build the joint density. Kurowicka and Cooke [2006] define regular vines in the following way:

Definition 2.31 (Regular vines.) *$\mathcal{V}$ is a regular vine on n elements if*

(i) $\mathcal{V} = (T_1, ..., T_{n-1})$.

(ii) $T_1 = (N_1, E_1)$ is a tree with nodes $N_1 = \{1, ..., n\}$. For $i = 2, ..., n-1, T_i = (N_i, E_i)$ is a tree with nodes $N_i = E_{i-1}$.

(iii) Two nodes in tree T_{i+1} can be joined by an edge only if the corresponding edges in tree T_i share a node, for $i = 1, ..., n-2$. (Proximity condition)

An example of a regular vine on 7 nodes is shown in Figure 2.10 (cp. Czado [2012]).

Since the number of possible regular vines on n nodes is still very large[6], two special cases of regular vines were recently studied, named *canonical vines* and *D-vines*[7] (see Aas et al. [2009]). They are defined as follows (due to Kurowicka and Cooke [2006]):

[6] Morales-Nápoles [2010] shows: $\binom{n}{2} \times (n-2)! \times 2^{\binom{n-2}{2}}$ possible regular vines on n nodes.

[7] D-vines were originally called "drawable" vines. "Canonical" vines are named due to the fact that sampling from such vines is most natural. (see Kurowicka and Cooke [2006])

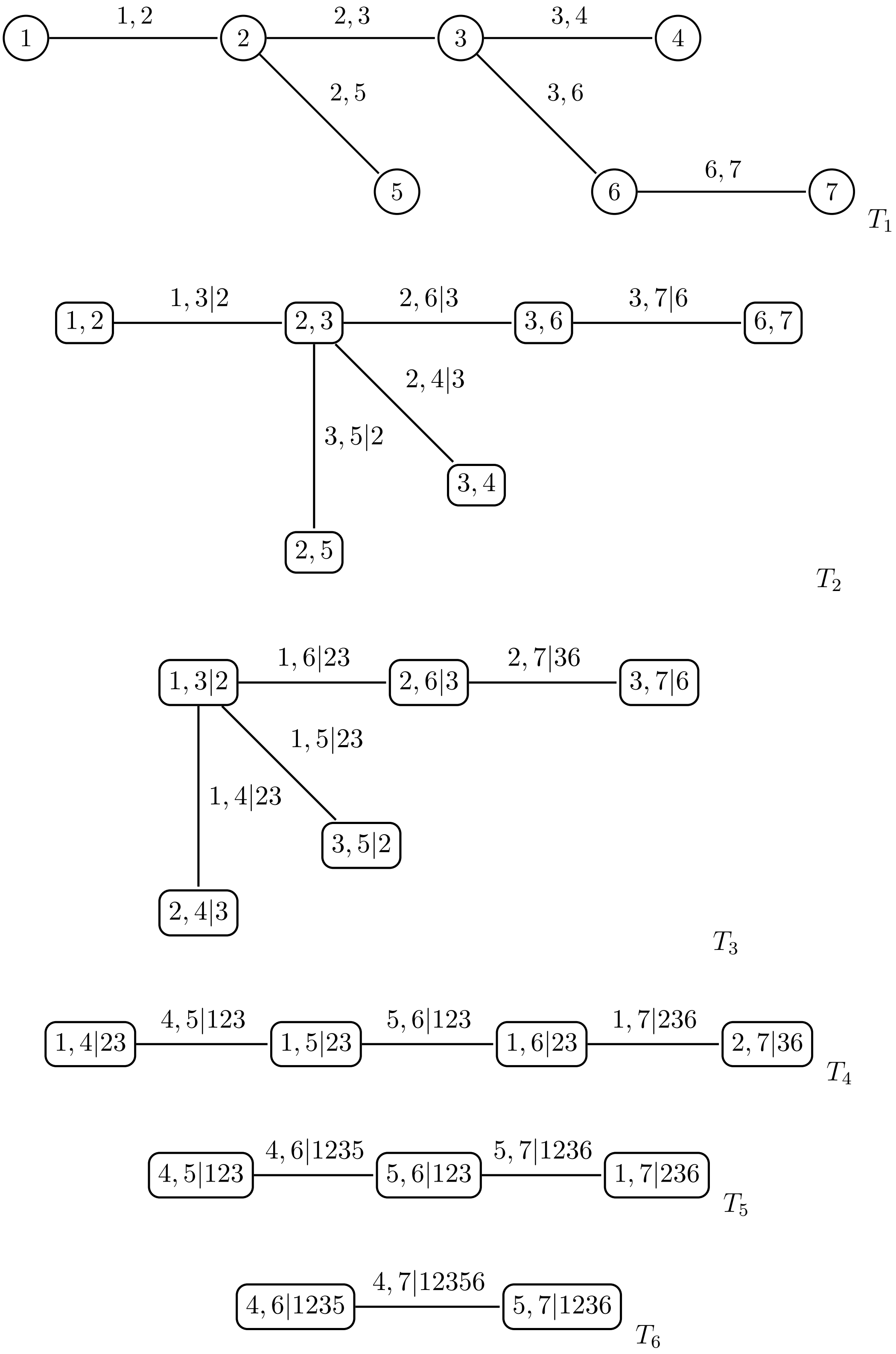


Figure 2.10: Example of a seven-dimensional R-vine.

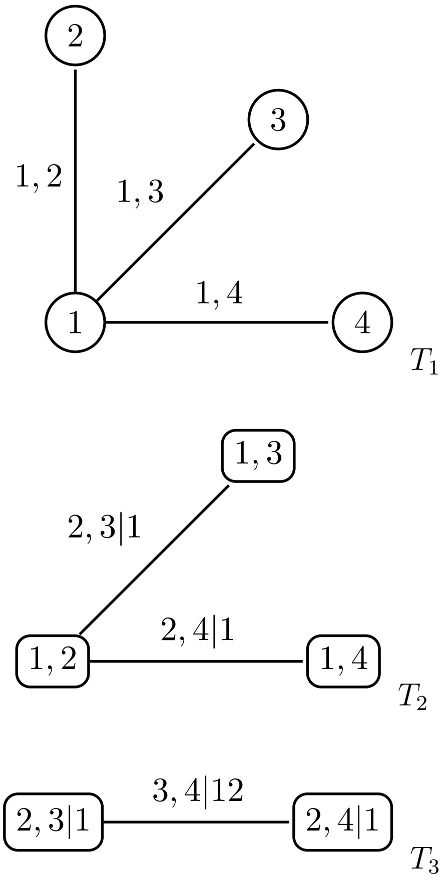


Figure 2.11: Example of a four-dimensional C-vine.

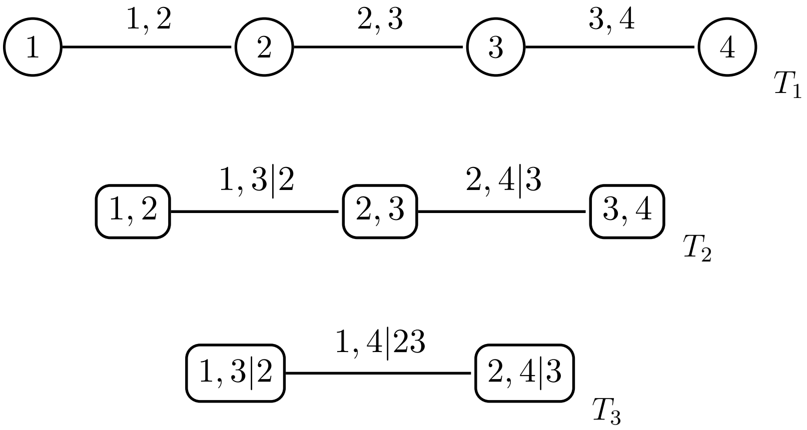


Figure 2.12: Example of a four-dimensional D-vine.

Definition 2.32 (Canonical vine, D-vine.) *A regular vine is called a*

(i) canonical vine if each tree T_i, $i = 1, ..., n-1$, is a star, i.e., if each tree T_i has a unique node, the root node, of degree $n-i$ (root node is connected to $n-i$ edges).

(ii) D-vine, if T_1 is a path, i.e., if each node in T_1 has a degree of at most 2.

Due to the proximity condition (iii) in Definition 2.31, it holds, that the first tree T_1 of a D-vine still determines all higher order trees $T_2, ..., T_{n-1}$ uniquely. The additional restrictions limit the number of different canonical or D-vines on n nodes to $n!/2$. In the following regular vines and canonical vines will be denoted as *R-vines* and *C-vines*, respectively. Examples of a C- and D-vine on 4 nodes respectively, are given in Figures 2.11 and 2.12.

Bedford and Cooke [2002] and Kurowicka and Cooke [2006] show that the edges of an R-vine can be uniquely identified by two nodes, called conditioned nodes and a set of conditioning nodes. The edges in tree T_i are identified by $jk|D$, where $j < k$ and D is the conditioning set. If $D = \varnothing$, then the corresponding edge is denoted by jk. The conditioned nodes $\{j, k\}$ are ordered here to propose an unique order of the arguments of the bivariate copluas, identified by the edges (see Czado [2010]).

Due to the proximity condition in Definition 2.31, the notation of an edge e in T_i will depend on the two edges in T_{i-1}, which have a common node in T_{i-1}. Denote these edges by $a = j(a), k(a)|D(a)$ and $b = j(b), k(b)|D(b)$ with $V(a) := \{j(a), k(a), D(a)\}$ and $V(b) := \{j(b), k(b), D(b)\}$, respectively. The nodes a and b in tree T_i are therefore joined by edge $e = j(e), k(e)|D(e)$, where

$$\begin{aligned} j(e) &:= \min\{i : i \in (V(a) \cup V(b)) \backslash D(e)\}, \\ k(e) &:= \max\{i : i \in (V(a) \cup V(b)) \backslash D(e)\}, \\ D(e) &:= V(a) \cap V(b). \end{aligned} \tag{2.62}$$

However, this unique order of the conditioned nodes is not necessary. It is made out of convenience.

For example the edge $e = 1,4|23$ in tree T_3 of Figure 2.10 is derived from edges $a = 1,3|2$ with $V(a) = \{1,2,3\}$ and $b = 2,4|3$ with $V(b) = \{2,3,4\}$. It holds, that $D(e) = \{2,3\}, j(e) = 1$ and $k(e) = 4$.

We build up a statistical model on a regular vine tree with node set $\mathcal{N} = \{N_1, ..., N_{n-1}\}$ and edge set $\mathcal{E} = \{E_1, ..., E_{n-1}\}$ by associating each edge $e = j(e), k(e)|D(e)$ in E_i with a bivariate copula density $c_{j(e),k(e)|D(e)}$. Let $\boldsymbol{X}_{D(e)}$ be the sub random vector of $\boldsymbol{X}$, indicated by the indices contained in $D(e)$. An *R-vine distribution* is defined as the distribution of the random vector $\boldsymbol{X} := (X_1, ..., X_n)$ with marginal densities f_k, $k = 1, ..., n$ and the conditional density of $(X_{j(e)}, X_{k(e)})$ given the variables $\boldsymbol{X}_{D(e)}$ specified as $c_{j(e),k(e)|D(e)}$ for the R-vine tree with node set $\mathcal{N}$ and edge set $\mathcal{E}$. Kurowicka and Cooke [2006] prove that the joint density of $\boldsymbol{X}$ is uniquely determined and given by

$$f(x_1, ..., x_n) = \prod_{r=1}^{n} f(x_r) \times \prod_{i=1}^{n-1} \prod_{e \in E_i} c_{j(e),k(e)|D(e)}(F(x_{j(e)}|\boldsymbol{x}_{D(e)}), F(x_{k(e)}|\boldsymbol{x}_{D(e)})), \tag{2.63}$$

where $\boldsymbol{x}_{D(e)}$ denotes the subvector of $\boldsymbol{x}$ indicated by the indices contained in $D(e)$. The joint density (2.63) is called *R-vine density*. Thus the corresponding *R-vine copula specification* is defined as

Definition 2.33 (R-vine copula specification.) $(\boldsymbol{F}, \mathcal{V}, B)$ *is an R-vine copula specification if* $\boldsymbol{F} = (F_1, ..., F_n)$ *is the vector of the continuous invertible marginal distribution functions of* $\boldsymbol{X} = (X_1, ..., X_n)$, $\mathcal{V}$ *is an n-dimensional R-vine and* $B = \{B_e | i = 1, ..., n-1; e \in E_i\}$ *is the set of the corresponding pair copulas.*

For the special case of C-vines, the conditioned set only depends on the tree level, i.e. $D(e) \equiv D_i \forall e \in E_i$. Assuming the order $1, ..., n$ and hence $D_i = \{1, ..., i-1\}$, a *C-vine density* can be written as

$$f(x_1, ..., x_n) = \prod_{r=1}^{n} f(x_r) \times \prod_{i=1}^{n-1} \prod_{j=1}^{n-i} c_{i,i+j|1:(i-1)} = \prod_{r=1}^{n} f(x_r) \times \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} c_{i,j|1:(i-1)}, \tag{2.64}$$

where $c_{j,k|i_1,...,i_m} = c_{j,k|i_1,...,i_m}(F(x_j|x_{i_1}, ..., x_{i_m}), F(x_k|x_{i_1}, ..., x_{i_m}))$.

Again assuming the order $1, ..., n$, in D-vines the conditioning sets of edges $e = (a, b)$ are always those nodes which lie between the nodes a and b in the first tree T_1. A *D-vine density* is given by

$$f(x_1, ..., x_n) = \prod_{r=1}^{n} f(x_r) \times \prod_{i=1}^{n-1} \prod_{j=1}^{n-i} c_{j,j+i|(j+1):(j+i-1)}. \tag{2.65}$$

Statistical inference for C- and D-vines has been discussed in Aas et al. [2009]. R-vines specifications were recently explored by Dißmann [2010] and Dißmann et al. [2011] which provides us with a basis of our model among meteorological variables later. In a next step, we introduce the concept of a so called *R-vine matrix* that gives a convenient representation of an R-vine, as we will use it to display the results of our model later.

2.4.3 Regular vine matrices

R-vine matrices (RVM) were introduced by Morales-Nápoles [2008]. However, we will use the notation introduced by Dißmann [2010].
Let $M = (m_{i,j})_{i,j=1,...,n} \in \{0, ..., n\}^{n \times n}$ be a lower triangular matrix.

(i) We denote the set of the non-zero entries in the i-th column of M by

$$L_M(i) = \{m_{i,i}, ..., m_{n,i}\}.$$

(ii) Further we define the following two sets

$$\begin{aligned} B_M(i) &= \{(m_{i,i}, D) : k = i+1, ..., n, D = \{m_{k,i}, ..., m_{n,i}\}\}, \\ \tilde{B}_M(i) &= \{(m_{k,i}, D) : k = i+1, ..., n, D = \{m_{i,i}\} \cup \{m_{k+1,i}, ..., m_{n,i}\}\}. \end{aligned}$$

With these notations, we can define an RVM.

Definition 2.34 (R-vine matrix.) *Let $M \in \{0, ..., n\}^{n \times n}$ be a lower triangular matrix. $M = (m_{i,j})_{i,j=1,...,n}$ is called R-vine matrix if it satisfies the following conditions:*

(i) $L_M(i) \subset L_M(j)$ *for* $1 \leq j < i \leq n$,

(ii) $m_{i,i} \notin L_M(i+1)$ *for* $i = 1, ..., n-1$, *and*

(iii) for $i = 1, ..., n-1$ *and for all* $k = i+1, ..., n-1$,

$$\begin{aligned}(m_{k,i}, \{m_{k+1,i}, ..., m_{n,i}\}) \in & \; B_M(i+1) \cup ... \cup B_M(n-1) \\ & \cup \tilde{B}_M(i+1) \cup ... \cup \tilde{B}_M(n-1).\end{aligned} \tag{2.66}$$

Condition (i) states that all entries of a column have to be contained in all columns on the left of this column. The second condition ensures that there is a new entry on the diagonal in each column. Together, it results that the variables are added to the RVM sequentially from the right to the left. Condition (iii) is caused by the proximity condition in Definition 2.31 and is however rather laborious to check for a given matrix. Nevertheless it is the critical condition of Definition 2.34 and it can be shown, that conditions (i) and (ii) follow from condition (iii). More Details can be found in Dißmann [2010]. But two simple properties of an R-vine matrix can be seen directly from the definition:

(i) All elements in a column are different.

(ii) Deleting the first row and column from an n-dimensional R-vine matrix gives an $(n-1)$-dimensional R-vine matrix.

But how to read from an RVM? RVMs are not unique, there are 2^{n-1} different RVM's possible which correspond to the same R-vine [Dißmann, 2010, Theorem 3.20]. We can construct an R-vine from a given RVM in the following way:

(i) The nodes of T_1 are given by $1, ..., n$.

(ii) The edges of T_1, and hence the nodes of T_2, are given by

$$\{\{m_{i,i}, m_{n,i}\} : i = 1, ..., n-1\},$$

(iii) The edges of T_2 (and nodes of T_3) are given by

$$\{\{m_{i,i}, m_{n-1,i} | m_{n,i}\} : i = 1, ..., n-2\}, \tag{2.67}$$

i.e. by the diagonal element and the second last element conditioned on the last element of columns $i = 1, ..., n-2$.

(iv) Generally, the edges of T_j (and node of T_{j+1}) are given by

$$\{\{m_{i,i}, m_{n-j+1,i} | m_{n-j+2,i}, ..., m_{n,i}\} : i = 1, ..., n-2\},$$

i.e. by the diagonal element and the element and the element in row $n-j+1$ conditioned on the last elements of columns $i = 1, ..., n-j$.

The best way to explore this, is to give an illustrative example:

Example 2 (Five-dimensional RVM.) *Let M an RVM, given by (the zero entries in the upper triangle are omitted for simplicity)*

$$M = \begin{pmatrix} 4 & & & & \\ 5 & 3 & & & \\ 3 & 5 & 5 & & \\ 2 & 1 & 2 & 2 & \\ 1 & 2 & 1 & 1 & 1 \end{pmatrix}. \tag{2.68}$$

It can be checked that the conditions of Definition 2.34 are satisfied. Following our procedure described above, we construct the following R-vine:

(i) Edges of tree T_1 and nodes of Tree T_2: $\{m_{1,1}, m_{5,1}\} = \{1,4\}$, $\{m_{2,2}, m_{5,2}\} = \{2,3\}$, $\{m_{3,3}, m_{5,3}\} = \{1,5\}$ and $\{m_{4,4}, m_{5,4}\} = \{1,2\}$.

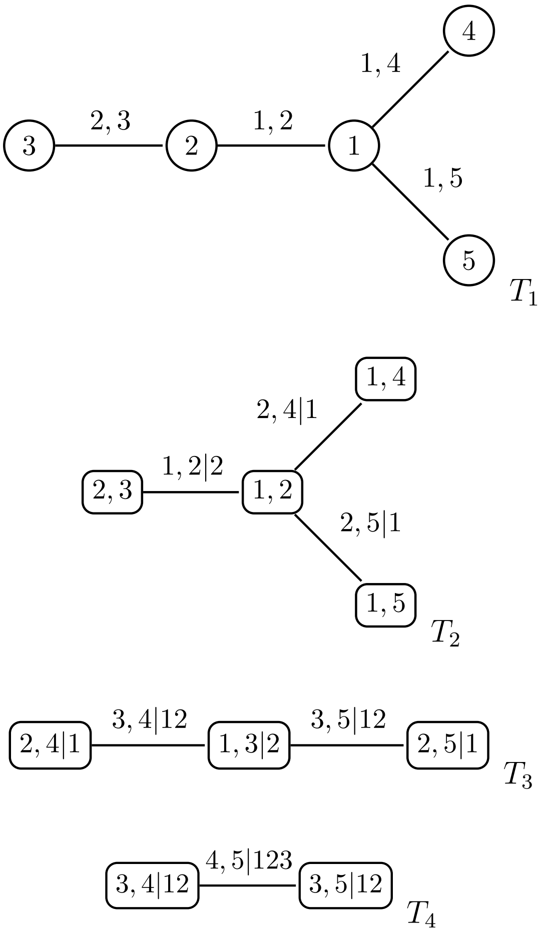


Figure 2.13: Five-dimensional R-vine corresponding to Example 2.

(ii) *According to (2.68), the edges of tree T_2 and nodes of tree T_3 are given by*

$$\{m_{1,1}, m_{4,1} | m_{5,1}\} = \{2, 4|1\} \left\{ \begin{array}{lllll} \boxed{4} & & & & \\ 5 & 3 & & & \\ 3 & 5 & 5 & & \\ \boxed{2} & 1 & 2 & 2 & \\ \hline \boxed{1} & 2 & 1 & 1 & 1 \end{array} \right.$$

and similarly for the 2nd and 3rd column: $\{m_{2,2}, m_{4,2} | m_{5,2}\} = \{1, 3|2\}$ *and* $\{m_{3,3}, m_{4,3} | m_{5,3}\} = \{2, 5|1\}$.

(iii) *The edges of tree T_3 and the nodes of tree T_4 follows*

$$\{m_{1,1}, m_{3,1} | m_{4,1}, m_{5,1}\} = \{3, 4|1, 2\} \left\{ \begin{array}{lllll} \boxed{4} & & & & \\ 5 & 3 & & & \\ \boxed{3} & 5 & 5 & & \\ \hline \boxed{2} & 1 & 2 & 2 & \\ \boxed{1} & 2 & 1 & 1 & 1 \end{array} \right.$$

and $\{m_{2,2}, m_{3,2} | m_{4,2}, m_{5,2}\} = \{3, 5|1, 2\}$.

(iv) *Finally, the edge of tree T_4 is given by* $\{m_{1,1}, m_{2,2} | m_{3,1}, m_{4,1}, m_{5,1}\} = \{4, 5|1, 2, 3\}$.

RVMs of C- and D-vines can be represented by particularly well-structured RVMs. For details, see Brechmann [2010].

The chosen copula types and parameters belonging to an R-Vine, constructed by pair copulas and expressed by an RVM, can easily be denoted in matrix form as well. The corresponding pair copula types and parameters can be set in the corresponding off-diagonal entries, since the diagonal entry in each column defines one element of the conditioned sets of the edges corresponding to this column uniquely. This means, that copula type and parameter(s) corresponding to $\{m_{i,i}, m_{k,i} | m_{k+1,i}, ..., m_{n,i}\}$, $k > i$, are stored in the (k, i)-th entry. We illustrate that in the following example:

Example 3 (R-vine copula type and parameter matrices.) *Consider the R-vine of Example 2, defined by the RVM in (2.68). Further, we have the following R-vine copula type and parameter matrices T and P_1:*

$$T = \begin{pmatrix} & & & & \\ N & & & & \\ \Pi^2 & G & & & \\ C & N & \Pi^2 & & \\ J & C & N & F & \end{pmatrix}, P_1 = \begin{pmatrix} & & & & \\ 0.23 & & & & \\ 0.01 & 1.54 & & & \\ 0.77 & 0.35 & 0.00 & & \\ 0.90 & 2.87 & 0.77 & 1.28 & \end{pmatrix},$$

where Π^2 denotes a bivariate independence copula, N a Gaussian, C a Clayton, G a Gumbel, F a Frank and J a Joe copula. Then we can identify, e.g., the copula type and *parameter of the edge* $1,3|2$ *as described above:*

$$M = \begin{pmatrix} 4 & & & & \\ 5 & \mathbf{3} & & & \\ 3 & 5 & 5 & & \\ 2 & \boxed{1} & 2 & 2 & \\ 1 & \mathbf{2} & 1 & 1 & 1 \end{pmatrix}, T = \begin{pmatrix} & & & \\ N & & & \\ \Pi^2 & G & & \\ C & \boxed{N} & \Pi^2 & \\ J & C & N & F \end{pmatrix}, P_1 = \begin{pmatrix} & & & \\ 0.23 & & & \\ 0.01 & 1.54 & & \\ 0.77 & \boxed{0.35} & 0.00 & \\ 0.90 & 2.87 & 0.77 & 1.28 \end{pmatrix},$$

i.e. the copula of edge $1,3|2$ *is a Gaussian copula with parameter* 0.35*. Other copula types and parameters are identified similarly.*

Note: If a copula belongs to a two-parametric family such as the t or Clayton-Gumbel copulas, one needs a second copula parameter matrix P_2 *to specify the corresponding second parameters.*

2.4.4 Selecting regular vine distributions

According to Dißmann et al. [2011], fitting an R-vine copula specification to a given dataset, requires three separate tasks:

(1.) Selection of the R-vine (structure), i.e. selecting which unconditioned and conditioned pairs to use.

(2.) Choice of a bivariate copula family for each pair selected in (1.).

(3.) Estimation of the corresponding parameter(s) for each copula.

The most intuitive way of finding the "best" model would be to accomplish steps (2.) and (3.) for all possible R-vine constructions. But this is not feasible, since the number of possible R-vines on n variables increases very rapidly with n, as we mentioned above. Other approaches are based on manual interpretation of plots, e.g. K- or Chi-Plots (cp. Genest and Favre [2007]), to decide which bivariate copula family to use. But this is again not feasible to do for every possible copula in every possible R-vine decomposition, especially in higher dimensions. In addition, such methods do not guarantee objectivity.

Therefore, Dißmann et al. [2011] developed a sequential, heuristic method to select the tree structure of the R-vine, the so called *sequential method.* They start by defining the first tree $T_1 = (N_1, E_1)$ for the R-vine, continuing with the second tree and so on. So the proposed method for (1.) depends on the copulas selected in (2.) and estimated in (3.) one step before. In detail, it looks as follows.

Sequential method to select an regular vine copula specification based on Kendall's τ

For selecting one possible R-vine for a given dataset it is necessary to decide for which pairs of variables you want to specify appropriate copulas. As discussed, therefore the trees are selected sequentially in such a way that the chosen pairs model the strongest pairwise dependencies present [Dißmann et al., 2011, p. 13–14]. One uses Kendall's τ as instrument to measure the strongest dependence, since it measures dependence independently of the assumed distribution and hence, is especially useful when combining different

Algorithm 1 Sequential method to select an R-vine model based on Kendall's τ.

```
Input: Data $(x_{l1}, ..., x_{ln})$, $l = 1, ..., N$ (realizations of i.i.d. random vectors).
Output: R-vine copula specification, i.e. $\mathcal{V}$, $B$.
Calculate the empirical Kendall's tau $\hat{\tau}_{j,k}$ for all possible variables pairs $\{j,k\}, 1 \leq j < k \leq n$.
Select the spanning tree that maximizes the sum of absolute empirical Kendall's taus, i.e.
    $$\max \sum_{e=\{j,k\} \text{ in spanning tree}} |\hat{\tau}_{j,k}| .$$
For each edge $\{j,k\}$ in the selected spanning tree, select a copula and estimate the corresponding parameter(s). Then transform $\widehat{F}_{j|k}(x_{lj}|x_{lk})$ and $\widehat{F}_{k|j}(x_{lk}|x_{lj})$, $l = 1, ..., N$, using the fitted copula $\widehat{C}_{jk}$ (see (2.53)).
for $i = 2, ..., n-1$ do
    Calculate the empirical Kendall's tau $\hat{\tau}_{j,k|D}$ for all conditional variable pairs $\{j,k|D\}$ that can be part of tree $T_i$, i.e. all edges fulfilling the proximity condition (see Definition 2.31).
    Among these edges, select the spanning tree that maximizes the sum of absolute empirical Kendall's tau, i.e.
        $$\max \sum_{e=\{j,k|D\} \text{ in spanning tree}} \left|\hat{\tau}_{j,k|D}\right| .$$
    For each edge $\{j,k|D\}$ in the selected spanning tree, select a conditional copula and estimate the corresponding parameter(s). Then transform $\widehat{F}_{j|k\cup D}(x_{lj}|x_{lk}, \boldsymbol{x}_{lD})$ and $\widehat{F}_{k|j\cup D}(x_{lk}|x_{lj}, \boldsymbol{x}_{lD})$, $l = 1, ..., N$, using the fitted copula $\widehat{C}_{jk|D}$ (see (2.69)).
end for
```

(non-Gaussian) copula families.[8] The sequential method is described by Algorithm 1.

Note: Since one examines every tree seperately, it is not guaranted to find a global optimum.[9] Nevertheless Dißmann et al. [2011] think their approach is reasonable, since the copulas specified in the first tree of the R-vine often have the greatest influence on the model fit. Further, it is more important to model the dependence structure between random variables that have high dependencies correctly, because most copulas can model independence and the copulas distribution functions for parameters close to independence are very similar. In addition, their approach minimizes the influence of rounding errors in later trees, which pairs with strong pairwise dependence are most prone to and for real applications it is natural to assume that randomness is driven by the dependence of only some variables and not all. For further detail we refer to [Dißmann et al., 2011, p. 13].

[8] However, Brechmann [2010] discusses that the described method also works in the same way for every other measure of dependence.

[9] Global optimum is meant in terms of model fit, e.g., higher likelihood, smaller AIC/BIC or superior in terms of the likelihood-ratio based test for comparing non-nested models proposed by Vuong [1989].

We use a *maximum spanning tree (MST) algorithm*, such as the Algorithm of Prim [Cormen et al., 2009, Section 23.1], to select the tree that maximizes the sum of absolute empirical Kendall's taus (Steps 2 and 6 in Algorithm 1).[10]
So in Steps 2 and 6 of Algorithm 1 we are looking for a tree, hence we could look also for a star or a path instead, to obtain a C- or a D-vine structure, respectively.

A proof, that this algorithm creates an R-vine can be found in [Dißmann et al., 2011, p. 15]. In the next subsection, we will explore how to select the appropriate pair-copula families with corresponding parameters sequentially in Steps 3 & 7.

But how to calculate $\widehat{F}_{j|k\cup D}(x_{lj}|x_{lk}, \boldsymbol{x}_{lD})$ and $\widehat{F}_{k|j\cup D}(x_{lk}|x_{lj}, \boldsymbol{x}_{lD})$, $l = 1, ..., N$ in Steps 3 & 7 in Algorithm 1? Let $E_i^{'}$ be the set of all possible edges in Tree T_i due to the proximity condition. For all $e \in E_i^{'}$ we have to calculate the value of Kendall's τ (Steps 1 & 5 in Algorithm 1). So if $e \in E_i^{'}, e = \{a, b\} = j(e), k(e)|D(e)$, as defined in (2.62) with $a = j(a), k(a)|D(a)$ and $b = j(b), k(b)|D(b)$ respectively, connects variables $x_{j(e)}$ with $x_{k(e)}$ given the variables $\boldsymbol{x}_{D(e)}$, we hence need the transformed variables $\widehat{F}_{j(e)|D(e)}(x_{j(e)}|\boldsymbol{x}_{D(e)})$ and $\widehat{F}_{k(e)|D(e)}(x_{k(e)}|\boldsymbol{x}_{D(e)})$ (cp. to Steps 3 & 7 in Algorithm 1). They are calculated as described in Equation (2.53), i.e, w.l.o.g. $j(e) = j(a)$, then

$$\begin{aligned}\widehat{F}_{j(e)|D(e)}(x_{j(e)}|\boldsymbol{x}_{D(e)}) &= \frac{\partial C_{j(a),k(a)|D(a)}\left(\widehat{F}_{j(a)|D(a)}(x_{j(a)|D(a)}|\boldsymbol{x}_{D(a)}), \widehat{F}_{k(a)|D(a)}(x_{k(a)|D(a)}|\boldsymbol{x}_{D(a)})\right)}{\partial \widehat{F}_{k(a)|D(a)}(x_{k(a)|D(a)}|\boldsymbol{x}_{D(a)})} \\ &=: h\left(\widehat{F}_{j(a)|D(a)}(x_{j(a)|D(a)}|\boldsymbol{x}_{D(a)}), \widehat{F}_{k(a)|D(a)}(x_{k(a)|D(a)}|\boldsymbol{x}_{D(a)})\right), \qquad (2.69)\end{aligned}$$

where $\widehat{F}_{j(a)|D(a)}(x_{j(a)|D(a)}|\boldsymbol{x}_{D(a)})$ and $\widehat{F}_{k(a)|D(a)}(x_{k(a)|D(a)}|\boldsymbol{x}_{D(a)})$ were obtained before recursively in the same way by Algorithm 1.

For these it is then straightforward to calculate the empirical Kendall's τ. Then we proceed with Step 2 and for the edges selected in the MST, we need to fit a copula based on two conditioned variables again. The latter point is outlined in the following section. An exemplification of Algorithm 1 is given in Table 2.5.

Selecting pair-copula families sequentially

Due to Algorithm 1, we need to select a copula family for every pair of variables, tree by tree. The choice of families therefore is based on the presented pair copulas in Section 2.2.3 (except the two-parameter Archimedean copulas, such as Clayton-Gumbel copulas). Then we proceed as follows:

1. In case of positive dependence, one can select among the Gaussian, t-, (survival) Gumbel, (survival) Clayton, Frank and (survival) Joe copulas. If one models negative

[10] Typically such algorithms are described to find a *minimal* spanning tree. But the algorithms work in both ways. Further, an MST algorithm does not depend on the actual values of the edges, instead it only uses their ranks. Therefore we would get the same tree even if we took other weights, i.e. transformed edge values by a monotone increasing function, like squared taus [Dißmann et al., 2011, p. 14].

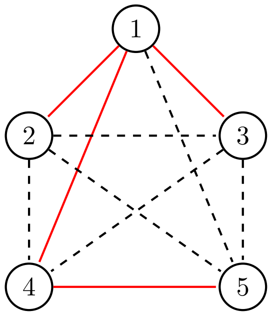
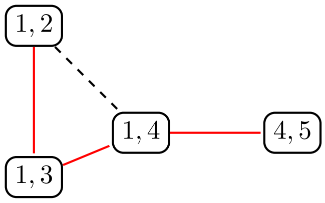
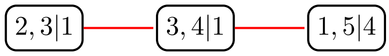
i	Graph	Description
1		Assume that we have 5 variables $N_1 = \{1,2,3,4,5\}$. The first graph is always a complete graph, where we can connect every node with every other node. Let us assume the Algorithm of Prim selects the solid edges. The concrete edge values (Kendall's taus) are not of interest in this example.
2		All edges from the previous step are now nodes. An edge is drawn (dashed and solid) whenever the nodes share a common node in the previous tree (proximity condition). The graph is again connected and the now selected tree is indicated by the solid edges.
3		Here we need all edges to form a tree, therefore there are no options in this step. Interesting: As soon as a graph has a D-vine structure, there are no more options in the following trees, since it uniquely determines all following conditioned and conditioned sets (cp. Definition 2.32).

Table 2.5: Exemplification of the model selection Algorithm 1.

dependence, we have the choice among Gaussian, t-, rotated Gumbel (by 90 or 270 degrees), rotated Clayton (by 90 or 270 degrees), Frank and rotated Joe copulas (by 90 or 270 degrees) respectively.

2. Then the corresponding parameters are estimated by maximum likelihood estimation. Note: If the estimation of the degrees of freedom in case of a t-copula leads to a value of 30 or higher, then the t-copula is very close to the Gaussian, which can be used instead.

3. To find the copula, which fits "best", we use the AIC[11] (Akaike [1973]). It corrects the log likelihood of a copula for the number of parameters. Hence, the use of the t-copula is penalized compared to the other ones, since it is the only two parameter family under consideration.

[11] Bivariate copula selection using AIC has been investigated in Manner [2007] and [Brechmann, 2010, Section 5.4] who found that it is a quite reliable criterion, in particular, in comparison to alternative criteria, such as copula goodness-of-fit tests.

Evaluation of the joint R-vine density and simulation of an R-vine specification

Given an R-vine matrix, introduced in Section 2.4.3, we are interested in building the corresponding R-vine density of the specified R-vine distribution. Further, in a next step, we would like to simulate from this distribution. However, these are non-trivial tasks, since the order of the conditioning variables required is not obvious. But Dißmann et al. [2011] introduce two efficient algorithms which can handle our purposes. Nevertheless, we skip any details here, since the extensive technical code is not relevant for our work. So we rather refer to [Dißmann et al., 2011, pp. 8-12] for a deeper insight.

For our application we will use the R-package `VineCopula` implemented by Schepsmeier et al. [2012]. It is based on the presented theory in the previous sections and models, selects and samples from an appropriate R-vine specification for a given dataset.

2.5 Regression models

To build an R-vine model for a given dataset, as it will be the set of meteorological measurements in our case, we have to find out the marginal distributions of each variable first. This is necessary in the context of fitting the appropriate pair copulas of the selected pairs of variables in an R-vine.

The behavior of meteorological variables at a point in time t depends on the behavior of the corresponding meteorological variables in the past, i.e. at time $t-1$, $t-2$, and so on. So the variables include *autoregressive* parts, which we have to model. Further, each variable will exhibit an seasonal pattern, hence, e.g., the daily mean temperature will likely be higher in summer than in winter. We will capture these properties with several *regression models* for the different variables to model their marginal distributions (in detail, see Chapter 3).

In this section we introduce the definitions of different regression models, their theoretical properties, their methods for parameter estimation and their goodness of fit measures. We start with linear models, needed to model the distribution of temperature and air pressure variables. Then we introduce the beta regression (for modeling humidity) and end with the introduction of generalized linear models such as the binomial and gamma regression, needed to fit the behavior of precipitation.

2.5.1 Linear models

Certainly the most famous regression model is presented by the *linear model*. Here we want to model a random variable Y, called *response*, in terms of k known predictors $x_1, ..., x_k$, denoted as *covariates*.

The linear model explains the response as a linear function of the predictors, i.e.

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon, \tag{2.70}$$

where we have $p := k+1$ regression parameters: β_0 (intercept parameter) and parameters $\beta_1, ..., \beta_k$ according to the covariates; ϵ is a random error variable. The unknown regression parameters have to be estimated from n observations

$$(y_i, x_{i1}, ..., x_{ik}), i = 1, ..., n,$$

where y_i are the observed values of the random variables Y_i. Then, our linear model is build on the following assumptions:

Definition 2.35 (Assumptions of the linear model.)

1. ***Linearity:*** *There is a linear relationship between the random response Y_i and the covariate vector $\boldsymbol{x_i}$ of the form*

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \ i = 1, ..., n, \tag{2.71}$$

   where ϵ_i is a random variable with $E[\epsilon_i] = 0, \ \forall i \in \{1, ..., n\}$.

2. ***Independence:*** *The random variables ϵ_i are independent.*

3. ***Variance homogeneity:*** *The random variables ϵ_i have constant variances for all i, i.e.*

$$Var(Y_i) = Var(\epsilon_i) = \sigma^2, \ \forall i \in \{1, ..., n\}. \tag{2.72}$$

4. ***Normality:*** *The random variables ϵ_i are normal distributed, i.e. together with 2. and 3., ϵ_i are iid with*

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2), \ \forall i \in \{1, ..., n\}.$$

The linear regression model of Definition 2.35 is often represented in matrix-vector notation. Therefore we define the *design matrix* X, containing the n observations of the covariates in its rows

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ & & \vdots & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

The vectors of random variables Y_i and ϵ_i as well as the regression coefficients are represented by

$$\begin{aligned} \boldsymbol{Y} &= (Y_1, Y_2, ..., Y_n)' \in \mathbb{R}^n, \\ \boldsymbol{\epsilon} &= (\epsilon_1, \epsilon_2, ..., \epsilon_n)' \in \mathbb{R}^n, \\ \boldsymbol{\beta} &= (\beta_0, \beta_1, ..., \beta_k)' \in \mathbb{R}^p. \end{aligned}$$

Then the model can be formulated as

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \text{ with } \boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 I_n), \tag{2.73}$$

where $\boldsymbol{0}$ is the n-dimensional null vector and I_n the $n \times n$ identity matrix; $\mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ denotes the n-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ (cp. to Section 2.1.3).

Thus, it follows that

$$E[\boldsymbol{Y}] = X\boldsymbol{\beta} \text{ and } Var(\boldsymbol{Y}) = \sigma^2 I_n,$$

and hence

$$\boldsymbol{Y} \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 I_n). \tag{2.74}$$

The expectation of Y_i is a linear function of the unknown regression parameters, i.e.

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} \text{ for all } i.$$

Parameter estimation

So we are interessted in estimating the unknown model parameters $\boldsymbol{\beta}$ and σ, i.e obtaining estimates

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k) \in \mathbb{R}^p, \text{ and } \hat{\sigma},$$

based on n observations. Usually one uses either the *least-squares estimation* or *maximum likelihood estimation.* In the case of linear models under the assumption of independence, homogeneity and normality, both techniques yield the same estimate for $\boldsymbol{\beta}$ (Czado and Schmidt [2011]). The only difference results in the fact that the maximum likelihood estimation provides also an estimate for σ and additional statistical analysis in form of prediction intervals and tests [McCulloch and Searle, 2001, p. 116].

In the **least squares estimation**, one does not make any distributional assumptions on the response variable Y_i. Our goal is to find regression coefficients $\hat{\boldsymbol{\beta}}$ such that the sum of *raw residuals*

$$r_i = y_i - \hat{y}_i \tag{2.75}$$

is minimized, where $\hat{y}_i$ are the *fitted values*, i.e.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}, \tag{2.76}$$

for all $i \in \{1, ..., n\}$.

The method of least squares minimizes the sum of squared residuals, i.e.

$$\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{y}) := \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (r_i)^2. \tag{2.77}$$

Then,

$$\min_{\boldsymbol{\beta}} \mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{y}) \Leftrightarrow \frac{\partial \mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \boldsymbol{0} \Leftrightarrow X^{'} X \boldsymbol{\beta} = X^{'} \boldsymbol{y}. \tag{2.78}$$

The right hand side of (2.78) is called *normal equation.* Thus, if the matrix X is of full rank p, the minimum of $\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{y})$ and so the estimate of $\boldsymbol{\beta}$[12], is obtained by

$$\hat{\boldsymbol{\beta}} = (X^{'} X)^{-1} X^{'} \boldsymbol{y}. \tag{2.79}$$

[12] The least squares solution has the following geometric interpretation: The vector of fitted values $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = X(X^{'}X)^{-1}X^{'}\boldsymbol{y}$ is the projection of $\boldsymbol{y}$ onto the linear space, that is spanned by the columns of X. $H := X(X^{'}X)^{-1}X^{'}$ denotes the corresponding projection matrix [Czado and Schmidt, 2011, pp. 201-203].

In addition, we define

$$\hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\boldsymbol{Y}. \tag{2.80}$$

In contrast, the **maximum likelihood estimation** uses the assumptions that the random variables Y_i are normally distributed (see (2.73)). So the likelihood of $(\boldsymbol{\beta}, \sigma)$ given $\boldsymbol{y}$ is given by

$$L(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left\|\boldsymbol{y} - X\boldsymbol{\beta}\right\|^2\right\}. \tag{2.81}$$

One directly sees, that (2.81) is maximized when $\left\|\boldsymbol{y} - X\boldsymbol{\beta}\right\|^2$ is minimized. Hence, the *maximum likelihood estimate (MLE)* of the regression parameter $\boldsymbol{\beta}$ under the normality assumption is equal to the least squares estimate in (2.79).

The MLE for variance σ^2 equals

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{2.82}$$

However, an unbiased estimator for σ^2 is given by

$$s^2 := \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{2.83}$$

Goodness of fit

We assume, that the assumptions of Definition 2.35 are fulfilled. We are now interested in the goodness of fit of our linear model (2.71). Therefore we define the *multiple coefficient of determination.*

Definition 2.36 (Multiple coefficient of determination.) *We define the multiple coefficient of determination R^2 as*

$$R^2 := \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

where

$$\begin{aligned} SST &:= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \text{ total sum of squares,} \\ SSR &:= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \text{ regression sum of squares,} \\ SSE &:= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \text{ sum of squares error,} \end{aligned}$$

with $\bar{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Further, we define the adjusted multiple coefficient of determination R^2_{adj} as

$$R^2_{adj} := 1 - \frac{SSE/(n-p)}{SST/(n-1)}. \tag{2.84}$$

One can show, that $SST = SSR + SSE$ [Czado and Schmidt, 2011, p. 217]. For the interpretation of R^2 we have to look at the total sum of squares SST, since it is an estimator of the variance of the random responses Y_i (only the factor $1/(n-1)$ is missed). Thus, R^2 explains the proportion of the response variability, which is explained by our regression model. It holds that $R^2 \in (0, 1)$ and the closer it is to 1, the better the model explains the variability of the response. But, it can be shown that if one adds additional covariates to the model, then R^2 always increases. In contrast to that fact, R^2_{adj} also consider the number of estimated regression parameters and we can compare the goodness of fit of models with different numbers of covariates.

Further, we are often interested in statistical inference for the model parameters. Especially to test the following hypothesis for a single parameter:

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0, \text{ for a fixed } j \in \{0, ..., k\}.$$

This yields to the following test statistic (see, e.g., McCulloch and Searle [2001])

$$T_j := \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)} \overset{H_0}{\sim} t_{n-p}, \tag{2.85}$$

where $\hat{se}(\hat{\beta}_j) := s\sqrt{((X'X)^{-1})_{jj}}$ is the estimated standard error of $\hat{\beta}_j$. So we reject the null hypothesis H_0 at a level of significance α, if $|T_j| > t^{-1}_{n-p,1-\alpha/2}$, where $t^{-1}_{n-p,1-\alpha/2}$ is the $100(1-\alpha/2)\%$ quantile of a t-distribution with $n-p$ degrees of freedom (cp. Definition 2.5 of a t-distribution). In this case we can assume that $\beta_j \neq 0$ significantly, i.e. covariate x_{ij} has an significant influence on response Y_i.
For further goodness of fit measures, such as the analysis of variance (ANOVA) or diagnostic plots, we refer to McCulloch and Searle [2001] or Czado and Schmidt [2011].

2.5.2 Linear skew normal and skew t regression

Our definition of *linear skew normal and skew t regressions* differs only in point 4 of Definition 2.35 for linear models. Here, the random error term follows either a skew normal or a skew t distribution instead of a normal distribution. This yields to the following definition:

Definition 2.37 (Assumptions of the linear skew normal/skew t regression.)

1. ***Linearity:*** *There is a linear relationship between the random response Y_i and the covariate vector $\boldsymbol{x_i}$ of the form*

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \; i = 1, ..., n, \tag{2.86}$$

where ϵ_i is a random variable with $E[\epsilon_i] = 0$, $\forall i \in \{1, ..., n\}$.

2. ***Independence:*** *The random variables ϵ_i are independent.*

3. ***Variance homogeneity:*** *The random variables ϵ_i have constant variances for all i, i.e.*

$$Var(Y_i) = Var(\epsilon_i) = \sigma^2, \; \forall i \in \{1, ..., n\}. \tag{2.87}$$

*4. **Skew normal/skew** t: The random variables ϵ_i are either skew normal distributed for all $i = 1, ..., n$ or skew t distributed for all $i = 1, ..., n$. Hence, together with 2. and 3., ϵ_i are iid with*

$$\epsilon_i \sim \mathcal{SN}(\xi, \omega, \alpha), \ \forall i \in \{1, ..., n\}, \ \textit{or}$$

$$\epsilon_i \sim \textit{skewt}(\xi, \omega, \alpha, \nu), \ \forall i \in \{1, ..., n\}.$$

The skew normal and skew t distribution correspond to their definitions in Section 2.1.

Parameter estimation

To estimate our parameters $\boldsymbol{\beta}, \xi, \omega, \alpha$, respectively, $\boldsymbol{\beta}, \xi, \omega, \alpha, \nu$, we proceed in 2 steps:

1. Similar to the linear models we are using the method of least squares to estimate our parameter vector $\hat{\boldsymbol{\beta}}$. Since this method does not depend on any distributional assumptions we get similar to (2.77)

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}.$$

2. We take the raw residuals $r_i = y_i - \hat{y}_i$ and use the method of maximum likelihood to estimate $\hat{\xi}, \hat{\omega}, \hat{\alpha}$ and $\hat{\xi}, \hat{\omega}, \hat{\alpha}, \hat{\nu}$ respectively. Since there exist no closed-form expressions for the estimates, except for $\alpha = 0$, the maximum likelihood estimation has to be computed numerically. Details can be found in [Azzalini and Capitanio, 1999, pp. 12-14] and Bowman and Azzalini [1997] as well as in [Azzalini and Capitanio, 2003, pp. 19-20].

We will use the corresponding R-functions `sn.mle` and `st.mle` from R-library `sn`, see Azzalini [2011], to fit appropriate distributions.

Goodness of fit

It is clear that we cannot use the goodness of fit measures of linear models, since they depend on the assumption of normally distributed error terms. Thus, in case of our skew normal/skew t regression, we will restrict the goodness of fit analysis to two measures:

1. Ljung-Box tests to test the assumption 2 of independent error terms in Definition 2.37 (details for that kind of testing will be presented in Section 2.6) and

2. Quantile-quantile (Q-Q) plots to validate the distributional assumption 4 in Definition 2.37, i.e. comparing the empirical quantiles of the raw residuals with the theoretical ones of either the fitted $\mathcal{SN}$ or fitted *skewt* distribution. Deviations from a straight line indicate a violation from the distributional assumption.

Note: We will also use both measures in case of linear models, since they can also be applied for them.

Weighted least squares

In practice, the assumption of variance homogeneity (3.) in Definition 2.37 is often violated. The variance is then rather dependent on i, i.e. we assume

$$Var(\epsilon_i) = \sigma^2 \cdot w_i, \tag{2.88}$$

with different $w_i > 0 \ \forall i \in \{1, ..., n\}$. The w_i's are called *weights* and one assumes, that they are known. This kind of regression is called *heteroskedastic*. If we set

$$Z_i := \frac{Y_i}{\sqrt{w_i}} = \frac{1}{\sqrt{w_i}}(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}) + \epsilon_i^*,$$

where $\epsilon_i^* = \epsilon_i w_i^{-1/2}$ it follows, that

$$Var(\epsilon_i^*) = \frac{1}{w_i} Var(\epsilon_i) = \frac{1}{w_i} w_i \sigma^2 = \sigma^2.$$

Thus, all assumptions of Definition 2.37 are fulfilled. We define

$$W := \begin{pmatrix} w_1 & 0 & 0 & \dots & 0 \\ 0 & w_2 & 0 & \dots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \dots & 0 & w_n \end{pmatrix} \in \mathbb{R}_+^{n \times n}, \tag{2.89}$$

and $\boldsymbol{Z} := (\frac{Y_1}{\sqrt{w_1}}, ..., \frac{Y_n}{\sqrt{w_n}})' = W^{-\frac{1}{2}} \boldsymbol{Y}$. Then we can find the estimator $\hat{\boldsymbol{\beta}}_{WLS}$ by using the method of least squares from (2.77). We get

$$\hat{\boldsymbol{\beta}}_{WLS} = (X' W^{-1} X)^{-1} X' W^{-1} \boldsymbol{y}, \tag{2.90}$$

where $\boldsymbol{y} = (y_1, ..., y_n)'$ is the sample of n observations of the response. This method is called *weighted least squares*[13]. We then consider the residuals $\boldsymbol{r}_W = W^{-\frac{1}{2}}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{WLS})$ to fit the appropriate distribution.

2.5.3 Beta regression

Now we want to define a regression model for beta distributed random variables. Typically for such an regression analysis it is useful to model the mean of the response [Ferrari and Cribari-Neto, 2004, pp. 802-803]. Therefore one uses the parameterization (2.28) of the beta density, introduced in Section 2.1.8. So if $Y \sim Beta(\mu, \phi)$, it follows that

$$E[Y] = \mu \text{ and } Var(Y) = \frac{\mu(1-\mu)}{1+\phi}. \tag{2.91}$$

The parameter ϕ can be interpreted as a *precision parameter*, i.e., for fixed μ, the larger the value of ϕ, the smaller the variance of Y; $1/\phi$ is called *dispersion parameter*.

[13] This definition is based on Czado and Schmidt [2011].

Note also that a beta regression model based on the standard parameterization $Beta(a, b)$ (cp. (2.26)) was proposed by Vasconcellos and Cribari-Neto [2005].

We assume in Section 2.1.8 that the beta distributed response is constrained to the standard unit interval $(0, 1)$. Hence, if a response is restricted to the interval (l, u) with known l, u and $l < u$, one simply models $(Y - l)/(u - l)$ instead.

So let $Y_1, ..., Y_n$ be independent random variables, each with $Y_i \sim Beta(\mu_i, \phi)$ $\forall i \in \{1, ..., n\}$. The model is obtained by assuming that the mean for every Y_i can be expressed as[14]

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j = \boldsymbol{x_i}'\boldsymbol{\beta} =: \eta_i, \tag{2.92}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)' \in \mathbb{R}^{k+1}$ is the vector of unknown regression parameters and $\boldsymbol{x_i} = (1, x_{i1}, ..., x_{ik})' \in \mathbb{R}^{k+1}$ are the observations on k covariates $(k < n)$; η_i is called *linear predictor.*

$g : (0, 1) \to \mathbb{R}$ is strictly monotonic and twice differentiable, called *link function.* A particularly useful link function, that we will also use in our case, is the standard *logit link*, i.e.

$$g(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) \Leftrightarrow \mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \tag{2.93}$$

However, there are several possible choices for other link functions. For details see Ferrari and Cribari-Neto [2004].

Note: Since the variance of Y_i is a function of μ_i and hence of the covariates, nonconstant response variances are naturally accomodated into the model [Ferrari and Cribari-Neto, 2004, p. 805].

Parameter estimation

We estimate the paramaters $(\boldsymbol{\beta}, \phi)$ by using the method of maximum likelihood. The log-likelihood function based on a sample of n independent observations $y_1, ..., y_n$ is given by

$$l(\boldsymbol{\beta}, \phi) := \log\left(\prod_{i=1}^{n} f(y_i; \mu_i, \phi)\right) = \sum_{i=1}^{n} l_i(\mu_i, \phi),$$

where

$$\begin{aligned} l_i(\mu_i, \phi) &= \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1 - \mu_i)\phi) + (\mu_i\phi - 1)\log y_i \\ &\quad +\{(1 - \mu_i)\phi - 1\}\log(1 - y_i), \end{aligned}$$

with μ_i defined from (2.93).

However, the maximum likelihood estimators of $\boldsymbol{\beta}$ and ϕ do not have a closed-form. One needs a numerical maximization of the log-likelihood function using a nonlinear

[14] cp. to Ferrari and Cribari-Neto [2004].

optimization algorithm, such as a *Newton algorithm* or a *quasi-Newton algorithm* (for details, see Ferrari and Cribari-Neto [2004]). These algorithms require the specification of initial values as start points for their iterations. Ferrari and Cribari-Neto [2004] suggest to use a least squares estimation from a linear regression of the transformed responses $g(y_1), ..., g(y_n)$ on covariates X to get an initial point estimate for $\boldsymbol{\beta}$, i.e. according to (2.77), $\hat{\boldsymbol{\beta}}_{start} = (X'X)^{-1}X'\boldsymbol{z}$ where $\boldsymbol{z} = (g(y_1), ..., g(y_n))'$. For an initial guess for ϕ, we use from (2.91) that $\phi = \mu_i(1-\mu_i)/Var(Y_i) - 1$ and

$$Var(g(Y_i)) \overset{Taylor}{\approx} Var\left(g(\mu_i) + (Y_i - \mu_i)g'(\mu_i)\right) = Var(Y_i)\left(g'(\mu_i)\right)^2.$$

Hence, $Var(Y_i) \approx Var(g(Y_i))/\left(g'(\mu_i)\right)^2$ and the initial guess for ϕ is given by

$$\hat{\phi}_{start} = \frac{1}{n}\sum_{i=1}^{n}\frac{\breve{\mu}_i(1-\breve{\mu}_i)}{\breve{\sigma}_i^2} - 1,$$

where $\breve{\mu}_i = g^{-1}\left(\boldsymbol{x_i}'(X'X)^{-1}X'\boldsymbol{z}\right)$ is the i-th fitted value from the linear regression of $g(y_1), ..., g(y_n)$ on X. $\breve{\sigma}_i^2 = \breve{\boldsymbol{e}}'\breve{\boldsymbol{e}}/[(n-k)\left(g'(\mu_i)\right)^2]$ where $\breve{\boldsymbol{e}} = \boldsymbol{z} - X(X'X)^{-1}X'\boldsymbol{z}$ is the vector of least squares residuals from the above mentioned linear regression.

Further, under the usual regularity conditions for maximum likelihood estimation when the sample size is large, one can show that

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\phi}\end{pmatrix} \sim \mathcal{N}_{k+1}\left(\begin{pmatrix}\boldsymbol{\beta}\\ \phi\end{pmatrix}, \Sigma_{beta}\right),$$

with an appropriate covariance matrix Σ_{beta}. The calculation of Σ_{beta} can be found in [Ferrari and Cribari-Neto, 2004, p. 806]. Thus with this property one can build confidence intervals to perform *asymptotic Wald tests*, i.e. testing whether the β_i's significantly differs from 0.

Goodness of fit

In case of beta regressions, a global measure of explained variation is denoted by the *pseudo* R^2 *(*R_p^2*)*. It is defined as the square of the sample correlation coefficient between the vector of the fitted linear predictors $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, ..., \hat{\eta}_n)'$ and $g(\boldsymbol{y}) = (g(y_1), ..., g(y_n))'$, i.e.

$$R_p^2 = \{\widehat{corr}(\hat{\boldsymbol{\eta}}, g(\boldsymbol{y}))\}^2, \tag{2.94}$$

where $\widehat{corr}$ is defined in (2.55) and $0 \leq R_p^2 \leq 1$. In case of $R_p^2 = 1$ we have a perfect agreement between $\hat{\boldsymbol{\eta}}$ and $g(\boldsymbol{y})$ and hence between $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{y}$.

Further graphical tools for detecting departures from the postulated model and influential observations are the following:

1. Define the standardized residuals

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var}(y_i)}},$$

where $\hat{\mu}_i = g^{-1}(\boldsymbol{x_i}'\hat{\boldsymbol{\beta}})$ and $\widehat{Var}(y_i) = (\hat{\mu}_i(1-\hat{\mu}_i))/(1+\hat{\phi})$. A plot of these residuals against the index of the observations i should show no detectable pattern.

2. There should be no trend detectable in the plot of r_i against $\hat{\eta}_i$.

3. The plot of observed values y_i against the predicted values $\hat{\mu}_i$ should follow a straight line.

4. Since the distribution of the residuals is not known, *half-normal plots* with simulated envelopes are a helpful diagnostic tool. Therefore we enhance the usual half-normal plot by adding a simulated envelope which can be used to decide whether the observed residuals are consistent with the fitted model. They are produced as follows:

   (i) Fit the model and generate a simulated sample of n independent observations using the fitted model as if it were the true model.

   (ii) Fit the model to the generated sample and compute the ordered absolute values of the residuals.

   (iii) Repeat steps (i) and (ii) k times. (Ferrari and Cribari-Neto [2004] consider $k = 19$ to be a good choice.)

   (iv) We consider n sets of k order statistics. For each set, compute its average, minimum and maximum values.

   (v) Plot these values and the ordered residuals of the original sample against the halfnormal scores $\Phi^{-1}\left((i+n-\frac{1}{8})/(2n+\frac{1}{2})\right)$.

   The envelope is formed by the minimum and maximum values of the k order statistics. Observations of absolute residuals from the original sample outside the limits of the envelope should need further investigation. Additionally, if a considerable proportion of points falls outside the envelope, then one has evidence against the adequacy of the fitted model [Ferrari and Cribari-Neto, 2004, p. 809].

5. There are further plots to detect influential points on the regression parameter by a high *Cook's distance* or to identify leverage points by a corresponding high leverage parameter. However, we skip any details here how to calculate them. For further information we refer to Ferrari and Cribari-Neto [2004].

We will use the R package `betareg`, described in Cribari-Neto and Zeileis [2010], to fit the appropriate beta regressions to our model in chapter 3.

2.5.4 Generalized linear models

Now we want to formulate regression models for further non-normally distributed responses, as for e.g. binomial, Poisson or Gamma distributed responses. The class of *generalized linear models (GLM)* formalizes a framework to model such problems. One can easily show that a linear model, introduced in Section 2.5.1, can be respresented as a GLM. The sections about GLM's are based on McCullagh and Nelder [1989].

As before we denote the independent response variables by Y_i, for $i = 1, ..., n$, and the covariates corresponding to Y_i by $\boldsymbol{x_i} = (1, x_{i1}, ..., x_{ik})' \in \mathbb{R}^p$, $p := k + 1$. y_i, $i = 1, ..., n$, represents the set of observation of Y_i. Then we can characterize a generalized linear model as follows:

Definition 2.38 (Components of a GLM.)

1. ***Random component:***
Response variables Y_i, $i = 1, ..., n$, are independent with density or probability mass function from the exponential family with canonical parameter θ and dispersion parameter ϕ given by

$$f(y; \theta, \phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\}. \tag{2.95}$$

The functions $a(\cdot), b(\cdot)$ and $c(\cdot, \cdot)$ are known. It holds that $\mu_i = b'(\theta_i)$ where μ_i is the mean of Y_i and $Var(Y_i) = b''(\theta_i)a(\phi)$.

2. ***Systematic component:***
The linear predictor is defined as

$$\eta_i(\boldsymbol{\beta}) := \boldsymbol{x_i}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{k1},$$

where $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)' \in \mathbb{R}^p$ is the vector of p unknown regression parameters which have to be estimated.

3. ***Parametric Link Component:***
The relationship between the linear predictor η_i and the mean μ_i of Y_i is explained by the link function

$$g(\mu_i) = \eta_i(\boldsymbol{\beta}) = \boldsymbol{x_i}'\boldsymbol{\beta}.$$

Points 2 and 3 of Definition 2.38 resemble the ones in the definition of the beta regression model in the previous section. In contrast, GLM's are only defined for response variables which are members of the exponential family. Many well known distributions are members of it, like the normal, binary, Poisson, Gamma and inverse Gaussian distributions. A table with the corresponding functions a, b and c as well as the corresponding parameters θ and ϕ can be found in [McCullagh and Nelder, 1989, p.30]. For a binomial random variable $Y_i \sim Bin(n, p)$ the distribution of $\frac{Y_i}{n}$ belongs to the exponential family. The above stated dispersion parameter can be known or unknown. In some GLMs such as the normal and Gamma regression we have an unknown dispersion parameter, which also has to be estimated.

We assume, that the link function g is monotone and differentiable. This is reasonable because, due to the monotonicity of g, μ_i is increasing if x_{ij} is increasing for $\beta_j > 0$. Some examples of link functions for binary/binomial responses are:

(i) *logit link*:
Using the logistic cdf $F(z) = \frac{e^z}{1+e^z}$, we get (if $\mu \in (0,1)$)

$$g(\mu) := F^{-1}(\mu) = log\left(\frac{\mu}{1-\mu}\right). \tag{2.96}$$

We also used this link function for the beta regression in the previous section. It is symmetric around 0.5, i.e. $g(0.5-\mu) = -g(0.5+\mu)$ for $0 < \mu < 0.5$. The corresponding GLM is called *logistic regression.*

(ii) *probit link*:
Using the standard normal cdf $F(z) = \Phi(z)$, we get (if $\mu \in (0,1)$)

$$g(\mu) = \phi^{-1}(\mu).$$

The corresponding GLM is called *probit regression.*

(iii) *complementary log-log link*:
Using the the complementary log-log cdf $F(z) = 1 - \exp\{-\exp\{z\}\}$, we get (if $\mu \in (0,1)$)

$$g(\mu) = \log(-\log(1-\mu)).$$

The corresponding GLM is called *complementary log-log regression.*

GLM's with such link function ((i)-(iii)) are called *binomial GLM's*, since one assumes that the mean $\mu \in (0,1)$ and then uses the inverse of cdf's to define the corresponding link functions, i.e. $g(\mu) = F^{-1}(\mu)$ where F is a cdf.
A link function $g(\cdot)$ is called *canonical* if $\theta_i = \eta_i$ for all i.

Parameter estimation

To estimate the regression parameters $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$ we use again the method of maximum likelihood. Thus, we want to maximize the log-likelihood. For the observed data $\boldsymbol{y} = (y_1, ..., y_n)'$ of $\boldsymbol{Y} = (Y_1, ..., Y_n)'$ the log-likelihood in a GLM is given by

$$l(\boldsymbol{\beta}, \phi, \boldsymbol{y}) = \sum_{i=1}^{n} \log(f(y_i, \theta_i, \phi)) = \sum_{i=1}^{n} l_i(\mu_i, \phi, y_i), \tag{2.97}$$

where $l_i(\mu_i, \phi, y_i)$ is the log-likelihood for observation y_i, i.e.

$$l_i(\mu_i, \phi, y_i) = \frac{(y_i\theta_i - b(\theta_i))}{a(\phi)} + c(y_i, \phi).$$

However, we do it numerically, using the algorithm of *iterated weighted least squares (IWLS)* which is geared to the Fisher scoring algorithm. The IWLS is stated in Algorithm 2. In that Z_i^r can be identified as the first order Taylor approximation of g around $\widehat{\mu}_i^r$. A rough approximation to the variance of Z_i^r is given by $(W_i^r)^{-1}$.

Algorithm 2 Iterative weighted least squares algorithm for GLM estimation of $\boldsymbol{\beta}$.

```
1: Choose an initial value $\widehat{\boldsymbol{\beta}}^{\mathbf{0}}$ and $\varepsilon > 0$.
2: Let $\widehat{\boldsymbol{\beta}}^{\boldsymbol{r}}$ be the current estimate of $\boldsymbol{\beta}$, determine
       - $\widehat{\eta}_i^r := \boldsymbol{x_i}'\widehat{\boldsymbol{\beta}}^{\boldsymbol{r}}$ $i = 1, ..., n$ (current linear predictors)
       - $\widehat{\mu}_i^r := g^{-1}(\widehat{\eta}_i^r)$ (current fitted means)
       - $\widehat{\theta}_i^r := h(\widehat{\mu}_i^r)$ (current canonical parameters)
         where $h$ is the inverse function of $b'$
       - $Z_i^r := \widehat{\eta}_i^r + (y_i - \widehat{\mu}_i^r)\left(\left.\frac{d\eta_i}{d\mu_i}\right|_{\mu_i=\widehat{\mu}_i^r}\right)$ (adjusted dependent variable)
       - $W_i^r := \left[ b''(\theta_i)|_{\theta_i=\widehat{\theta}_i^r}\left(\left.\frac{d\eta_i}{d\mu_i}\right|_{\mu_i=\widehat{\mu}_i^r}\right)^2\right]^{-1}$
3: Regress $Z_i^r$ on $x_{i1}, ..., x_{ik}$ with weights $(W_i^r)^{-1}$ (Weighted least squares) to obtain new
   estimate $\widehat{\boldsymbol{\beta}}^{\boldsymbol{r+1}}$ and continue with step 2 until $\left\|\widehat{\boldsymbol{\beta}}^{\boldsymbol{r}} - \widehat{\boldsymbol{\beta}}^{\boldsymbol{r+1}}\right\| < \varepsilon$.
```

Further one can show, that under some regularity conditions the MLE of $\boldsymbol{\beta}$ is asymptotically normal distributed, i.e

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} \to \boldsymbol{\beta} \text{ in probability as } n \to \infty \text{ and}$$

$$\left[Var\left(\frac{\partial l_n(\boldsymbol{\beta}, \boldsymbol{Y})}{\partial \boldsymbol{\beta}}\right)\right]^{\frac{1}{2}} (\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}) \to \mathcal{N}_p(\mathbf{0}, I_p).$$

Thus, the asymptotic normality can be used to construct asymptotic *Wald tests* for testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. For further details, see McCullagh and Nelder [1989].

The estimation of an unknown dispersion parameter will be treated in the next section, because we need the goodness of fit measure *Pearson's statistic* to be able to define it.

Goodness of fit

In the context of GLM's the major tools of assessing the goodness of fit are the *deviance* and the Pearson's statistic.

Now the regression parameters $\boldsymbol{\beta}$ are estimated and with the *fitted means*

$$\widehat{\mu}_i = g^{-1}(\boldsymbol{x_i'}\widehat{\boldsymbol{\beta}}),$$

one estimates $\mu_i = E[Y_i]$. For a good model we want to have the fitted means $\widehat{\mu}_i$ close to the observations y_i. Therefore we need a discrepancy measure between y_i and $\widehat{\mu}_i$, called deviance. It is defined as the difference between the log-likelihood of the fitted model and the log-likelihood of the *saturated model*. The saturated model is the largest well defined model for n responses which allows for n parameters, i.e. one estimates μ_i by the observation y_i. Clearly, this model fits perfectly but it is completely non-informative about the relationship between the covariates and the responses. Then, the deviance is defined as follows:

Definition 2.39 (Deviance in a GLM.) *The scaled deviance in a GLM is given by*

$$\begin{aligned} D_s(\widehat{\boldsymbol{\mu}}, \boldsymbol{y}, \phi) &:= -2[l(\widehat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - l(\boldsymbol{y}, \phi, \boldsymbol{y})] \\ &= 2\sum_{i=1}^{n} \frac{y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)}{a(\phi)}, \end{aligned} \tag{2.98}$$

where $l(\cdot, \phi, \boldsymbol{y})$ is the corresponding log-likelihood, defined in (2.97), $\widehat{\theta}_i := h(\widehat{\mu}_i)$ and $\widetilde{\theta}_i := h(y_i)$; $h(\cdot)$ denotes the inverse function of $b'(\cdot)$.

If we assume that $a(\phi) = \phi/w$ holds, we define the (unscaled) deviance in a GLM as

$$D(\widehat{\boldsymbol{\mu}}, \boldsymbol{y}) := \phi D_s(\widehat{\boldsymbol{\mu}}, \boldsymbol{y}, \phi). \tag{2.99}$$

The unscaled deviance in (2.99) eliminate the influence of the dispersion parameter ϕ.

For a random variable Y, with distribution from the exponential family, one can show that $Var(Y) = b''(\theta)a(\phi)$. Therefore, we define $V(\mu) := b''(\theta)$ as the variance function in a GLM with $h(\mu) = \theta$. A further discrepancy measure for assessing the goodness of fit of a GLM is given by the *generalized Pearson χ^2 statistic* and it is defined as follows:

Definition 2.40 (Generalized Pearson χ^2 statistics in a GLM.)

$$\chi^2(\widehat{\boldsymbol{\mu}}, \boldsymbol{y}) := \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{2.100}$$

where $V(\widehat{\mu}_i)$ is the estimated variance function for the i-th observation.

Under some regularity conditions, which can be found in McCullagh and Nelder [1989], the deviance and Pearson χ^2 statistic have the following distribution for large n.

$$\begin{aligned} D(\widehat{\boldsymbol{\mu}}, \boldsymbol{y}) &\overset{\mathcal{L}}{\approx} \phi\chi^2_{n-p}, \\ \chi^2(\widehat{\boldsymbol{\mu}}, \boldsymbol{y}) &\overset{\mathcal{L}}{\approx} \phi\chi^2_{n-p}, . \end{aligned} \tag{2.101, 2.102}$$

$\mathcal{L}$ denotes approximation in distribution.

A moment estimator of ϕ can now be constructed based on the asymptotic distribution. It is defined as

$$\widehat{\phi} := \frac{\chi^2(\widehat{\boldsymbol{\mu}}, \boldsymbol{Y})}{n-p}. \tag{2.103}$$

Further, we are using the distributional properties (2.101) and (2.102) to test asymptotically at a level α the goodness of fit of a specified GLM. Therefore we construct two test scenarios, i.e. *the residual deviance test* and *the partial deviance test for nested GLM models*. The residual deviance test for a GLM model is a asymptotic level α test, where we reject

$$H_0\text{: The specified GLM is true}$$

$$\text{versus}$$

$$H_1\text{: not } H_0,$$

if

$$\frac{D(\widehat{\boldsymbol{\mu}}, \boldsymbol{y})}{\widehat{\phi}} > \chi^2_{n-p,1-\alpha} \ . \tag{2.104}$$

Here, $\chi^2_{r,1-\alpha}$ denotes the $100(1-\alpha)\%$ quantile of a χ^2 distribution with r degrees of freedom and $\widehat{\phi}$ is an estimate of the dispersion parameter ϕ.

If we want to compare the fit of two nested GLM's, we have to consider a partial deviance test. Hence, we fit two models:

1. Model F with linear predictor: $\boldsymbol{\eta} = X_1\boldsymbol{\beta_1} + X_2\boldsymbol{\beta_2}$ and deviance D_F,
2. Model R with linear predictor: $\boldsymbol{\eta} = X_1\boldsymbol{\beta_1}$ and deviance D_R,

where $\boldsymbol{\beta_1} \in \mathbb{R}^{p_1}$ and $\boldsymbol{\beta_2} \in \mathbb{R}^{p_2}$ with $p := p_1 + p_2$. Then the partial deviance test for nested GLM models is a asymptotic level α test, where we reject

$$H_0 : \boldsymbol{\beta_2} = \mathbf{0}$$

$$\text{versus}$$

$$H_1 : \boldsymbol{\beta_2} \neq \mathbf{0},$$

if

$$\frac{D_R - D_F}{\widehat{\phi}_F} > \chi^2_{p_2,1-\alpha}, \tag{2.105}$$

where $\widehat{\phi}_F$ is the estimate of the dispersion parameter ϕ based on the full data, i.e. the Model F is assumed to hold.

For our modeling of daily precipitation we will use two GLM's, the binomial and the gamma regression. They are presented in the next section.

2.5.5 Binomial regression

For example, the number of days on which it rains can be modeled by binomial responses. So we consider the data given by $(y_i, \boldsymbol{x_i}), i = 1, ..., n$, where y_i are the realizations of independent binomial distributed random variables $Y_i \sim Bin(n_i, p_i(\boldsymbol{x_i}))$ and $\boldsymbol{x_i}$ are known covariates. $p_i(\boldsymbol{x_i})$ denotes the success probability of Y_i that is dependent on covariates $\boldsymbol{x_i}$. As mentioned before the distribution of $\frac{Y_i}{n_i}$ belongs to the class of exponential families in contrast to the binomial distribution. This holds due to the fact that we can express the corresponding probability mass function as

$$P(Y_i/n_i = k_i) = P(Y_i = n_i k_i) = \binom{n_i}{n_i k_i} p_i(\boldsymbol{x_i})^{n_i k_i}(1 - p_i(\boldsymbol{x_i}))^{n_i - n_i k_i}$$

$$= \exp\left[\log\left(\binom{n_i}{n_i k_i}\right) + n_i k_i \log\left(\frac{p_i(\boldsymbol{x_i})}{1 - p_i(\boldsymbol{x_i})}\right) + n_i \log(1 - p_i(\boldsymbol{x_i}))\right]. \tag{2.106}$$

Thus, it follows $c(\phi_i, y_i) = \log\left(\binom{\phi_i}{\phi_i y_i}\right), a(\phi_i) = \frac{1}{\phi_i}, \theta_i = \log\left(\frac{p_i(\boldsymbol{x_i})}{1-p_i(\boldsymbol{x_i})}\right)$ and $b(\theta_i) = \log(1 + \exp(\theta_i))$. In addition $\phi_i = n_i$ is known and therefore has not to be estimated.

The mean of $\frac{Y_i}{n_i}$ equals the succes probability, i.e.

$$\mu_i = E\left[\frac{Y_i}{n_i}\right] = \frac{n_i p_i(\boldsymbol{x_i})}{n_i} = p_i(\boldsymbol{x_i}).$$

Since we are interested in modeling the success probability we fit a GLM on $\frac{Y_i}{n_i}$ instead of Y_i. If we choose the logit link function $g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) = \boldsymbol{x_i}'\boldsymbol{\beta} = \eta_i$ from (2.96), we get

$$\mu_i = p_i(\boldsymbol{x_i}) = \frac{e^{\boldsymbol{x_i}'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x_i}'\boldsymbol{\beta}}}. \tag{2.107}$$

In this case the logit link is the canonical link since $g(\mu_i) = \theta_i$ (cp. (2.106)). The estimation of the unknown regression parameters $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)' \in \mathbb{R}^{k+1}$ as well as corresponding goodness of fit measures are described in the previous section about GLM's.

2.5.6 Gamma regression

Modeling of positive continuous responses like e.g. positive rain amount on rain days can be obtained by a gamma regression model since the gamma distribution is only defined for positive random variables. Additionally in practice one often observes that the variances among different responses are not constant but are increasing with the mean. This property can also be captured by gamma distributed variables.

Let us consider the data $(y_i, \boldsymbol{x_i}), i = 1, ..., n$, where y_i are realizations of independent gamma distributed random variables, i.e. $Y_i \sim Gamma(\mu_i, \kappa)$ defined in (2.31) in Section 2.1; $\boldsymbol{x_i}$ are the known covariates. The mean and variance of Y_i are given by

$$E[Y_i] = \mu_i \text{ and } Var(Y_i) = \frac{\mu_i^2}{\kappa}.$$

The gamma distribution belongs to the class of exponential families, since

$$\begin{aligned} f(y_i;\mu_i,\kappa) &= \exp\left[(\kappa-1)\log(y_i) - \frac{\kappa}{\mu_i}y_i + \kappa\log(\kappa) - \kappa\log(\mu_i) - \log\Gamma(\kappa)\right] \\ &= \exp\left[\kappa\left(\frac{-y_i}{\mu_i} - \log(\mu_i)\right) + c(y_i,\kappa)\right], \end{aligned} \tag{2.108}$$

where c is independent of μ_i, $\theta_i = -1/\mu_i, b(\theta_i) = \log(\mu) = \log(-1/\theta_i) = -\log(-\theta_i)$ and the dispersion parameter is given by $a(\phi) = \phi = 1/\kappa$. Thus, we can set up a GLM using the *log link* function, i.e.

$$g(\mu_i) = \log(\mu_i) = \boldsymbol{x_i}'\boldsymbol{\beta} = \eta_i, \tag{2.109}$$

where $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)' \in \mathbb{R}^{k+1}$ has to be estimated. Clearly, the log link is not the canonical link but it is often used since it does not impose any restrictions on $\boldsymbol{\beta}$ and it is convenient for modeling and for interpretation of the parameters.

In addition, in the case of the gamma regression one has to estimate the corresponding dispersion parameter. It is done by the Pearson's χ^2 statistic described in Section 2.5.4. The estimation for $\boldsymbol{\beta}$ follows by an IWLS algorithm, also described in the section about GLM's.

Note, often the sum of a group of responses has a specific meaning. For example, if we observe the positive rain amount on day i of the year, $i = 1, ..., 365$ and we have data over 5 years, each year denoted by $j = 1, .., 5$. We express our responses as Y_{ij} and $Y_i := \sum_{j=1}^{5} Y_{ij}$ stands for the total rain amount on day i of the year, measured over 5 years. $Y_i^s := Y_i/5$ denotes the average rain amount on day i of the year. More formally, we have responses Y_{ij} with $j = 1, ..., n_i$ and $i = 1, ..., n$. We set $Y_i := \sum_{j=1}^{n_i} Y_{ij}$ and $Y_i^s := \frac{Y_i}{n_i}$. If $Y_{ij} \sim Gamma(\mu_i, \kappa)$ are independent one can show that

$$Y_i^s = \frac{Y_i}{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \sim Gamma(\mu_i, n_i\kappa). \tag{2.110}$$

Thus, we model $Y_i^s, i = 1, ..., n$ instead of Y_{ij} by a gamma regression, called *gamma regression with weights n_i*.

2.6 Ljung-Box test

A further goodness of fit measure to investigate whether the chosen linear models or linear skew normal/skew t regressions are appropriate is given by the test of Ljung and Box [1978]. In detail we want to examine whether the assumption of independent error terms in the Definitions 2.35 and 2.37 is obtained. So we take the raw residuals $\widehat{r}_i = y_i - \boldsymbol{x_i}'\widehat{\boldsymbol{\beta}}, i = 1, ..., n$ and perform the *Ljung-Box test*, i.e. we test the hypothesis

$$H_0 : (\widehat{r}_i)_{i=1,...,n} \text{ are independent} \quad \text{versus} \quad H_1 : \text{not } H_0. \tag{2.111}$$

Note that the Ljung-Box test is generally performed to check for autocorrelated residuals which we modify here for our purpose.

The test statistic of the Ljung-Box test is constructed by considering the sample autocorrelation of $(\widehat{r}_i)_{i=1,...,n}$. It is defined as

$$\widehat{\rho}_h = \frac{\sum_{i=h+1}^{n} \widehat{r}_i \widehat{r}_{i-h}}{\sum_{i=1}^{n} \widehat{r}_i^2},$$

for lags $h = 1, ..., n-1$. The test of Ljung and Box [1978] jointly considers the autocorrelations of the first $1 \leq m \leq n-1$ lags and has the test statistic

$$\widehat{Q}(\widehat{\rho}) = n(n+2) \sum_{h=1}^{m} \frac{\widehat{\rho}_h^2}{n-h}.$$

Under H_0 (cp. 2.111), $\widehat{Q}(\widehat{\rho})$ is asymptotically χ^2 distributed with m degrees of freedom. Thus, we reject the null hypothesis at a significant level α if $\widehat{Q}(\widehat{\rho}) > \chi^2_{m,1-\alpha}$ where $\chi^2_{m,1-\alpha}$ denotes the $100(1-\alpha)\%$ quantile of a χ^2 distribution with m degrees of freedom.

2.7 Markov chains

We still want to mention a brief definition of *Markov chains* because we will need them in case of modeling daily total precipitation. According to Georgii [2007] a Markov chain is defined as follows:

Definition 2.41 (Markov chain.) *A sequence of discrete random variables $Y_1, Y_2, Y_3, ...$ with possible values in a countable set E is called (first order) Markov chain if the variables follow the Markov property, i.e.*

$$P(Y_{n+1} = y | Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) = P(Y_{n+1} = y | Y_n = y_n). \tag{2.112}$$

Thus, the distribution of Y_{n+1} depends only on the present state y_n and does not depend on the past.

Markov chains are called *stationary* if the conditional probability distribution in (2.112) is independent of n, i.e.

$$P(Y_{n+1} = y | Y_n = y_n) = P(Y_n = y | Y_{n-1} = y_n),$$

for all n. Hence, the conditional probability distribution of a non stationary Markov chain depends on n. For further details we refer to Georgii [2007], pp. 153-182.

Chapter 3

Marginal models

After completing the preliminaries we now want to model the multivariate dependence structure of meteorological variables by an R-vine distribution. Therefore, as Section 2.4 suggests, we need to model the marginal behavior, i.e. the marginal distributions of our variables first. In order to do this, data of meteorological observations are needed.

Our model will be built on data coming from the meteorological observatory in *Hohenpeissenberg*, southern Germany. It is based at mount Hoher Peissenberg which is located about 80km southwest of Munich in the foothills of the Alps at an altitude of 1000 m a.s.l. (see Figure 3.1). Meteorological data has been collected since 1781 and thus it represents the oldest mountain observatory in the world.

After the very first observations had been made around 1758/1759, it all began with the planning of an academic observatory at Hohenpeissenberg in 1772. The constitution of a meteorological station followed 1780 by abbot Johann Jakob on behalf of Karl Theodor, Elector of the Palatinate. And thus one started the daily weather observation under the rules of the Meteorological Society of the Palatinate (Societas Meteorologica Palatina) on January, 1st 1781. The standard hours of observation of that time, the so called "Mannheim hours" (at 7:00, 14:00 and 21:00 hours local mean time) are still used as a standard for today's climatological observations. Because of several political changes and wars the corresponding competences for the wheather station Hohenpeissenberg were often changed too. Hence, the station was attended by priests, teachers and vicars until the Meteorological Service of the Third Reich ("Reichswetterdienst") took over the observatory in 1934. In 1952 after the Second World War, the German Weather Service ("Deutsche Wetterdienst"), DWD, decided to incorporate the station into their meteorological network because of its meaningful location for ecology and climate research.

Today the wheather station is part of the meteorological observatory Hohenpeissenberg. The observatory became a global station within the "Global Atmosphere Watch (GAW)" programme which the World Organisation for Meteorology (WMO) had initiated in the beginning of the nineties. Beside the regular duties from before, the emphasis is placed on the permanent monitoring of trace gasses, the identification of physical, chemical and optical characteristics of aerosols as well as the determination of the chemical composition of precipitation. In addition, continuous long-term measurements of volatile hydrocarbons, such as OH and H_2SO_4, are made. The DWD additionally uses the located radar device for their applied research which improves weather forecasts and warnings not

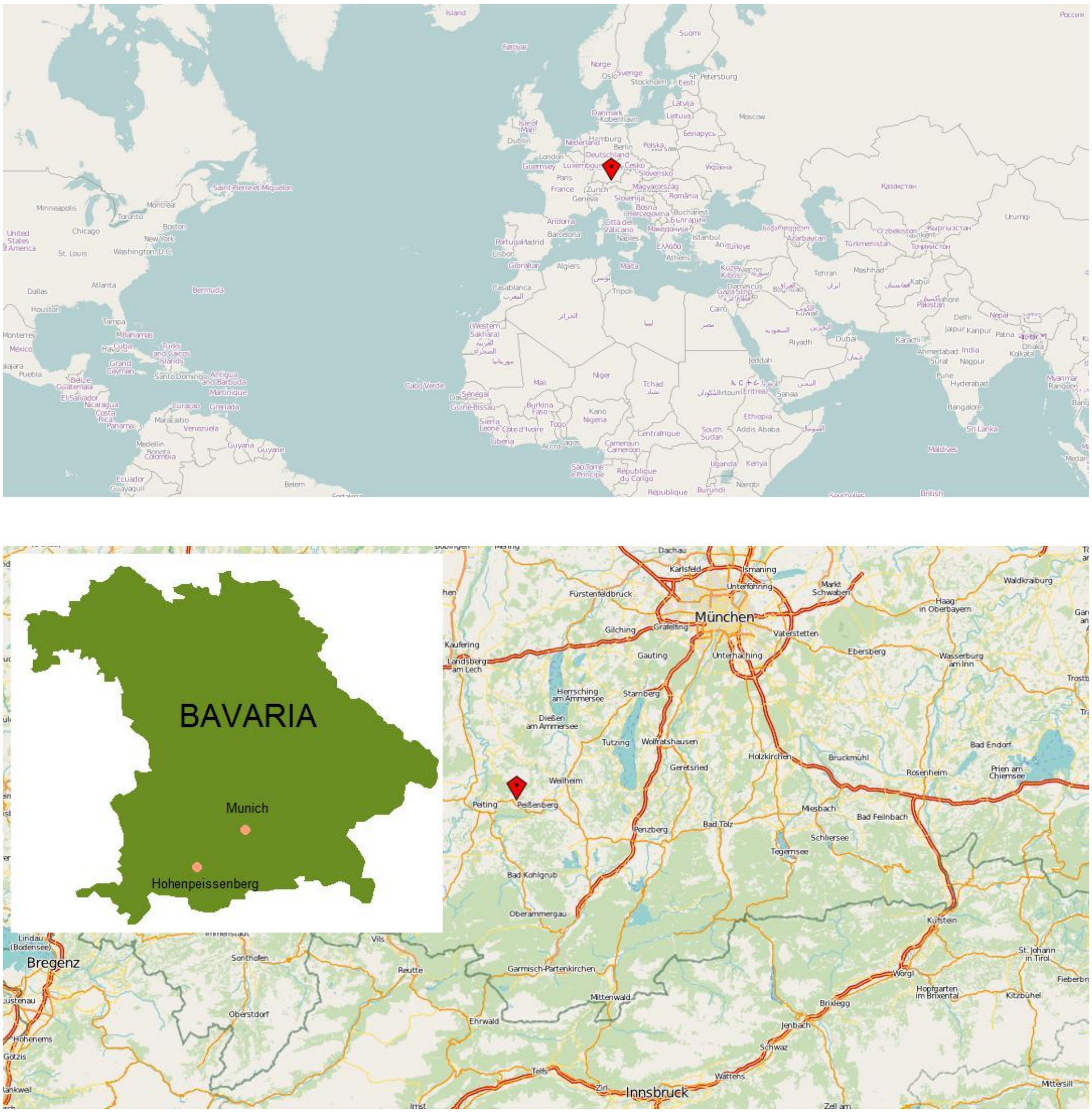


Figure 3.1: Location of the meteorological observatory Hohenpeissenberg in Bavaria, southern Germany. It is based about 80km southwest of Munich in the foothills of the Alps at an altitude of 1000 m a.s.l. © OpenStreetMap and contributors, CC-BY-SA.

only for the region. Interesting climate data from Hohenpeissenberg of the last about 230 years can be found in Table 3.1. More about the history of the meteorological observatory Hohenpeissenberg can be found at Deutsche Wetterdienst DWD [2012].

The German Weather Service was founded in 1952 as National Meteorological Service of the Federal Republic of Germany. It is responsible for providing services for the protection of life and property in the form of weather and climate information and therefore maintains today about 2200 measurement stations all over Germany. The core of this

Results	Time frame	Date	Value
Warmest month	1781 – Jan. 2012	August 2003	20,7 °C
Coldest month	1781 – Jan. 2012	February 1956	-12,4 °C
Highest temperature	1879 – Jan. 2012	July 29, 1947	33,8 °C
Lowest temperature	1879 – Jan. 2012	February 11, 1929	-29,1 °C
Highest 24-hour precipitation total	1879 – Jan. 2012	May 21, 1999	138,5 mm
Highest monthly precipitation	1879 – Jan. 2012	June 1979	366,6 mm
Lowest monthly precipitation	1879 – Jan. 2012	November 2011	0,2 mm
Sunniest month	1937 – Jan. 2012	July 2006	332 Hours
Month with least sunshine	1937 – Jan. 2012	December 1947	31 Hours
Strongest wind gust	1949 – Jan. 2012	November 27,1983	177 km/h
Deepest snow cover	1901 – Jan. 2012	March 10, 1931	145 cm
Highest fresh snow depth in 24 hrs	1947 – Jan. 2012	November 23, 1972	48 cm
Highest amount of new snow within a month	1947 – Jan. 2012	January 1968	178 cm
Highest air pressure (sea level)	1879 – Jan. 2012	January 16, 1882	1047 hPa
Lowest air pressure (sea level)	1879 – Jan. 2012	February 25, 1989	968 hPa
Annual mean temperature	1781 – Jan. 2012		6,2 °C
Mean cloud cover	1879 – Jan. 2012		66%
Mean relative humidity	1879 – Jan. 2012		77%

Table 3.1: Climate data for Hohenpeissenberg from Deutsche Wetterdienst DWD [2012].

network are twelve so called climate reference stations and the meteorological observatory Hohenpeissenberg belongs to this kind of stations. They should detect possible climate changes with common measuring techniques and well-trained wheather observers over the next years. At these climate reference stations the DWD measures meteorological data on the half hour over the whole year in order to doing their research. In this connection, air pressure, temperature, wet-bulb temperature, maximum and minimum temperature, minimum temperature at ground, temperature at ground in 5, 10, 20, 50 and 100 cm depth, precipitation height, relative humidity and sunshine duration are measured by a sensor system *and* by man. In contrast, wheather phenomena (e.g. snow, rain, fog, and so on) as well as the meteorological quantities of cloud forms, depth of fresh snowfall and the state of the ground (e.g. wet, frozen, watery, and so on) are observed only by man. Wind direction and wind speed as well as global radiation are measured only by a sensor system.[1]

[1] cp. website http://www.dwd.de.

Note however, for our dependence model we will concentrate on *daily* measurements of the following six fundamental meteorological variables:

1. Daily mean air temperature [measured in °C]
2. Daily minimum air temperature [measured in °C]
3. Daily maximum air temperature [measured in °C]
4. Daily mean relative humidity [measured in %]
5. Daily mean air pressure [measured in mbar]
6. Daily total precipitation [measured in mm]

Cleary, one could investigate and add more meteorological variables like daily mean windspeed or daily mean temperature at ground. But we think for our purpose it will be sufficient to start with these stated ones to build a common 6-dimensional R-vine distribution and hence a non-trivial statistical dependence structure of these six fundamental meteorological quantities.

The observatory Hohenpeissenberg provides data over the last 230 years as we mentioned above, but not for all meteorological quantities. Some observations of, e.g., wind speed started first in the 1940s and some long-term observations provides incomplete data over long periods. However, we will use the data of observations of the six variables stated above from a time span of $\mathbf{1950-2009}$. Then we divide these sixty years into 12 subperiods, i.e. each subperiod represents 5 years (**12 5-years subperiods**). For all subperiods we will fit the marginal distributions for 5 of our six meteorological variables by different regression models (see Sections 3.1 - 3.6). Due to robustness, the marginal distribution of daily total precipitation will fit by a binomial/gamma regression based on data of the whole time span 1950-2009 (Section 3.7). So for the remaining 5 variables we take three periods, i.e. one period of the beginning, one period in the middle and one period of the end of the whole time span, and present their corresponding results as demonstration. These three periods are

1. Period 1955-1959 (with 1826 daily observations a variable),
2. Period 1980-1984 (with 1827 daily observations a variable),
3. Period 2005-2009 (with 1826 daily observations a variable).

For these three periods we will then select appropriate R-vine distributions respectively and compare whether possible structural differences have occured over time. Note that we have 1826 daily observations for each variable in periods 1955-1959 & 2005-2009 as well as 1827 data points for each variable in period 1980-1984 (since it contains two leap years). The described schedule of the used time periods is illustrated in Figure 3.2.

One further important assumption is the *homogeneity* of our data. In this connection homogeneity means that our data series is not affected by instrumentation changes and station moves of the DWD. One can do several homogeneity tests for raw time series to

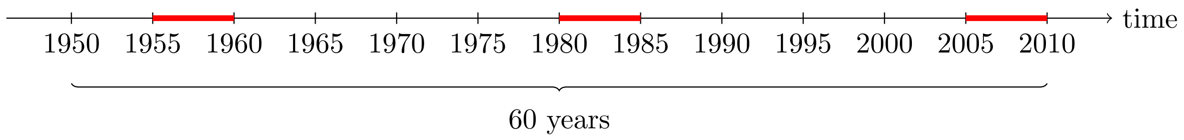


Figure 3.2: Observations from time span 1950-2009 (60 years) which is divided into 12 subperiods a 5 years. We present the results of the marginal models and build corresponding R-vines for the three red marked periods (1955-1959, 1980-1984 and 2005-2009). Note that the regression model of daily total precipitation is based on data of the whole time span 1950-2009.

detect any abnormalities. For details we refer, e.g., to Herzog and Müller-Westermeier [1996]. However, we skip any homogeneity tests for our data at this point since we got the data directly from the DWD with property to be already a homogenized station series.[2].

A lot of work were done in stastical modeling the marginal behavior of temperatures. In fact most of them using time series or linear regression models including e.g. *ARMA (Auto Regressive–Moving Average)* or *GARCH (Generalized Autoregressive Conditional Heteroscedasticity)* models to fit the distribution and variability of temperature variables most precisely. Their intentions, however, differ among the sighted studies. Zheng et al. [1997] use a systematic statistical approach that selects the optimum statistical model (with respect to serial correlation, linearity, etc., i.e. an ARMA model) to detect any trend in regional-mean temperature series. As well, Visser and Molenaar [1995] use structural time series models, i.e. *ARIMA (Auto Regressive–Integrated–Moving Average)* models, to estimate trends and do regression analysis in climatological series. The estimation of current temperature trends are also studied by Mills [2009] using a variety of statistical signal extraction and filtering techniques and their extrapolations. Gil-Alana [2005] models long-term temperature series by fractional integration techniques (including ARMA processes) with long memory behavior.
Pricing wheather derivates correctly is the aim of Cao and Wei [1999] and Campbell and Diebold [2005]. Therefore they need to model conditional mean and variance dynamics in daily average temperatures by approximation of the seasonal volatility component using a Fourier series and by approximation of the cyclical volatility component using a GARCH process to be able to give out-of-sample weather forecasts. Anastasiadou and López-Cabrera [2012] tie in with their work to model temperature risk to investigate the statistical evidence of global warming by identifying shifts in seasonal mean of daily average temperatures over time and in seasonal variance of temperature residuals. Further temperature forecasting approaches can be found, e.g., in Harvey [1989] and Kleiber et al. [2011].

However, the above listed models are based on temperature series data measured at

[2] At all climate reference stations, measurements of newly installed climatological sensors are compared with at least ten years old comparative measurements of conventional sensor systems to avoid misinterpretations in climate series (Deutsche Wetterdienst DWD [2012]).

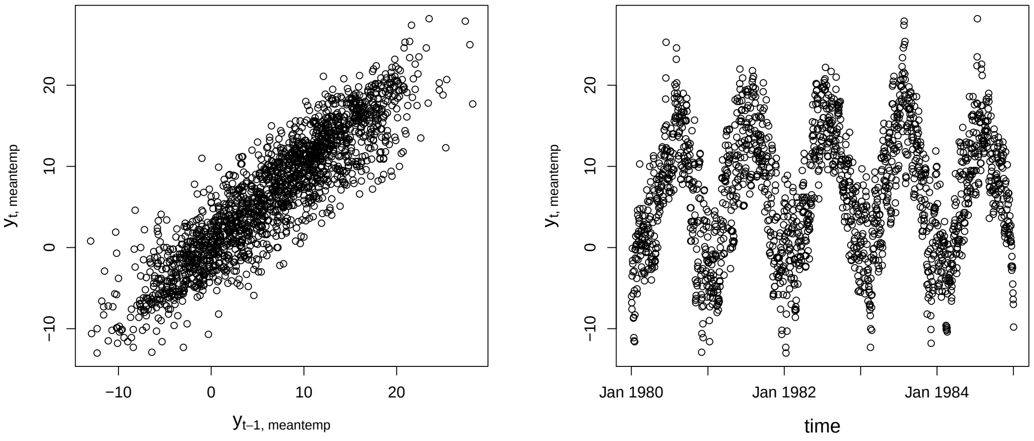


Figure 3.3: Example of the strong positive autocorrelation (left panel: data $y_{t,meantemp}$ against $y_{t-1,meantemp}$) and the seasonal behavior (right panel: $y_{t,meantemp}$ against time t) of daily mean air temperature measured in Hohenpeissenberg in the period 1980-1984.

different and fixed places around the world. Thus their approaches are locally justified. But all models present normal or slightly skewed distributed residuals after capturing autoregressive parts and seasonal dynamics. Therefore we prefer linear regressions to model the temperature variables daily mean air temperature $Y_{t,meantemp}$ at time t, daily minimum air temperature $Y_{t,mintemp}$ at time t and daily maximum air temperature $Y_{t,maxtemp}$ at time t. Clearly, these variables are each strongly autocorrelated (i.e. the temperature today is strongly influenced by the temperature of yesterday and by the temperature of the day before yesterday and so on) and exihibt seasonal patterns (i.e. the temperature in summer is probably higher than in winter) which is illustrated for an example in Figure 3.3. These properties can then be modeled by implementing appropriate covariates in the predictor parts of the linear regressions (see Sections 3.1 - 3.3). Adding a trend component does not seem to be relevant since we model data "only" over 5 years periods. Aditionally, we will see that this approach also fits the behavior of daily mean air pressure, $Y_{t,press}$ at time t, respectively well (Section 3.6).

The variable of daily relative humidity $Y_{t,humidity}$ at time t takes values between 0% and 100%. Thus, it seems to be reasonable that its distribution follows a beta distribution. Yao [1974] enters this approach, but we like to extend it in order to model autoregression and seasonal effects additionally by a beta regression model which was introduced by Ferrari and Cribari-Neto [2004] (see Section 3.5).

Modeling the behavior of daily total precipitation $Y_{t,prec}$ at time t is a bit more demanding since it often takes the value equal to zero (in our dataset of Hohenpeissenberg, $Y_{t,prec} = 0$ in 48% of the cases between 1950 and 2009). A wide range of literature can be found on this topic with different intentions. A regression model using spline functions for generating time series of daily precipitation amounts is presented by Buishand and Klein Tank [1996] in order to study climate change impacts. Sloughter et al. [2007]

Variable	No. of observations Period 1955-1959	No. of observations Period 1980-1984	No. of observations Period 2005-2009
$Y_{t,meantemp}$	1826	1827	1826
$Y_{t,mintemp}$	1826	1827	1826
$Y_{t,maxtemp}$	1826	1827	1826
$Y_{t,humidity}$	1826	1827	1826
$Y_{t,press}$	1826	1827	1826
$Y_{t,prec}$	modeled on data over whole time span 1950-2009: 21915 observations		

Variable	Regression model	Covariates of the linear predictor				
		$Y_{t-1,\cdot}$	$Y_{t-2,\cdot}$	$Y_{t-3,\cdot}$	$Y_{t-7,\cdot}$	$x_{t,season}$
$Y_{t,meantemp}$	Linear model/ Linear skew normal regression	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
$Y_{t,mintemp}$	Linear skew t regression	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
$Y_{t,maxtemp}$	Linear model/ Linear skew normal regression	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
$Y_{t,humidity}$	Beta regression	$\surd$				$\surd$
$Y_{t,press}$	Linear skew t regression	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
$Y_{t,prec}$	Binomial/Gamma regression					

Variable	Covariates of the linear predictor			
	$\sin(x_{t,winddirection})$	$\cos(x_{t,winddirection})$	$\sin(\frac{k\pi d(t)}{366}), k = 2, 4$	$\cos(\frac{k\pi d(t)}{366}), k = 2, 4$
$Y_{t,meantemp}$				
$Y_{t,mintemp}$				
$Y_{t,maxtemp}$				
$Y_{t,humidity}$	$\surd$	$\surd$		
$Y_{t,press}$				
$Y_{t,prec}$			$\surd$	$\surd$

Table 3.2: Framework of our marginal models for the six meteorological variables. Note that $x_{t,season} \in \{1, 2, 3, 4\}$ denotes the meteorological season at time t, i.e. 1=winter, 2=spring, 3=summer and 4=fall. The daily mean wind direction at time t is given by $x_{t,winddirection}$ (needed to explain the "Fön"-effect, i.e. dry and hot winds in the foothills region of the Alps) and $d(t)$ stands for the number of the day in the corresponding year at time t.

use Bayesian model averaging as a statistical way of postprocessing forecast ensembles to create predictive probability density functions of precipitation variables, while Little et al. [2009] use generalized linear models for forecast the density of daily rainfall. In contrast, Kim and Mallick [2004] build a model based on the skew normal distribution which is applied in the spatial prediction of weekly rainfall. On the other hand a general linear regression is used by Turlapaty et al. [2009] to merge precipitation data. Berrocal et al. [2008] present a statistical model that is a spatial version of a two-stage model that represents the distribution of precipitation by a mixture of a point mass at zero and a Gamma density for the continuous distribution of precipitation accumulation. A simultaneous simulation of daily precipitation at multiple locations is discussed by Wilks [1998]. It is based on the work of Stern and Coe [1984] which provides a chain-dependent-process stochastic model of daily precipitation: First they model occurrences of rain by a two-state, first-order Markov chain using an binomial regression and secondly they model nonzero amounts of rain by a gamma distribution using a gamma regression. We will also use the model of Stern and Coe [1984] for our data since it provides an appropriate instrument of modeling the behavior of total daily precipitation over the whole year at Hohenpeissenberg. We already mentioned above that due to robustness our model will be based on observations over the whole time period 1950-2009 as Stern and Coe [1984] suggest (Section 3.7).

The whole framework of the modeling the marginal behavior of our six meteorological variables is summerized in Table 3.2.

3.1 Daily mean air temperature

Now we start to model daily mean air temperature $Y_{t,meantemp}$ in Hohenpeissenberg at time t using linear models and linear skew normal regressions respectively. We need both types of regressions here, since we detect some skewness in the distribution of the residuals from period 1985-1989 to last period 2005-2009 which can be better fit by a skew normal distribution as we will see immediately.

Note that we decided to regress on standardized variables, i.e. on

$$\widetilde{y}_{t,meantemp} = \frac{y_{t,meantemp} - \bar{y}_{meantemp}}{\widehat{s}_{meantemp}}, \tag{3.1}$$

where

$$\bar{y}_{meantemp} := \frac{1}{1826} \sum_{t=1}^{1826} y_{t,meantemp}$$

is the mean of observations[3] and

$$\widehat{s}_{meantemp} := \sqrt{\frac{1}{1825} \sum_{t=1}^{1826} (y_{t,meantemp} - \bar{y}_{meantemp})^2}$$

denotes the empirical standard deviation[4]. This standardization might be useful if the variables are on very different scales and/or the magnitude of coefficients for variables with small values may not indicate their relative importance influencing the response variable (Quinn and Keough [2002]).

[3] $\bar{y}_{meantemp} := \frac{1}{1827} \sum_{t=1}^{1827} y_{t,meantemp}$, when the period contains two leap years.

[4] $\widehat{s}_{meantemp} := \sqrt{\frac{1}{1826} \sum_{t=1}^{1827} (y_{t,meantemp} - \bar{y}_{meantemp})^2}$, when the period contains two leap years.

This yields to the following model specification describing the behavior of standardized daily mean temperature observations at time t in Hohenpeissenberg for each of the 12 periods:

$$\begin{aligned}\widetilde{y}_{t,meantemp} &= \beta_0 + \gamma_1 \widetilde{y}_{t-1,meantemp} + \gamma_2 \widetilde{y}_{t-2,meantemp} + \gamma_3 \widetilde{y}_{t-3,meantemp} \\ &\quad + \gamma_7 \widetilde{y}_{t-7,meantemp} + \beta_{season} x_{t,season} + \epsilon_{t,meantemp}, \end{aligned} \tag{3.2}$$

for $t \in \{1, ..., 1826\}$ ($t \in \{1, ..., 1827\}$ when the period contains two leap years), where

(i) Covariable $x_{t,season} = 1$ for winter days, $x_{t,season} = 2$ for spring days, $x_{t,season} = 3$ for summer days and $x_{t,season} = 4$ for fall days, for all t (thus, covariable $x_{t,season}$ as factors). This division results from the meteorological seasons:

  - winter = December, January, February
  - spring = March, April, May
  - summer = June, July, August
  - fall = September, October, November

(ii) We assume:

  - From period 1950-1954 to period 1980-1984:

$$\epsilon_{t,meantemp} \sim N(0, \sigma^2_{meantemp}) \text{ (i.i.d.)}.$$

  - From period 1985-1989 to period 2005-2009:

$$\epsilon_{t,meantemp} \sim \mathcal{SN}(\xi_{meantemp}, \omega_{meantemp}, \alpha_{meantemp}) \text{ (i.i.d.)}.$$

Thus, the strong autoregressive behavior is modeled by the linear part $\gamma_1 \widetilde{y}_{t-1,meantemp} + \gamma_2 \widetilde{y}_{t-2,meantemp} + \gamma_3 \widetilde{y}_{t-3,meantemp} + \gamma_7 \widetilde{y}_{t-7,meantemp}$ in (3.2) and the seasonal behavior is captured by $\beta_{season} x_{t,season}$. Alternatively one could also use Fourier series instead for modeling seasonal effects as Campbell and Diebold [2005] and Anastasiadou and López-Cabrera [2012] do. However, it does not result in any signficant differences in our outcomes.

Due to the model we use the method of least squares to calculate the estimates of the coefficients $(\beta_0, \gamma_1, \gamma_2, \gamma_3, \gamma_7, \beta_{season=spring}, \beta_{season=summer}, \beta_{season=fall})' \in \mathbb{R}^8$ (cp. Sections 2.5.1 and 2.5.2); $\beta_{season=spring}, \beta_{season=summer}$ and $\beta_{season=fall}$ here describe the seasonal difference of the standardized daily mean temperature between winter and spring, winter and summer or winter and fall, respectively, corresponding to the season at time t (hence we are using a dummy coding for $x_{t,season}$ here). The estimates of the coefficients for the three periods 1955-1959, 1980-1984 and 2005-2009 as well as the corresponding p-values of Wald tests (testing $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ or $H_0 : \gamma_i = 0$ vs. $H_1 : \gamma_i \neq 0$) and the R^2_{adj} can be found in Table 3.3. Note that there exists no Wald tests and R^2_{adj} for the last period since we fit a linear skew normal regression. All covariates seem to have an significant influence on the standardized daily mean temperature (p-values $<< 0.05$) and the $R^2_{adj} > 0.8$ indicate an appropriate explanation of the variability of the response by our models. One can show that further autoregressive covariates do not improve the model here.

Coefficient	Period 1955-1959		Period 1980-1984		Period 2005-2009
	Estimate	p-value of Wald test	Estimate	p-value of Wald test	Estimate
β_0	-0.13	0.00	-0.20	0.00	-0.13
γ_1	0.96	0.00	0.88	0.00	0.99
γ_2	-0.24	0.00	-0.23	0.00	-0.27
γ_3	0.08	0.00	0.13	0.00	0.10
γ_7	0.06	0.00	0.03	0.03	0.06
$\beta_{season=spring}$	0.13	0.00	0.18	0.00	0.14
$\beta_{season=summer}$	0.26	0.00	0.39	0.00	0.26
$\beta_{season=fall}$	0.11	0.00	0.21	0.00	0.12
R^2_{adj}	0.85		0.83		/

Table 3.3: Summary of the coefficient estimations and the adjusted R^2_{adj} for the daily mean air temperature models of the different periods. Note that there exists no Wald tests and R^2_{adj} for the last period, since we use a linear skew normal regression.

With the resulted fitted values $\widehat{\widetilde{y}}_{t,meantemp}$ we calculate the raw residuals

$$r_{t,meantemp} := \widetilde{y}_{t,meantemp} - \widehat{\widetilde{y}}_{t,meantemp}.$$

Using the Ljung-Box test, i.e.

$$H_0: \text{ residuals are independent vs. } H_1: \text{ residuals are not independent,}$$

the assumption (ii) of independent errors seems plausible (see Table 3.4: only for period 1980-1984 we would have to reject the null hypothesis at lag 365. However we think that we can ignore this single result, since the large lag of 365 may be not very meaningful.).

Results of the maximum likelihood estimation of the distribution parameters (see Sections 2.5.1 and 2.5.2) can be found in Table 3.5. Three goodness of fit diagnostic plots for each period are then shown in Figure 3.4. First we plot the raw residuals of our models against their observation numbers and thus against their point in time. We do not detect any systematic pattern for any season here which underlines the assumption

Period	lag 1	lag 5	lag 365
1955-1959	0.88	0.95	0.25
1980-1984	0.86	0.94	0.01
2005-2009	0.96	0.84	0.57

Table 3.4: p-values of the Ljung-Box tests (with lag 1,5 and 365) with the raw residuals of the daily mean air temperature models for the different periods. As a result we cannot reject the null hypothesis of a Ljung Box Test at a 5% significance level (except for period 1980-1984 at lag 365).

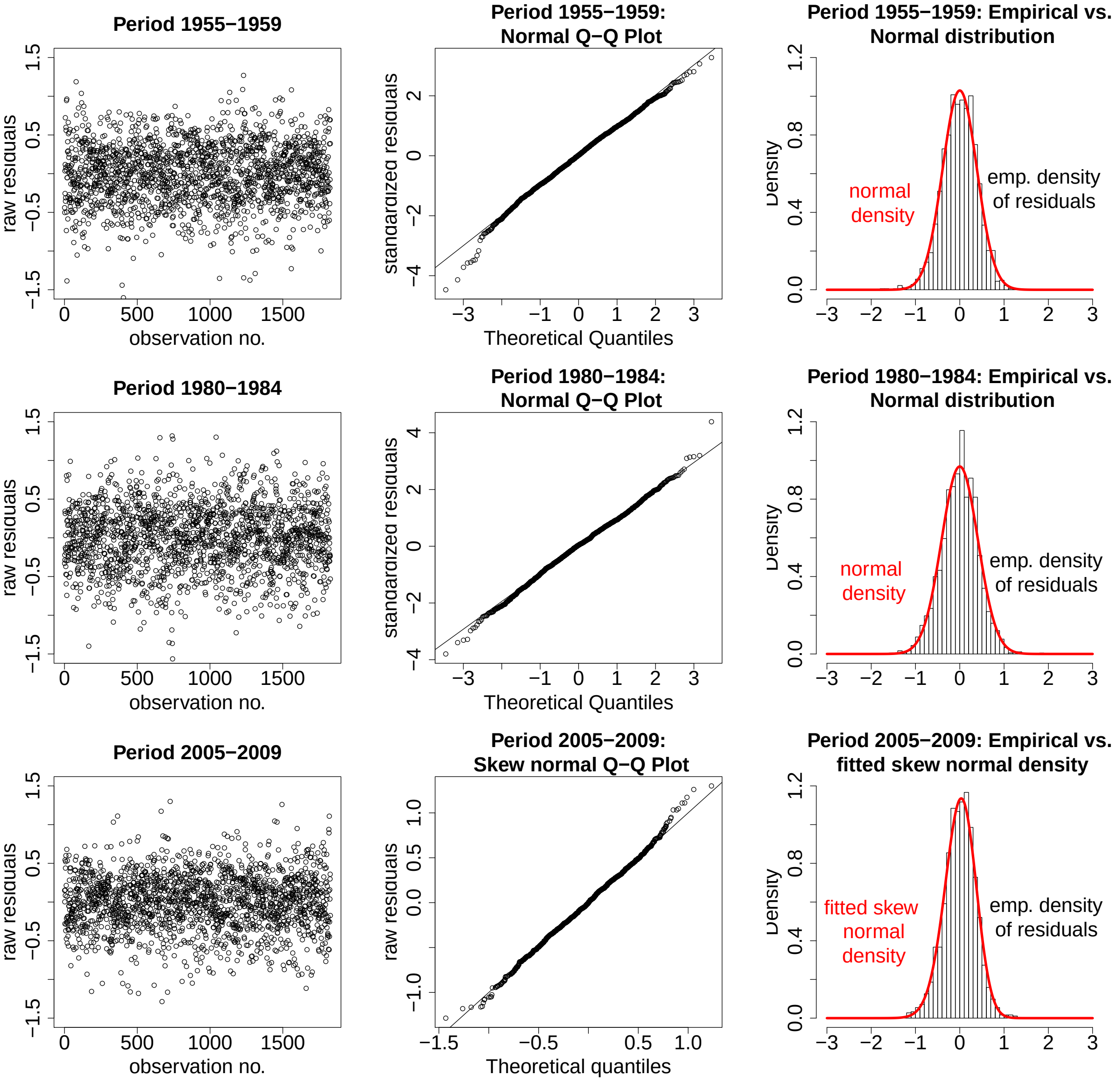


Figure 3.4: Goodness of fit plots for our fitted daily mean air temperature models. The left panel describes the raw residuals plotted against their observation numbers. The middle panel shows Q-Q plots of the corresponding fitted distributions which are compared to the empirical ones in the third panel. For all periods the plots do not exhibit any severe objections against our assumed models.

of independent error terms. The corresponding quantile-quantile plots (Q-Q plots) are all following nearly straight lines which attest the goodness of our fits. The fitted distributions compared to the empirical ones are then illustrated in the third plots.

Note that the result of normal or slightly skewed distributions of the (standardized) daily mean air temperature error terms coincide with the outcomes of Campbell and Diebold [2005]. The estimated shape parameter $\widehat{\alpha}_{meantemp} = -1.14$ for the last period

Parameter	Period 1955-1959	Period 1980-1984	Period 2005-2009
$\sigma^2_{meantemp}$	0.15	0.17	0.13
$\xi_{meantemp}$	/	/	0.27
$\omega_{meantemp}$	/	/	0.44
$\alpha_{meantemp}$	/	/	-1.14

Table 3.5: Estimates of the fitted distribution parameters in the (standardized) daily mean air temperature model for every period.

2005-2009 indicates a slightly negatively skewed normal distribution around zero. That means that the variability of negative error terms is a bit higher than the variability of positve error terms. In other words, there is a slightly greater variability for "wheather surprises" from our fitted mean of (standardized) daily mean air temperature downwards than upwards. In addition, the fitted distributions have nearly the same estimated variance over the periods.

3.2 Daily minimum air temperature

Our modeling of daily minimum air temperature is similar to the previous one. We also regress on standardized variables like in the daily mean air temperature case before. The only difference results in the distribution of the error terms which can be best fitted by a skew t distribution (defined in Section 2.1.7). Hence, the model specification is given by:

$$\begin{aligned}\widetilde{y}_{t,mintemp} &= \beta_0 + \gamma_1 \widetilde{y}_{t-1,mintemp} + \gamma_2 \widetilde{y}_{t-2,mintemp} + \gamma_3 \widetilde{y}_{t-3,mintemp} \\ &\quad + \gamma_7 \widetilde{y}_{t-7,mintemp} + \beta_{season} x_{t,season} + \epsilon_{t,mintemp}, \end{aligned} \tag{3.3}$$

for $t \in \{1, ..., 1826\}$ ($t \in \{1, ..., 1827\}$ when the period contains two leap years), where

(i) $\widetilde{y}_{t,mintemp} = \frac{y_{t,mintemp} - \bar{y}_{mintemp}}{\widehat{s}_{mintemp}}$, where $\widehat{s}_{mintemp}$ is the empirical standard deviation and $\bar{y}_{mintemp}$ the empirical mean of the time series.

(ii) Covariable $x_{t,season} = 1$ for winter days, $x_{t,season} = 2$ for spring days, $x_{t,season} = 3$ for summer days and $x_{t,season} = 4$ for fall days, for all t (as factors). This division results from the meteorological seasons:

- winter = December, January, February
- spring = March, April, May
- summer = June, July, August
- fall = September, October, November

(iii) We assume:

$$\epsilon_{t,mintemp} \sim skewt(\xi_{t,mintemp}, \omega_{t,mintemp}, \alpha_{t,mintemp}, df_{t,mintemp})$$

(independent but with different variances at time t).

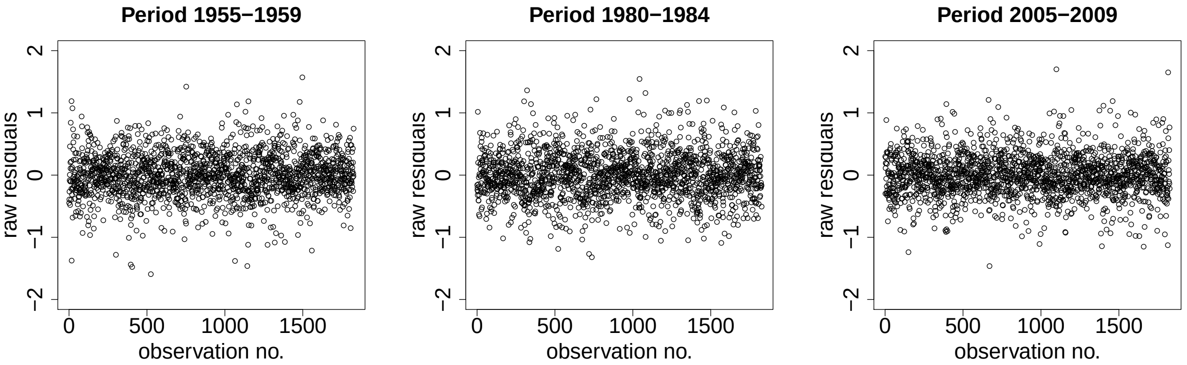


Figure 3.5: Plots of the raw residuals from our daily minimum air temperature model against the observation numbers for every period. One might detect different variances of the residuals at different observations numbers t and thus at different point in time t.

The last assumption (iii) is manifested by looking at the plots of the raw residuals $r_{t,mintemp} := \widetilde{y}_{t,mintemp} - \widehat{\widetilde{y}}_{t,mintemp}$ against the observation numbers (Figure 3.5), where $\widehat{\widetilde{y}}_{t,mintemp}$ denotes the fitted values of model (3.3). They are suggestive of different variances at different points in time. For that reason we use the method of weighted least squares (described in Section 2.5.2) to estimate the parameters in (3.3) with appropriate weights $w_{t,mintemp}$ such that

$$\frac{1}{\sqrt{w_{t,mintemp}}}\epsilon_{t,mintemp} \sim skewt(\xi_{mintemp}, \omega_{mintemp}, \alpha_{mintemp}, df_{mintemp}) \text{ (i.i.d)},$$

for all t. Suitable weights, as we will see, are given by the empirical standard deviation over all standardized observations at day d of the year in the period, i.e.

$$w_{d,mintemp} := sd(\widetilde{y}_{d,mintemp}, \widetilde{y}_{d+365,mintemp}, \widetilde{y}_{d+2\times365,mintemp}, \widetilde{y}_{d+3\times365,mintemp}, \widetilde{y}_{d+4\times365,mintemp}),$$

where d=1,...366. Then it follows that $w_{t,mintemp} = w_{d(t),mintemp}$, where $d(t) \in \{1, ..., 366\}$ denotes the corresponding day of the year at oberservation t.

The estimates of the coefficients in (3.3) by the method of weighted least squares (WLS) for the different periods are presented in Table 3.6. Note again, since we do a skew t regression we cannot perform any Wald tests to test the significance of the parameters (cp. Section 2.5.2). The raw residuals $r^{WLS}_{t,mintemp} := \widetilde{y}_{t,mintemp} - \widehat{\widetilde{y}}^{WLS}_{t,mintemp}$ after weighted least squares seem to be independent (compare Table 3.7) which underlines the assumption of independent error terms. Note that $\widehat{\widetilde{y}}^{WLS}_{t,mintemp}$ here denotes the fitted values of the weighted least square regression.

The fitted distribution parameters are estimated by the method of maximum likelihood and can also be found in Table 3.6. For the period 1955-1959 we have nearly no skewness in the fitted distribution of the error terms after WLS since $\widehat{\alpha}_{mintemp} \approx 0$. For the

Coefficient	Period 1955-1959	Period 1980-1984	Period 2005-2009
	Estimate	Estimate	Estimate
β_0	-0.15	-0.19	-0.16
γ_1	0.93	0.88	0.94
γ_2	-0.21	-0.22	-0.21
γ_3	0.07	0.10	0.05
γ_7	0.05	0.05	0.07
$\beta_{season=spring}$	0.14	0.16	0.16
$\beta_{season=summer}$	0.30	0.39	0.31
$\beta_{season=fall}$	0.13	0.22	0.17
$\sigma^2_{mintemp}$	0.25	0.28	0.23
$\xi_{mintemp}$	0.00	-0.13	-0.08
$\omega_{mintemp}$	0.42	0.48	0.39
$\alpha_{mintemp}$	-0.01	0.33	0.21
$df_{mintemp}$	7.13	10.67	6.02

Table 3.6: Summary of the coefficient estimations after weighted least square regression of (standardized) daily minimum air temperature as well as the estimated parameters for the fitted distributions of the corresponding residuals for the different periods. Note that there exists no Wald tests in case of skew t regressions.

further periods we then fit slight positively skewed t distributions around mean 0 which means that the variability of positive error terms ("temperature up surprises from the mean") is slight higher than the variability of negative error terms ("temperature down surprises from the mean"). The estimated parameters $\widehat{\xi}_{mintemp}$, $\widehat{\omega}_{mintemp}$ and $\widehat{\alpha}_{mintemp}$ as well as the estimated degrees of freedom $\widehat{df}_{mintemp}$ for all of the 12 periods (all periods are marked as gray dashed lines in the background) are illustrated in Figure 3.6. Note that there is no significant trend detectable over time since we did a simple linear regression of the parameters against time (black dashed line) and so these distributions have almost the same estimated variance over the periods.

Period	lag 1	lag 5	lag 365
1955-1959	0.13	0.55	0.01
1980-1984	0.22	0.53	0.06
2005-2009	0.08	0.59	0.12

Table 3.7: p-values of the Ljung-Box tests (with lag 1,5 and 365) with the raw residuals after weighted least square regression of the daily minimum air temperature for the different periods. As a result we cannot reject the null hypothesis of a Ljung Box Test (H_0 : residuals are independent vs. H_1 : residuals are not independent) at a 5% significance level (except for period 1955-1959 at lag 365).

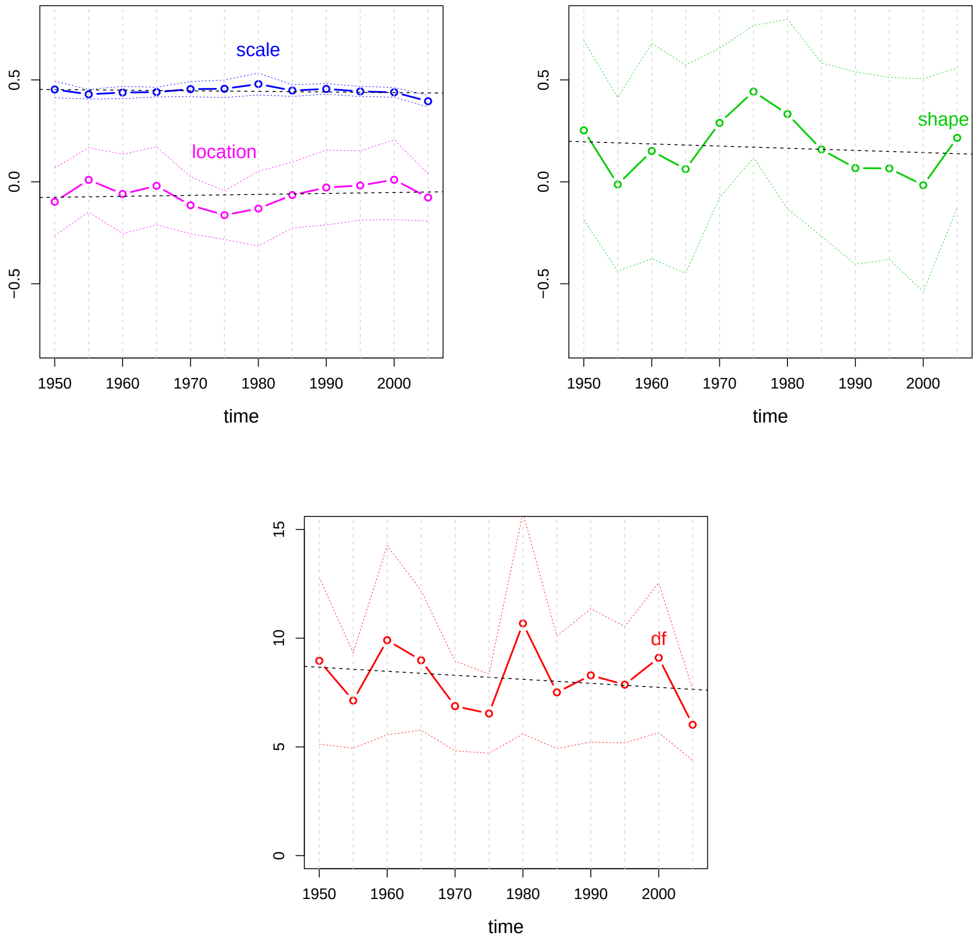


Figure 3.6: Estimated scale $\widehat{\omega}_{mintemp}$ and location $\widehat{\xi}_{mintemp}$ parameters (top left panel), estimated shape parameters $\widehat{\alpha}_{mintemp}$ (top right panel) as well as estimated degrees of freedom $\widehat{df}_{mintemp}$ (bottom panel) of the fitted distributions of the residuals in the (standardized) daily minimum air temperature model for all periods. The colored dashed lines represents corresponding 95%-confidence intervals and the black dashed lines correspond to a simple linear regression of the parameters against time. Note that there is no significant trend detectable. The gray dashed lines in the background mark the 12 5-years periods.

The diagnostic plots in Figure 3.7 underline our model assumptions. In detail, the raw residuals after WLS regression seem to be independent, they do not exihibit any systematical pattern against their observation numbers t (and thus against time). Further they now seem to have a common variance at all points in time. The corresponding Q-Q plots of the fitted distributions of the residuals are following straight lines which can be compared by looking in the plots of the empirical distributions against the fitted ones.

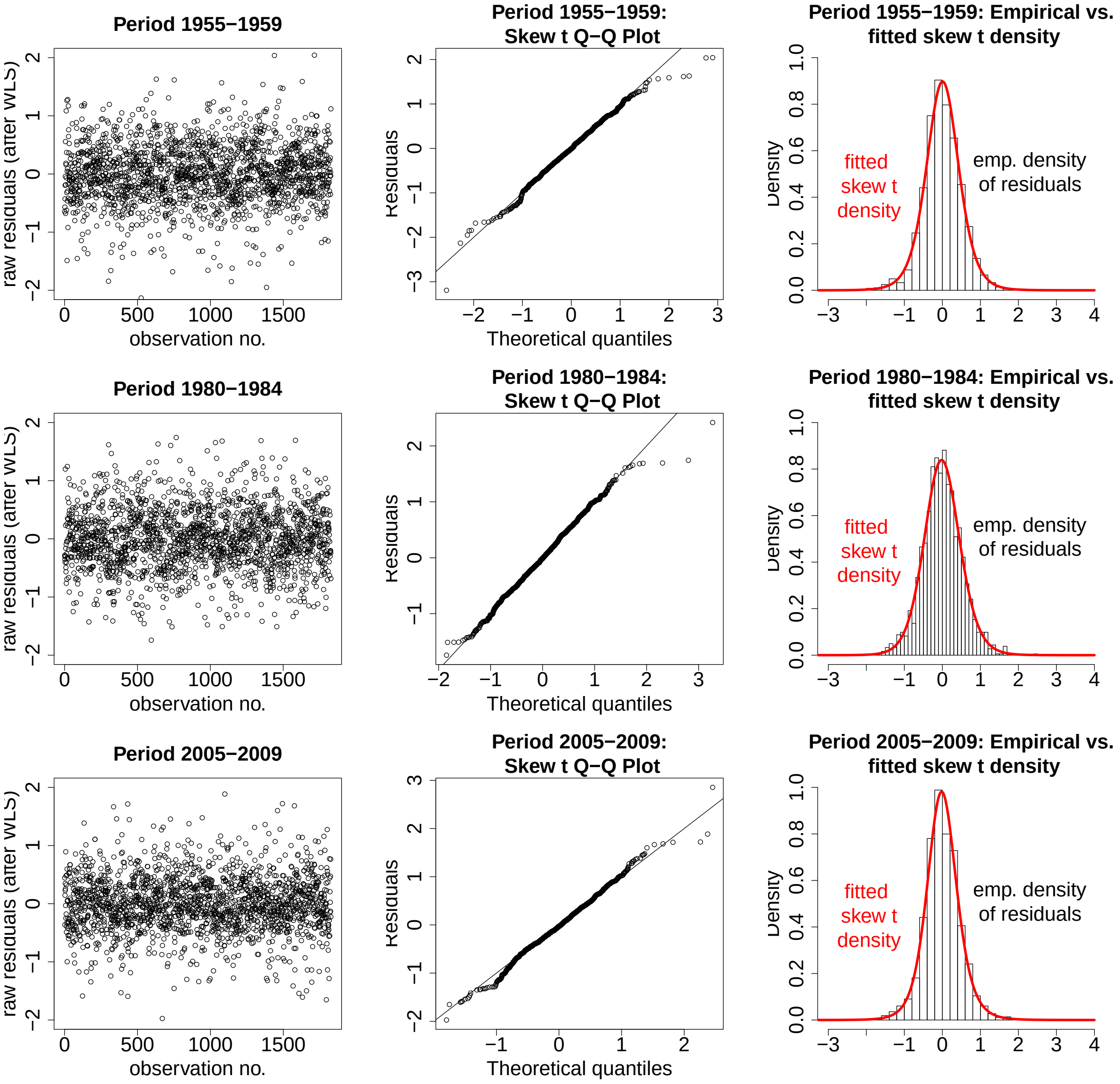


Figure 3.7: Goodness of fit plots for our fitted daily minimum air temperature models. The left panel describes the raw residuals after weighted least squares regression plotted against their observation numbers. The middle panel shows Q-Q plots of the corresponding fitted distributions which are compared to the empirical ones in the third panel. For all periods, the plots do not exhibit any severe objections against our assumed models.

3.3 Daily maximum air temperature

The daily maximum air temperature model resembles basically the model of daily mean air temperature. Again we are using standardized variables for our regression and the error terms follow a skew normal distribution in the last periods in contrast to the ones in the early periods. Thus, our observations of daily maximum temperature are explained by the following model:

$$\begin{aligned}\widetilde{y}_{t,maxtemp} &= \beta_0 + \gamma_1 \widetilde{y}_{t-1,maxtemp} + \gamma_2 \widetilde{y}_{t-2,maxtemp} + \gamma_3 \widetilde{y}_{t-3,maxtemp} \\ &\quad + \gamma_7 \widetilde{y}_{t-7,maxtemp} + \beta_{season} x_{t,season} + \epsilon_{t,maxtemp}, \end{aligned} \tag{3.4}$$

for $t \in \{1, ..., 1826\}$ ($t \in \{1, ..., 1827\}$ when the period contains two leap years), where

(i) $\widetilde{y}_{t,maxtemp} = \frac{y_{t,maxtemp} - \bar{y}_{maxtemp}}{\widehat{s}_{maxtemp}}$, where $\widehat{s}_{maxtemp}$ is the empirical standard deviation and $\bar{y}_{maxtemp}$ the empirical mean of the time series.

(ii) Covariable $x_{t,season} = 1$ for winter days, $x_{t,season} = 2$ for spring days, $x_{t,season} = 3$ for summer days and $x_{t,season} = 4$ for fall days, for all t (as factors). This division results from the meteorological seasons:

- winter = December, January, February
- spring = March, April, May
- summer = June, July, August
- fall = September, October, November

(iii) We assume:

- From period 1950-1954 to period 1980-1984:

$$\epsilon_{t,maxtemp} \sim N(0, \sigma^2_{maxtemp}) \text{ (i.i.d.)}.$$

- From period 1985-1989 to period 2005-2009:

$$\epsilon_{t,maxtemp} \sim \mathcal{SN}(\xi_{maxtemp}, \omega_{maxtemp}, \alpha_{maxtemp}) \text{ (i.i.d.)}.$$

We estimate the coefficients of our model described in (3.4) by the method of least squares. The results are given in Table 3.8. As before, the coefficients $\beta_{season=spring}$, $\beta_{season=summer}$ and $\beta_{season=fall}$ describe the seasonal difference of the standardized daily maximum air temperature between spring and winter, summer and winter or fall and winter, respectively, depending on the season at time t (dummy coding of $x_{t,season}$). All coefficients seem to have a significant influence on the response ((nearly) all p-values of the Wald tests ≤ 0.05). Also the adjusted $R^2_{adj} = 0.81$ attest a relatively broad explanation of the response variabilty by our model. Two facts results from comparing the coefficient estimates of all temperature models: Firstly the standardized temperature one day before has the most influence on the standardized temperature today (estimates $\widehat{\gamma}_{t-1,\cdot} \geq 0.8$ in all temperature models). Secondly all temperature models exhibit that the standardized temperature the day before yesterday has an negatively influence on the standardized temperature today ($\widehat{\gamma}_{t-2,\cdot} < 0$ in all temperature models).

Also in this case of the raw residuals $r_{t,maxtemp} := \widetilde{y}_{t,maxtemp} - \widehat{\widetilde{y}}_{t,maxtemp}$ ($\widehat{\widetilde{y}}_{t,maxtemp}$ are the fitted values from our model) underline the assumption of independent error terms, since we cannot reject the nullhypothesis, H_0 : residuals are independent against H_1 : they are not, of a Ljung-Box test for different lags (cp. Table 3.9).

Coefficient	Period 1955-1959		Period 1980-1984		Period 2005-2009
	Estimate	p-value of Wald test	Estimate	p-value of Wald test	Estimate
β_0	-0.15	0.00	-0.22	0.00	-0.17
γ_1	0.85	0.00	0.83	0.00	0.79
γ_2	-0.14	0.00	-0.18	0.00	-0.07
γ_3	0.07	0.00	0.12	0.00	0.06
γ_7	0.06	0.00	0.03	0.05	0.06
$\beta_{season=spring}$	0.16	0.00	0.21	0.00	0.19
$\beta_{season=summer}$	0.30	0.00	0.43	0.00	0.34
$\beta_{season=fall}$	0.13	0.00	0.23	0.00	0.16
R^2_{adj}	0.81		0.81		/
$\sigma^2_{maxtemp}$	0.19		0.19		0.19
$\xi_{maxtemp}$	/		/		0.40
$\omega_{maxtemp}$	/		/		0.59
$\alpha_{maxtemp}$	/		/		-1.59

Table 3.8: Summary of the coefficient estimations and the adjusted R^2_{adj} for the daily maximum air temperature models as well as the estimated parameters of the fitted distributions of the corresponding error terms for the different periods. Note that there exists no Wald tests and R^2_{adj} for the last period since we use a linear skew normal regression.

The parameters of the fitted distributions of the error terms are again estimated by the method of maximum likelihood and the results are given in Table 3.8. These distributions have mean zero and the same estimated variance over the periods. Similarly to the daily mean temperature model the fitted distribution of the last period is slightly negatively skewed around mean 0, described by the estimated shape parameter $\widehat{\alpha}_{maxtemp} = -1.59$. Thus, negative error terms have a little wider variability than postive error terms. The further diagnostic plots in Figure 3.8 certify the goodness of our model fits.

Period	lag 1	lag 5	lag 365
1955-1959	0.90	0.87	0.03
1980-1984	0.82	0.65	0.01
2005-2009	0.85	0.84	0.71

Table 3.9: p-values of the Ljung-Box tests (with lag 1,5 and 365) with the raw residuals of the daily maximum air temperature model for the different periods. As a result we cannot reject the null hypothesis of a Ljung Box Test (H_0 : residuals are independent vs. H_1 : residuals are not independent) at a 5% significance level (except for periods 1955-1959 and 1980-1984 at lag 365).

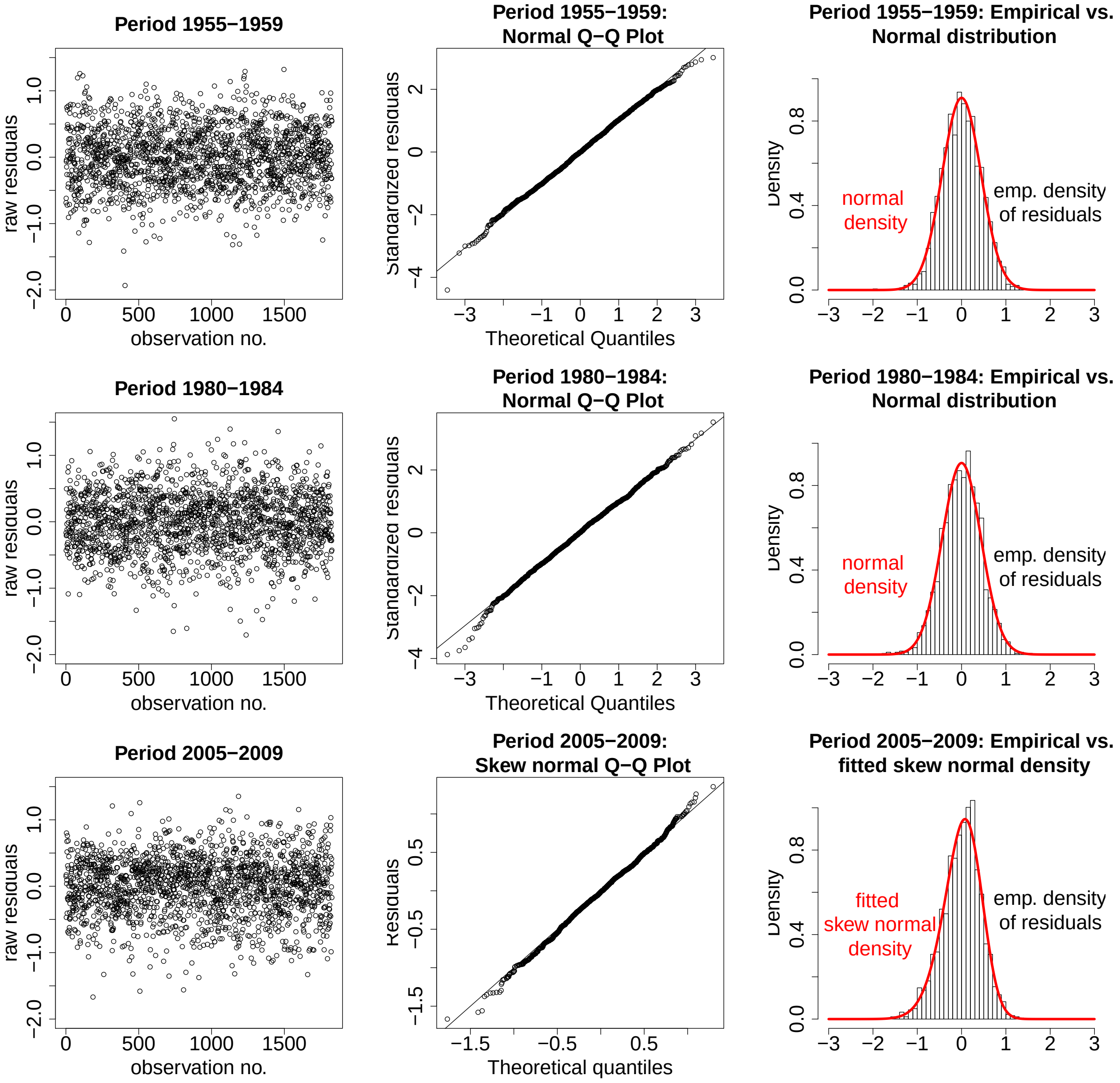


Figure 3.8: Goodness of fit plots for our fitted daily maximum air temperature models. The left panel describes the raw residuals plotted against their observation numbers. The middle panel shows Q-Q plots of the corresponding fitted distributions which are compared to the empirical ones in the third panel. For all periods the plots do not exhibit any severe objections against our assumed models.

3.4 Fitted values of the temperature models

After building the marginal models of our three (standardized) temperature variables we would still like to have a look at the fitted values of our models. Clearly, the standardized variables are not much of our interest. Therefore we rather have to restandardize our fitted values to be able to compare them to real observations.

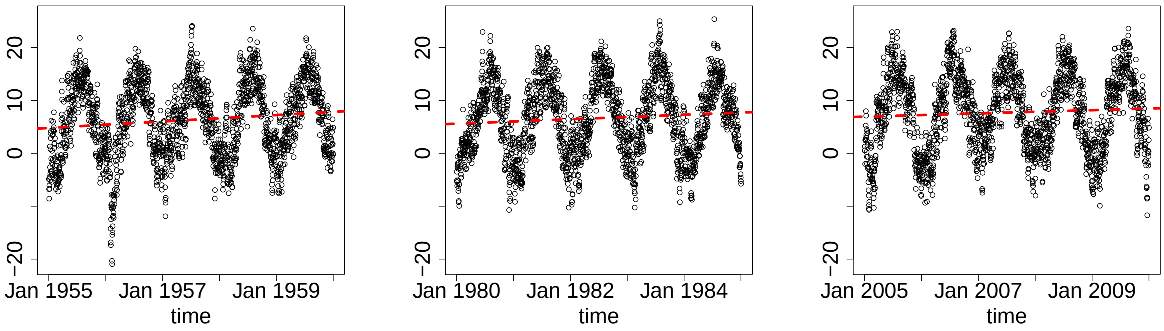


Figure 3.9: Restandardized fitted values of the daily mean temperature model for three periods (Period 1955-1959 in the left plot, period 1980-1985 in the middle plot and period 2005-2009 in the right plot). In all three periods we detect an significant increase of the fitted values over time. The red dashed lines correspond to simple linear regressions of the fitted mean against time.

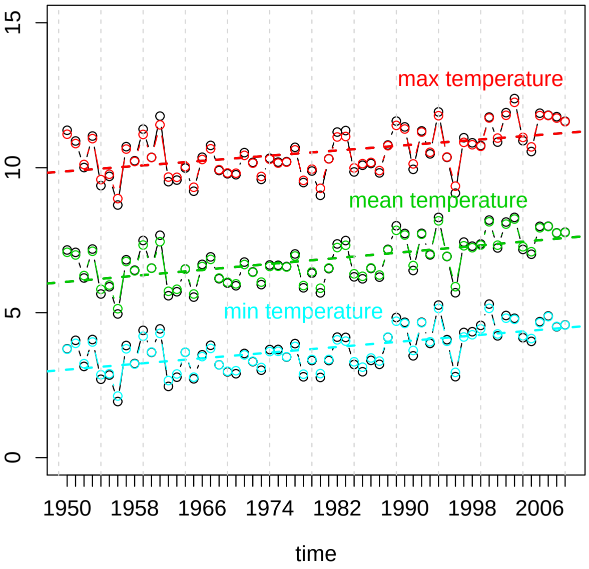


Figure 3.10: Yearly means of the fitted (restandardized) values of daily mean, minimum and maximum air temperature (colored plots) compared to the observed yearly means of the corresponding variables (black lines). The colored dashed lines represent simple linear regressions of the yearly means against time. In all cases we observe an significant increase of the mean over time. The gray dashed lines in the background mark the 12 5-years periods, i.e. the points in time when our models switch.

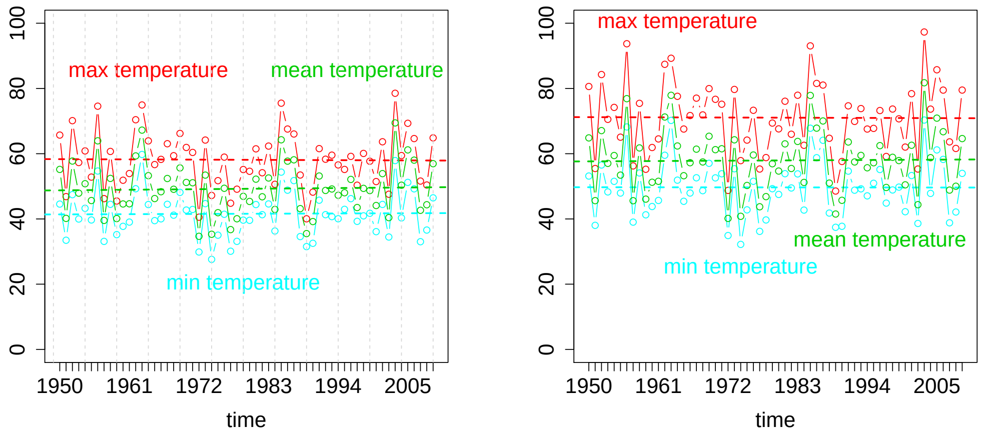


Figure 3.11: Yearly variances of the fitted (restandardized) values of daily mean, minimum and maximum air temperature (left plot) compared to the observed yearly variances of the corresponding variables (right plot). The colored dashed lines represent simple linear regressions of the yearly variances against time. In all cases there is no significant trend detectable. The gray dashed lines in the background mark the 12 5-years periods, i.e. the points in time when our models switch.

In this connection restandardization of the fitted values, i.e. in contrast to Equation (3.1), means

$$\widehat{y}_{t,temp} := \widehat{\widetilde{y}}_{t,temp} \times \widehat{s}_{temp} + \bar{y}_{temp}$$

where $t \in \{1, ..., 1826\}$ ($t \in \{1, ..., 1827\}$ when the period contains two leap years) and $\widehat{\widetilde{y}}_{t,temp}$ are the fitted values out of one of our standardized temperature models, $\widehat{s}_{temp}$ the empirical standard deviation and $\bar{y}_{temp}$ the empirical mean of the underlying variable observations as defined in (3.1). A comparison of, e.g., restandardized daily mean air temperature fitted values for three periods is given in Figure 3.9. Confirmed by simple linear regressions (red dashed lines) the fitted means increase significantly in all three periods. In detail we detect an increase on average by 2.5°C per 5 years in the period 1955-1959, by 2°C per 5 years in the period 1980-1984 and by 1.5°C per 5 years in the last period 2005-2009. However, the considered periods follow only a timespan over five years. In other five years periods the fitted values could theoretically descrease significantly again. Thus it is more interesting to look at the yearly means of the fitted restandardized values. We can calculate the yearly means of the fitted values since we modeled the temperature variables for all 12 five years periods over a whole time span of sixty years (1950-2009). These are illustrated in Figure 3.10 for our three temperature variables by the colored plots. For comparison we plot also the corresponding yearly means of the corresponding variable oberservations, i.e. the ”real” values (black lines). According to our goodness of fit diagnostics one cannot detect any large differences between the fitted values and the observed ones for all three temperature variables. Simple linear regression (dashed lines) of the fitted values against time show an significant increase of the yearly mean of daily mean temperatures on average by 0.025°C p.a. (1.5°C in 60 years), an significant increase of the yearly mean of daily maximum temperatures on average by 0.022°C p.a. (1.32°C in 60 years) as well as an significant increase of the yearly mean of daily minimum temperatures on average by 0.024°C p.a. (1.44°C in 60 years). The gray dashed lines in the background of Figure 3.10 represent the points in time where our models switch every five years. Note, to place emphasis on it, that we detect a significant increase of the yearly means of temperatures only over the last 60 years. We cannot say anything about the overall view of temperature developments over the last and next centuries.

Finally, we also take a look at the yearly variances of the fitted (restandardized) values of three temperature variables over the last 60 years (left plot in Figure 3.11). No significant trend is detectable. Clearly the observed yearly variances of the temperature variables (”real” variances in the right plot of Figure 3.11) are higher compared to the variances of the fitted values, since these differences are explained by the error terms in our models.

3.5 Daily mean relative humidity

We are using a beta regression model to model the distributional behavior of the daily mean humidity variable $Y_{t,humidity} \in (0\%, 100\%)$ at t in Hohenpeissenberg. The beta regression model is defined in Section 2.5.3 for modeling continuous variables Y that assume values in the open standard unit interval $(0,1)$. Thus in case of our daily mean humidity variable, we instead model

$$\widetilde{Y}_{t,humidity} := \frac{Y_{t,humidity}}{100\%}$$

with values in $(0,1)$ for every period, based on the corresponding observations $\widetilde{y}_{t,humidity} := \frac{y_{t,humidity}}{100\%}$, $t \in \{1, ..., 1826\}$ ($t \in \{1, ..., 1827\}$ when the period contains two leap years).

Using the parametrization of (2.28) in Section 2.1.8, we assume that $\widetilde{Y}_{t,humidity} \sim B(\widetilde{\mu}_{t,humidity}, \phi_{humidity})$ with $E\left[\widetilde{Y}_{t,humidity}\right] = \widetilde{\mu}_{t,humidity}$ and appropriate precision parameter $\phi_{humidity}$. Hence, due to the regression model, we assume that

$$g(\widetilde{\mu}_{t,humidity}) = \widetilde{\eta}_{t,humidity}$$

where $g : (0,1) \to \mathbb{R}$ is the link function and $\widetilde{\eta}_{t,humidity}$ is the linear predictor at t. It is defined as

$$\begin{aligned}\widetilde{\eta}_{t,humidity} &= \beta_0 + \gamma_1 \widetilde{y}_{t-1,humidity} + \beta_{sinwinddirection} \sin\left(x_{t,winddirection} \times \tfrac{\pi}{180}\right) + \\ &\quad \beta_{coswinddirection} \cos\left(x_{t,winddirection} \times \tfrac{\pi}{180}\right) + \beta_{season} x_{t,season},\end{aligned} \tag{3.5}$$

for $t \in \{1, ..., 1826\}$ ($t \in \{1, ..., 1827\}$ when the period contains two leap years), where

(i) Covariable $x_{t,winddirection}$, which is the daily mean wind direction at t,
($x_{t,winddirection} \in [0°, 360°]$)
captures the ”Fön”-effect (dry and hot winds in the foothills region of the Alps).

(ii) Covariable $x_{t,season} = 1$ for winter days, $x_{t,season} = 2$ for spring days, $x_{t,season} = 3$ for summer days and $x_{t,season} = 4$ for fall days, for all t (as factors). This division results from the meteorological seasons:

- winter = December, January, February
- spring = March, April, May
- summer = June, July, August
- fall = September, October, November

Note that we choose the usual logit link as our link function g in the model, i.e. $g(\mu) := \log(\mu/(1-\mu))$ (cp. to Section 2.5.3). We checked that other (asymmetric) link functions as the log-log, complementary log-log, probit or Cauchy link do not improve our model. Additional further autoregressive variables in 3.5 also did not result in better model fits. However, an improvement can be achieved by adding the covariable of daily mean wind direction. It explains an additional effect on humdity resulting from dry and hot winds in the foothills region of the Alps, called "Fön" winds. This effect does not appear in other lowland regions therefore the adding of such a covariable seems to be reasonable. A method how to calculate the daily mean of hourly measured wind directions can be found in Appendix B.
Notice that there is no available data of daily winddirection for period 1950-1954 to period 1970-1974. Thus we implement the above mentioned linear predictor without the two winddirection covariates for these time spans, i.e.

$$\widetilde{\eta}_{t,humidity} = \beta_0 + \gamma_1 \widetilde{y}_{t-1,humidity} + \beta_{season} x_{t,season}.$$

The maximum likelihood estimates of the coefficients in (3.5) and the p-values of corresponding asymptotic Wald tests (i.e. testing $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ and $H_0 : \gamma_1 = 0$ vs. $H_1 : \gamma_1 \neq 0$ respectively) are given in Table 3.10. Note as before, that the coefficients $\beta_{season=spring}$, $\beta_{season=summer}$ and $\beta_{season=fall}$ describe the seasonal difference of $\widetilde{\eta}_{t,humidity}$ between spring and winter, summer and winter or fall and winter, respectively, depending on the season at time t (i.e. dummy coding for factor variable $x_{t,season}$). Almost all coefficients seem to have an significant influence on the linear predictor $\widetilde{\eta}_{t,humidity}$ except some seasonal differences. The pseudo R^2 raises from 0.17 to the level of 0.32 by adding the daily mean winddirection covariates.

The resulted fitted values (× 100) are plotted together with the observations in Figure 3.12. The diagnostic plots for our model fits of period 1955-1959, period 1980-1984 and period 2005-2009 can be found in Figures 3.13 - 3.15. They include the plot of residuals against their observation numbers (top left panel: should show no systematic pattern), the plot of the corresponding Cook's distance (middle panel at the top: should be small for all observations), the plot of generalized leverage vs. predicted values (top right panel: should be small for all predicted values), the plot of residuals against the values of the linear predictor (bottom left panel: no trend should be detectable), the half-normal plot of the residuals (bottom middle panel: points should lie inside the envelope) and the plot of predicted vs. observed values (bottom right panel: should follow a straight line). When we compare the half-normal plots (their constructions are described in Section 2.5.3) we

Coefficient	Period 1955-1959		Period 1980-1984		Period 2005-2009	
	Estimate	p-value of Wald test	Estimate	p-value of Wald test	Estimate	p-value of Wald test
β_0	-1.21	0.00	-0.56	0.00	-0.84	0.00
γ_1	3.30	0.00	2.53	0.00	3.00	0.00
$\beta_{sinwinddirection}$	/		0.52	0.00	0.60	0.00
$\beta_{coswinddirection}$	/		-0.41	0.00	-0.38	0.00
$\beta_{season=spring}$	-0.03	0.55	-0.23	0.00	-0.34	0.00
$\beta_{season=summer}$	-0.03	0.57	-0.31	0.00	-0.39	0.00
$\beta_{season=fall}$	0.24	0.00	-0.05	0.30	-0.04	0.44
pseudo R^2	0.17		0.29		0.32	
$\widehat{\phi}_{humidity}$	5.30		7.49		7.43	

Table 3.10: Summary of the coefficient estimations together with the p-values of the corresponding asymptotic Wald tests, the pseudo R^2 and the estimated precision parameters $\widehat{\phi}_{humidity}$ in the beta regression model for daily mean relative humidity of the three periods.

detect that a proportion of residuals fall outside the envelope lines in all periods. This proportion is getting smaller in both later periods compared to the period 1955-1959 probably due to the fact of including the wind direction variables into the model. The other plots offer no severe noticeable problems. However, the basic assumption of a beta distributed relative humidity variable seems to be fundamental. Maybe for a "future work", one should use a kind of weighted beta regression to improve the model fit here.
Note that the variance of $\widetilde{y}_{t,humidity}$ is a function of $\widetilde{\mu}_{t,humidity}$ and thus our regression model based on this parameterization is naturally heteroskedastic.

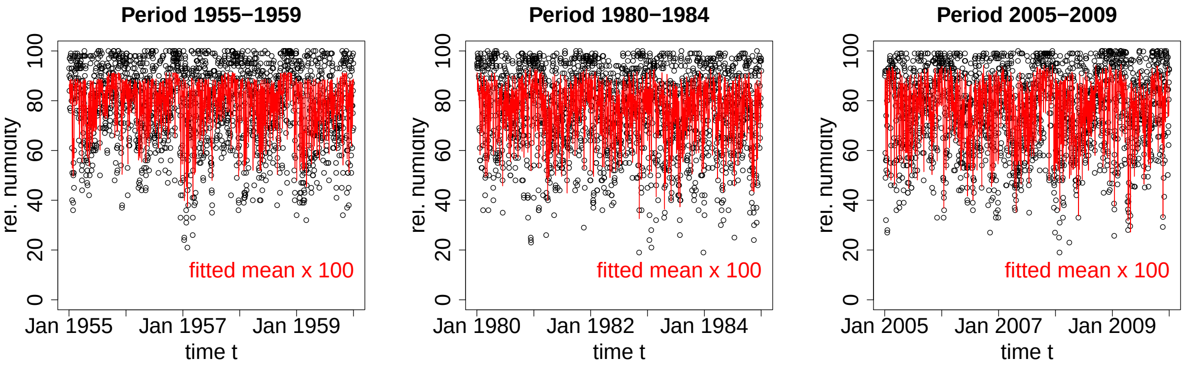


Figure 3.12: Observations of daily mean relative humidity against the fitted values (×100) of the corresponding regression models (red points) for the three periods.

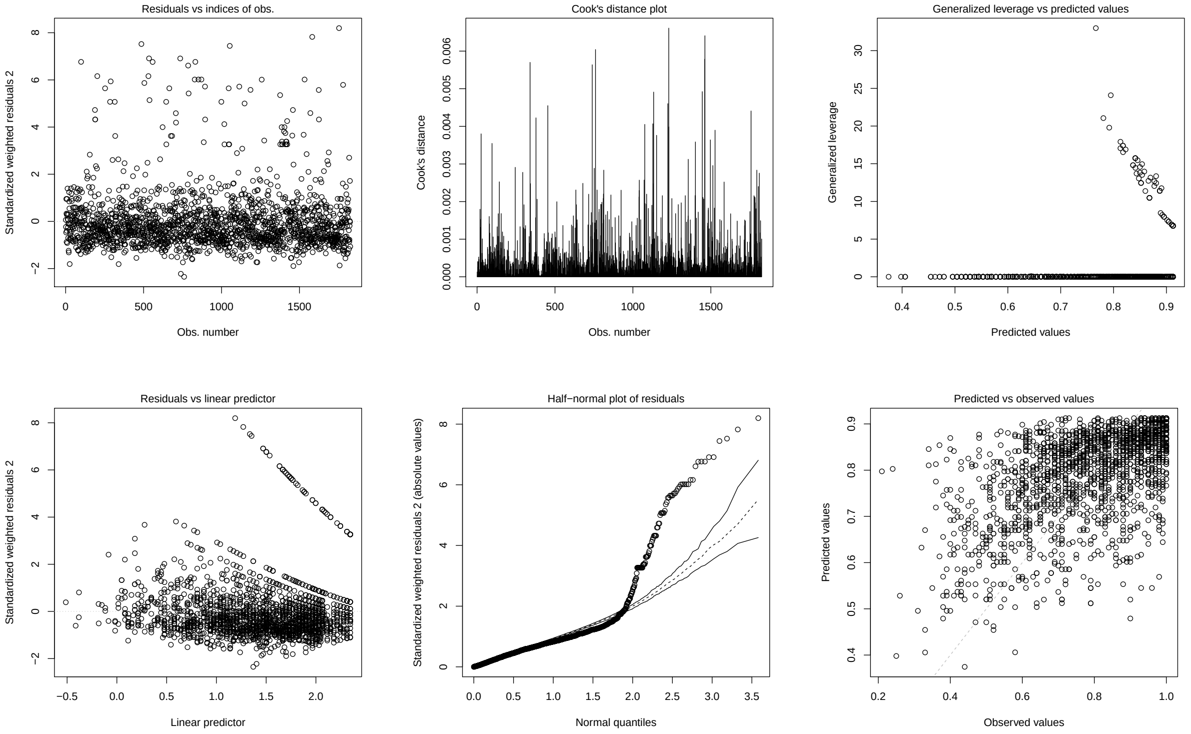


Figure 3.13: Diagnostic plots of daily mean relative humidity model - period 1955-1959.

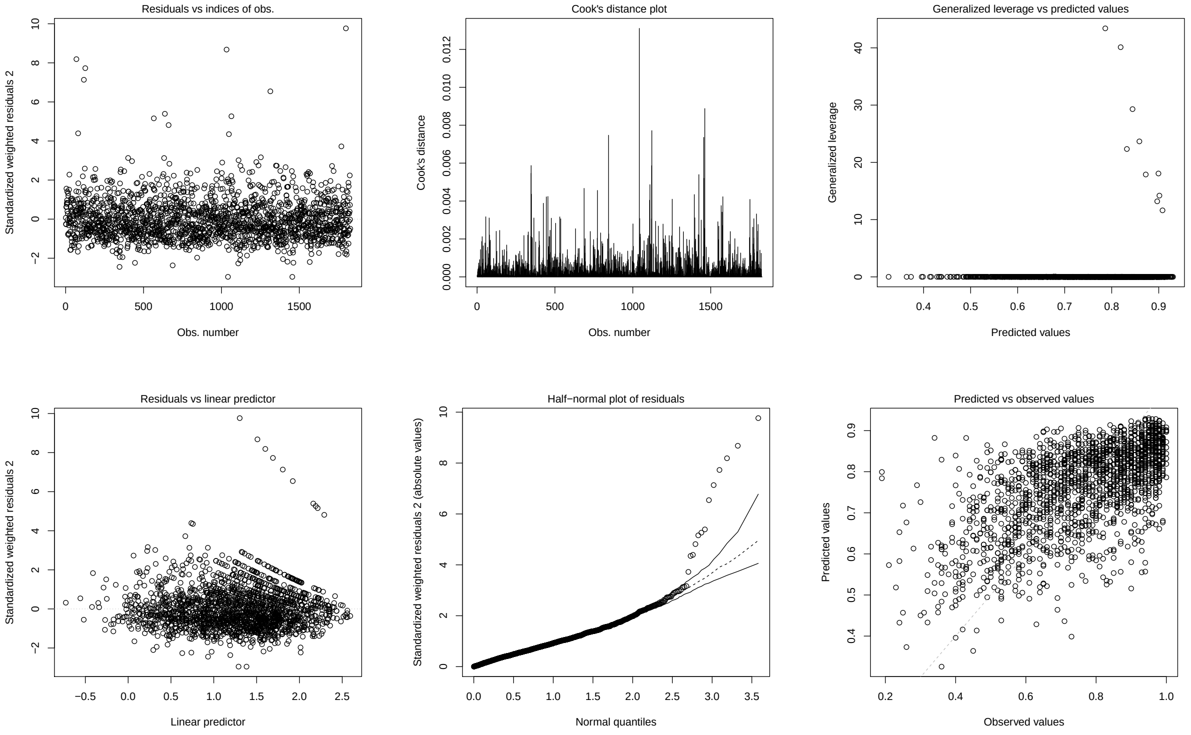


Figure 3.14: Diagnostic plots of daily mean relative humidity model - period 1980-1984.

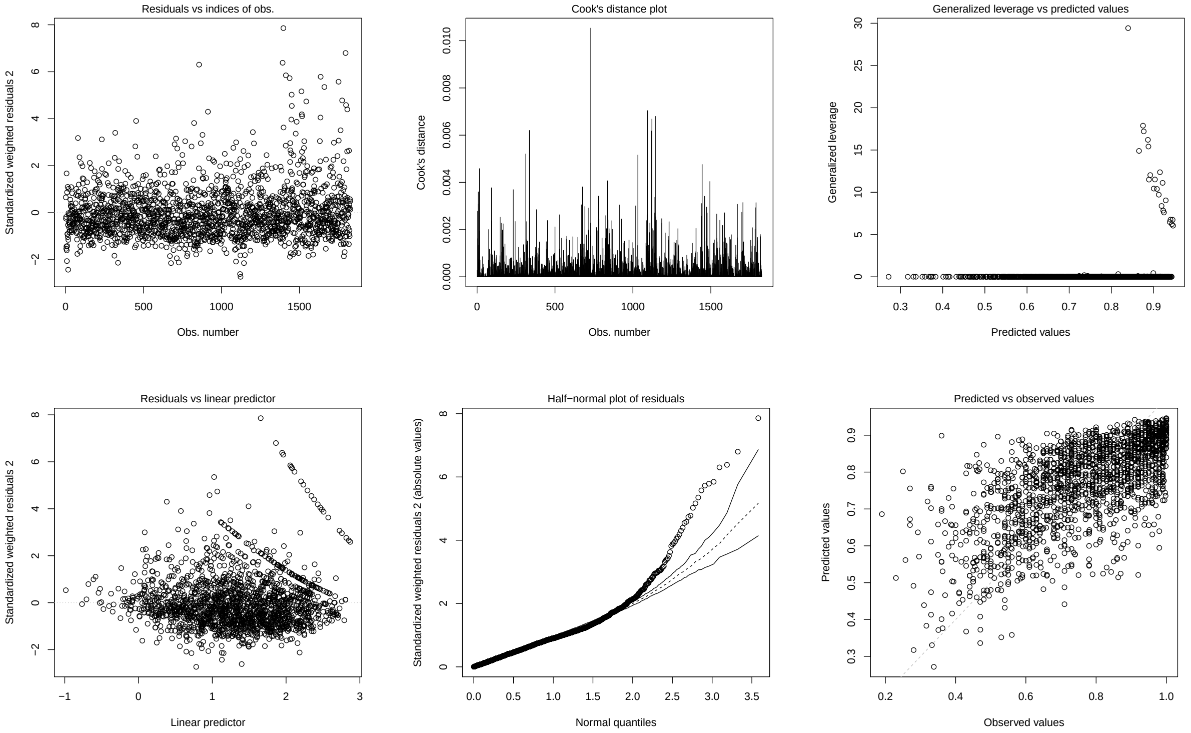


Figure 3.15: Diagnostic plots of daily mean relative humidity model - period 2005-2009.

3.6 Daily mean air pressure

The distributional modeling of the daily mean air pressure variable basically follows the modeling of the temperature variables. More precisely it follows the daily minimum air temperature model, since we also decided to fit a heteroskedastic regression model with skew t distributed residuals in case of air pressure. Note that we again regress on standardized variables and hence we assume the following model for the daily mean air pressure observations:

$$\begin{aligned}\widetilde{y}_{t,press} &= \beta_0 + \gamma_1 \widetilde{y}_{t-1,press} + \gamma_2 \widetilde{y}_{t-2,press} + \gamma_3 \widetilde{y}_{t-3,press} \\ &\quad + \gamma_7 \widetilde{y}_{t-7,press} + \beta_{season} x_{t,season} + \epsilon_{t,press},\end{aligned} \tag{3.6}$$

for $t \in \{1, ..., 1826\}$ ($t \in \{1, ..., 1827\}$ when the period contains two leap years), where

(i) $\widetilde{y}_{t,press} = \frac{y_{t,press} - \bar{y}_{press}}{\widehat{s}_{press}}$, where $\widehat{s}_{press}$ is the empirical standard deviation and $\bar{y}_{press}$ the empirical mean of the time series.

(ii) Covariable $x_{t,season} = 1$ for winter days, $x_{t,season} = 2$ for spring days, $x_{t,season} = 3$ for summer days and $x_{t,season} = 4$ for fall days, for all t (as factors). This division results from the meteorological seasons:

- winter = December, January, February

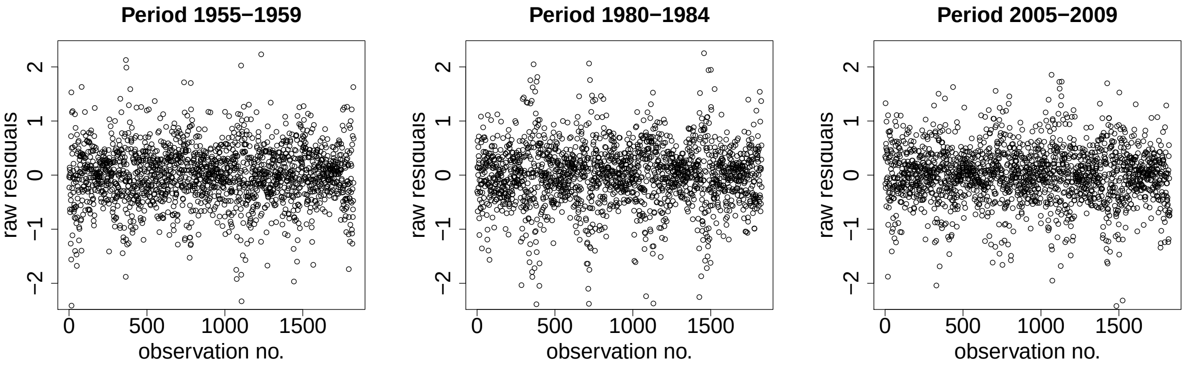


Figure 3.16: Plots of the raw residuals from our daily mean air pressure model against the observation numbers for every period. One might detect different variances of the residuals at different observations numbers t and thus at different point in time t.

- spring = March, April, May
- summer = June, July, August
- fall = September, October, November

(iii) We assume:

$$\epsilon_{t,press} \sim skewt(\xi_{t,press}, \omega_{t,press}, \alpha_{t,press}, df_{t,press})$$

(independent but with different variances at time t).

The last assumption (iii) is detected by looking at the plots of the raw residuals $r_{t,press} := \widetilde{y}_{t,press} - \widehat{\widetilde{y}}_{t,press}$ against the observation numbers (Figure 3.16), where $\widehat{\widetilde{y}}_{t,press}$ are the fitted values of model (3.6). They show that the residuals might have different variances at different points in time. For that reason we use the method of weighted least squares (described in Section 2.5.2) to estimate the parameters in (3.6) with appropriate weights $w_{t,press}$ such that

$$\frac{1}{\sqrt{w_{t,press}}}\epsilon_{t,press} \sim skewt(\xi_{press}, \omega_{press}, \alpha_{press}, df_{press}) \text{ (i.i.d)},$$

for all t. Suitable weights, as already in the minimum air temperature model, are given by the empirical standard deviation over all standardized observations at day d of the year in the period, i.e.

$$w_{d,press} := sd(\widetilde{y}_{d,press}, \widetilde{y}_{d+365,press}, \widetilde{y}_{d+2\times365,press}, \widetilde{y}_{d+3\times365,press}, \widetilde{y}_{d+4\times365,press}),$$

where d=1,...366. Then it follows that $w_{t,press} = w_{d(t),press}$, where $d(t) \in \{1, ..., 366\}$ denotes the corresponding day of the year at oberservation t.

The estimates of the coefficients in (3.6) by the method of weighted least squares (WLS) are presented in Table 3.11 for the different periods. We see that the standardized

Coefficient	Period 1955-1959	Period 1980-1984	Period 2005-2009
	Estimate	Estimate	Estimate
β_0	-0.08	-0.01	-0.01
γ_1	0.99	0.99	1.06
γ_2	-0.35	-0.38	-0.45
γ_3	0.07	0.10	0.14
γ_7	0.01	0.03	0.00
$\beta_{season=spring}$	0.05	-0.08	-0.02
$\beta_{season=summer}$	0.16	0.11	0.08
$\beta_{season=fall}$	0.19	0.09	0.10
σ^2_{press}	0.38	0.39	0.32
ξ_{press}	0.19	0.23	0.09
ω_{press}	0.55	0.57	0.49
α_{press}	-0.48	-0.56	-0.24
df_{press}	9.16	12.10	7.61

Table 3.11: Summary of the coefficient estimations after weighted least square regression of (standardized) daily mean air pressure as well as the estimated parameters for the fitted distributions of the corresponding residuals for the different periods. Note that there exists no Wald tests in case of skew t regressions.

variable of daily mean air pressure at t strongly depends on the value of the standardized variable at $t-1$ (estimates $\widehat{\gamma}_1 \geq 0.99$). Note again, since we do a skew t regression we cannot perform any Wald tests to test the significance of the parameters (cp. Section 2.5.2). As before, the coefficients $\beta_{season=spring}$, $\beta_{season=summer}$ and $\beta_{season=fall}$ describe the seasonal difference of the standardized daily mean air pressure between spring and winter, summer and winter or fall and winter, respectively, depending on the season at time t (i.e. dummy coding for factor variable $x_{t,season}$).

The raw residuals $r^{WLS}_{t,press} := \widetilde{y}_{t,press} - \widehat{\widetilde{y}}^{WLS}_{t,press}$ after weighted least squares seem to be independent (compare Table 3.12 after Ljung-Box tests) which underlines the assumption of

Period	lag 1	lag 5	lag 365
1955-1959	0.11	0.12	0.19
1980-1984	0.09	0.17	0.48
2005-2009	0.05	0.03	0.64

Table 3.12: p-values of the Ljung-Box tests (with lag 1,5 and 365) with the raw residuals after weighted least square regression of the daily mean air pressure for the different periods. As a result we cannot reject the null hypothesis of a Ljung Box Test (H_0 : residuals are independent vs. H_1 : residuals are not independent) at a 5% significance level (except for period 2005-2009 at lag 5).

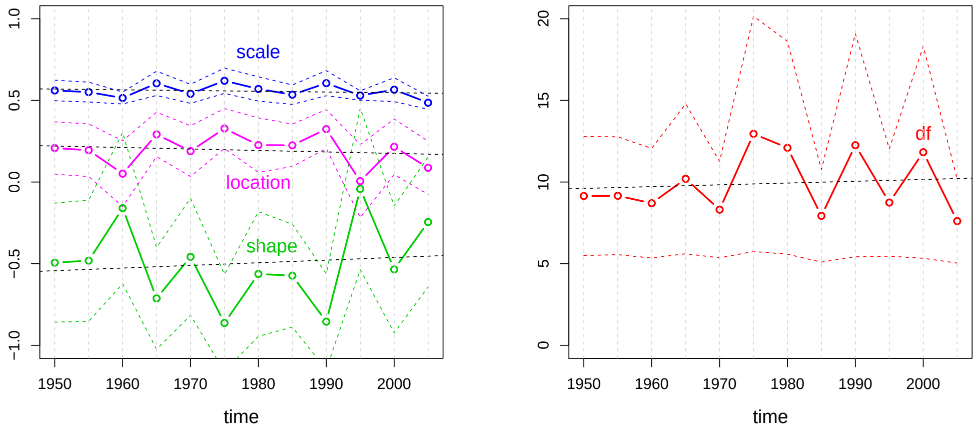


Figure 3.17: Estimated scale $\widehat{\omega}_{press}$, location $\widehat{\xi}_{press}$ and shape $\widehat{\alpha}_{press}$ parameters (left panel) as well as estimated degrees of freedom $\widehat{df}_{press}$ (right panel) of the fitted distributions of the residuals in the (standardized) daily mean air pressure model for all periods. The colored dashed lines represents corresponding 95%-confidence intervals and the black dashed lines correspond to a simple linear regression of the parameters against time. Note that there is no significant trend detectable. The gray dashed lines in the background mark the 12 5-years periods.

independent error terms. Note that $\widehat{y}^{WLS}_{t,press}$ here denotes the fitted values of the weighted least square regression.

The distribution parameters are estimated by the method of maximum likelihood and can also be found in Table 3.11. For all regarded periods we fitted negatively skewed t distributions of the error terms after WLS since $\widehat{\alpha}_{press} < 0$. This means that the variability of negative error terms (”air pressure down surprises from the mean”) is slight higher than the variability of positive error terms (”air pressure up surprises from the mean”). The estimated parameters $\widehat{\xi}_{press}$, $\widehat{\omega}_{press}$ and $\widehat{\alpha}_{press}$ as well as the estimated degrees of freedom $\widehat{df}_{press}$ for all of the 12 periods (all periods are marked as gray dashed lines in the background) are illustrated in Figure 3.17. Note that there is no significant trend detectable over time since we did a simple linear regression of the parameters against time (black dashed line) and so these distributions have almost the same estimated variance and a mean close to zero over all periods.

The diagnostic plots in Figure 3.18 underline our model assumptions. In detail, the raw residuals after WLS regression seem to be independent, they do not exihibit any systematical pattern against their observation numbers t (and thus against time). Further they now seem to have a common variance at all points in time. The corresponding Q-Q plots of the fitted distributions of the residuals are following straight lines which can be compared by looking in the plots of the empirical distributions against the fitted ones.

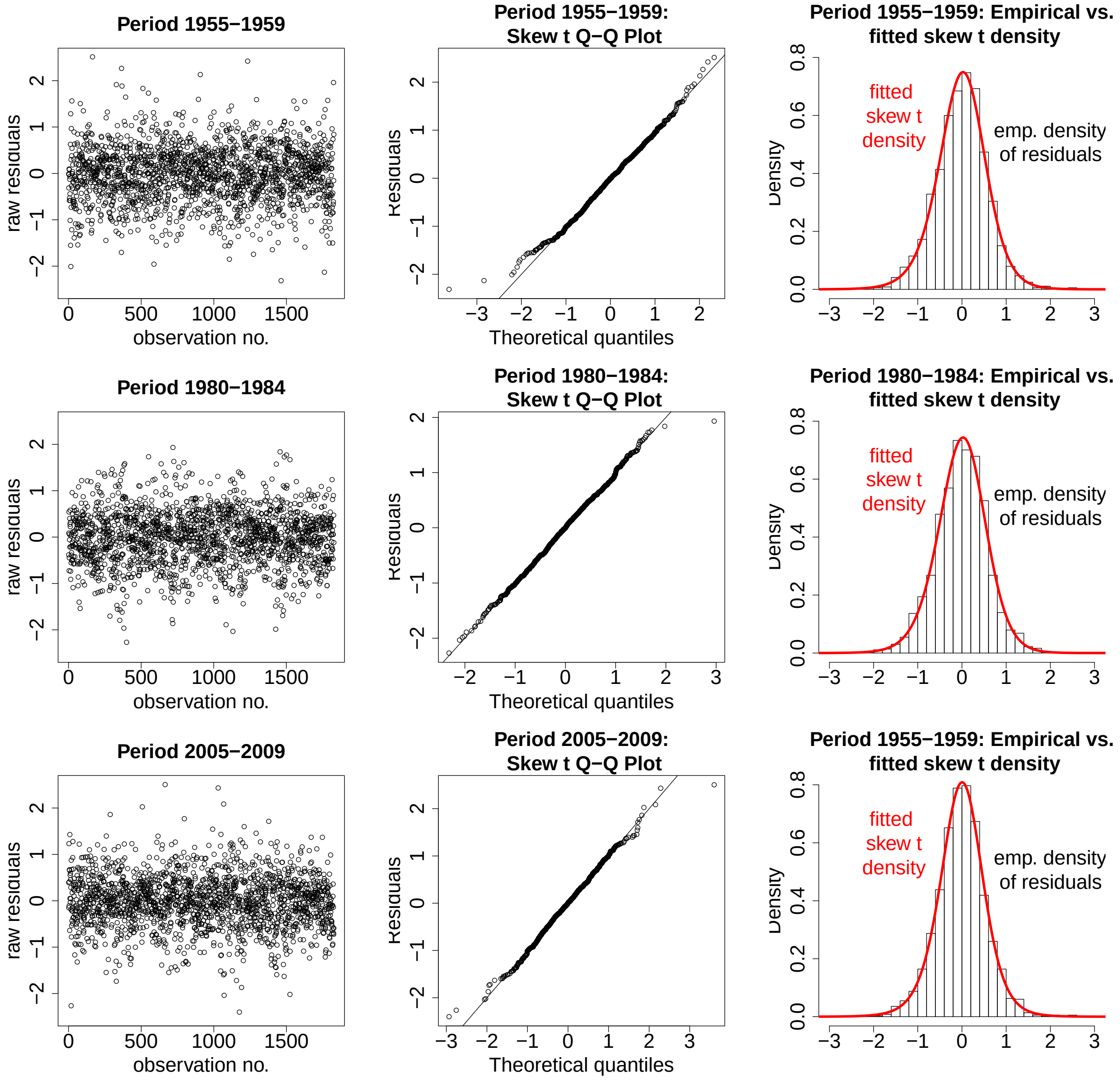


Figure 3.18: Goodness of fit plots for our fitted daily mean air pressure models. The left panel describes the raw residuals after weighted least squares regression plotted against their observation numbers. The middle panel shows Q-Q plots of the corresponding fitted distributions which are compared to the empirical ones in the third panel. For all periods, the plots do not exhibit any severe objections against our assumed models.

We take a look at the restandardized fitted values, i.e.

$$\widehat{y}_{t,press} := \widehat{\widetilde{y}}_{t,press} \times \widehat{s}_{press} + \bar{y}_{press}$$

where $t \in \{1, ..., 1826\}$ ($t \in \{1, ..., 1827\}$ when the period contains two leap years) and $\widehat{\widetilde{y}}_{t,press}$ are the fitted values from our above described standardized mean air pressure

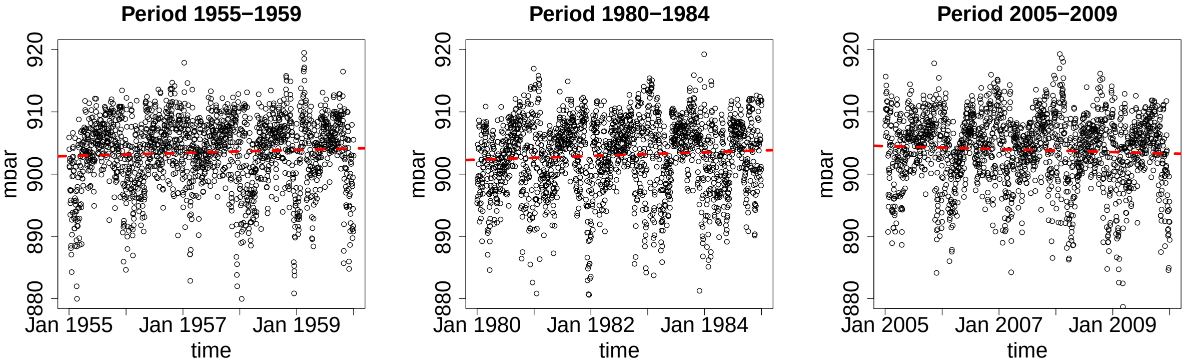


Figure 3.19: Restandardized fitted values of the daily mean air temperature model for three periods (Period 1955-1959 in the left plot, period 1980-1985 in the middle plot and period 2005-2009 in the right plot). The fitted means increase significantly in periods 1955-1959 and 1980-1984 and decrease significantly in the last period 2005-2009 over time. The red dashed lines correspond to simple linear regressions of the fitted mean against time.

model, $\widehat{s}_{press}$ the empirical standard deviation and $\bar{y}_{press}$ the empirical mean of the underlying air pressure observations. A comparison of restandardized daily mean air pressure fitted values for three periods is given in Figure 3.19. Confirmed by simple linear regressions (red dashed lines) the fitted daily means increase significantly in periods 1955-1959 and 1980-1984 and decrease significantly in the last period 2005-2009. In detail we detect an increase on average by 1.2 mbar per 5 years in the period 1955-1959 and by 1.48 mbar per 5 years in the period 1980-1984. In the last period 2005-2009 we have an siginficant decrease of the fitted daily values on average by 1.2 mbar per 5 years.

The look at the yearly means of the fitted restandardized values of daily mean air pressure (red line in the left plot in Figure 3.20) over sixty years (1950-2009) indicates an significant increase on average by 0.02 mbar p.a. (1.2 mbar in 60 years). For comparison we also plot the corresponding yearly means of the daily oberservations of mean air pressure at Hohenpeissenberg, i.e. the "real" values of measured air pressure (black line). They are almost falling on the same line. The gray dashed lines in the background of Figure 3.20 represent the points in time where our models switch every five years. Note again, that we detect an significant increase of the yearly means of daily mean air pressure only over the last 60 years. However, we cannot say anything about the overall view of air pressure developments over the last and next centuries.

The look at the yearly variances of the fitted (restandardized) values of daily mean air pressure over the whole period 1950-2009 (red line in the right plot in Figure 3.20) detects no significant trend. Clearly the yearly variances of the air pressure observations ("real" variances; the black line in the right plot of Figure 3.20) are higher compared to the variances of the fitted values, since these differences are explained by the error terms in our model.

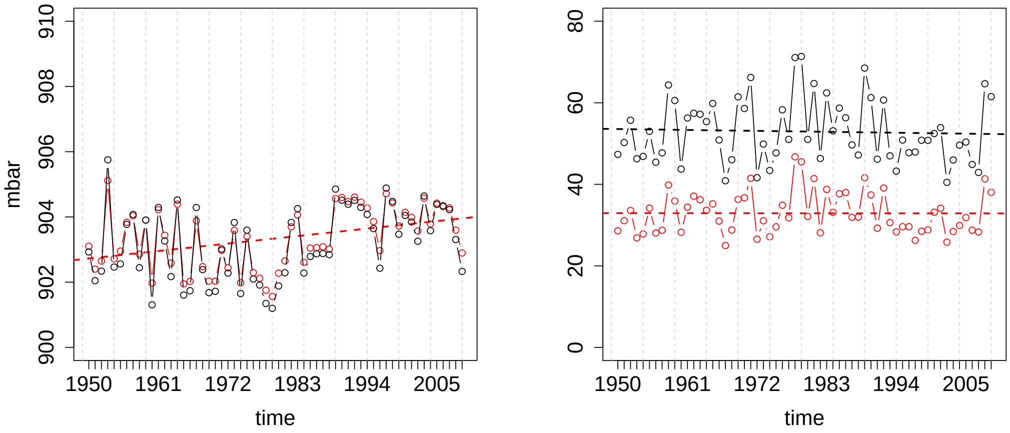


Figure 3.20: Yearly means of the fitted restandardized values of daily mean air pressure (red line in the left plot) compared to the yearly means of the "real" air pressure observations (black line in the left plot). One detects a significant increase in yearly means of daily mean air pressure based on simple linear regressions of the means against time (dashed lines). The yearly variances of the fitted (restandardized) values of daily mean air pressure are presented by the red line in the right plot compared to the observed yearly variances of the corresponding variable (black line in the right plot). The dashed lines represent simple linear regressions of the yearly variances against time. In the variance case there is no significant trend detectable. The gray dashed lines in the background mark the 12 5-years periods, i.e. the points in time when our models switch.

3.7 Daily total precipitation

Modeling daily total precipitation, i.e. $Y_{t,prec}$ at time t in Hohenpeissenberg, is a special case because the variable is often equal to zero. Stern and Coe [1984] introduced a two step method to model such a behavior:

1. We model the rain occurrence on day d of the year, i.e. day d of the year is dry or has rain with an appropriate rain probability following a binomial regression and

2. the positive rain amount when day d of the year is rainy is modeled by a gamma regression,

where $d = 1, ..., 366$. Hence, we model the distributional behavior of $Y_{d,prec}$ for every day d of the year and assume that the distribution of $Y_{t,prec}$ equals the distribution of $Y_{d(t),prec}$, where $d(t)$ denotes the corresponding calendar day d of the year at observation t. Due to Stern and Coe [1984] we implement the model based on observations measured from the whole period 1950-2009 (all in all 21915 observations) to have a suitable number of data.

3.7.1 Modeling rain occurrence

We begin our model with defining the random variable $J(d)$ as

$$J(d) = \begin{cases} 0, & \text{if day } d \text{ is dry,} \\ 1, & \text{if day } d \text{ has rain,} \end{cases} \tag{3.7}$$

where $d = 1, .., 366$.

We assume that $J(d)$ is a first-order, non stationary Markov chain, i.e. it holds that

$$P(J(d) = 1|J(d-1), J(d-2), J(d-3), ...) = P(J(d) = 1|J(d-1)).$$

Thus we are interested in modeling this probability, i.e. the rain success probability "day d has rain", conditional on whether the day before had rain or not. Note, we checked that the assumption of higher ordered Markov chains (i.e. conditioning on $J(d-1)$, $J(d-2)$, $J(d-3)$ and so on) would not improve our model fit here. So let us define

$$\begin{aligned} p_0(d) &:= P(J(d) = 1|J(d-1) = 0) \\ p_1(d) &:= P(J(d) = 1|J(d-1) = 1), \end{aligned} \tag{3.8}$$

where $d = 1, ..., 366$.

Based on our observations we further define

$$\begin{aligned} n_{01}(d) &:= \text{Number of days with } J(d) = 1 \text{ and } J(d-1) = 0, \\ n_{11}(d) &:= \text{Number of days with } J(d) = 1 \text{ and } J(d-1) = 1, \end{aligned} \tag{3.9}$$

$d = 1, ..., 366$.

It is reasonable to assume that the in (3.9) defined numbers of success (success $\widehat{=}$ rain) on day d of the year, depending on what happened one day before, are binomial distributed, i.e.

$$n_{i1}(d) \sim Bin(n_{i+}(d), p_i(d)), \; i = 0, 1, \tag{3.10}$$

where $p_i(d)$ denotes the corresponding rain success probability as defined in (3.8) and $n_{i+}(d)$ is the whole number of days conditional on whether the day before had rain or not $(i = 0, 1)$, $d = 1, ..., 366$, i.e.

$$\begin{aligned} n_{0+}(d) &:= \text{Number of days with } J(d) = 0, 1 \text{ and } J(d-1) = 0, \\ n_{1+}(d) &:= \text{Number of days with } J(d) = 0, 1 \text{ and } J(d-1) = 1. \end{aligned}$$

Consequently, using the usual logit link function, we model the two rain success probabilities of (3.8) by a binomial regression (described in Section 2.5.4) in the following way:

$$\begin{aligned} \mu_i(d) &:= E\left[\frac{n_{i1}(d)}{n_{i+}(d)}\right] = p_i(d) = \frac{\exp(\eta_i(d))}{[1 + \exp(\eta_i(d))]} \\ \Leftrightarrow \; \eta_i(d) &= g(\mu_i(d)) = \log\left(\frac{\mu_i(d)}{(1 - \mu_i(d))}\right), \; i = 0, 1, \end{aligned}$$

Harmonics	Residual degrees of freedom	Residual deviance	p-value of res. deviance test	Partial degrees of freedom	Partial deviance	p-value of part. deviance test
0	365	471.53	0.00	/	/	/
1	363	391.50	0.15	2	80.03	0.00
2	361	374.96	0.30	2	16.54	0.00
3	359	372.38	0.30	2	2.58	0.28

Table 3.13: Analysis of deviance in our binomial regression model for $n_{01}(d)$ ("there was no rain yesterday on day $d-1$"). 2 harmonics are sufficient here, since the p-value of the partial deviance test is equal to 0.28, i.e. the model is not improved by an extending on 3 harmonics. The corresponding residual deviance test (p-value = 0.30) underlines an appropriate model choice with 2 harmonics.

where $\eta_i(d)$ is the linear predictor. In our case according to Stern and Coe [1984], $\eta_i(d)$ should be a Fourier series with k harmonics to capture seasonality effects, i.e.

$$\eta_i(d) = a_{i0} + \sum_{l=1}^{k} [a_{il} \sin(ld') + b_{il} \cos(ld')], \ i = 0, 1, \tag{3.11}$$

where $d' := \frac{2\pi d}{366}$ with $d = 1, ..., 366$.

With the aid of partial deviance tests (i.e. testing H_0 : the model extension improves the model fit vs. H_1 : it does not, see Section 2.5.4) we notice that 2 harmonics are sufficient in the linear predictor (3.11) for both models of $n_{01}(d)$ and $n_{11}(d)$. Further residual deviance tests (i.e. testing H_0 : the described model is true vs. H_1 : it is not, see Section 2.5.4) emphasize the goodness of fit. The detailed values of the deviances and the p-values of the corresponding tests can be found in Tables 3.13 and 3.14. The estimates

Harmonics	Residual degrees of freedom	Residual deviance	p-value of res. deviance test	Partial degrees of freedom	Partial deviance	p-value of part. deviance test
0	365	394.70	0.14	/	/	/
1	363	372.68	0.35	2	22.02	0.00
2	361	363.56	0.45	2	9.12	0.01
3	359	363.27	0.43	2	0.29	0.87

Table 3.14: Analysis of deviance in our binomial regression model for $n_{11}(d)$ ("there was rain yesterday on day $d-1$"). 2 harmonics are also sufficient here, since the p-value of the partial deviance test is equal to 0.87, i.e. the model is not improved by an extending on 3 harmonics. The corresponding residual deviance test (p-value = 0.45) underlines an appropriate model choice with 2 harmonics.

Coefficient	$n_{01}(d)$ model ($i = 0$)		$n_{11}(d)$ model ($i = 1$)	
	Estimate	p-value of Wald test	Estimate	p-value of Wald test
a_{i0}	-0.69	0.00	0.83	0.00
a_{i1}	0.06	0.02	0.14	0.00
b_{i1}	-0.26	0.00	-0.01	0.78
a_{i2}	-0.04	0.23	-0.02	0.39
b_{i2}	0.12	0.00	0.08	0.00

Table 3.15: Coefficient estimations in case of modeling $n_{01}(d)$ and $n_{11}(d)$ respectively with p-values of the corresponding asymptotic Wald tests.

of the coefficients in the linear predictor (3.11) are calculated by an iterative weighted least squares algorithm (IWLS) and the results are presented together with the p-values of corresponding asymptotic Wald tests in Table 3.15. The comparison of the empirical estimations of $p_0(d)$ and $p_1(d)$ in contrast to the fitted ones of our model in Figure 3.21 shows that almost all empirical estimates lie inside the 95%-confidence interval of our fitted probabilities for each day of the year ($d = 1, ..., 366$). Note that empirical estimations of $p_0(d)$ and $p_1(d)$, i.e. $\widehat{p}_0(d)$ and $\widehat{p}_1(d)$ respectively, are calculated here as

$$\begin{aligned}
\widehat{p}_0(d) &:= \frac{\#\text{ observations with rain at day } d \text{ and no rain at day } d-1}{\#\text{ observations with no rain at day } d-1}, \\
\widehat{p}_1(d) &:= \frac{\#\text{ observations with rain at day } d \text{ and at day } d-1}{\#\text{ observations with rain at day } d-1}.
\end{aligned}$$

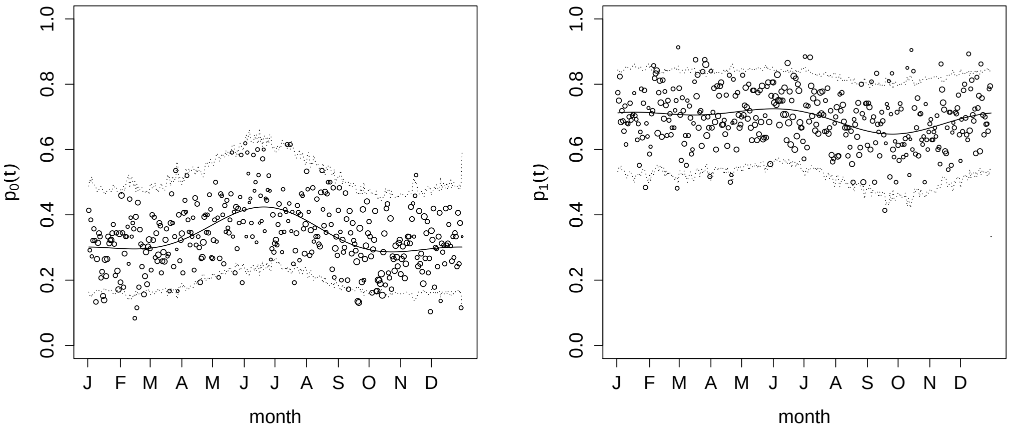


Figure 3.21: Plots of the empirical estimations defined in Equation (3.12) of $p_0(d)$ (points in left plot) and $p_1(d)$ (points in right plot) together with the fitted probabilities from the model (solid lines in both plots) with corresponding 95%-confidence intervals (dashed lines) for every day d of the year, $d = 1, ..., 366$.

The plots in Figure 3.21 present a significant difference between the values of $p_0(d)$ and the values of $p_1(d)$ over the whole year. While the empirical probabilities of rain conditioned on no rain one day before ($\widehat{p}_0(d)$) ranges between 0.1 and 0.61 over the year, the empirical probabilities of rain conditioned on rain occurrence also one day before ($\widehat{p}_1(d)$) lie between 0.42 and 0.92 over the year. Further, the probabilities of rain conditioned on no rain one day before ($p_0(d)$) might be higher in summer than in winter. In contrast, the empirical probabilities of rain conditioned on rain occurrence also one day before ($p_1(d)$) rather stay the same level (only a slight break in fall is illustrated).

3.7.2 Modeling positive rain amount

Now when $J(d) = 1$, i.e. it rains on day d of the year, we are interested in modeling this positive amount of rain on day d for $d = 1, ..., 366$. As before we are modeling the positive rain amount $Y_i^+(d)$ of day d conditional on what happened one day before, i.e. whether day $d-1$ was dry ($i = 0$) or not ($i = 1$). More formally:

$$Y_i^+(d) = \text{ amount of rain at day } d, \text{ when } J(d) = 1 \text{ and } J(d-1) = i,\ i = 0, 1. \tag{3.12}$$

According to Stern and Coe [1984] studies have shown that it is reasonable to assume $Y_i^+(d)$ to be gamma distributed, i.e.

$$Y_i^+(d) \sim Gamma(\kappa_i, \mu_i(d)),\ i = 0, 1.$$

With this parametrization introduced in Section 2.1.9, it holds that

$$E[Y_i^+(d)] = \mu_i(d),\ i = 0, 1,$$

where $d = 1, ..., 366$. Therefore it seems to be reasonable to implement again a GLM, i.e. a gamma regression with log link funtion (defined in Section 2.5.4) to model that mean $E[Y_i^+(d)] = \mu_i(d)$ on day d of the year. Hence, we assume that

$$\log(\mu_i(d)) = \eta_i(d) \Leftrightarrow \mu_i(d) = \exp(\eta_i(d)),\ i = 0, 1,$$

where $\eta_i(d)$ denotes the linear predictor. Again we use a Fourier series as linear predictor to capture the seasonality:

$$\eta_i(d) = a_{i0} + \sum_{l=1}^{k} [a_{il}\sin(ld') + b_{il}\cos(ld')],\ i = 0, 1, \tag{3.13}$$

where $d' = \frac{2\pi d}{366}$, with $d = 1, ..., 366$.

Note since we have 60 independent observations for each day of the year d (except for February, 29th, there are only 15 observations), i.e. $y_{t,prec}$ with $d(t) = d$ for $d = 1, ..., 366$, we perform a gamma regression with weights (introduced in Section 2.5.4) to model $Y_i^+(d)$ based on averaged observations for each day d of the year (depending on what happend one day before at $d-1$), i.e.

$$\widetilde{y}_0^+(d) := \frac{1}{\widetilde{n}_{01}(d)} \sum_{\{t:\ d(t)=d\ \cap\ y_{t-1,prec}=0\}} y_{t,prec}, \tag{3.14}$$

Harmonics (with estimated $\widehat{\phi}_k$)	Residual degrees of freedom	Residual deviance	p-value of res. deviance test	Partial degrees of freedom	Partial deviance	p-value of part. deviance test
0 ($\widehat{\phi}_0$=2.96)	365	1052.08	0.63	/	/	/
1 ($\widehat{\phi}_1$=1.43)	363	522.50	0.45	2	529.58	0.00
2 ($\widehat{\phi}_2$=1.43)	361	521.22	0.44	2	1.28	0.64

Table 3.16: Analysis of deviance in our gamma regression model for $Y_0^+(d)$ ("there was no rain yesterday on day $d-1$"). One harmonic is sufficient here, since the p-value of the partial deviance test is equal to 0.64, i.e. the model is not improved by an extending on 2 harmonics. The corresponding residual deviance test (p-value = 0.45) shows no lack of fit with 1 harmonic. Note: $\widehat{\phi}_k$ denotes the estimated dispersion parameter for the gamma regression with weights with k harmonics, needed to perform the residual and partial deviance tests here.

$$\widetilde{y}_1^+(d) := \frac{1}{\widetilde{n}_{11}(d)} \sum_{\{t:\ d(t)=d\ \cap\ y_{t-1,prec}>0\}} y_{t,prec}, \tag{3.15}$$

with weights $\widetilde{n}_{i1}(d)$, $i = 1, 2$. Here $\widetilde{n}_{01}(d)$ corresponds to the number of observations with positive rain amount on day d of the year and no rain amount on day $d-1$ of the year. Similarly, $\widetilde{n}_{11}(d)$ denotes the number of observations with positive rain amount on day d and positive rain amount on day $d-1$ of the year.

As described in Tables 3.16 and 3.17, one harmonic in the linear predictor (3.13) is sufficient in case of modeling $Y_0^+(d)$ and two harmonics are required in case of modeling $Y_1^+(d)$. The estimated coefficients of the corresponding models can then be found in Table 3.18. A look at the plots of the averaged observations (points in both plots of Figure

Harmonics (with estimated $\widehat{\phi}_k$)	Residual degrees of freedom	Residual deviance	p-value of res. deviance test	Partial degrees of freedom	Partial deviance	p-value of part. deviance test
0 ($\widehat{\phi}_0$=3.69)	365	1303.39	0.66	/	/	/
1 ($\widehat{\phi}_1$=1.50)	363	573.01	0.24	2	730.38	0.00
2 ($\widehat{\phi}_2$=1.48)	361	558.20	0.27	2	14.81	0.01
3 ($\widehat{\phi}_3$=1.49)	359	557.93	0.28	2	0.27	0.91

Table 3.17: Analysis of deviance in our gamma regression model for $Y_1^+(d)$ ("there was rain yesterday on day $d-1$"). Two harmonics are sufficient here, since the p-value of the partial deviance test is equal to 0.91, i.e. the model is not improved by an extending on 3 harmonics. The corresponding residual deviance test (p-value = 0.27) shows no lack of fit with 2 harmonics. Note: $\widehat{\phi}_k$ denotes the estimated dispersion parameter for the gamma regression with weights with k harmonics, needed to perform the residual and partial deviance tests here.

Coefficient	$Y_0^+(d)$ model ($i = 0$)		$Y_1^+(d)$ model ($i = 1$)	
	Estimate	p-value of Wald test	Estimate	p-value of Wald test
a_{i0}	1.61	0.00	1.80	0.00
a_{i1}	-0.17	0.00	-0.14	0.00
b_{i1}	-0.53	0.00	-0.39	0.00
a_{i2}	/	/	0.01	0.56
b_{i2}	/	/	0.06	0.00
$\widehat{\phi}$	1.43		1.48	

Table 3.18: Coefficient estimations in case of modeling $Y_0^+(d)$ and $Y_1^+(d)$ respectively with p-values of the corresponding asymptotic Wald tests. $\widehat{\phi}$ denotes the estimated dispersion parameter of the corresponding gamma regressions with weights.

3.22) together with the fitted values (solid lines in both plots of Figure 3.22) indicates that almost all averaged observations lie inside the 95%-confidence intervals of the fitted values. One can detect a slight but significant difference between the models of $Y_0^+(d)$ and $Y_1^+(d)$. However, both models show a higher rain amount in summer than in winter months. This phenomenon might occur due to the fact that the foothills of the Alps are a high precipitation region which reaches its distinctive maximum in summers (see Müller-Westermeier [2001]). It is because the portion of water vapor in the air is getting higher the higher the temperature is in that region. Therefore higher rain amounts in summer coming from convective precipitations are the consequence.

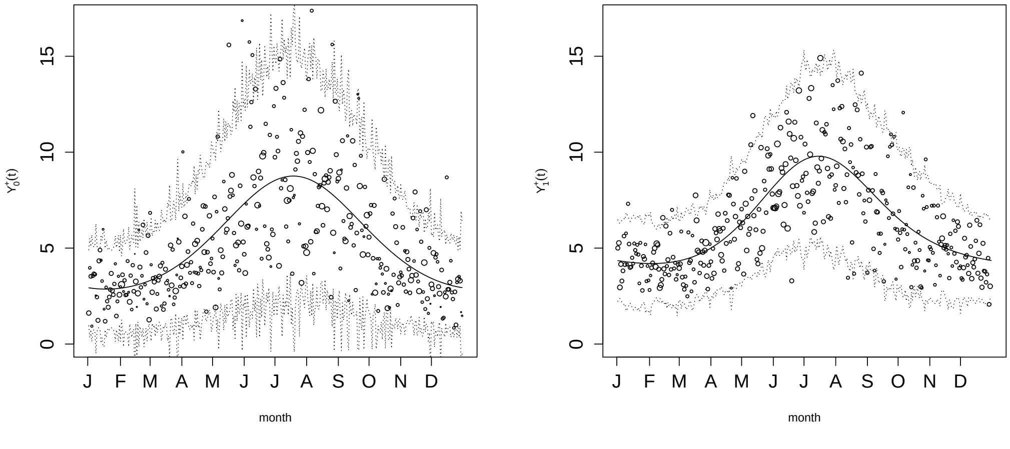


Figure 3.22: Plots of the averaged observations defined in Equation (3.14) and (3.15) of $Y_0^+(d)$ (points in left plot) and $Y_1^+(d)$ (points in right plot) together with the fitted values from the model (lines in both plots) with corresponding 95%-confidence intervals (dashed lines) for every day d of the year, $d = 1, ..., 366$.

Chapter 4

R-vine specification

Now we are ready to model the dependencies among our six meteorological variables by an R-vine distribution. After we have fitted the marginal distributional behaviors in the previous chapter we would like to specify the common distribution of the random vector

$$\boldsymbol{Y}_t := (Y_{t,meantemp}, Y_{t,mintemp}, Y_{t,maxtemp}, Y_{t,humidity}, Y_{t,press}, Y_{t,prec})' \in \mathbb{R}^6, \tag{4.1}$$

at time t for the three periods 1955-1960, 1980-1984 and 2005-2009 ($t \in \{1, ..., 1826\}$ for periods 1955-1960 and 2005-2009, and $t \in \{1, ..., 1827\}$ for the period 1980-1984 that contains two leap years). Note, because we are interested in the common dependency structure among the variables, it is sufficient to examine the distribution of the i.i.d. error terms $\epsilon_{t,\cdot}$ from our standardized (WLS) regression models instead of the quantities $Y_{t,meantemp}, Y_{t,mintemp}, Y_{t,maxtemp}$ and $Y_{t,press}$ respectively, as well as the distribution of the standardized daily humidity variable $\widetilde{Y}_{t,humidity}$ instead of $Y_{t,humidity}$. This is reasonable since the differences are only constrained to rescalings in the mean but the marginal distributional behaviors remain the same. Then we take the marginal distribution functions $\widehat{F}_{t,\cdot}$ (which we have estimated in the previous chapter) and calculate the corresponding margins $\widehat{u}_{t,\cdot} = \widehat{F}_{t,\cdot}(y_{t,\cdot})$ for every variable and every point in time t (here $y_{t,\cdot}$ denotes the observation of a variable at time t) to get (hopefully) uniform on $[0, 1]$ distributed margins to be able to apply Sklar's Theorem.

We have to consider that the distribution of daily total precipitation $Y_{t,prec}$ at time t has an additional point mass at zero as we have seen in Section 3.7. Thus, the corresponding probability function is not continuous at zero and therefore violates the assumption of Sklar's Theorem in the continuous case that we need here (cp. assumptions of Theorem 2.14). As a consequence, our six dimensional R-vine approach based on pairwise copula constructions is not straightforward anymore. Erhardt and Czado [2012] developed a model to handle such a scenario. Although they apply their approach to insurance data, i.e. yearly claim totals, we can easily transfer it to our application. We will specify this immediately in the next subsection, followed by a description how to construct this kind of model by appropriate R-vines and we will conclude this chapter by showing how to simulate from the model.

4.1 Model formulation

First of all we can rewrite $\boldsymbol{Y}_t$ at time t from (4.1) as

$$\boldsymbol{Y}_t = (\boldsymbol{Y}_{t,-prec}, Y_{t,prec})' \in \mathbb{R}^6, \tag{4.2}$$

where $\boldsymbol{Y}_{t,-prec} := (Y_{t,meantemp}, Y_{t,mintemp}, Y_{t,maxtemp}, Y_{t,humidity}, Y_{t,press}) \in \mathbb{R}^5$ is the random vector without the variable of daily total amount of precipitation $Y_{t,prec}$ at time t. Further, we can express $Y_{t,prec}$ as

$$Y_{t,prec} := [1 - J\,(d(t))] \times 0 + J\,(d(t))\, Y^{+}_{t,prec} \geq 0, \tag{4.3}$$

where $J\,(d(t))$ denotes the binary indicator random variable for the rain event on day $d(t)$ as defined in (3.7) in Section 3.7. As a reminder, $J\,(d(t)) = 1$ when day $d(t)$ has rain with "rain success" probability $p_i(d(t)) := p\,(J\,(d(t)) = 1 | J\,(d(t-1)) = i)$, $i = 0, 1$, i.e. depending on whether it was also rainy at day $d(t-1)$ or not. Otherwise $J\,(d(t)) = 0$ when day $d(t)$ is dry with probability $1 - p_i(d(t))$, $i = 0, 1$, also depending on what happened one day before on $d(t-1)$. Note that again $d(t) \in \{1, ..., 366\}$ stands for the calendar day at time t. The random variable $Y^{+}_{t,prec}$ represents the positive amount of rain when $J\,(d(t)) = 1$ at time t. As we have modeled in Section 3.7, its continuous gamma distribution also depends on $J\,(d(t-1))$ at time $t-1$. But the level of positive precipitation amount is independent of $J\,(d(t))$ at time t since $Y^{+}_{t,prec}$ is observable when $J\,(d(t)) = 1$. Therefore the common distribution of $\boldsymbol{Y}_{t,-prec}$, $Y_{t,prec}$ and $J\,(d(t))$ at time t can be calculated as

$$\begin{aligned}
& F_{\boldsymbol{Y}_{t,-prec}, Y_{t,prec}, J(d(t))}(\boldsymbol{y}_{t,-prec}, y_{t,prec}, j_{d(t)}) \\
&= P\left(\boldsymbol{Y}_{t,-prec} \leq \boldsymbol{y}_{t,-prec}, Y_{t,prec} \leq y_{t,prec}, J\,(d(t)) = j_{d(t)}\right) \\
&= p\left(J\,(d(t)) = j_{d(t)} | J\,(d(t-1))\right) P\left(\boldsymbol{Y}_{t,-prec} \leq \boldsymbol{y}_{t,-prec}, Y_{t,prec} \leq y_{t,prec} | J\,(d(t)) = j_{d(t)}\right) \\
&= p\left(J\,(d(t)) = j_{d(t)} | J\,(d(t-1))\right) \times \\
&\quad \begin{cases} P\left(\boldsymbol{Y}_{t,-prec} \leq \boldsymbol{y}_{t,-prec}, 0 \leq y_{t,prec} | J\,(d(t)) = 0\right), & \text{if } j_{d(t)} = 0, \\ P\left(\boldsymbol{Y}_{t,-prec} \leq \boldsymbol{y}_{t,-prec}, Y^{+}_{t,prec} \leq y_{t,prec} | J\,(d(t)) = 1\right), & \text{if } j_{d(t)} = 1. \end{cases} \\
&= p\left(J\,(d(t)) = j_{d(t)} | J\,(d(t-1))\right) \times \\
&\quad \begin{cases} P\left(\boldsymbol{Y}_{t,-prec} \leq \boldsymbol{y}_{t,-prec}\right), & \text{if } j_{d(t)} = 0, \\ P\left(\boldsymbol{Y}_{t,-prec} \leq \boldsymbol{y}_{t,-prec}, Y^{+}_{t,prec} \leq y_{t,prec}\right), & \text{if } j_{d(t)} = 1, \end{cases}
\end{aligned} \tag{4.4}$$

for $y_{t,prec} \geq 0$ and $\boldsymbol{y}_{t,-prec} := (y^{(1)}_{t,-prec}, ..., y^{(5)}_{t,-prec})' \in \mathbb{R}^5$.

The last line of Equation (4.4) suggests that our common distribution of $\boldsymbol{Y}_t$ (i.e. including $Y_{t,prec}$) equals either the common five dimensional distribution of $\boldsymbol{Y}_{t,-prec}$ or the common six dimensional distribution of $(\boldsymbol{Y}_{t,-prec}, Y^{+}_{t,prec}) \in \mathbb{R}^6$, depending on the value of $J\,(d(t))$, times the distribution of $J\,(d(t))$ at time t. Clearly, these common five and six dimensional distributions can be modeled by two R-vines.

We can then determine the common densitiy of $\boldsymbol{Y}_t$ (i.e. including $Y_{prec,t}$) and $J\left(d(t)\right)$ at time t (cp. Erhardt and Czado [2012]) as follows:

$$\begin{aligned} f_{\boldsymbol{Y_t},J(d(t))}(\boldsymbol{y}_t, j_{d(t)}) &= p\left(J\left(d(t)\right) = j_{d(t)} | J\left(d(t-1)\right)\right) \times \Big[\mathbf{1}_{\{J(d(t))=0\}} f_{\boldsymbol{Y}_{t,-prec}}(\boldsymbol{y}_{t,-prec}) \\ &\quad + \mathbf{1}_{\{J(d(t))=1\}} f_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}(\boldsymbol{y}_{t,-prec}, y_{t,prec})\Big] \\ &\overset{\text{Sklar}}{=} p\left(J\left(d(t)\right) = j_{d(t)} | J\left(d(t-1)\right)\right) \times \qquad (4.5) \\ &\quad \Big[\mathbf{1}_{\{J(d(t))=0\}} c_{\boldsymbol{Y}_{t,-prec}} \left(F_{Y_{t,meantemp}}(y^{(1)}_{t,-prec}), ..., F_{Y_{t,press}}(y^{(5)}_{t,-prec})\right) \times \\ &\quad f_{Y_{t,meantemp}}(y^{(1)}_{t,-prec}) \cdots f_{Y_{t,press}}(y^{(5)}_{t,-prec}) \\ &\quad + \mathbf{1}_{\{J(d(t))=1\}} \times \\ &\quad c_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})} \left(F_{Y_{t,meantemp}}(y^{(1)}_{t,-prec}), ..., F_{Y^+_{t,prec}}(y_{t,prec})\right) \times \\ &\quad f_{Y_{t,meantemp}}(y^{(1)}_{t,-prec}) \cdots f_{Y^+_{t,prec}}(y_{t,prec})\Big], \end{aligned}$$

where $\boldsymbol{y_t} := (\boldsymbol{y}_{t,-prec}, y_{t,prec})' = (y^{(1)}_{t,-prec}, ..., y^{(5)}_{t,-prec}, y_{t,prec})' \in \mathbb{R}^6$ with $y_{t,prec} \geq 0$. The functions $F_{\cdot}(y^{(j)}_{t,-prec})$, $F_{Y^+_{t,prec}}(y_{t,prec})$, $f_{\cdot}(y^{(j)}_{t,-prec})$ and $f_{Y^+_{t,prec}}(y_{t,prec})$ represent the marginal distribution and density functions respectively of the corresponding variables.

Due to the Theorem of Sklar (cp. Th. 2.14) the Equation (4.5) shows the common distribution of $\boldsymbol{Y_t}$ and $J\left(d(t)\right)$ in terms of marginal distributions and copulas $C_{\boldsymbol{Y}_{t,-prec}}$ and $C_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}$. Here for the copulas it holds

$$C_{\boldsymbol{Y}_{t,-prec}}(u_1, ..., u_5) = C_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}(u_1, ..., u_5, 1).$$

That is, while $C_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}$ describes the dependence between all meteorological variables at time t (when $J\left(d(t)\right) = 1$), $C_{\boldsymbol{Y}_{t,-prec}}$ models that of all variables except precipitation.

By looking at the log-likelihood of the model, it becomes clear that the modeling of $(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})$ and $J\left(d(t)\right)$ can proceed independently:

$$\log f_{\boldsymbol{Y_t},J(d(t))}(\boldsymbol{y}_t, j_{d(t)}) \overset{(4.5)}{=} \log p\left(J\left(d(t)\right) = j_{d(t)} | J\left(d(t-1)\right)\right) + \log[...].$$

Thus, our way of modeling, i.e. first modeling the behavior of $J\left(d(t)\right)$ in Section 3.7 and afterwards modeling the common dependence of $(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})$, is substantiated.

Thereby, according to the model built in Equations (4.4) and (4.5), our approach will be the following :

$$\begin{gathered} C_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})} \text{ will be modeled as vine copula with} \\ C_{\boldsymbol{Y}_{t,-prec}} \text{ as } \textit{5-dimensional subvine.} \end{gathered} \qquad (4.6)$$

The corresponding implementation will be now presented in the following section.

4.2 Model implementation

The idea behind our approach (4.6) is simple: We model the dependence structure of the first five variables (except precipitation) by a 5-dimensional R-vine copula specification $C_{\boldsymbol{Y}_{t,-prec}}$ (”vine without rain”) and connect the positive rain amount variable $Y^{+}_{t,prec}$ to the established ”vine without rain” to get a 6-dimensional R-vine copula specification $C_{(\boldsymbol{Y}_{t,-prec},Y^{+}_{t,prec})}$ (”vine with rain”) when it rains. It then explains the dependence among the variables $(\boldsymbol{Y}_{t,-prec}, Y^{+}_{t,prec})$ on rainy days. At the same time the dependence among our first five variables $\boldsymbol{Y}_{t,-prec}$ should not change and therefore our ”vine without rain” remains as 5-dimensional subvine in the ”vine with rain”.

The general detailed ”docking” procedure is described in Algorithm 3 and illustrated in Table 4.1. In the Algorithm, $(y_{t,1}, ..., y_{t,n})$ for $t = 1, ..., N$, represents N realizations of n i.i.d. random vectors. In our case it corresponds to realisations of the vector $\boldsymbol{Y}_{t,-prec}$ and

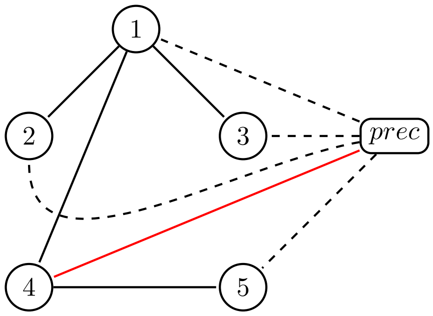
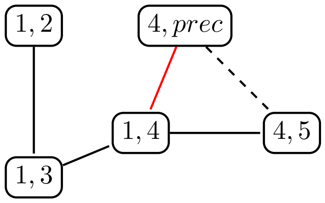
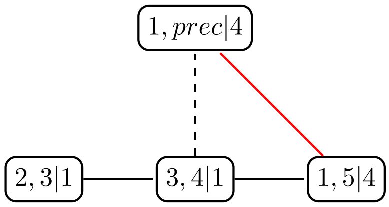
i	Graph	Description
1		Assume that we have the selected R-vine of 5 variables $N_1 = \{1, 2, 3, 4, 5\}$ (solid black lines) from the example in Table 2.5 in Section 2.4.4. The additional variable *prec* can now be connected with every other node (dashed lines). Let us assume the variable pair $(4, prec)$ has the maximum empirical Kendall's tau of these pairs including *prec* and thus it will be connected (red line). Note: The original R-vine tree remains unchanged.
2		All edges from the previous step are now nodes. Due to the proximity condition the node $(4, prec)$ has two connection possibilities (dashed lines). Assume that the conditional pair $(1, prec\|4)$ has the maximum empirical Kendall's tau of both possibilities, $(4, prec)$ will be then connected with $(1, 4)$ (red line). Note again: The original R-vine tree T_2 remains also unchanged here (solid black lines).
3		Again the conditional node $(1, prec\|4)$ has two possibilities of nodes to connect with due to the proximity condition (dashed lines). The connection to node $(1, 5\|4)$ will be selected based on maximum empirical Kendall's tau (red line). The initial R-vine tree T_3 remains as D-vine (solid black lines). Further steps will proceed in the same way.

Table 4.1: Exemplification of Algorithm 3 for the selection of the model with rain based on the example from Table 2.5 in Section 2.4.4.

Algorithm 3 Sequential method to select an R-vine model with rain based on Kendall's τ.

```
Input: Data $(y_{t,1},...,y_{t,n})$, $t = 1,...,N$, and $(y^{+}_{r,prec})$, $r = 1,...,M$ (realizations of i.i.d. random vectors). Here $M$ denotes the number of observations whith positive rain amount. Clearly, $M < N$.
Output: $n+1$-dim. R-vine copula specification, where the $n$-dimensional R-vine copula specification of $(y_{t,1},...,y_{t,n})$ is a subvine of it.
1: Determine the $n$-dimensional R-vine copula specification based on $(y_{t,1},...,y_{t,n})$ ("vine without rain") as described in Algorithm 1 in Section 2.4.4.
2: Take only the variables at time $t$ when it has rained ($M$ observations) and calculate the empirical Kendall's tau $\widehat{\tau}_{j,prec}$ for all possible variables pairs including the positive precipitation amount variable $\{j, prec\}, 1 \leq j \leq n$.
3: Select the pair $\{j, prec\}$ with the maximum empirical Kendall's tau, select a copula and estimate the corresponding parameter(s). The pair $\{j, prec\}$ becomes an additional edge in tree 1 and thus an additional node in tree 2 in the in Step 1 selected "vine without rain". Then transform $\widehat{F}_{prec|j}(y_{l,prec}|y_{l,j})$, $l = 1,...,M$, using the fitted copula $\widehat{C}_{j,prec}$ (see Algorithm 1).
4: for $i = 2,...,n-1$ do
5:     Calculate the empirical Kendall's tau $\widehat{\tau}_{j,prec|D}$ for all possible conditional variables pairs $\{j, prec|D\}$ that can be part of tree $T_i$, i.e. all edges $\{j, prec|D\}, 1 \leq j \leq n$ fulfilling the proximity condition (cp. Definition 2.31).
6:     Among these edges, select the conditional variable pair $\{j, prec|D\}$ with maximum empirical Kendall's tau, select a conditional copula and estimate the corresponding parameter(s). The conditional pair $\{j, prec|D\}$ becomes an additional edge in tree $i$ and thus an additional node in tree $i+1$ (for $i = n-1$ node in the additional tree $T_n$) in the in Step 1 selected "vine without rain". Then transform $\widehat{F}_{prec|j \cup D}(x_{l,prec}|x_{l,j}, \boldsymbol{x}_{l,D})$, $l = 1,...,M$, using the fitted copula $\widehat{C}_{j,prec|D}$ (see (2.69)).
7: end for
8: Take conditional variables pairs $\{j, prec|D\}$ and $\{j', k'|D'\}$ from the two edges of tree $T_{n-1}$ which become the nodes in the additional tree $T_n$. Then select a conditional copula and estimate the corresponding parameter(s) for the last conditional variables pair, i.e. w.l.o.g. $\{j', prec|k', D'\}$ to define the last edge in tree $T_n$.
```

thus $n = 5$ and $N = 1826$ or $N = 1827$, depending on the considered season. Further in the algorithm, M denotes the number of observations with positive rain amount which corresponds in our case to $M = 960$ in period 1980-1984 and $M = 928$ in case of periods 1955-1959 and 2005-2009.

A summary of our model specification is given in Figure 4.1. We then continue with the simulation from the model.

Model summary

1. Model dependence among $\boldsymbol{Y}_t = (\boldsymbol{Y}_{t,-prec}, Y_{t,prec})' =$ $= (Y_{t,meantemp}, Y_{t,mintemp}, Y_{t,maxtemp}, Y_{t,humidity}, Y_{t,press}, Y_{t,prec})' \in \mathbb{R}^6$ at time t, $t = 1, ..., 1826(1827)$ for the three periods 1955-1959 (1826 observations), 1980-1984 (1827 observations) and period 2005-2009 (1826 observations) respectively.

2. $Y^+_{t,prec}$: Precipitation amount when $J\left(d(t)\right) = 1$, where $d(t) \in \{1, ..., 366\}$ denotes the calendar year at time t.

3. Then:
$$Y_{t,prec} := [1 - J\left(d(t)\right)] \times 0 + J\left(d(t)\right) Y^+_{t,prec} \geq 0.$$

4. Joint distribution of $\boldsymbol{Y}_t$
$$\begin{aligned} f_{\boldsymbol{Y_t},J(d(t))}(\boldsymbol{y}_t, j_{d(t)}) \;&=\; p\left(J\left(d(t)\right) = j_{d(t)} | J\left(d(t-1)\right)\right) \\ &\quad \times \left[\mathbf{1}_{\{J(d(t))=0\}} f_{\boldsymbol{Y}_{t,-prec}}(\boldsymbol{y}_{t,-prec}) \right. \\ &\quad \left. + \, \mathbf{1}_{\{J(d(t))=1\}} f_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}(\boldsymbol{y}_{t,-prec}, y_{t,prec})\right] . \end{aligned}$$

5. Theorem by Sklar: $f_{\boldsymbol{Y}_{t,-prec}}$ and $f_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}$ in terms of marginal distributions and copulas $C_{\boldsymbol{Y}_{t,-prec}}$ and $C_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}$, where
$$C_{\boldsymbol{Y}_{t,-prec}}(u_1, ..., u_5) = C_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}(u_1, ..., u_5, 1), \text{ with}$$
   (i) $C_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}$: dependence between all meteorological variables at t.
   (ii) $C_{\boldsymbol{Y}_{t,-prec}}$ dependence between all variables except precipitation.

6. Here $C_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}$ as vine copula with $C_{\boldsymbol{Y}_{t,-prec}}$ as 5-dim. subvine.
   (i) Construct 5-dimensional R-vine $C_{\boldsymbol{Y}_{t,-prec}}$.
   (ii) Obtain $C_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}$ by connecting $Y^+_{t,prec}$ appropriately to each R-vine tree of $C_{\boldsymbol{Y}_{t,-prec}}$, i.e. for tree T_1 as example:

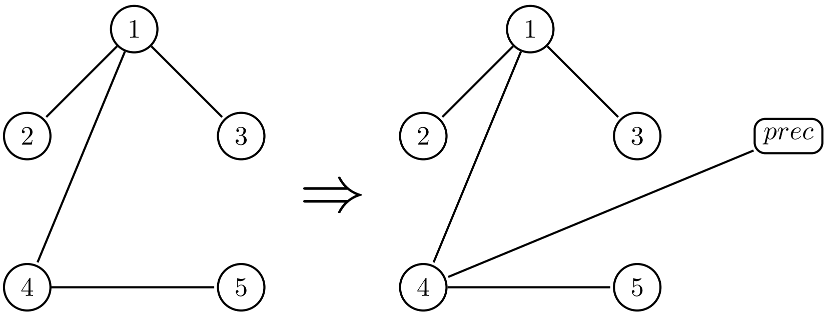


Figure 4.1: Model summary.

4.3 Simulation from the model

The simulation from our model is then straightforward. Therefore we proceed as follows:

(i) Draw sample J_t from $p\left(J\left(d(t)\right) = j_{d(t)}|J\left(d(t-1)\right)\right)$.

(ii) If $J_t = 0$:

a) Set $y_{t,prec} = 0$,

b) Draw sample $(u_{t,meantemp}, ..., u_{t,press})$ from $C_{\boldsymbol{Y}_{t,-prec}}$ (”vine without rain”) and transform back via the corresponding inverse distribution functions at time t to get simulations of $(Y_{t,meantemp}, ..., Y_{t,press})$.

(iii) If $J_t = 1$:

a) Draw sample $(u_{t,meantemp}, ..., u_{t,prec})$ from $C_{(\boldsymbol{Y}_{t,-prec}, Y^+_{t,prec})}$ (”vine with rain”) as constructed described in Algorithm 3 and transform back via the corresponding inverse distribution functions at time t to get simulations of $(Y_{t,meantemp}, ..., Y^+_{t,prec})$.

The results of our model implementation for the Hohenpeissenberg data as well as simulations from it are now presented in the next chapter.

Chapter 5

Results

In this chapter we will present the results of our dependence modeling among the Hohenpeissenberg variables using the model described in the previous chapter. As mentioned before, we evaluate the model for three periods (1955-1960, 1980-1984 and 2005-2009) and compare the results over time. Following our model building process we begin with the selection and estimation of appropriate R-vines for every period (”vines without rain”) and then construct the corresponding ”vines with rain”. Afterwards we will check the goodness of the model by comparing the log-likelihoods of the selected models applied to the data of the different periods. In addition we will also compare them to the log-likelihoods of Gauss-models, i.e. R-vine constructions using only Gaussian copulas, for each period. A further goodness of fit measure is to compare the empirical Kendall's taus of variables pairs with the simulated ones out of the model for every period. In the end of the chapter we would also like to consider the probabilites of several scenarios over the different periods, like warm and dry wheather in summer, extreme precipitation or high pressure, which can be estimated by our models.

5.1 Selected R-vines

We consider the 1826 (1827) Hohenpeissenberg observations of the random vector $\boldsymbol{Y}_t = (\boldsymbol{Y}_{t,-prec}, Y_{t,prec})' = (Y_{t,meantemp}, Y_{t,mintemp}, Y_{t,maxtemp}, Y_{t,humidity}, Y_{t,press}, Y_{t,prec})'$ and the corresponding fitted marginal distribution functions $\widehat{F}_{t,\cdot}$ to select appropriate R-vines. Note that in our case the observations of the error terms $\epsilon_{t,\cdot}$ are the residuals from the regression models in Chapter 3.

5.1.1 R-vine model estimation - without precipitation

We first compare the observations of the variables except precipitation in pairs plots to illustrate their pairwise relationship and dependence structure. Therefore we are using bivariate copula contour plots of standard normal quantiles $\widehat{z}_{t,i} = \Phi^{-1}(\widehat{u}_{t,i})$ calculated from our uniform margins $\widehat{u}_{t,i} = \widehat{F}_{t,i}(y^{(i)}_{t,-prec})$, where $i \in \{meantemp, mintemp, maxtemp, humidity, press\}$ and $y^{(i)}_{t,-prec}$ denotes the observations of the i-th component of $\boldsymbol{Y}_{t,-prec} = (Y_{t,meantemp}, Y_{t,mintemp}, Y_{t,maxtemp}, Y_{t,humidity}, Y_{t,press})'$ for $t = 1, ..., 1826(1827)$.

Variable	*meantemp*	*mintemp*	*maxtemp*	*humidity*	*pressure*
meantemp	1	0.47	0.66	-0.51	-0.09
mintemp	0.47	1	0.41	-0.25	-0.07
maxtemp	0.66	0.41	1	-0.50	-0.12
humidity	-0.51	-0.25	-0.50	1	0.07
pressure	-0.09	-0.07	-0.12	0.07	1

Table 5.1: Empirical Kendall's taus for pairs of margins $\widehat{u}_{t,i} = \widehat{F}_{t,i}(y_{t,-prec}^{(i)})$, where $i \in \{meantemp, mintemp, maxtemp, humidity, press\}$, for all t of period 1955-1959.

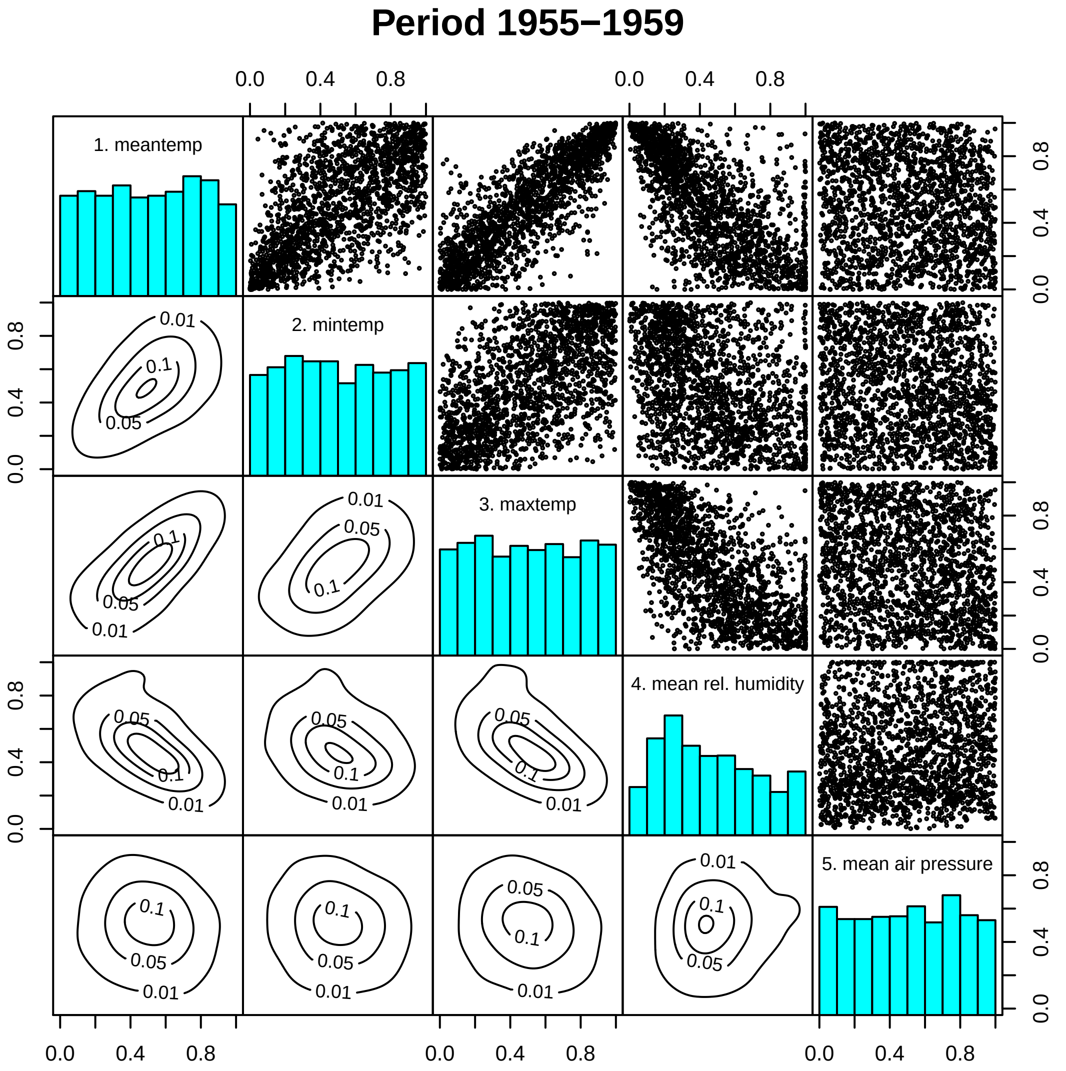


Figure 5.1: Pairs plot of the uniform margins $\widehat{u}_{t,i} = \widehat{F}_{t,i}(y_{t,-prec}^{(i)})$ (upper panels) and corresponding bivariate copula contour plots of standard normal quantiles $\widehat{z}_{t,i} = \Phi^{-1}(\widehat{u}_{t,i})$ (lower panels) for $i \in \{meantemp, mintemp, maxtemp, humidity, press\}$ except precipitation and for all t of period 1955-1959. The middle panel shows the histograms of the margins $\widehat{u}_{t,i}$ (blue bars). Note that the scaling of axes for the contour plots naturally ranges between $[-3, 3]$ instead of $[0, 1]$.

Variable	*meantemp*	*mintemp*	*maxtemp*	*humidity*	*pressure*
meantemp	1	0.50	0.65	-0.43	-0.11
mintemp	0.50	1	0.42	-0.26	-0.08
maxtemp	0.65	0.42	1	-0.42	-0.16
humidity	-0.43	-0.26	-0.42	1	0.02
pressure	-0.11	-0.08	-0.16	0.02	1

Table 5.2: Empirical Kendall's taus for pairs of margins $\widehat{u}_{t,i} = \widehat{F}_{t,i}(y_{t,-prec}^{(i)})$, where $i \in \{meantemp, mintemp, maxtemp, humidity, press\}$, for all t of period 1980-1984.

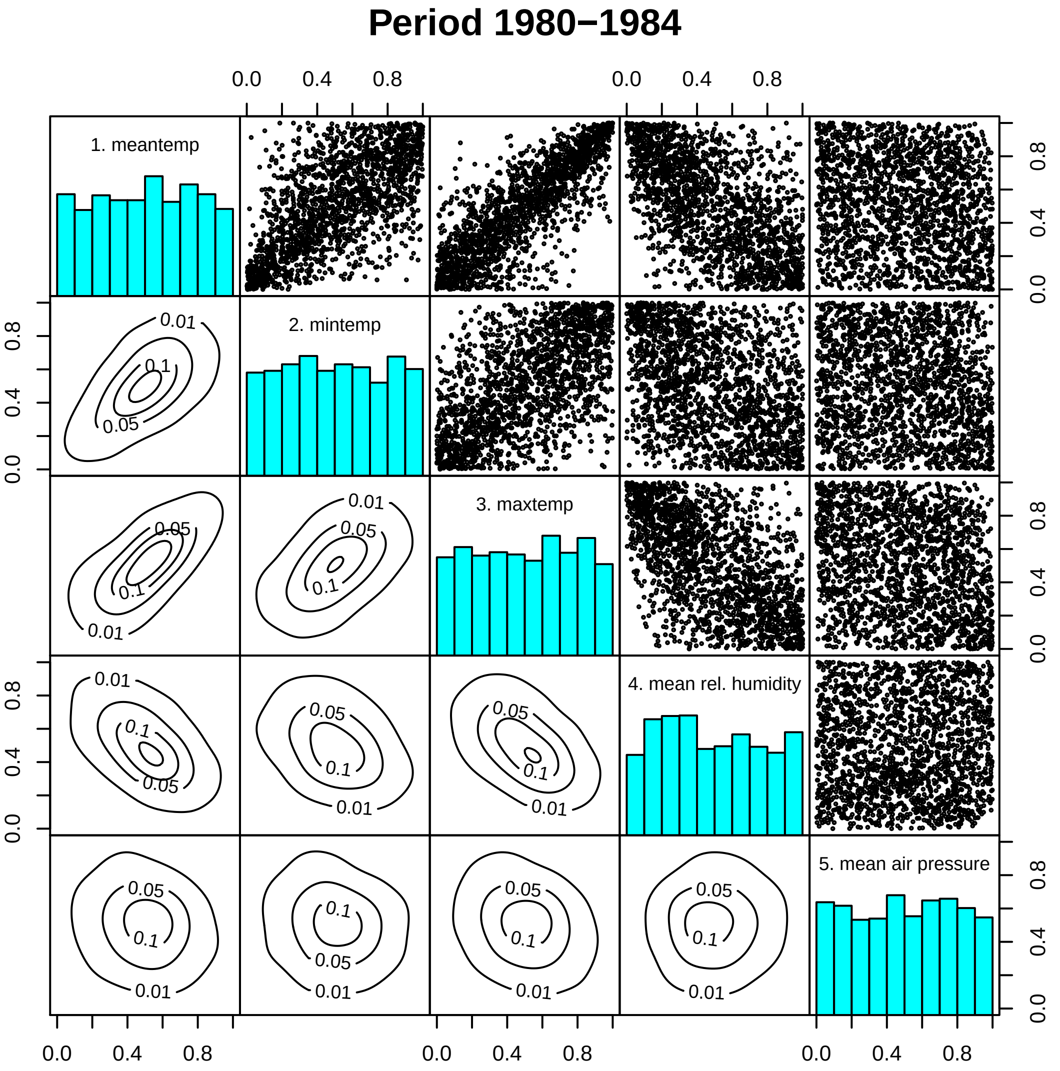


Figure 5.2: Pairs plot of the uniform margins $\widehat{u}_{t,i} = \widehat{F}_{t,i}(y_{t,-prec}^{(i)})$ (upper panels) and corresponding bivariate copula contour plots of standard normal quantiles $\widehat{z}_{t,i} = \Phi^{-1}(\widehat{u}_{t,i})$ (lower panels) for $i \in \{meantemp, mintemp, maxtemp, humidity, press\}$ except precipitation and for all t of period 1980-1984. The middle panel shows the histograms of the margins $\widehat{u}_{t,i}$ (blue bars). Note that the scaling of axes for the contour plots naturally ranges between $[-3, 3]$ instead of $[0, 1]$.

Variable	*meantemp*	*mintemp*	*maxtemp*	*humidity*	*pressure*
meantemp	1	0.56	0.72	-0.42	-0.11
mintemp	0.56	1	0.41	-0.27	-0.01
maxtemp	0.72	0.41	1	-0.42	-0.13
humidity	-0.42	-0.27	-0.42	1	0.02
pressure	-0.11	-0.01	-0.13	0.02	1

Table 5.3: Empirical Kendall's taus for pairs of margins $\widehat{u}_{t,i} = \widehat{F}_{t,i}(y_{t,-prec}^{(i)})$, where $i \in \{meantemp, mintemp, maxtemp, humidity, press\}$, for all t of period 2005-2009.

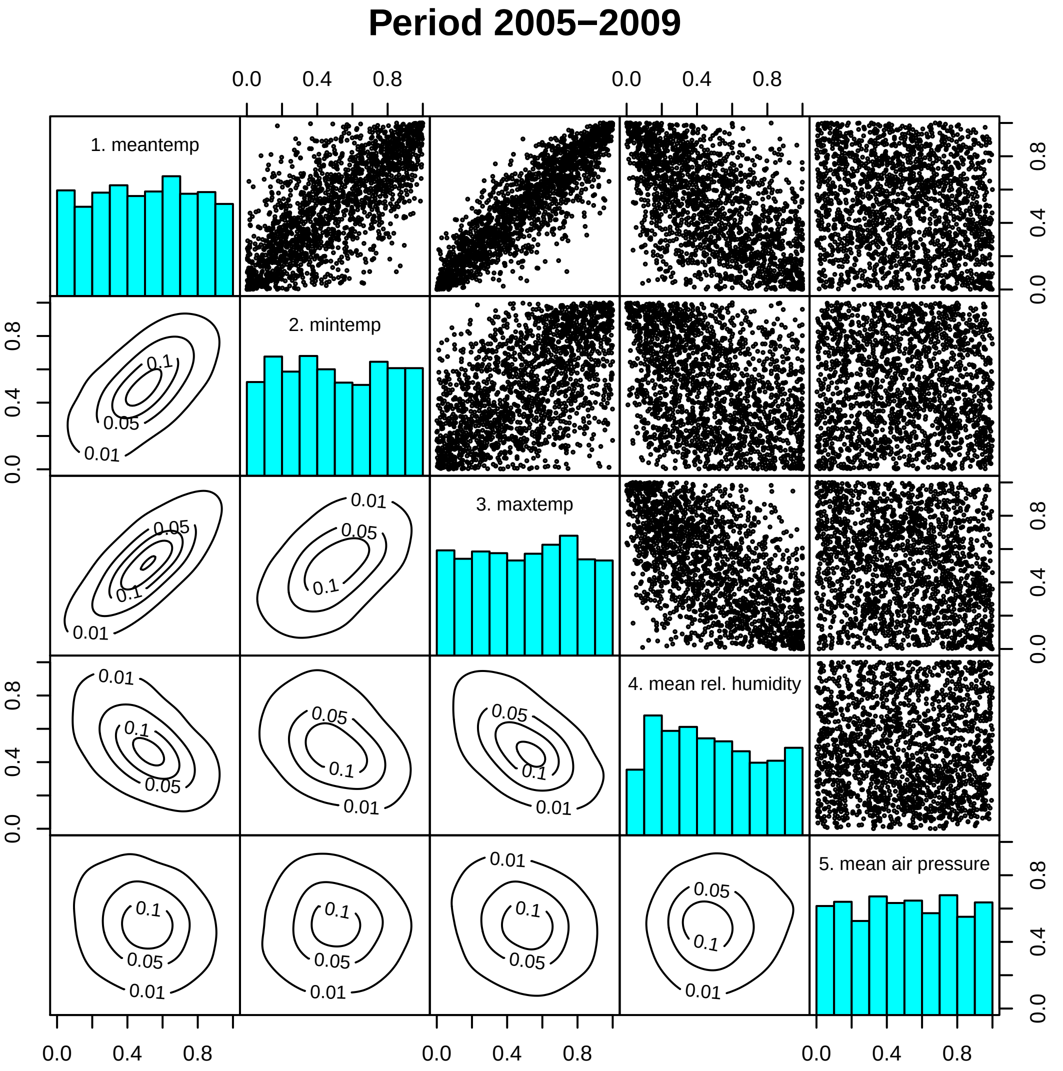


Figure 5.3: Pairs plot of the uniform margins $\widehat{u}_{t,i} = \widehat{F}_{t,i}(y_{t,-prec}^{(i)})$ (upper panels) and corresponding bivariate copula contour plots of standard normal quantiles $\widehat{z}_{t,i} = \Phi^{-1}(\widehat{u}_{t,i})$ (lower panels) for $i \in \{meantemp, mintemp, maxtemp, humidity, press\}$ except precipitation and for all t of period 2005-2009. The middle panel shows the histograms of the margins $\widehat{u}_{t,i}$ (blue bars). Note that the scaling of axes for the contour plots naturally ranges between $[-3, 3]$ instead of $[0, 1]$.

The plots of the margins and the pairwise contour plots of the corresponding standard normal quantiles are given in Figures 5.1 - 5.3 for the three periods. Additionally, the pairwise empirical Kendall's taus of the margins can be found in the above Tables 5.1 - 5.3 where we detect naturally a strong positive pairwise dependence among the temperature margins. In contrast to it, the margins of daily mean humidity indicate a negative dependence on the temperature variables while the air pressure margins nearly seem to be independent of humidity and minimum temperature margins. The look at the contour plots illustrate unusual small peaks like "bubbles" in the pairwise dependence structure among the humidity margins and the other ones in the period 1955-1959. However, this conspicuity does not occur in the other both periods. One can show that these peaks will also arise in the other periods when one removes the covariates of wind direction (explain the "Fön-effect") in the marginal modeling of daily mean relative humidity. Thus, since there was no daily wind direction data available for period 1955-1959, it shows that the more information is included the better our model. The blue bars in the middle panels of the pairs plots represent the empirical distribution of our margins.[1] Due to Sklar's theorem, they all should be uniform on $[0, 1]$ to match our model assumption. All variables more or less seem to fulfill this property, only the humdity margins offer a slightly non-uniform spread on $[0, 1]$ which becomes a bit better over the periods as a result of adding the wind direction covariable in our marginal regression modeling of humidity in Section 3.5.

All in all a multivariate Gaussian distribution of our variables does not seem to be appropriate for all periods since the contour plots indicate pairwise tail dependencies, in parts asymmetric tail dependencies especially among the temperature variables and among humidity and the other variables.

The selected R-vines (due to Algorithm 1 in Section 2.4.4) and the corresponding theoretical Kendall's taus based on the estimated copula parameters are presented in Figures 5.4 - 5.6. The first vine trees T_1 of each period can be deduced from our pairs plots and show the strongest dependencies among its nodes over all trees. The edges in T_1 remain the same over all periods but the estimated pair copulas change. The only difference in the first tree between period 1955-1959 and 1980-1984 results in a change of pair copula of mean temperature and humidity convenient to the change in the marginal humidity model between both periods. The Frank copula then also remains in the last period describing the bivariate behavior of both variables. Further we observe that the asymmetric pair copulas among the temperature variables (i.e. Gumbel and survival Gumbel) in the first both periods change to symmectric pair copulas (Frank and t) in the last period 2005-2009 in tree T_1. The other trees $T_i, i = 2, 3, 4$ still indicate only small dependencies among their conditional pair variables (except $(min\_t, max\_t|mean\_t)$ in tree T_2 of period 2005-2009 with $\tau = -0.26$ of a t copula). There is also a change in the tree structure of the second trees T_2 from the first period to the last ones. But from the third trees on, T_3, the vines correspond as D-vines in all periods. The R-vines with corresponding parameters are described by regular vine matrices (RVMs) and corresponding parameter matrices which are given afterwards.

[1] We are here using the R function `panel.hist` of the package `BioStatR` (implemented by Bertrand and Maumy-Bertrand [2012]) to plot the histograms in the middle panels.

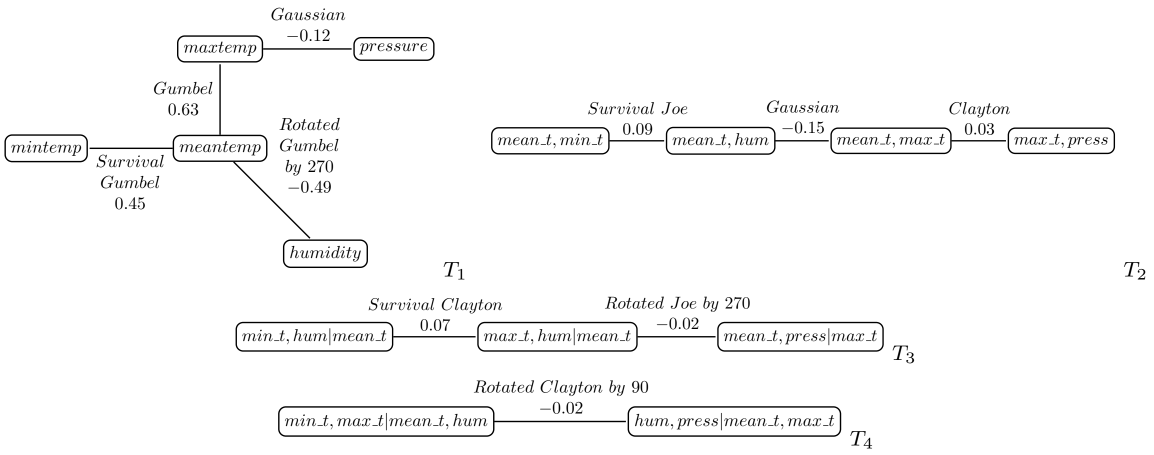


Figure 5.4: Selected R-vine with theoretical Kendall's τ - Period 1955-1959

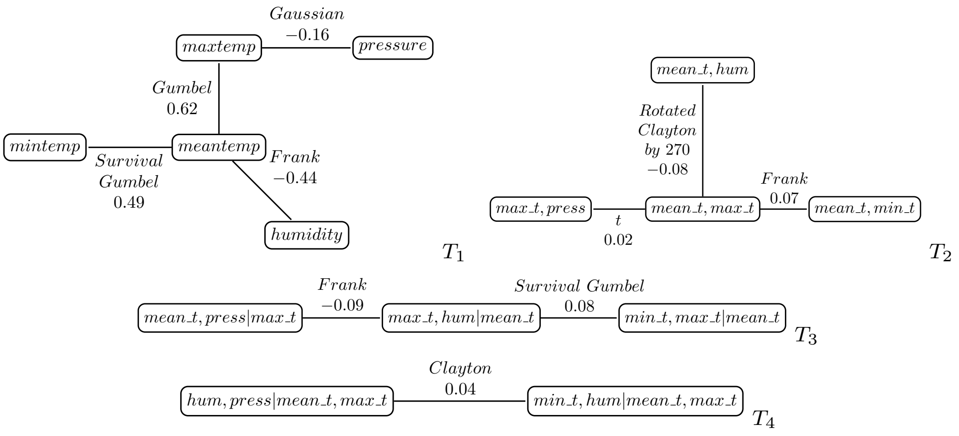


Figure 5.5: Selected R-vine with theoretical Kendall's τ - Period 1980-1984

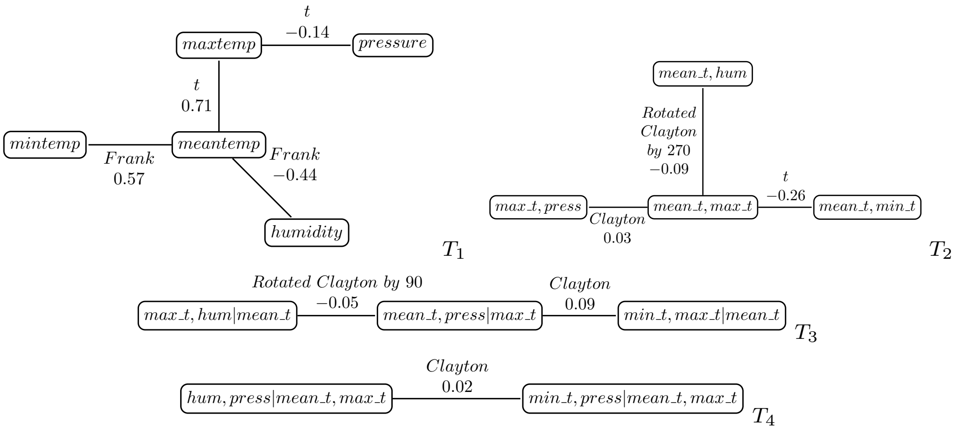


Figure 5.6: Selected R-vine with theoretical Kendall's τ - Period 2005-2009

The R-vines are described by the following regular vine matrices (RVMs) (cp. Section 2.4.3). For the period 1955-1959, we have

$$M_{period\ 1955-1959} = \begin{pmatrix} 2 & & & & \\ 5 & 4 & & & \\ 3 & 5 & 1 & & \\ 4 & 3 & 5 & 5 & \\ 1 & 1 & 3 & 3 & 3 \end{pmatrix}, \; T_{period\ 1955-1959} = \begin{pmatrix} & & & & \\ C90 & & & & \\ SC & J270 & & & \\ SJ & N & C & & \\ SG & G270 & G & N & \end{pmatrix},$$

$$P_{1,period\ 1955-1959} = \begin{pmatrix} & & & & \\ -0.04 & & & & \\ 0.15 & -1.04 & & & \\ 1.17 & -0.23 & 0.05 & & \\ 1.82 & -1.96 & 2.72 & -0.18 & \end{pmatrix}.$$

Period 1980-1984:

$$M_{period\ 1980-1984} = \begin{pmatrix} 2 & & & & \\ 5 & 4 & & & \\ 4 & 5 & 1 & & \\ 3 & 3 & 5 & 5 & \\ 1 & 1 & 3 & 3 & 3 \end{pmatrix}, \; T_{period\ 1980-1984} = \begin{pmatrix} & & & & \\ C & & & & \\ SG & F & & & \\ F & C270 & t & & \\ SG & F & G & N & \end{pmatrix},$$

$$P_{1,period\ 1980-1984} = \begin{pmatrix} & & & & \\ 0.08 & & & & \\ 1.09 & -0.81 & & & \\ 0.68 & -0.17 & 0.03 & & \\ 1.96 & -4.72 & 2.66 & -0.25 & \end{pmatrix}, \; P_{2,period\ 1980-1984} = \begin{pmatrix} & & & & \\ 0 & & & & \\ 0 & 0 & & & \\ 0 & 0 & 17.88 & & \\ 0 & 0 & 0 & 0 & \end{pmatrix}.$$

Period 2005-2009:

$$M_{period\ 2005-2009} = \begin{pmatrix} 2 & & & & \\ 4 & 4 & & & \\ 5 & 5 & 1 & & \\ 3 & 3 & 5 & 5 & \\ 1 & 1 & 3 & 3 & 3 \end{pmatrix}, \; T_{period\ 2005-2009} = \begin{pmatrix} & & & & \\ C & & & & \\ C & C90 & & & \\ t & C270 & C & & \\ F & F & t & t & \end{pmatrix},$$

$$P_{1,period\ 2005-2009} = \begin{pmatrix} & & & & \\ 0.04 & & & & \\ 0.19 & -0.10 & & & \\ -0.40 & -0.21 & 0.06 & & \\ 7.11 & -4.67 & 0.90 & -0.21 & \end{pmatrix},$$

$$P_{2,period\ 2005-2009} = \begin{pmatrix} & & & & \\ 0 & & & & \\ 0 & 0 & & & \\ 11.56 & 0 & 0 & & \\ 0 & 0 & 12.29 & 16.40 & \end{pmatrix},$$

where $1 = meantemp$, $2 = mintemp$, $3 = maxtemp$, $4 = humidity$ and $5 = pressure$.

5.1.2 R-vine model estimation - including positive precipitation amount

We now only consider the margins $\widehat{u}_{t,i} = \widehat{F}_{t,i}(y^{(i)}_{t,-prec})$ and $\widehat{u}_{t,prec} = \widehat{F^+}_{t,prec}(y_{t,prec})$ at time t when it rains, i.e. at time t when the observations of total precipitation amount are $y_{t,prec} > 0$. As mentioned before the number of rain days is equal to 960 in the period 1980-1984 and we compare 928 rain days observations in the periods 1955-1959 and 2005-2009. Again $\widehat{F}_{t,i}$ represents the fitted marginal distribution functions and $\widehat{F^+}_{t,prec}$ corresponds to the fitted distribution function of the positive rain amount, i.e. it is gamma distributed according to our modeling in Section 3.7. $y^{(i)}_{t,-prec}$ denotes the observations of $\boldsymbol{Y}_{t,-prec} = (Y_{t,meantemp}, Y_{t,mintemp}, Y_{t,maxtemp}, Y_{t,humidity}, Y_{t,press})'$ for $t = 1, ..., 1826(1827)$ and $i \in \{meantemp, mintemp, maxtemp, humidity, press\}$.

We start comparing the observations of the variables when it rains again in pairwise scatter and contour plots of the margins and the correspondig standard normal quantiles respectively to illustrate their pairwise relationship and dependence structure on rain days (Figures 5.7 - 5.9). Note that the scaling of axes for the contour plots then naturally ranges between $[-3, 3]$ instead of $[0, 1]$. Afterwards we will select the appropriate R-vines including positive rain amount variable (Figures 5.10 - 5.12) specified in Algorithm 3 in Section 4.2. The R-vines are described by corresponding regular vine matrices (RVMs) and parameter matrices. Again we detect small peaks like "bubbles" in the pairwise dependence structure among the humidity margins and the other ones in the period 1955-1959 by looking at the contour plots. As expected such "bubbles" do not occur in the other both periods anymore but one can show that they will also occur in the other periods when one would remove the "Fön"-effect explaining wind direction variable out of the marginal humidity regression model.

The empirical Kendall's taus of pairs of margins on rain days are given in Tables 5.4 - 5.6. They indicate the strongest pairwise dependence including the positive precipitation amount variable by the pair humidity and precipitation in all regarded periods ($0.18 \leq \widehat{\tau}_{humidity,prec} \leq 0.23$) which makes sense naturally. Thus, the positive rain amount variable will be connected with humidity in the first trees T_1 of our R-vines in all three periods (due to Algorithm 3). However, the corresponding estimated pair copula of both variables varies from survival Gumbel over Gaussian to a Frank copula over time, thus becomes more tail symmetric.

A further look at the Tables 5.4 - 5.6 indicates that other pairwise dependencies among the variables do not change noticeable when we only consider margins on rain days. This fact is in line with our model assumption that the common dependence structure of the variables without precipitation does not change when it rains. As we have just modeled before, the R-vines of the variables without precipitation are now 5-dimensional subvines in our selected vines including precipitation (compare Figures 5.10 - 5.12). Consequently, the RVMs and parameter matrices from the previous Section 5.1.1 are now 5×5 submatrices in the RVMs and parameter matrices describing the R-vines including postive precipitation amount here in this section.

Variable	*meantemp*	*mintemp*	*maxtemp*	*humidity*	*pressure*	*prec* > 0
meantemp	1	0.50	0.57	-0.39	-0.10	-0.15
mintemp	0.50	1	0.40	-0.19	-0.09	-0.06
maxtemp	0.57	0.40	1	-0.42	-0.17	-0.08
humidity	-0.39	-0.19	-0.42	1	0.13	0.23
pressure	-0.10	-0.09	-0.17	0.13	1	-0.15
precipitation > 0	-0.15	-0.06	-0.08	0.23	-0.15	1

Table 5.4: Empirical Kendall's taus for pairs of margins when it rains - period 1955-1959.

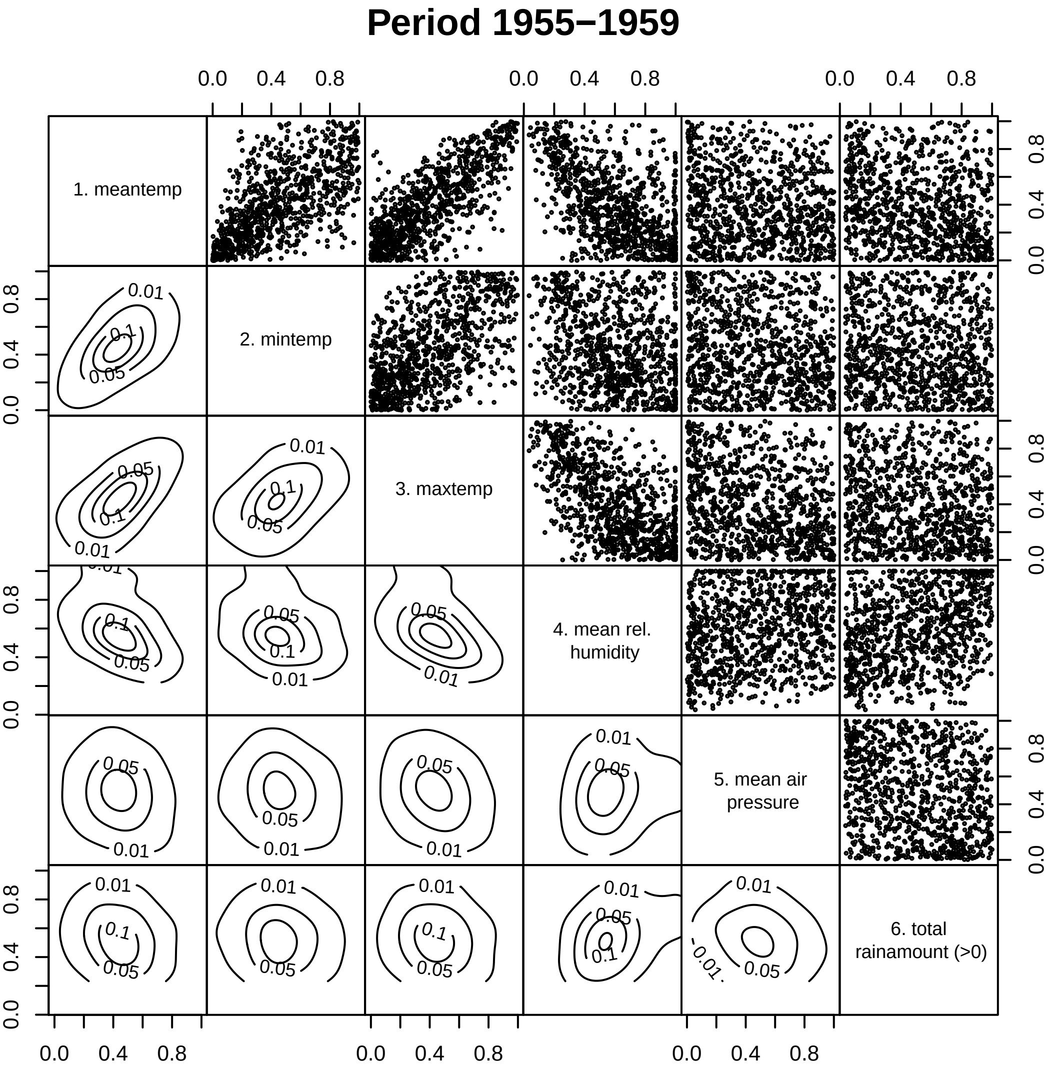


Figure 5.7: Pairwise scatter and contour plots (note: axes between [-3,3]) of the margins and corresponding standard normal quantiles resp. when it rains in period 1955-1959.

Variable	*meantemp*	*mintemp*	*maxtemp*	*humidity*	*pressure*	*prec* > 0
meantemp	1	0.53	0.56	-0.39	-0.17	-0.11
mintemp	0.53	1	0.40	-0.23	-0.12	-0.07
maxtemp	0.56	0.40	1	-0.38	-0.22	-0.05
humidity	-0.39	-0.23	-0.38	1	0.09	0.18
pressure	-0.17	-0.12	-0.22	0.09	1	-0.13
precipitation > 0	-0.11	-0.07	-0.05	0.18	-0.13	1

Table 5.5: Empirical Kendall's taus for pairs of margins when it rains - period 1980-1984.

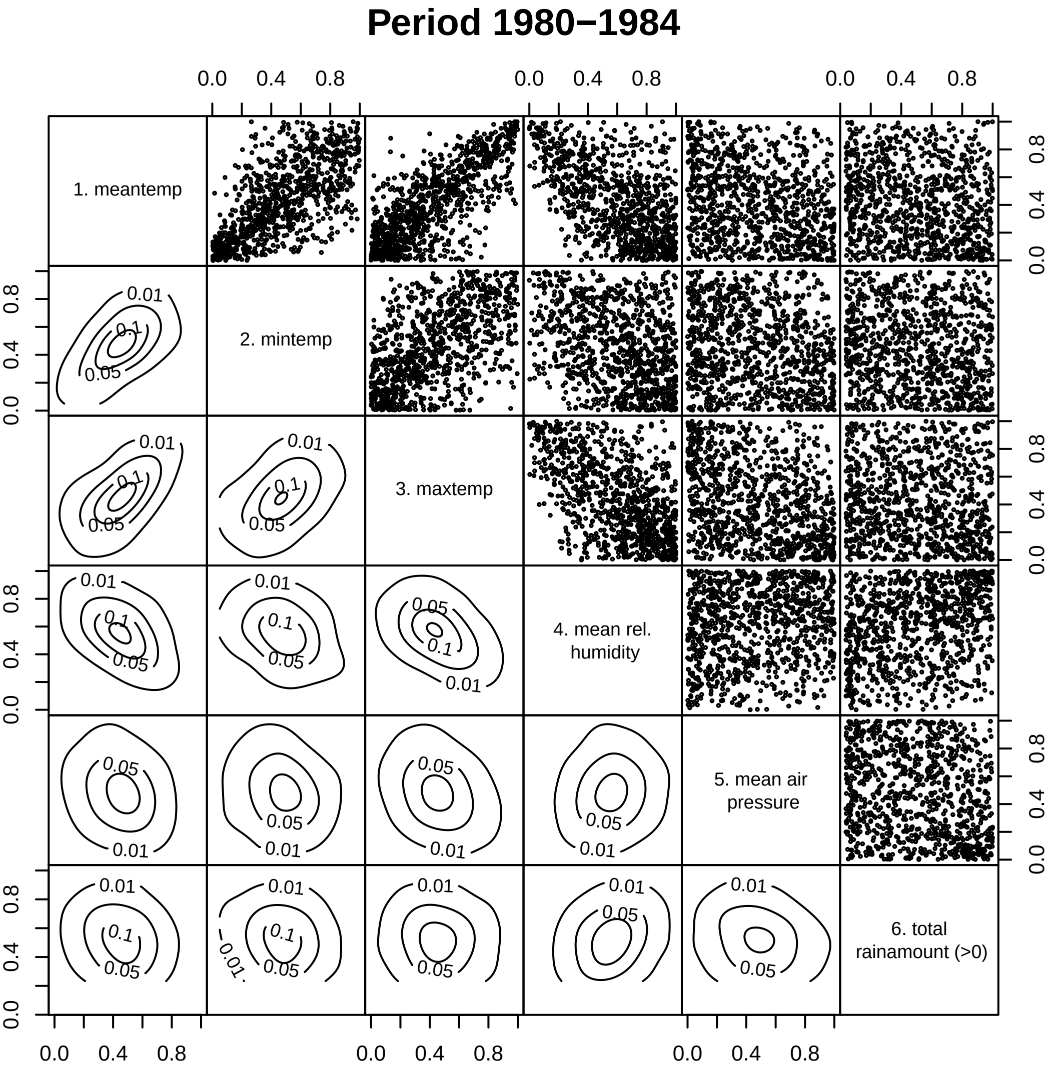


Figure 5.8: Pairwise scatter and contour plots (note: axes between [-3,3]) of the margins and corresponding standard normal quantiles resp. when it rains in period 1980-1984.

Variable	*meantemp*	*mintemp*	*maxtemp*	*humidity*	*pressure*	*prec* > 0
meantemp	1	0.53	0.70	-0.39	-0.15	-0.09
mintemp	0.53	1	0.36	-0.19	-0.01	-0.10
maxtemp	0.70	0.36	1	-0.41	-0.20	-0.08
humidity	-0.39	-0.19	-0.41	1	0.11	0.19
pressure	-0.15	-0.01	-0.20	0.11	1	-0.14
precipitation > 0	-0.09	-0.10	-0.08	0.19	-0.14	1

Table 5.6: Empirical Kendall's taus for pairs of margins when it rains - period 2005-2009.

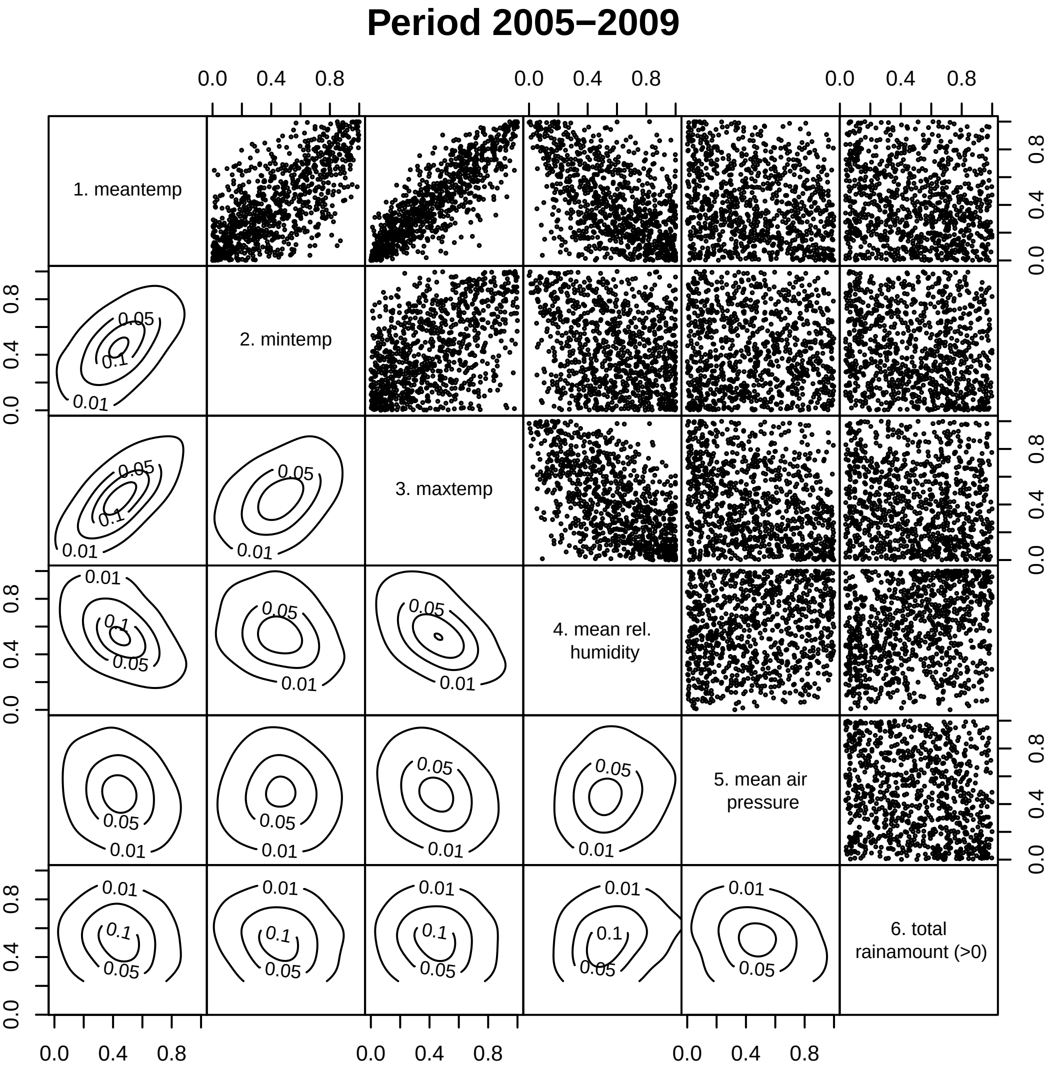


Figure 5.9: Pairwise scatter and contour plots (note: axes between [-3,3]) of the margins and corresponding standard normal quantiles resp. when it rains in period 2005-2009.

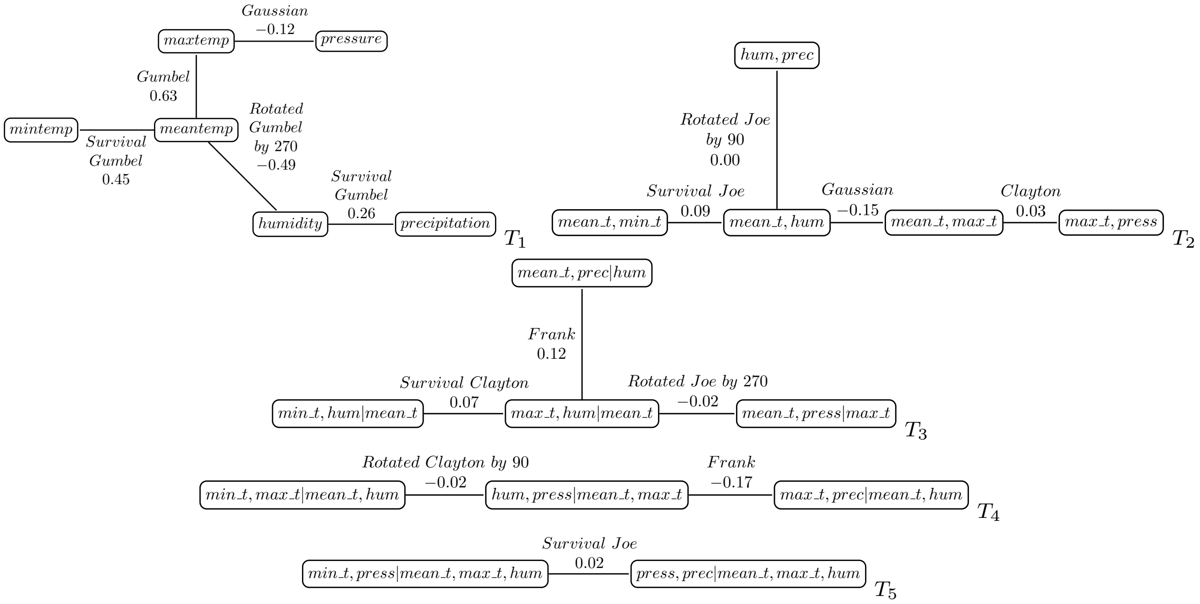


Figure 5.10: Selected R-vine including positive precipitation amount variable with theoretical Kendall's τ - Period 1955-1959

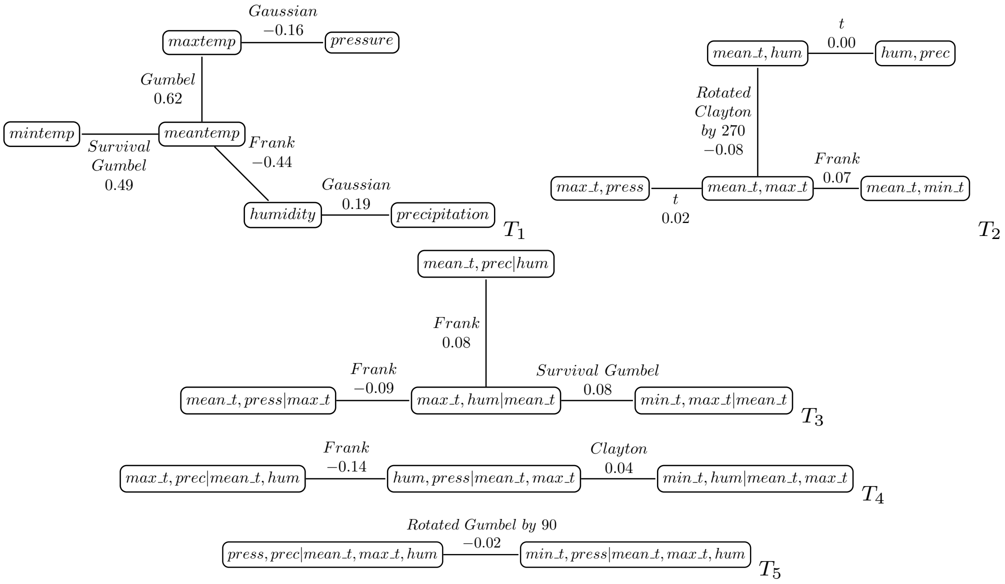


Figure 5.11: Selected R-vine including positive precipitation amount variable with theoretical Kendall's τ - Period 1980-1984

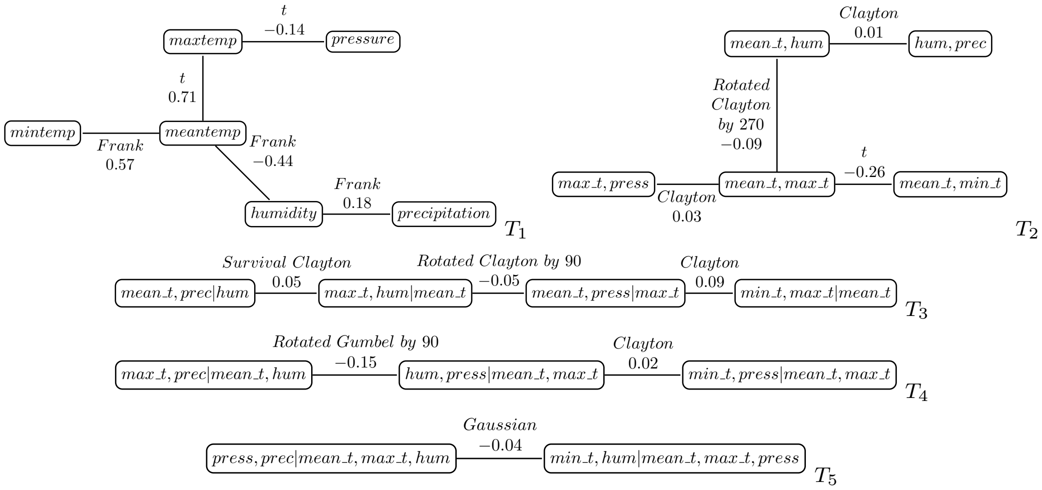


Figure 5.12: Selected R-vine including positive precipitation amount variable with theoretical Kendall's τ - Period 2005-2009

The R-vines including positive precipitation amount are described by the following regular vine matrices (RVMs) (cp. Section 2.4.3). For the period 1955-1959, we have

$$M_{period\ 1955-1959} = \begin{pmatrix} 6 & & & & & \\ 2 & 2 & & & & \\ 5 & 5 & 4 & & & \\ 3 & 3 & 5 & 1 & & \\ 1 & 4 & 3 & 5 & 5 & \\ 4 & 1 & 1 & 3 & 3 & 3 \end{pmatrix}, \; T_{period\ 1955-1959} = \begin{pmatrix} & & & & \\ SJ & & & & \\ F & C90 & & & \\ F & SC & J270 & & \\ J90 & SJ & N & C & \\ SG & SG & G270 & G & N \end{pmatrix},$$

$$P_{1,period\ 1955-1959} = \begin{pmatrix} & & & & \\ 1.03 & & & & \\ -1.56 & -0.04 & & & \\ 1.08 & 0.15 & -1.04 & & \\ -1.00 & 1.17 & -0.23 & 0.05 & \\ 1.34 & 1.82 & -1.96 & 2.72 & -0.18 \end{pmatrix}.$$

Period 1980-1984:

$$M_{period\ 1980-1984} = \begin{pmatrix} 6 & & & & & \\ 2 & 2 & & & & \\ 5 & 5 & 4 & & & \\ 3 & 4 & 5 & 1 & & \\ 1 & 3 & 3 & 5 & 5 & \\ 4 & 1 & 1 & 3 & 3 & 3 \end{pmatrix}, \; T_{period\ 1980-1984} = \begin{pmatrix} & & & & \\ G90 & & & & \\ F & C & & & \\ F & SG & F & & \\ t & F & C270 & t & \\ N & SG & F & G & N \end{pmatrix},$$

$$P_{1,period\ 1980-1984} = \begin{pmatrix} -1.02 & & & & \\ -1.25 & 0.08 & & & \\ 0.76 & 1.09 & -0.81 & & \\ -0.01 & 0.68 & -0.17 & 0.03 & \\ 0.29 & 1.96 & -4.72 & 2.66 & -0.25 \end{pmatrix},$$

$$P_{2,period\ 1980-1984} = \begin{pmatrix} 0 & & & & \\ 0 & 0 & & & \\ 0 & 0 & 0 & & \\ 21.48 & 0 & 0 & 17.88 & \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Period 2005-2009:

$$M_{period\ 2005-2009} = \begin{pmatrix} 6 & & & & & \\ 2 & 2 & & & & \\ 5 & 4 & 4 & & & \\ 3 & 5 & 5 & 1 & & \\ 1 & 3 & 3 & 5 & 5 & \\ 4 & 1 & 1 & 3 & 3 & 3 \end{pmatrix}, \ T_{period\ 2005-2009} = \begin{pmatrix} N & & & & \\ G90 & C & & & \\ SC & C & C90 & & \\ C & t & C270 & C & \\ F & F & F & t & t \end{pmatrix},$$

$$P_{1,period\ 2005-2009} = \begin{pmatrix} -0.08 & & & & \\ -1.18 & 0.04 & & & \\ 0.11 & 0.19 & -0.10 & & \\ 0.01 & -0.40 & -0.21 & 0.06 & \\ 1.67 & 7.11 & -4.67 & 0.90 & -0.21 \end{pmatrix},$$

$$P_{2,period\ 2005-2009} = \begin{pmatrix} 0 & & & & \\ 0 & 0 & & & \\ 0 & 0 & 0 & & \\ 0 & 11.56 & 0 & 0 & \\ 0 & 0 & 0 & 12.29 & 16.40 \end{pmatrix},$$

where $1 = meantemp$, $2 = mintemp$, $3 = maxtemp$, $4 = humidity$, $5 = pressure$ and $6 = precipitation > 0$.

5.2 Log-likelihood comparison

We would like to substantiate our model by comparing the log-likelihoods of the selected models applied to the data from the three periods. Additionally it would be interesting to compare these results with the log-likelihoods of an R-vine model using only Gaussian pair copulas. The outcome is presented in Figure 5.13 and the corresponding log-likelihood

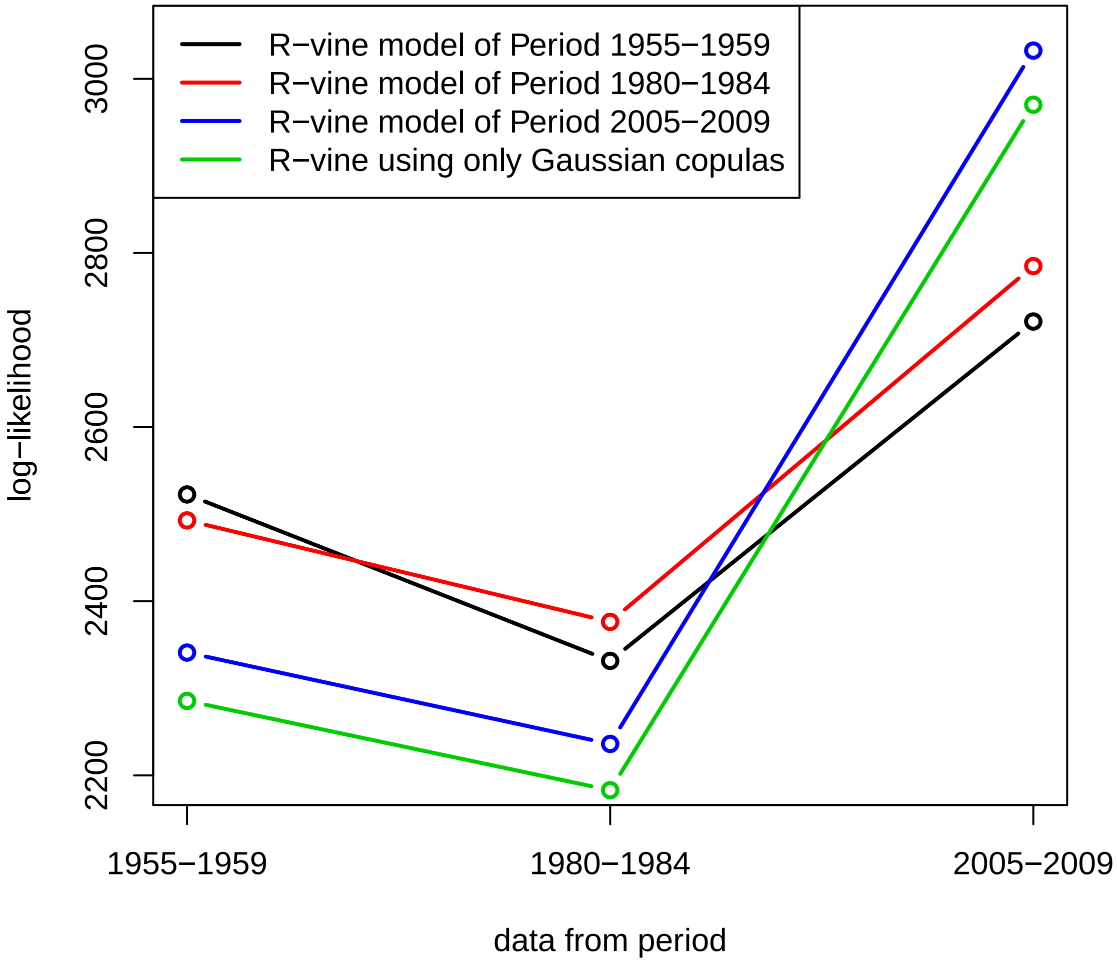


Figure 5.13: Log-likelihood comparison between the models of each period and a model using only Gaussian copulas applied to the data from the different periods.

Data from period	Model of period 1955-1959	Model of period 1980-1984	Model of period 2005-2009	Gaussian model
1955-1959	2522.70	2492.89	2341.13	2285.65
1980-1984	2331.59	2376.38	2236.19	2183.04
2005-2009	2721.35	2784.97	3032.48	2970.31

Table 5.7: Log-likelihoods of the selected models and of a Gaussian vine model applied to the data from the different periods.

Data from period	Model of period 1955-1959		Model of period 1980-1984		Model of period 2005-2009		Gaussian model	
	vine without prec	vine incl. prec	vine without prec	vine incl. prec	vine without prec	vine incl. prec	vine without prec	vine incl. prec
1955-1959	1316.99	1205.71	1298.14	1194.75	1263.26	1077.87	1223.63	1062.02
1980-1984	1144.99	1186.60	1181.26	1195.12	1123.11	1113.08	1096.92	1086.12
2005-2009	1317.53	1403.82	1357.73	1427.24	1468.04	1564.44	1443.87	1526.44

Table 5.8: Log-likelihoods of the selected models and of a Gaussian vine model applied to the data when it is not raining and to the data on rain days from the different periods. Note that the sum of both log-likelihoods of each model is given in Table 5.7 and illustrated in Figure 5.13.

values can be found in Table 5.7. Note that in this connection the log-likelihood of a model means the log-likelihood of the vine without precipitation applied to the data when it is not raining plus the log-likelihood of the vine including positive precipitation amount applied to the data on rain days, i.e.

$$l_{vine\ without\ prec}(\text{data}[prec = 0]) + l_{vine\ including\ prec}(\text{data}[prec > 0]),$$

where $l_{vine\ without\ prec}(\cdot)$ and $l_{vine\ including\ prec}(\cdot)$ denote the log-likelihood functions of the corresponding vines. The single values of the summands $l_{vine\ without\ prec}(\text{data}[prec = 0])$ and $l_{vine\ including\ prec}(\text{data}[prec > 0])$ are presented in Table 5.8 for the three periods.

The first observation we notice from Table 5.7 is that the resulting summed log-likelihood values are in the range between 2183.04 and 3032.48. Further, the model of a specific period applied to the corresponding data has the highest log-likelihood in contrast to the models of the other periods and the Gaussian model applied to the same data. An interesting fact is the large increase in the log-likelihood values applying data from the second period to the third period 2005-2009. In addition, there is still a difference of almost 300 in the values of the last period vine model and the Gaussian model in contrast to the other both models applied to data from period 2005-2009. It suggests, however, that a considerable change in the dependencies among the variables occured from the second to the last period. In particular, the difference is as large that even Gaussian copulas are better fit than the no longer suitable selected vines from before. One could say "our model became more Gaussian" in the last period. It corresponds to the previous observations we have made so far, namely the change from pair copulas with asymmetric tail dependendence in the selected R-vines to ones with symmetric tail dependendence among the temperature variables. In detail, as our selected vines in Figures 5.5 and 5.6 show, the pair copulas among the temperature variables shift from survival Gumbel between *meantemp* and *mintemp* to a Frank copula and from Gumbel between *meantemp* and *maxtemp* to a t copula in the last period respectively. The look at the corresponding tail dependencies (illustrated in Figure 5.14 by Kendall's taus of the "upper" and "lower" 20% regions between the margins of the variables) indicates an decrease of the lower tail dependence

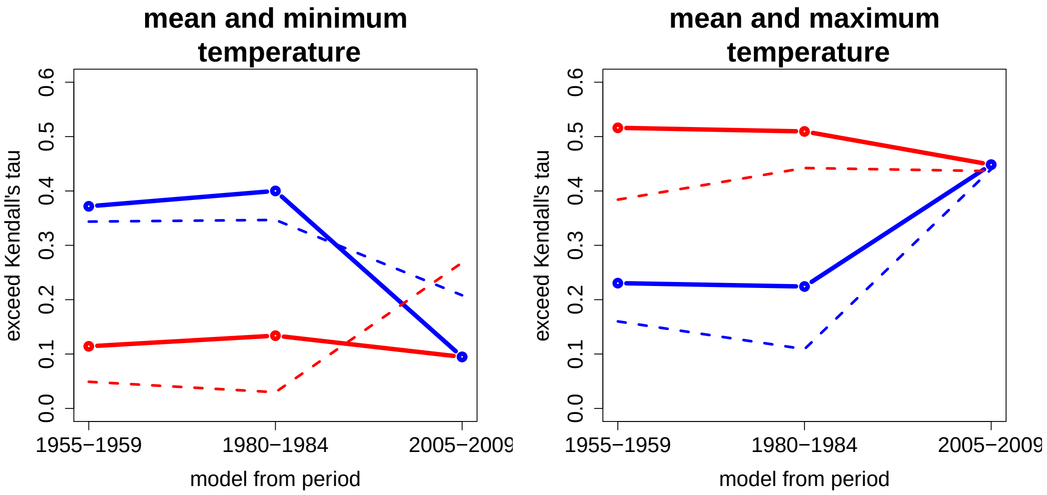


Figure 5.14: Theoretical exceedance Kendall's taus, i.e. Kendall's taus of the "upper" (solid red lines) and "lower" (solid blue lines) 20% regions based on selected copulas of the three models. The empirical counterparts of the margins are illustrated as dashed lines.

among mean and minimum temperature and contemporaneously an increase of the lower tail dependence among mean and maximum temperature to the level of the upper tail dependecies. These remain nearly unchanged among the variables in all models. However, to be precise, the empirical upper tail dependence among mean and minimum temperature seems to increase from the second to the last period. Thus our models suggest that large deviations downwards from the expected daily mean temperature strongly depend on large downward deviations from the expected daily maximum temperature in the last period. In the previous two periods we modeled instead that large downward deviations from the expected daily mean temperature rather depended on large downward deviations from the expected daily minimum temperature. So this tail dependence becomes more variable in the last period. We can compare two examples out of our data exemplarily to make things more concrete. Table 5.9 presents data of two summer days from periods 1980-1984 and 2005-2009. In both cases we detect a large downward deviation from our expected daily minimum temperature, the corresponding margins of the transformed residuals, i.e.

Variable	06/04/1981			06/25/2007		
	observation	fitted expectation	margin of residual	observation	fitted expectation	margin of residual
Mean temp.	8.8°C	17.8°C	< 0.01	16.3°C	16.8°C	0.42
Min. temp.	7.5°C	15.5°C	0.01	7.1°C	12.4°C	0.02
Max. temp.	21.2°C	21.7°C	0.44	22.8°C	21°C	0.67

Table 5.9: Observations vs. fitted expectations of two days from different periods.

both values of $\widehat{u}_{t,mintemp}$ are ≤ 0.02. While we also see a large downward deviation from the fitted daily mean temperature (margin of the residual < 0.01) on June, 4th 1981, the observation of daily mean temperature is close to our expectation (margin of the residual $= 0.42$) on June, 25th 2007. The observations of daily maximum temperature are also close to the fitted expectation (corresponding margins of the residuals 0.44 and 0.67) in both cases. This exemplification should show what the change to more variable (lower) tail dependence among mean and minimum temperature in the last period does mean: large downward deviations from the expected daily minimum temperature do not cause usually large downward deviations from the expected daily mean temperature anymore in the last period. In the same way one could also present an example to illustrate the change to stronger lower tail dependence among daily mean and maximum temperature we have modeled in the last period. However, one should manifest our modeling by studying the dependence structures over all periods of the last decades as well as studying the dependence structure among future observations as next steps.

5.3 Model simulations

As described in Section 4.3, we now simulate 100 × 5-years periods from each R-vine model and so get $100 \times 5 \times 366 = 100 \times 1830 = 183000$ simulated values on copula level (all $\in [0, 1]$). Out of our simulations we then calculate 100 Kendall's taus for each variable pair and compare them with the empirical ones. The results are given in Figure 5.15. Nearly all empirical values are falling inside the 95%-confidence intervals of our simulated Kendall's taus. There are no large deviations from our simulated values detectable. Only in cases of mean and maximum temperature as well as maximum temperature and humidity the empirical taus lie slightly above of the 97.5%- respectively under the 2.5%-quantiles in the first two periods. In case of minimum and maximum temperature the empirical observation lie slightly outside the corresponding confidence interval only in the first period. Nevertheless based on this comparison our models seem to fit the pairwise dependencies relatively sufficient. We notice an upward movement of the pairwise dependence among mean and minimum as well as among mean and maximum temperature by nearly 0.1 over our modeled periods. Thereby the dependence among mean and maximum temperature seems to be higher ($\tau > 0.6$ in all three periods) in contrast to the dependence among mean and minimum temperature (Kendall's tau ranges between 0.4 and 0.6 in all three periods). Another interesting fact is that the negative pairwise dependence among mean temperature and positive precipitation amount as well as among maximum temperature and positive precipitation amount rather seems to decrease over our modeled periods, while the pairwise dependence among minimum temperature and positive precipitation seems to increase over time. Thus the positive rain amount on raindays becomes more variable from the values of daily mean and maximum temperatures in the last period while in contrast our model of this period suggests rather fewer rain amount on rain days when the daily minimum temperature gets higher. In addition there are still more slight movements in the pairwise dependencies among the variables. However, further studies are needed here to prove their significance since we modeled only three 5-years periods of the last 60 years.

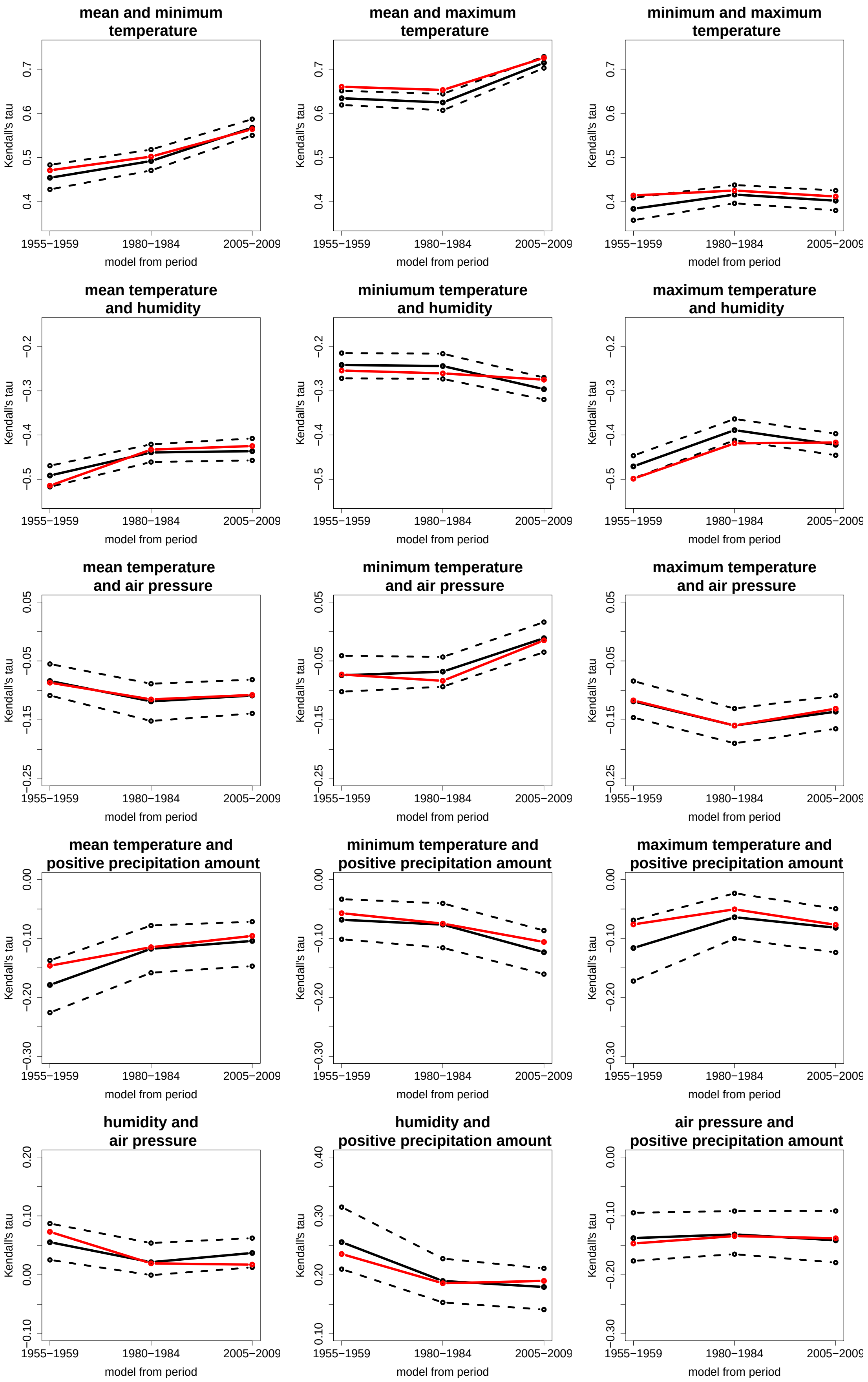


Figure 5.15: [

Mean, 97.5% and 2.5%-quantiles of pairwise simulated Kendall's taus from our R-vine models compared to the empirical ones (red lines).

5.4 Simulated probabilities of different scenarios

A further interesting feature that we obtain using our model simulations is to calculate probabilities of different scenarios, i.e. probabilities of occurances of several (extreme) events at the same time. We decide to calculate the probabilities of the following scenarios:

1: Probability of extreme high daily maximum temperature, extreme high daily mean temperature and extreme high daily minimum temperature at the same time.

2: Probability of extreme daily high maximum temperature, extreme high daily mean temperature and extreme high daily minimum temperature at the same time in winter.

3: Probability of extreme high daily maximum temperature, extreme high daily mean temperature and extreme high daily minimum temperature at the same time in summer.

4: Probability of extreme high daily maximum temperature and no rain at the same time in summer.

5: Probability of extreme high daily maximum temperature, extreme low humidity and no rain at the same time in summer.

6: Probability of extreme high daily maximum temperature, extreme low humidity, extreme high daily mean air pressure and no rain at the same time in summer.

7: Probability of extreme high daily mean temperature and extreme high daily mean air pressure at the same time.

8: Probability of extreme high daily maximum temperature and extreme high daily mean air pressure at the same time in summer.

9: Probability of extreme high daily maximum temperature, extreme high daily mean air pressure and no rain at the same time in summer.

10: Probability of extreme high daily maximum temperature and extreme high rain amount on raindays at the same time in summer.

11: Probability of extreme high daily mean temperature and extreme high rain amount on raindays at the same time in summer.

Note that in this connection extreme high and extreme low values mean margins higher than 0.85 and lower than 0.15 respectively. We calculate 100 probabilities (corresponding to every simulated 5-years period) out of our simulations simply by dividing the number of our event occurrences by the whole number of regarded days. We compare these results with the empirical probabilities of the events for all three periods. Additionally we calculate corresponding 95%-confidence intervals around our empirical values, based on the assumption that they are succes probabilities of Bernoulli experiments. The outcomes are presented in Figure 5.16.

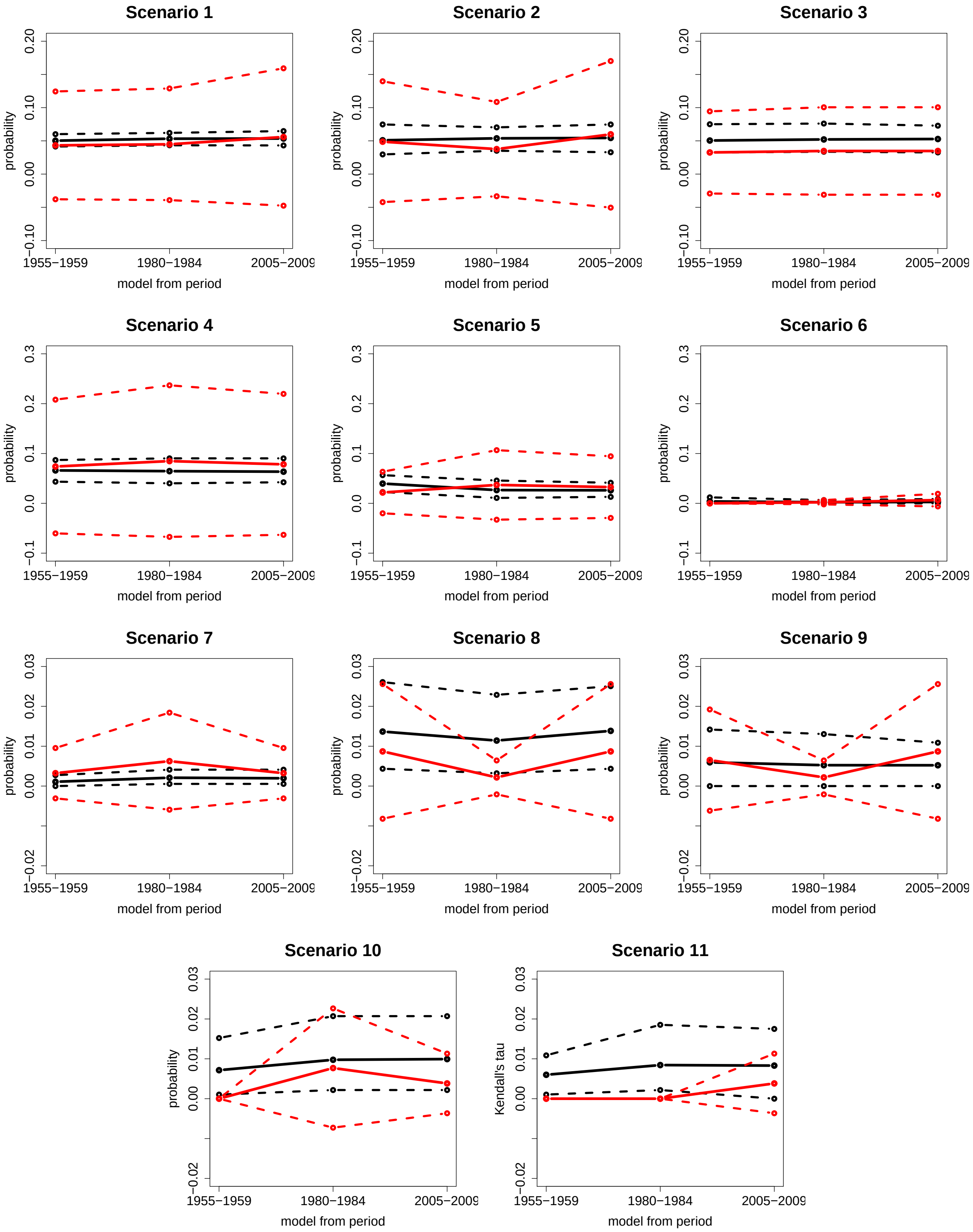


Figure 5.16: Simulated probabilities of different scenarios with 97.5%- and 2.5%-quantiles compared to the empirical ones with their corresponding 95%-confidence intervals (red (dashed) lines).

Naturally the more events should occur at the same time the smaller the probability will be. From Scenario 1 to Scenario 7 our simulated 95%-confidence intervals lie inside the empirical ones in contrast to Scenarios 8 to 11. Large changes in the probabilities of our scenarios do not occur in the regarded three periods. However, in our case it is relatively difficult to draw conclusions out of our results here, since the probabilites are very small and we are again considering only 3 different periods over the last sixty years. Nevertheless this exemplification shows which possibilities we obtain by using such R-vines to model dependence structures among different meteorological variables.

Part II

Modeling tree ring data by linear mixed models

Chapter 6

Preliminaries - Part II

In a second part of this work we would like to model the distributional behavior of tree ring data from Hohenpeissenberg and their dependence on different meteorological quantities. In this connection we think that using *linear mixed models (LMM)* will be the best approach to reach our targets and thus will be our method of choice. On these grounds we present some theoretical basics in the field of linear mixed models in this chapter to be sufficiently prepared for our tasks in the further work. However, beforehand, we will also need some knowledge in smoothing spline interpolation as well as in *generalized additive models (GAM)* to detrend some time effects in the raw tree ring data. Hence, we start with these latter topics and end with the theory of LMMs afterwards in this chapter.

6.1 Smoothing spline interpolation

In our analysis we want to standardize our raw tree ring data in a way that we detrend and transform the measured ring widths into dimensionless indices to equalize the growth variations between tree ring series over time regardless of tree age or size (cp. Cook and Peters [1997]). Thus, one wants to remove non-climatic variances of the data.
The process of detrending can be done by fitting a smooth growth curve to the ring widths, like the modified negative exponential curve or the *smoothing spline* as we will use in our case. Smoothing splines are based on cubic spline functions, i.e. given a subdivided interval $\{x_0 < x_1 < ... < x_n\}$, a cubic spline function corresponds to a real function $s : [x_0, x_n] \mapsto \mathbb{R}$ with the following properties (cp. Freund and Hoppe [2007]):

a) $s \in C^2[x_0, x_n]$, i.e. function s has continuous first and second derivatives on interval $[x_0, x_n]$.

b) On each sub interval $[x_i, x_{i+1}]$, for $i = 0, 1, ..., n-1$, function s corresponds with a cubic polynomial.

The continuity of the first and second derivatives assures that the different cubic polynomial segments are joined in a very smooth way. Hence a smoothing spline is a series of piecewise cubic polynomials with a knot at each data point (e.g. at each point in time t as in our case of tree ring detrending). Due to Reinsch [1967] the smoothing spline minimizes the total squared curvature of the spline function, i.e.

$$\min \int_{x_0}^{x_n} [s''(x)]^2 \, \mathrm{d}x \tag{6.1}$$

under the constraint

$$\sum_{i=0}^{n} \left[\frac{s(x_i)y_i}{\delta y_i} \right]^2 \leq S, \tag{6.2}$$

where y_i is the input series, δy_i is a series of weights and S is a scaling parameter. The quantities δy_i control the extent of smoothing and are implicitly rescaled by varying S. While Reinsch [1967] suggests to use a standard deviation associated with y_i for weights δy_i, Cook and Peters [1981] found out that in case of detrending year ring widths it would be better to weight all measurements equally by $\delta y_i = 1.0$ (for details see [Cook and Peters, 1981, p. 3]). Therefore expression (6.2) reduces to an unweighted residual sum-of-squares criterion and the spline fit is determined by parameter S which Cook and Peters [1981] scaled to be a fraction $s^{'}$ of the variance of the data about the mean. However, another parameter for spline selection is given by examining the standard methods of the calculus of variations. Taking an auxiliary variable z together with the Lagrangian parameter p, one has to minimize the following functional

$$\int_{x_0}^{x_n} [s''(x)]^2 \, \mathrm{d}x + p \left\{ \sum_{i=0}^{n} \left[\frac{s(x_i)y_i}{\delta y_i} \right]^2 + z^2 - S \right\} \tag{6.3}$$

to find a solution of the minimizing problem described in (6.1) and (6.2). Reinsch [1967] shows that if the value of the Lagrangian parameter p is given, one can obtain all other remaining parameters and coefficients (due to the corresponding Euler-Lagrange equations) which then describe the fitted smoothing spline completely. Thus each spline can be defined uniquely also by the value of the Lagrangian multiplier p and so if p is known one can compute the spline directly rather than iteratively. Cook and Peters [1981] found out that a value of $\log_{10}(p) = -4.0$ (here the base 10 logarithm is meant) is a useful starting point for using the smoothing spline in case of detrending tree ring width series because otherwise climatic variance could be indistinguishable from the variance judged to be non climatic [Cook and Peters, 1981, p. 9].

The detrended tree ring data is then given by the ratio actual-to-expected ring width for each year which yields a set of dimensionless tree ring "indices" with a defined mean of 1.0 and a largely homogeneous variance (see Cook and Peters [1997]), i.e. we consider

$$d_t = \frac{y_t}{s(t)}, \tag{6.4}$$

where $s(t)$ denotes the fitted smoothing spline at time t.
Note that there are still other methods to detrend tree ring width data, for further information see, e.g., Cook and Peters [1997].

6.2 Generalized additive model (GAM)

Generalized additive models (GAMs) describe the smooth extension of linear regression models. Due to Hastie and Tibshirani [1986], we consider a response variable $\boldsymbol{Y}$ and a set of predictor variables X_1, X_2,...,X_p. A set of n independent realizations of these random variables will be denoted by $(y_1, x_{11}, ..., x_{1p}), ..., (y_n, x_{n1}, ..., x_{np})$. The additive model generalizes the linear regression such that

$$E[\boldsymbol{Y}|X_1, ..., X_p] = s_0 + \sum_{j}^{p} s_j(X_j), \tag{6.5}$$

where the s_j's are smooth functions standardized so that $E[s_j(X_j)] = 0$. These smoothers then have to be estimated. One simple class of estimates is denoted by, e.g., local average estimates, where (in case of a single predictor $p = 1$)

$$\widetilde{s}(x_i) = Ave_{j \in N_i}(y_j).$$

The function *Ave* represents some averaging operator like the mean and N_i denotes the set of indices of points whose x values are closed to x_i.
However, beside these simple smoothers, other estimates could be used such as kernel, running lines or spline smoother (as we have described in the previous section) to fit a GAM adequately. Therefore different estimation procedures are used, like local scoring or local likelihood procedures, including backfitting procedures to specify the significance of the smooth functions in GAMs. We will not go into any details here but refer to Hastie and Tibshirani [1986] for further information.

6.3 Linear mixed models

Studies with clustered data such as, e.g., math scores from students in different classrooms or studies with longitudinal or repeated-measures data where subjects are measured repeatedly over time or under different conditions, can be modeled by linear mixed models. Also for a combination of both types of data, LMMs provide a flexible analytic tool to model these kinds of continuous outcome variables in which the residuals are normally distributed but may not be independent or not have a constant variance in contrast to linear models. An LMM may include fixed-effect parameters associated with one or more covariates *and* random effects associated with one or more random factors. Thus this mix of fixed and random effects gives the linear mixed models its name. Note, we are here mainly following West et al. [2007] and Fahrmeir et al. [2007].

Before we give a general definition of an LMM, we still define the types and structures of data sets which can be modeled by LMMs:

- **Clustered data:** It means that the dependent variable is measured once for each subject (the *unit of analysis*) and each subject belongs to a group of subjects (cluster). An example beside the above mentioned math scores of students are birth weights of rat pups (the unit of analysis) nested within litters (cluster of units).

- **Repeated-measures data:** These are data sets in which the dependent variable is measured more than once on the same unit of analysis across levels of a repeated measures factor(s). These factors may be time or other experimental or observational conditions. An example is denoted by analyzing the activation of a chemical measured in response to two treatments across three brain regions within each rat (the unit of analysis). Both brain region and treatment are then repeated-measures factors.

- **Longitudinal data:** Data sets in which the dependent variable is measured at several points in time for each unit of analysis. An example is the analysis of socialization scores of a sample of autistic children (the units of analysis), who are each measured at up to five time points (at ages 2, 3, 5, 9 and 13 years).

- **Clustured longitudinal data:** We will focus on these kind of data sets since they combine features of both clustered and longitudinal data as we have in our tree ring data set. The units of analysis are nested within clusters and each unit is measured more than once. Tree ring widths measured for two different drilling directions (units of analysis) are nested within 10 different trees (clusters). The units of analysis (tree ring widths corresponding to the drilling direction) were then repeatedly measured each year for each cluster (tree number).

In some cases it may be difficult to classify data sets as either longitudinal or repeated-measures data. However, this distinction is not critical since the important characteristic of both types of data is that the dependent variable is measured more than once for each unit of analysis.

All here described data sets are hierarchical, because the observations can be placed into levels of a hierarchy in the data. Thus we have *multilevel data sets.* We concentrate on three levels and these levels are categorized in the following way:

- **Level 1:** This level denotes the observations at the most detailed level of the data. Thus the continuous dependent variable is always measured at level 1 of the data. In clustered data sets it corresponds to the units of analysis in the study while in repeated-measures or longitudinal data sets level 1 denotes the repeated measures made on the same unit of analysis.

- **Level 2:** It denotes the next level of hierarchy. In clustered data we find here the clusters of units and in repeated-measures and longitudinal data sets it corresponds to the units of analysis.

- **Level 3:** A further next level of hierarchy that refers to clusters of units in clustered longitudinal data sets or denotes clusters of clusters).

An examplification of multiple levels of different hierarchical data sets according to the above mentioned examples is given in Table 6.1.

In a next step we still like to clarify the distinction between fixed and random factors and their related effects on a dependent variable in the context of LMMs. Due to West

Data type		Clustered data		Repeated-measures/longitudinal data		
		Two-level	Three-level	Repeated-measures	Longitudinal	Clustered longitudinal
Data set example		Rat pup	Classroom	Rat brain	Autism	Tree ring
Level of hierarchy	Level 1	Rat pup	Student	Repeated measures (brain region and treatment)	Longitudinal measures (age in years)	Longitudinal measures (time in years)
	Level 2	Litter	Classroom	Rat	Child	Drilling direction
	Level 3		School			Tree number

Table 6.1: Examplification of multiple levels of different hierarchical data sets according to the above mentioned examples.

et al. [2007], **fixed factors** are defined as categorical or classification variables for which the investigator has included **all levels** that are of interest in the study. They might include qualitative covariates or classification variables such as, e.g., gender, season or age group.

In contrast **random factors** are denoted by classification variables with levels that can be thought of as being **randomly sampled** from a population of levels being studied. The data set does not provide all possible levels of the random factor, but one has the aim to make inferences about the entire population of levels. Thus the results of the data analysis can be generalized to a greater population of levels of the random factor.

Fixed effects in an LMM are unknown constant parameters, i.e. the regression coefficients, associated with either continuous covariates or the levels of the categorical fixed factors. On the other hand the effects associated with the levels of the random factors can be modeled as **random effects** in an LMM. These random effects, in contrast to fixed effects, are represented by unobserved random variables, which are assumed to follow a normal distribution.

The levels of a factor (random or fixed) are said to be **nested** within levels of another factor, when a certain level of a factor can only be measured within a single level of another factor and thus not across multiple levels. The corresponding effects are known as **nested effects**. For example levels of classroom are nested within levels of school since each classroom can only appear within one school.

Moreover, one factor is said to be **crossed** with another, when a given level of a factor (random or fixed) can be measured across multiple levels of another factor. However, we will not go into any details here.

6.3.1 Specification of LMMs

To simplify matters we start to specify an LMM for a two-level longitudinal data set for illustration. Thus in this specification Y_{ti} represents the measure of the continuous response variable Y taken on the t-th point in time for the i-th subject, i.e.

$$\begin{aligned} Y_{ti} &= \underbrace{\beta_1 \times X_{ti}^{(1)} + \beta_2 \times X_{ti}^{(2)} + \beta_3 \times X_{ti}^{(3)} + ... + \beta_k \times X_{ti}^{(p)}}_{\text{fixed}} \\ &\quad \underbrace{+u_{1i} \times Z_{ti}^{(1)} + ... + u_{qi} \times Z_{ti}^{(q)} + \epsilon_{ti}.}_{\text{random}} \end{aligned} \tag{6.6}$$

Here $t = 1, ..., n_i$, where n_i denotes the number of longitudinal observatios on the dependent variable for a given subject i (unit of analysis) with $i = 1, ..., m$, i.e. m is the number of subjects. Model (6.6) involves two sets of covariates, X and Z. The first set contains p covariates $X^{(1)}, ..., X^{(p)}$ associated with the fixed effects $\beta_1, ..., \beta_p$. Thus, the terms $X_{ti}^{(1)}, ..., X_{ti}^{(p)}$ represent the t-th observation of the corresponding covariate for the i-th subject. The second set contains q covariates $Z^{(1)}, ..., Z^{(q)}$ associated with the random effects $u_{1i}, ..., u_{qi}$ corresponding to subject i. Variable ϵ_{ti} represents the residual associated with the t-th observation in the i-th subject. Note as described before, the random effects and residuals in Equation (6.6) are random variables which are assumed to be normal distributed. We additionally assume that for a given subject, the residuals are independent of the random effects.

We can now generalize the above model for all longitudinal and clustered data sets considering a general matrix specification of an LMM.

Definition 6.1 (Linear mixed model for longitudinal and clustered data.) *A linear mixed model (LMM) for longitudinal and clustered data is defined by*

$$\boldsymbol{Y}_i = \underbrace{X_i \boldsymbol{\beta}}_{\textit{fixed}} + \underbrace{Z_i \boldsymbol{u}_i + \boldsymbol{\epsilon}_i}_{\textit{random}}, \qquad i = 1, ..., m, \tag{6.7}$$

where

$$\boldsymbol{Y}_i := \begin{pmatrix} Y_{1i} \\ Y_{2i} \\ \vdots \\ Y_{n_i i} \end{pmatrix}, X_i := \begin{pmatrix} X_{1i}^{(1)} & X_{1i}^{(2)} & \dots & X_{1i}^{(p)} \\ X_{2i}^{(1)} & X_{2i}^{(2)} & \dots & X_{2i}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n_i i}^{(1)} & X_{n_i i}^{(2)} & \dots & X_{n_i i}^{(p)} \end{pmatrix}, \boldsymbol{\beta} := \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

and

$$Z_i := \begin{pmatrix} Z_{1i}^{(1)} & Z_{1i}^{(2)} & \dots & Z_{1i}^{(q)} \\ Z_{2i}^{(1)} & Z_{2i}^{(2)} & \dots & Z_{2i}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n_i i}^{(1)} & Z_{n_i i}^{(2)} & \dots & Z_{n_i i}^{(q)} \end{pmatrix}, \boldsymbol{u}_i := \begin{pmatrix} u_{1i} \\ u_{2i} \\ \vdots \\ u_{qi} \end{pmatrix}, \boldsymbol{\epsilon}_i := \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \vdots \\ \epsilon_{n_i i} \end{pmatrix}.$$

Here, $\boldsymbol{Y}_i$ represents the n_i-dimensional vector of continuous responses for the i-th subject or cluster respectively and m denotes the number of subjects respectively cluster. Further, X_i and Z_i are the $(n_i \times p)$- and $(n_i \times q)$-dimensional design matrices that respresent the known values of the p and q covariates $X^{(1)}, ..., X^{(p)}$ and $Z^{(1)}, ..., Z^{(q)}$; $\boldsymbol{\beta}$ corresponds to the p-dimensional vector of the fixed effects while $\boldsymbol{u}_i$ describes the vector of q random effects that are specific to subject or cluster i. The remaining error term is given by *the n_i-dimensional vector $\boldsymbol{\epsilon}_i$ for a given subject or cluster i. We assume that $\boldsymbol{u}_i$ and $\boldsymbol{\epsilon}_i$, $i = 1, ..., m$, are multivariate normal distributed, i.e.*

$$\begin{aligned} \boldsymbol{u}_i &\sim \mathcal{N}(\mathbf{0}, D), \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}(\mathbf{0}, R_i), \end{aligned} \tag{6.8}$$

where $\boldsymbol{u}_1, ..., \boldsymbol{u}_m$ and $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_m$ are independent and D corresponds to a symmetric and positive definite $q \times q$ matrix with

$$D := Var(\boldsymbol{u}_i) = \begin{pmatrix} Var(u_{1i}) & Cov(u_{1i}, u_{2i}) & \dots & Cov(u_{1i}, u_{qi}) \\ Cov(u_{1i}, u_{2i}) & Var(u_{2i}) & \dots & Cov(u_{2i}, u_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(u_{1i}, u_{qi}) & Cov(u_{2i}, u_{qi}) & \dots & Var(u_{qi}) \end{pmatrix},$$

which is constant for all $i = 1, ..., m$. The positive definite symmetric covariance matrix R_i is given by

$$R_i := Var(\boldsymbol{\epsilon}_i) = \begin{pmatrix} Var(\epsilon_{1i}) & Cov(\epsilon_{1i}, \epsilon_{2i}) & \dots & Cov(\epsilon_{1i}, \epsilon_{n_i i}) \\ Cov(\epsilon_{1i}, \epsilon_{2i}) & Var(\epsilon_{2i}) & \dots & Cov(\epsilon_{2i}, \epsilon_{n_i i}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\epsilon_{1i}, \epsilon_{n_i i}) & Cov(\epsilon_{2i}, \epsilon_{n_i i}) & \dots & Var(\epsilon_{n_i i}) \end{pmatrix}.$$

Note, for clustered longitudinal data as in our tree ring modeling, we will have a further index j representing the cluster. Thus we model Y_{tij} where index t denotes the points in time (Level 1 units), index i is being used for subjects (Level 2 units) and index j stands for the different cluster (Level 3 units). However, the concrete model in the case of tree ring data is specified in the next chapter.

The variances and covariances of the R_i matrix in assumption (6.8) are defined as functions of another usually small set of covariance parameters stored in a vector denoted by $\boldsymbol{\theta}_R$ which has to be estimated. Many different covariance structures are possible here for the R_i matrix such as, e.g. autoregressive structures. However we will concentrate on the simplest covariance matrix for R_i, namely the diagonal one in which the residuals are assumed to be uncorreslated and to have equal variance. That means for each subject i

$$R_i = Var(\boldsymbol{\epsilon}_i) = \sigma^2 I = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} \in \mathbb{R}^{n_i \times n_i} \quad \text{and} \quad \boldsymbol{\theta}_R = (\sigma^2). \tag{6.9}$$

Also for the D matrix from (6.8) one could consider different covariance structures. However, a D matrix with no additional constraints on its elements (aside from positive definiteness and symmetry) implies that $(q \times (q + 1))/2$ covariance parameters have to be estimated, stored in the vector $\boldsymbol{\theta}_D$. In case of an LMM having only two random effects associated with the i-th subject, the matrix D looks as follows

$$D = Var(\boldsymbol{u}_i) = \begin{pmatrix} \sigma_{u1}^2 & \sigma_{u1,u2} \\ \sigma_{u1,u2} & \sigma_{u2}^2 \end{pmatrix} \in \mathbb{R}^{2\times 2} \quad \text{and} \quad \boldsymbol{\theta}_D = \begin{pmatrix} \sigma_{u1}^2 \\ \sigma_{u1,u2} \\ \sigma_{u2}^2 \end{pmatrix}. \tag{6.10}$$

The vector $\boldsymbol{\theta}$ then combines all covariance parameters contained in the vectors $\boldsymbol{\theta}_D$ and $\boldsymbol{\theta}_R$.

In Definition 6.1 we present a general matrix specification of the LMM for a given subject i. An alternative matrix specification for all subjects is then given by

$$\boldsymbol{Y} = \underbrace{X\boldsymbol{\beta}}_{\text{fixed}} + \underbrace{Z\boldsymbol{u} + \boldsymbol{\epsilon}}_{\text{random}}, \quad \text{where} \tag{6.11}$$

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \mathcal{N}_{mq+n}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} G & 0_{mq\times n} \\ 0_{n\times mq} & R \end{pmatrix} \right),$$

and $n := \sum_{i=1}^m n_i$. Additionally, we have

$$\boldsymbol{Y} := \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix} \in \mathbb{R}^n, \ X := \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix} \in \mathbb{R}^{n\times p}, \ \boldsymbol{\beta} \in \mathbb{R}^p \quad \text{and}$$

$$Z := \begin{pmatrix} Z_1 & 0_{n_1\times q} & \dots & \dots \\ 0_{n_2\times q} & Z_2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_m\times q} & \dots & \dots & Z_m \end{pmatrix} \in \mathbb{R}^{n\times mq} \quad \text{with} \quad 0_{n_i\times q} := \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{n_i\times q}.$$

$$\boldsymbol{u} := \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_m \end{pmatrix} \in \mathbb{R}^{mq}, \ \boldsymbol{\epsilon} := \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{pmatrix} \in \mathbb{R}^n, \quad \text{as well as} \quad G := \begin{pmatrix} D & 0_{q\times q} & \dots & 0_{q\times q} \\ 0_{q\times q} & D & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_{q\times q} & \dots & \dots & D \end{pmatrix} \in \mathbb{R}^{mq\times mq}$$

$$\text{and} \quad R := \begin{pmatrix} R_1 & 0_{n_1\times n_2} & \dots & \dots \\ 0_{n_2\times n_1} & R_2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_m\times n_1} & \dots & \dots & R_m \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

The LMM introduced in Equation (6.11) implies the following marginal linear model:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \quad \text{where} \quad \boldsymbol{\epsilon}^* := Z\boldsymbol{u} + \boldsymbol{\epsilon} = \underbrace{(Z, I_{n\times n})}_{=:A} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \tag{6.12}$$

and thus

$$\boldsymbol{\epsilon}^* \sim \mathcal{N}_n(\boldsymbol{0}, V), \quad \text{where}$$

$$V := A \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} A' = (Z, I_{n\times n}) \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} Z' \\ I_{n\times n} \end{pmatrix} = ZGZ' + R \tag{6.13}$$

Hence, the error term of (6.12) for subject i is given by

$$\boldsymbol{\epsilon}_i^* = Z_i \boldsymbol{u}_i + \boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, V_i), \quad \text{with} \;\; V_i = Z_i D Z_i' + R_i.$$

Note that Model (6.11) implies the linear model (6.12), but Model (6.12) does not imply the LMM from Defintion 6.1 and Equation (6.11). The concept of the implied marginal model is certainly important since the estimation of the fixed-effect and covariance parameters in the LMM is carried out in the framework of the implied marginal linear model. So, e.g., if one is only interested in estimating $\boldsymbol{\beta}$, one can use the linear model of (6.12) and the method of weighted least squares to estimate the corresponding parameters as long as V is known.

The parameter estimation in LMMs will now be focused on in the next subsection.

6.3.2 Estimation in LMMs

In an LMM, as described in Defintion 6.1, we want to estimate the fixed-effect parameters $\boldsymbol{\beta}$ and the covariance parameters $\boldsymbol{\theta}$ (i.e. $\boldsymbol{\theta}_D$ and $\boldsymbol{\theta}_R$ for the D and R_i matrices). Therefore the commonly used methods to estimate these parameters are the *maximum likelihood (ML)* and the *restricted maximum likelihood (REML)* estimation which we will describe in the following.

Maximum likelihood estimation

This method of obtaining estimates of the unknown parameters by optimizing the likelihood function is well known, since we already introduced it in Chapter 2. Given the observed data of $\boldsymbol{Y}_i$ by $\boldsymbol{y}_i$ the likelihood function for the i-th subject is defined as follows according to the assumption of a multivariate normal distribution of $\boldsymbol{Y}_i$ in our implied marginal linear model from (6.12), i.e.

$$L_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = (2\pi)^{-\frac{n_i}{2}} \, |V_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{y}_i - X_i\boldsymbol{\beta})' V_i^{-1} (\boldsymbol{y}_i - X_i\boldsymbol{\beta}) \right),$$

where $|V_i|$ denotes the determinant of V_i. Then the likelihood function $L(\boldsymbol{\beta}, \boldsymbol{\theta})$ is given by the product of the m independent contributions of the likelihood function for the i-th subject, i.e.

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^{m} L_i(\boldsymbol{\beta}, \boldsymbol{\theta}).$$

The corresponding log-likelihood function equals $l(\boldsymbol{\beta}, \boldsymbol{\theta}) = \log L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \log L_i(\boldsymbol{\beta}, \boldsymbol{\theta})$ which looks like

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = -0.5n \times \log(2\pi) - 0.5 \times \sum_i \log(|V_i|) - 0.5 \times \sum_i (\boldsymbol{y}_i - X_i\boldsymbol{\beta})' V_i^{-1} (\boldsymbol{y}_i - X_i\boldsymbol{\beta}). \tag{6.14}$$

Now we distinguish between two cases: $\boldsymbol{\theta}$ is known and $\boldsymbol{\theta}$ is unknown respectively.

- ***θ* is known:** As we already mentioned above, if $\boldsymbol{\theta}$ is known it follows that matrices V_i are knwon for all i and thus also matrix V. Hence one is only interested in estimating the regression parameters $\boldsymbol{\beta}$. According to our implied marginal linear model (6.12) we can use the method of weighted least squares (introduced in Section 2.5.2 in Chapter 2) and the optimal value to estimate $\boldsymbol{\beta}$ can be obtained analytically. We get
$$\widehat{\boldsymbol{\beta}} = \left(\sum_i X_i' V_i^{-1} X_i\right)^{-1} \sum_i X_i' V_i^{-1} \boldsymbol{y}_i. \tag{6.15}$$
One can show that $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

- ***θ* is unknown:** When we assume $\boldsymbol{\theta}$ to be unknown, we first obtain estimates for the covariance parameters in $\boldsymbol{\theta}$ by using a *profile log-likelihood function* $l_{ML}(\boldsymbol{\theta})$. This function $l_{ML}(\boldsymbol{\theta})$ is derived from $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ in (6.14) by replacing $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}$, defined in Equation (6.15), i.e.
$$l_{ML}(\boldsymbol{\theta}) = -0.5n \times \log(2\pi) - 0.5 \times \sum_i \log(|V_i|) - 0.5 \times \sum_i \boldsymbol{r}_i' V_i^{-1} \boldsymbol{r}_i, \tag{6.16}$$
where
$$\boldsymbol{r}_i = \boldsymbol{y}_i - X_i \left(\left(\sum_i X_i' V_i^{-1} X_i\right)^{-1} \sum_i X_i' V_i^{-1} \boldsymbol{y}_i \right).$$
The maximization of $l_{ML}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is a nonlinear optimization problem including constraints on $\boldsymbol{\theta}$ so that the requirements of positive definiteness on D and the R_i matrices are satisfied. However, there exists no closed-form solution for the optimal $\boldsymbol{\theta}$ and hence one has to resort to a numerical algorithm (like Newton-Raphson or Fisher scoring) to obtain an estimate of $\boldsymbol{\theta}$.

  After doing this, we define $\widehat{V_i} := Z_i \widehat{D} Z_i' + \widehat{R}_i$ by replacing D and R_i in Equation (6.13) by their ML estimates $\widehat{D}$ and $\widehat{R}_i$ for all i. Then again we use the method of weighted least squares to estimate $\boldsymbol{\beta}$ with V_i replaced by its estimate $\widehat{V_i}$ and thus we get
$$\widehat{\boldsymbol{\beta}} = \left(\sum_i X_i' \widehat{V}_i^{-1} X_i\right)^{-1} \sum_i X_i' \widehat{V}_i^{-1} \boldsymbol{y}_i. \tag{6.17}$$
Here, $\widehat{\boldsymbol{\beta}}$ is again an unbiased estimator of $\boldsymbol{\beta}$, i.e. $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

Note in both cases the variance of estimate $\widehat{\boldsymbol{\beta}}$, i.e. $Var(\widehat{\boldsymbol{\beta}})$ is a $p \times p$ variance-covariance matrix given by
$$Var(\widehat{\boldsymbol{\beta}}) = \left(\sum_i X_i' \widehat{V}_i^{-1} X_i\right)^{-1}. \tag{6.18}$$

REML estimation

The REML estimation is an alternative method to estimate the covariance parameters in $\boldsymbol{\theta}$. REML is often preferred to ML estimation, because it produces unbiased estimates of covariance parameters in contrast to ML since it takes into account the loss of degrees of freedom that results from estimating the fixed effects in $\boldsymbol{\beta}$.
In detail, one optimizes the following *REML log-likelihood function* to estimate $\boldsymbol{\theta}$, i.e.

$$\begin{aligned} l_{REML}(\boldsymbol{\theta}) &= -0.5 \times (n-p) \times \log(2\pi) - 0.5 \times \sum_i \log(|V_i|) \\ &\quad -0.5 \times \sum_i \boldsymbol{r}_i' V_i^{-1} \boldsymbol{r}_i - 0.5 \times \sum_i \log\left(\left|X_i' \widehat{V}_i^{-1} X_i\right|\right), \end{aligned} \tag{6.19}$$

where $\boldsymbol{r}_i$ is defined as before. Again the estimate $\widehat{V}_i$ is computed numerically and the REML estimates of the fixed-effect parameters $\widehat{\boldsymbol{\beta}}$ and $Var(\widehat{\boldsymbol{\beta}})$ are then computed using the Equations (6.17) and (6.18) as in the ML case. Note, although using the same equations, the REML estimate of $\widehat{\boldsymbol{\beta}}$ and the corresponding $Var(\widehat{\boldsymbol{\beta}})$ differ from the ML estimate since the matrix $\widehat{V}_i$ is different in each case. But, however, the estimated variances of the estimated fixed-effect parameters are biased downward in both ML and REML estimation because they do not take into account the uncertainty introduced by replacing V_i with $\widehat{V}_i$ [West et al., 2007, p. 29].

6.3.3 Tools for model selection

An important task in this connection is to select the "best" model, i.e. a model that have fewest number of parameters used and at the same time is best at predicting or explaining the variation in our response (dependent) variable. Beside the assumption that one considers the research objectives, i.e. previous knowledge about important predictors and important subject matter considerations, we also use analytic tools here. We will especially focus on hypothesis tests and information criteria to select the most appropriatest model.

Hypothesis tests

We want to test hypotheses about parameters in an LMM and therefore test a null (H_0) against an alternative (H_1) hypothesis about the parameters. We start with *likelihood ratio tests (LRTs)*:

- **Likelihood ratio tests:** LRTs are a class of tests that are based on comparing the values of likelihood functions for two models defining a hypothesis being tested. In our context these two models have a nesting relationship, which means that a more general model (reference model) encompasses the null and alternative hypotheses while a second simpler model model (nested model) satisfies the null hypothesis. So the nested model is a "special case" of the reference model, i.e. the only difference between these two models is that the reference model contains the parameters being tested but the nested model does not. LRTs can be used to test hypotheses about covariance parameters or fixed-effect parameters. The corresponding LRT statistic is calculated as follows

$$LR := -2\log\left(\frac{L_{nested}}{L_{reference}}\right) = -2\log\left(L_{nested}\right) - \left(-2\log\left(L_{reference}\right)\right) \sim \chi^2_{df}, \quad (6.20)$$

where L_{nested} and $L_{reference}$ denote to the values of the likelihood functions of the corresponding models evaluated at the ML or REML estimates of the parameters. Under mild regularity conditions the LRT statistic follows a χ^2 distribution where the number of degrees of freedom *df* is obtained by subtracting the number of parameters in the nested model from the number of parameters in the reference model (i.e., for example, if the reference model includes 5 parameters and the nested model 3, the number of degrees of freedom equals $df = 2$).

Note, due to West et al. [2007], the likelihood ratio tests used to test hypotheses about fixed-effect parameters in an LMM should be based on ML estimation because REML estimation is not appropriate in this context. In contrast, when we test hypotheses about covariance parameters in an LMM, REML estimation should be rather used for reference and nested models.

We still draw our attention to the case when we want to test whether a given random effect should be kept in a model or not. We are doing this by testing whether the corresponding variances and covariances are equal to zero and thus whether the random effect can be omitted. Note, that the covariance parameters satisfying the null hypothesis lie on the boundary of the parameter space. Therefore we have to distinguish two cases:

(i) In case of having only a single random effect in our reference model, the calculations of p-values are based on a χ^2_1 distribution weighted by 0.5 [West et al., 2007, p. 36], i.e.

$$p - value = 0.5 \times P(\chi^2_1 > LR)$$

(ii) In case in which we have two random effects in our reference model and we wish to test whether one of them can be omitted, i.e. we test whether the variance for the given random effect that we wish to test and the associated covariance of the two random effects are both equal to zero. Then the corresponding asymptotic null distribution (i.e. the under H_0 assumed distribution) of the test statistic LR is a mixture of χ^2_1 and χ^2_2 distributions with each having an equal weight of 0.5 (cp. West et al. [2007]), i.e.

$$p - value = 0.5 \times P(\chi^2_1 > LR) + 0.5 \times P(\chi^2_2 > LR) \quad (6.21)$$

- ***t*-test:** It is used for testing a single fixed-effect parameter in an LMM, i.e. testing

$$H_0 : \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_k \neq 0,$$

for $k \in \{1, ..., p\}$. The corresponding t-statistic is then given by

$$t = \frac{\widehat{\beta}_k}{\sqrt{Var(\widehat{\beta}_k)}}. \quad (6.22)$$

However, unlike the standard linear model, the test statistic (6.22) does not follow an exact t distribution, i.e. the number of degrees of freedom for the null distribution of the test statistic is not equal to $n - p$. One has to use approximate methods to estimate appropriate degrees of freedom. Further details can be found in West et al. [2007].

- F**-test:** An F-test can be used to test hypotheses about multiple fixed effects in an LMM, i.e. testing $H_0 : X\boldsymbol{\beta} = \mathbf{0}$ vs. $H_1 : X\boldsymbol{\beta} \neq \mathbf{0}$. We will not go into any details here, we rather refer to the mentioned references for more information.

Information criteria

Information criteria are another set of useful tools in model selection, i.e. assessing the fit of a model based on its optimum log-likelihood. They provide a way to compare any two models fitted to the same set of observations without being nested. For the criteria we follow the form "the smaller the better", i.e. a smaller value of the criterion indicates a "better" fit. We introduce two measures:

- **Akaike information criterion (AIC):** We have already introduced the AIC in Chapter 2. It is calculated (in our case) based on the ML or REML log-likelihood of a fitted model due to Akaike [1973], i.e.

$$AIC = -2 \times l(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) + 2(p + q). \tag{6.23}$$

- **Bayes information criterion (BIC):** The BIC is given by

$$BIC = -2 \times l(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) + (p + q) \times \log(n).$$

  In contrast to the AIC, the BIC penalizes models with a large number of parameters more than the AIC does since one multiplies the number of parameters being estimated $(p + q)$ by the logarithm of the total number of observations n.

However, we will concentrate on the AIC.

We are now ready to model our tree ring data by linear mixed models. For the implementation, parameter estimation and model selection we will use the R-library `nlme`, implemented by Pinheiro et al. [2012].

Chapter 7

Model selection

In contrast to the first part of our work, we now study *yearly* data, i.e. yearly tree ring width (in mm) data measured from trees in the region of Hohenpeissenberg (Oberammergau). In detail, the altitude of the tree location is about 910m above sea level, the trees exposition corresponds to circa north-north-west (NNW) and the slope angle measures about 10 degrees in all cases. We consider complete data over 55 years (1950-2004) for 10 trees of two different tree species, namely Norway spruce (*Picea abies* [L.] Karst.) and silver fir (*Abies alba* Mill.). For every tree we measure the corresponding year ring from two different drilling directions, i.e. from south-east and south-west, to have fundamental data. Thus, altogether we have

$$\underbrace{55}_{years} \times \underbrace{10}_{number\ of\ trees} \times \underbrace{2}_{number\ of\ drilling\ directions} = 1100$$

data points a each tree species.

Before we start to specify our model, we would like to *detrend* the raw tree ring data series for each tree first. In order to do not distort the pure influence of different meteorological quantities on our dependent year ring variable we are using smoothing splines as introduced in Section 6.1 by Cook and Peters [1981] to remove any non-climatic variance out of our series. Then we take the ratio described in (6.4) to yield our detrended data, i.e.

$$treering_{tij} := RWI_{tij} = \frac{y_{tij}}{s_{tij}}, \tag{7.1}$$

where y_{tij} corresponds to the values of the raw data series and s_{tij} describes the values of the fitted smoothing spline at time $t \in \{1950, ..., 2004\}$ for tree number $j = 1, ..., 10$ measured from drilling direction $i = 1, 2$. Thus, the values of RWI_{tij} (ring width index) correspond to the detrended tree ring data calculated as described in (7.1). Note, for simplicity, we will name RWI_{tij} by $treering_{tij}$ in the following which we will then model by linear mixed models. The method of detrending, i.e. removing any time and tree depending trend, is performed by the R function `detrend` of the R library `dplR`, implemented by Bunn et al. [2012]. An example of detrended raw data for two trees of different species is presented in Figure 7.1. Note that we here detrend all available data for a single tree (thus for some trees we have data from more than 55 years) but in the end we only consider (detrended) data over the mentioned period 1950-2004 for our model to have complete

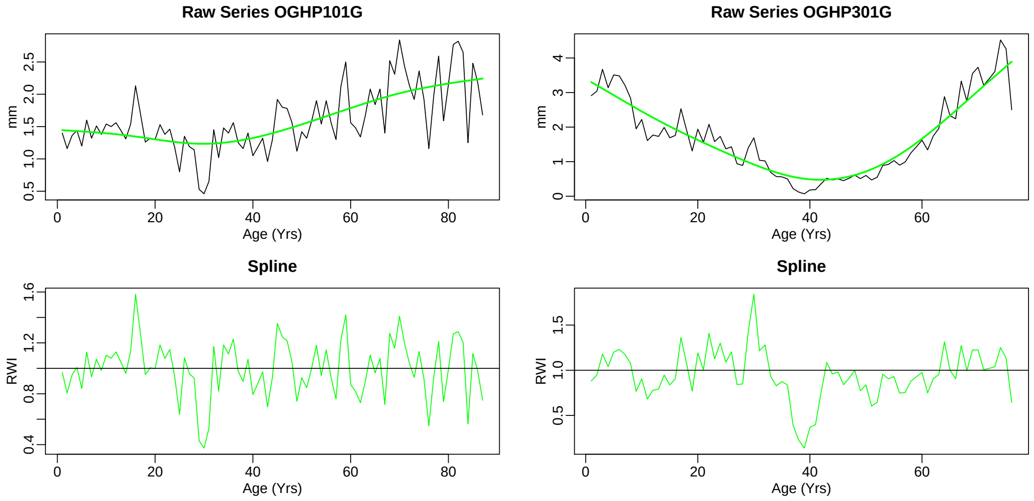


Figure 7.1: Two examples of detrending the raw tree ring series of a single tree. The left panel corresponds to the data of spruce tree (1) No. 1 (01) with drilling direction from south-east (G) while the right panel corresponds to data of fir tree (3) No. 1 (01) with drilling direction from south-east (G). Both trees are coming from the region Oberammergau/Hohenpeissenberg (OGHP).

data of all trees for comparison.

In our analysis we want to examine the (linear) relationship between the yearly tree ring width and several yearly meteorological quantities modeled by linear mixed models. As seen in the previous part we have daily data available of different meteorological variables from Hohenpeissenberg which we want to connect with the yearly tree ring width. Hence we will calculate *yearly means* of these daily data over different *periods or seasons* and use them as covariates in our model to explain the variation in the tree ring variables. In principle we will concentrate on the same quantities as in Chapter 3 but we will only regress on one temperature variable, i.e. we will calculate our yearly air temperature means only over daily mean air temperature. In addition we are interested in one further quantity here, namely the highest number of consecutive days without rain, i.e. the longest dry period in a regarded season. Thus, the following quantities will be implemented in our model:

1. Total amount of precipitation [measured in mm]

2. Relative humidity [measured in %]

3. Air pressure [measured in mbar]

4. Air temperature [measured in °C]

5. Longest dry period: Highest number of consecutive days without rain

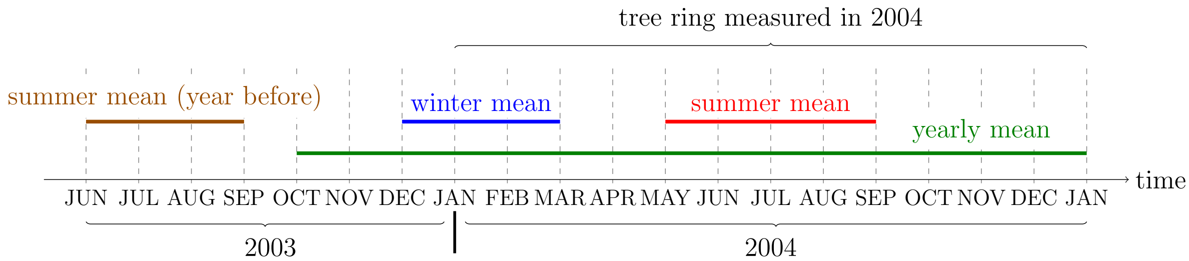

tree ring measured in 2004
summer mean (year before)
winter mean
summer mean
yearly mean
time
JUN JUL AUG SEP OCT NOV DEC JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC JAN
2003
2004

Figure 7.2: Illustration of the seasons or periods of the calculated means of our considered daily meteorological quantities in case of the year 2004.

A tree ring width measured in a specific year is influenced by climate occurrences during previous seasons and periods respectively (see, e.g., Fritts [1976]). Recently works, e.g. by Dittmar and Elling [1999] or Leal et al. [2008], have shown that year ring widths of fir and spruce trees are influenced by the climate of the different seasons defined in the following and hence we will calculate the means of our considered meteorological variables over these periods. We start by defining a ”yearly season” as the time period from October to December of the previous year plus the period January to December of the studied year. Then we calculate a **yearly mean** of the aboved mentioned daily data over the period

$$\begin{gathered}\text{October -- December of the previous year} + \\ \text{January -- December of the studied year.}\end{gathered} \tag{7.2}$$

Furthermore we want to examine the relationship between the yearly tree ring width and the mean of the aforementioned daily data only over summer months, resp. winter months. Therefore we calculate two **summer means** over the periods

$$\text{June -- August of the previous year} \tag{7.3}$$

and

$$\text{May -- August of the studied year.} \tag{7.4}$$

And the **winter mean** is calculated over the period

$$\begin{gathered}\text{December of the previous year} + \\ \text{January and February of the studied year.}\end{gathered} \tag{7.5}$$

An illustration of the used periods for the different mean calculations can be found in Figure 7.2 using the example of tree ring measured in year 2004.

7.1 Data analysis

Before we specify and start selecting appropriate linear mixed models for both tree species, we still take a look at the behavior of the detrended tree ring width series over the considered 55 years (1950-2004) for all 10 different trees of each species. Especially whether

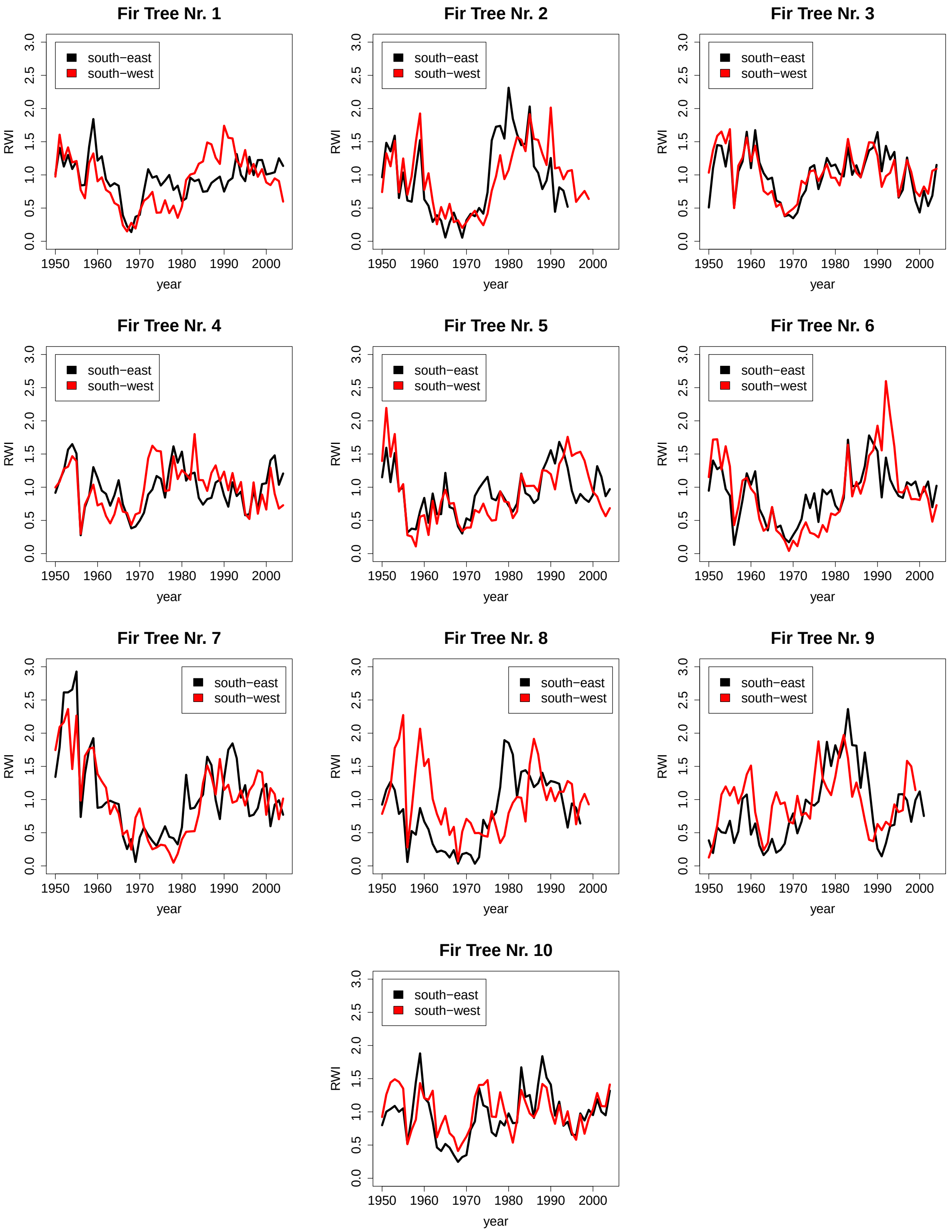
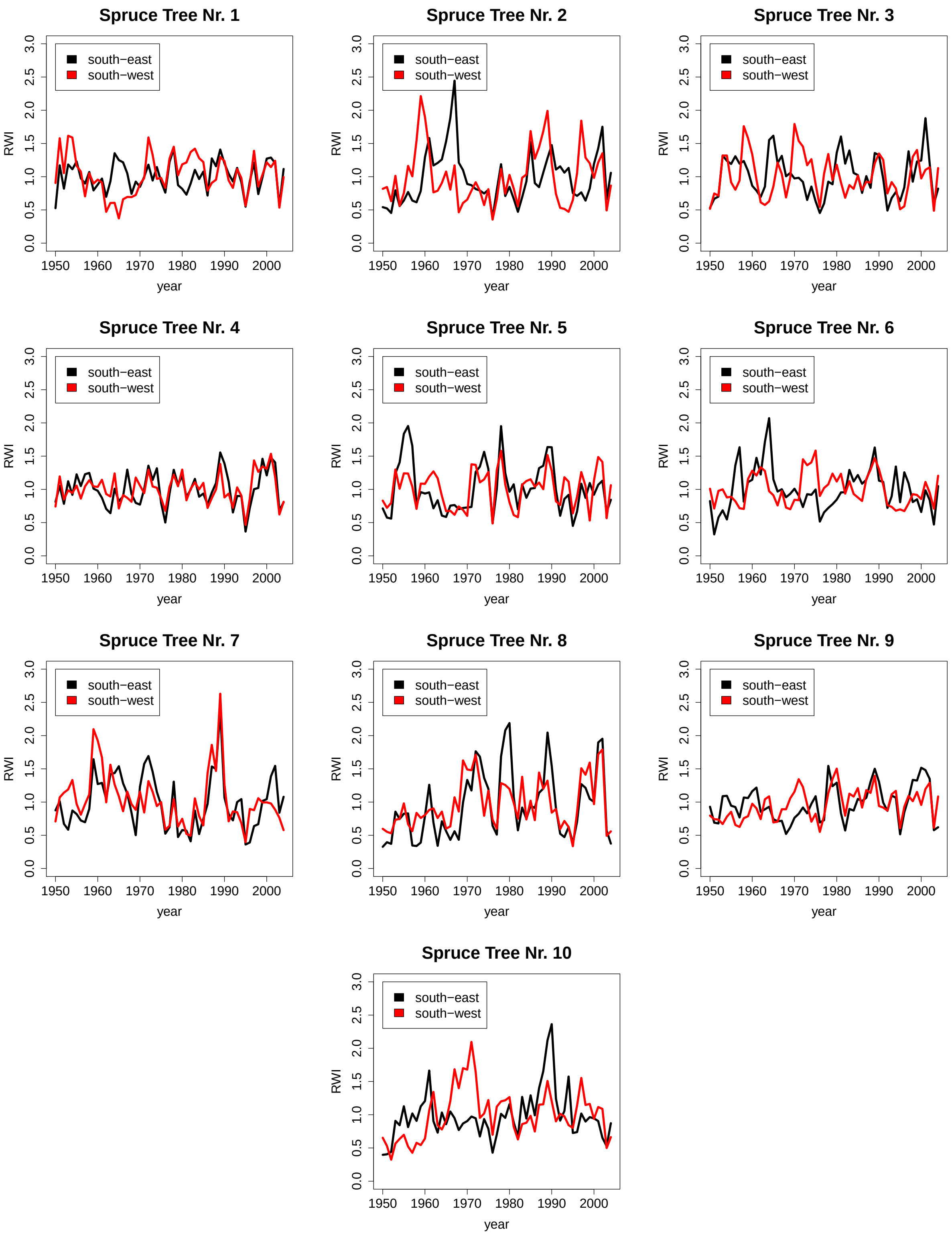


Figure 7.3: Detrended tree ring width series for each tree of the fir tree species.



Spruce Tree Nr. 1

Spruce Tree Nr. 2

Spruce Tree Nr. 3

Spruce Tree Nr. 4

Spruce Tree Nr. 5

Spruce Tree Nr. 6

Spruce Tree Nr. 7

Spruce Tree Nr. 8

Spruce Tree Nr. 9

Spruce Tree Nr. 10

Figure 7.4: Detrended tree ring width series for each tree of the spruce tree species.

Species	Tree number									
	1	2	3	4	5	6	7	8	9	10
Fir	0.32	0.56	0.67	0.47	0.54	0.59	0.54	0.28	0.35	0.53
Spruce	0.32	0.20	0.24	0.49	0.41	0.14	0.45	0.57	0.33	0.19

Table 7.1: Empirical Kendall's taus among the tree ring width series from two drilling directions for a single tree (for both species).

there might be some unusualness in the developments or non-dependencies among both drilling direction series corresponding to a single tree is a matter of interest. Table 7.1 presents the empirical Kendall's taus between the detrended ring series of two drilling directions for each tree of both species. As expected they are all quite strongly dependent among each other although one detects some outliers, e.g. in case of spruce tree No. 7 with a small dependence ($\widehat{\tau} = 0.14$) among the detrended tree ring series measured from drilling direction south-east and south-west respectively. Generally, the detrended series from both drilling directions seem less dependent among each other in case of spruce trees as in case of fir trees. However, the corresponding Figures 7.3 and 7.4 detect no further peculiarities among the series of each tree. The series are all lying in the range between 0 and 2.5 and the variances of the data seem to differ little from tree to tree. It underlines our approach of modeling the data by linear mixed models with random effects per tree. In case of all regarded fir trees (except fir tree No. 5), the corresponding detrended tree ring series seem to decrease between 1960 and 1970 and increase again from 1970. Actually one would expect that the detrended data would not show such a behavior anymore. But, however, as we have hinted in the Section 6.1, the used smoothing spline interpolation is relatively restricted in case of tree ring series in order to present as much low frequency climatic variance as possible in the data and remove only divergent non-climatic anomalies that could be wrongly interpreted in the time domain as exceptional climate events [Cook and Peters, 1981, p. 9]. Hence we will simply model this observed ”overall” time effect by an additional GAM model in the following before we set up linear mixed models based on our then twice detrended data. Note, for the sake of completeness, the pairwise relationships of the detrended ring width data of both tree species with the calculated seasonal means are illustrated in plots in Appendix C. For some pairs different functional relationships seem to be possible, however, we think modeling linear relationships among year rings and covariables by LMMs constitutes a good starting point in all cases for our analysis.

7.2 Model specification

Our (detrendend) tree ring data set for each tree species can be considered as **clustered longitudinal data**, in which **units of analysis** (drilling directions) are nested within **clusters** (tree numbers), and **repeated measures** are collected on the units of analysis every year. As already introduced in Chapter 6, our data set is structured into three levels: Level 3 represents the clusters of units (tree numbers), Level 2 the units of analysis

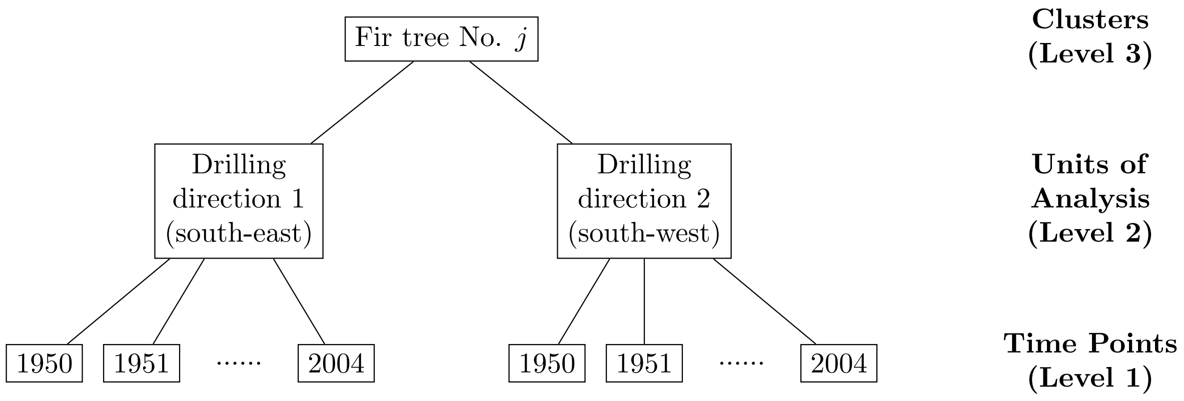

Fir tree No. j
Clusters (Level 3)
Drilling direction 1 (south-east)
Drilling direction 2 (south-west)
Units of Analysis (Level 2)
1950
1951
......
2004
1950
1951
......
2004
Time Points (Level 1)

Figure 7.5: Structure of the clustered longitudinal data for the j-th fir tree in the fir tree ring data set.

(drilling directions), and Level 1 represents the longitudinal (repeated) measures (tree ring widths) made over time. This level structure is illustrated in Figure 7.5 for the example of the j-th fir tree.

Altogether, we now have the following variables included in both tree ring data sets for each tree species, namely

- **Tree number (Level 3) variable**
  (i) **treeno.** $\widehat{=}$ Number of the fir/spruce tree ($j = 1, ..., 10$),
- **Drilling direction (Level 2) variable**
  (i) **drilldirection** $\widehat{=}$ drilling direction ($i = 1$ for south-east and $i = 2$ for south-west),
- **Time-varying (Level 1) variables**
  (i) **time** $\widehat{=}$ Time points of longitudinal measures (years $t = 1950, ..., 2004$),
  (ii) **treering** $\widehat{=}$ Detrended fir/spruce tree ring width (RWI), collected at each time point (dependent variable),
  (iii) **yearlyprec/winterprec/summerprec/summerprec_1** $\widehat{=}$ Yearly/winter/summer/summer one year before - mean of daily total precipitation, collected at each time point,
  (iv) **yearlyhum/winterhum/summerhum/summerhum_1** $\widehat{=}$ Yearly/winter/summer/summer one year before - mean of daily rel humidity, collected at each time point,

(v) **yearlypress/winterpress/summerpress/summerpress_1** $\widehat{=}$ Yearly/winter/summer/summer one year before - mean of daily air pressure, collected at each time point,

(vi) **yearlytemp/wintertemp/summertemp/summertemp_1** $\widehat{=}$ Yearly/winter/summer/summer one year before - mean of daily mean air temperature, collected at each time point,

(vii) **yeardry/winterdry/summerdry/summerdry_1** $\widehat{=}$ Longest yearly/winter/summer/summer one year before - dry period, collected at each time point.

The corresponding periods of the yearly/winter/summer/summer one year before means are described in (7.2) - (7.5).
Since the considered covariates are on very different scales, we are again using standardized variables here as in the marginal models in Chapter 3. So we indicate their relative influence on the response variable with our models. In detail, the standardization looks as follows:

$$\widetilde{treering}_{tij} = \frac{treering_{tij} - \overline{treering_{ij}}}{s_{treering_{ij}}}, \tag{7.6}$$

where

$$\begin{aligned} \overline{treering_{ij}} &= \frac{1}{55}\sum\nolimits_{t=1950}^{2004} (treering_{tij}) \text{ and} \\ s_{treering_{ij}} &= \sqrt{\frac{1}{54}\sum\nolimits_{t=1950}^{2004} \left(treering_{tij} - \overline{treering_{ij}}\right)^2}, \end{aligned}$$

for $t = 1950, ..., 2004$ corresponding to the years, winddirection i $(i = 1, 2)$ nested within fir tree j $(j = 1, ..., 10)$.

We do the same standardization for the different (seasonal) means, in example for the summer mean of daily precipitation:

$$\widetilde{summerprec}_t = \frac{summerprec_t - \overline{summerprec}}{s_{summerprec}}, \tag{7.7}$$

where

$$\begin{aligned} \overline{summerprec} &= \frac{1}{55}\sum\nolimits_{t=1950}^{2004} (summerprec_t) \text{ and} \\ s_{summerprec} &= \sqrt{\frac{1}{54}\sum\nolimits_{t=1950}^{2004} (summerprec_t - \overline{summerprec})^2}, \end{aligned}$$

for $t = 1950, ..., 2004$.

Attention: In case of spruce tree ring data, we use log-data to yield a better fit of our model, i.e.

$$\log(\widetilde{treering}_{tij}) = \frac{\log(treering_{tij}) - \overline{\log(treering_{ij})}}{s_{\log(treering_{ij})}}, \tag{7.8}$$

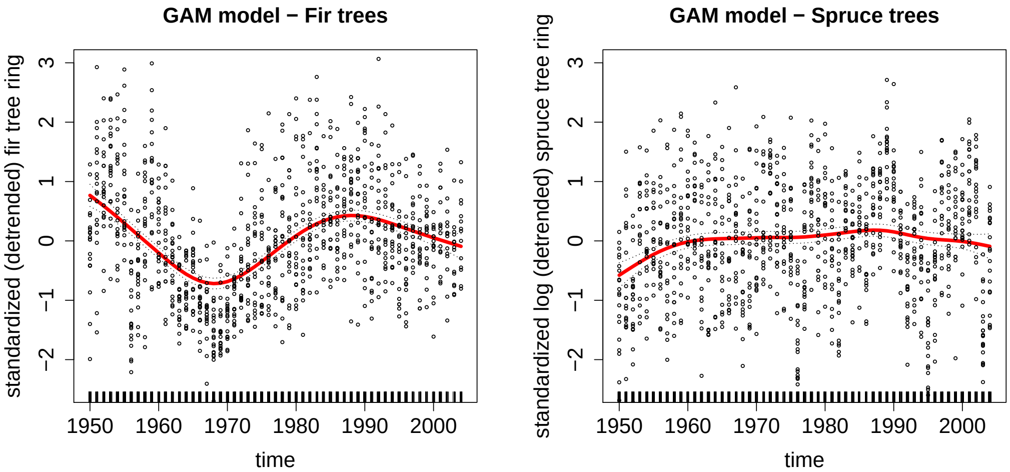


Figure 7.6: Modeling an overall time effect by GAMs for both tree species. Both smooth functions and thus time effects (red lines) are significant.

where

$$\begin{aligned}\overline{\log(treering_{ij})} &= \frac{1}{55}\sum\nolimits_{t=1950}^{2004}(\log(treering_{tij}))\ \text{ and}\\ s_{\log(treering_{ij})} &= \sqrt{\frac{1}{54}\sum\nolimits_{t=1950}^{2004}\left(\log(treering_{tij})-\overline{\log(treering_{ij})}\right)^2},\end{aligned}$$

for $t = 1950, ..., 2004$ corresponding to the years, winddirection i $(i = 1, 2)$ nested within fir tree j $(j = 1, ..., 10)$.

7.2.1 Modeling overall time effect by GAMs

In the data analysis in Section 7.1 we detect an obvious decreasing trend between 1960 and 1970 in actually all measured detrended fir tree widths. We used this detection as reason to model an additional *overall time effect* that still occurs in all ring series of both species although we already detrended the data series of every single tree by smoothing splines beforehand. Therefore we are using two *generalized additive models (GAMs)* (cp. Section 6.2) to explore the functional form of the relationship between covariate *time* and the (already detrended and then standardized) tree ring widths for both tree species.[1] Thus, we model

$$\widetilde{treering}_{tij} = s(time_t) + \widetilde{r}_{tij}, \tag{7.9}$$

where $s(\cdot)$ is the smooth function, chosen to be a cubic spline. Remember, in case of spruce tree ring data, we consider a modified model of (7.9) by log-data instead, i.e. $\log(\widetilde{treering}_{tij}) = s(time_t) + \widetilde{r}_{tij}$. The remainder $\widetilde{r}_{tij}$, i.e. the twice detrended tree ring data, is then modeled by a linear mixed model in both cases of the tree species. Figure

[1] We are here using the R library `gam`, implemented by Hastie [2011].

7.6 presents the resulted smooth functions of our models (red lines). The nonparametric effects (thus the overall time effect) are highly significant in both cases of our regarded tree species (the corresponding p-values are $\ll 0.01$). In case of fir trees, the detection of decreasing ring widths during the sixties and then again increasing widths in the seventies is captured by our model. But also in case of spruce trees there is a slight overall time effect detactable (since we have a small p-value) which is also modeled here.

7.2.2 General linear mixed models form

We now take the twice detrended and standardized tree ring data for both tree species, i.e. $\widetilde{r}_{tij}$, measured at time point t ($t = 1950, ..., 2004$, corresponding to the years), from drilling direction direction i ($i = 1, 2$) nested within tree number j ($j = 1, ..., 10$). Together with the above mentioned covariates, we specify the general form of a full linear mixed model from which we start our model selection for both tree species in a next step. This full model looks as follows:

$$
\begin{aligned}
\widetilde{r}_{tij} \quad & = \beta_0 + \beta_1 \times \widetilde{summerprec}_t + \beta_2 \times \widetilde{summerhum}_t + \\
& \beta_3 \times \widetilde{summerpress}_t + \beta_4 \times \widetilde{summertemp}_t + \beta_5 \times \widetilde{summerdry}_t + \\
& \beta_6 \times \widetilde{summerprec\_1}_t + \beta_7 \times \widetilde{summerhum\_1}_t + \beta_8 \times \widetilde{summerpress\_1}_t + \\
& \beta_9 \times \widetilde{summertemp\_1}_t + \beta_{10} \times \widetilde{summerdry\_1}_t + \beta_{11} \times \widetilde{winterprec}_t + \\
& \beta_{12} \times \widetilde{winterhum}_t + \beta_{13} \times \widetilde{winterpress}_t + \beta_{14} \times \widetilde{wintertemp}_t + \\
& \beta_{15} \times \widetilde{winterdry}_t + \beta_{16} \times \widetilde{yearlyprec}_t + \beta_{17} \times \widetilde{yearlyhum}_t + \\
& \underbrace{\beta_{18} \times \widetilde{yearlypress}_t + \beta_{19} \times \widetilde{yearlytemp}_t + \beta_{20} \times \widetilde{yeardry}_t +}_{\text{fixed}} \\
& \underbrace{u_{0j} + u_{1j} \times s(time_t) + u_{0i|j} + \epsilon_{tij}}_{\text{random}},
\end{aligned}
\tag{7.10}
$$

where

- β_0,...,β_{20} are the fixed effects associated with the intercept and the time-level standardized covariates (as defined before),
- u_{0j} and u_{1j} denote the random tree effects associated with the intercept and time slope,
- $u_{0i|j}$ represents an additional random effect associated with the drilling direction nested within a tree and
- ϵ_{tij} corresponds to the residual.

As in Section 6.3 of Chapter 6 described, we assume that the random effects u_{0j} and u_{1j} are joined normal distributed, i.e.

$$
\boldsymbol{u}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim \mathcal{N}_2(\boldsymbol{0}, D^{(2)}),
\tag{7.11}
$$

where the variance-covariance matrix $D^{(2)}$ is defined as

$$D^{(2)} := \begin{pmatrix} Var(u_{0j}) & Cov(u_{0j}, u_{1j}) \\ Cov(u_{0j}, u_{1j}) & Var(u_{1j}) \end{pmatrix} = \begin{pmatrix} \sigma^2_{int:tree} & \sigma_{int,time:tree} \\ \sigma_{int,time:tree} & \sigma^2_{time:tree} \end{pmatrix}, \tag{7.12}$$

for all $j = 1, ..., 10$. Thus one has to estimate the three covariance parameters $\sigma^2_{int:tree}$, $\sigma_{int,time:tree}$ and $\sigma^2_{time:tree}$ for the two random tree effects associated with intercept and time slope.
In addition, the third random effect associated with drilling direction i nested within tree j, i.e. $u_{0i|j}$, is also assumed to be normal distributed and independent of the other both random effects. Thus, it holds

$$u_{0i|j} \sim \mathcal{N}(0, D^{(1)}), \quad \text{for all } i \text{ and } j, \tag{7.13}$$

where

$$D^{(1)} = \sigma^2_{int:drilldirection(tree)},$$

which has to be estimated, if it is contained in the model.
Further in our case, we assume that the residuals are all i.i.d. normal distributed, i.e.

$$\epsilon_{tij} \sim \mathcal{N}(0, \sigma^2_{residual}), \quad \text{i.i.d. for all } t, i \text{ and } j. \tag{7.14}$$

Thus, only the parameter $\sigma^2_{residual}$ has to be estimated in this context.

Note that our general model from (7.10) does not contain an additional fixed time effect (i.e. the covariate *time*) anymore since we already detrended our series beforehand. We also checked that an additional covariate *time* would not improve the model fit of model (7.10) significantly in both tree species cases.

Now we start selecting which random and fixed effects need to remain in the model and which can be excluded to yield the best meaningful model fit.

7.3 Selecting random effects

In order to select which random effect might be omitted in our models, we are using likelihood ratio tests, described in Section 6.3.3. The corresponding test statistic LR is assumed to be chi-squared distributed as defined in (6.20). Note that in case of testing the random effects, we consider REML estimations for our regarded models with the corresponding restricted log-likelihood values (also for the LR calculation) since their estimates of the random effect parameters are unbiased in contrast to maximum likelihood estimates. The calculation of the p-values is described in Equation (6.21). For example, when we would like to test whether one of both random effects u_{0j} and u_{1j} can be omitted, we estimate the values of the restricted log-likelihoods for the reference and nested model respectively and then calculate the test statistic LR. In a next step, use Equation (6.21) to yield the p-value of our likelihood ratio test, i.e. $p-value = 0.5 \times P(\chi^2_1 > LR) + 0.5 \times P(\chi^2_2 > LR)$. Thereby we can neglect that there exists also a third random effect $u_{0i|j}$ since it is assumed to be independent of the effects u_{0j} and u_{1j}. The results are presented

Nullhypothesis	Fir trees			Spruce trees		
	LR (REML)	p-value	reject H_0	LR (REML)	p-value	reject H_0
H_0 : Random effect u_{1j} can be omitted ($vs.$ H_1 : It cannot.)	32.30 (ref: -1317.60 nest: -1333.75)	0.00	$\surd$	40.64 (ref: -1427.41 nest: -1447.73)	0.00	$\surd$
H_0 : Random effect $u_{0i\|j}$ can be omitted ($vs.$ H_1 : It cannot.)	0.00 (ref: -1317.60 nest: -1317.60)	0.50	×	0.00 (ref: -1427.41 nest: -1427.41)	0.50	×
H_0 : Random effect u_{1j} can be omitted in model without random effect $u_{0i\|j}$ ($vs.$ H_1 : It cannot.)	32.30 (ref: -1317.60 nest: -1333.75)	0.00	$\surd$	40.64 (ref: -1427.41 nest: -1447.73)	0.00	$\surd$
H_0 : All random effects can be omitted ($vs.$ H_1 : They cannot.)	32.30 (ref: -1317.60 nest: -1333.75)	0.00	$\surd$	40.64 (ref: -1427.41 nest: -1447.73)	0.00	$\surd$

Table 7.2: Selecting significant random effects based on likelihood ratio tests (cp. Section 6.3.3 in Chapter 6). The test statistic LR (due to REML estimations) is defined in (6.20). The corresponding (restricted) log-likelihood values can also be found in the table for the considered reference (ref:) and nested (nest:) model due to the different test scenarios. The p-values are calculated as shown in Equation (6.21).

in Table 7.2.

In both tree species models we can omit the random effect associated with the drilling direction nested within a tree $u_{0i|j}$ due to our likelihood ratio tests. The other both effects seem to be significant and thus will remain in our models. Therefore from now on we consider the models

$$\widetilde{r}_{tij} = fixed_t + u_{0j} + u_{1j} \times s(time_t) + \epsilon_{tij}, \tag{7.15}$$

for both tree species (in which we study (standardized) log data in case of spruce trees) and continue with selecting the significant fixed effects in the next section.

7.4 Selecting fixed effects

In contrast to the random effects, we estimate the fixed effect parameters of the full model (7.10) by the method of maximum likelihood (ML). We are using these unbiased estimates to calculate the corresponding AIC as defined in Equation (6.23) and compare its value to the AIC when we have removed one fixed parameter out of the full model. This procedure is repeated for every fixed parameter and in the end one selects that model with the smallest AIC. We then take this selected model and start again the described procedure of removing every single fixed parameter respectively that is contained in that model, estimate the rest of the fixed parameters by ML and calculate the corresponding AICs. Then select again that resulted model with the smallest AIC. This backward model selection by AIC ends when an AIC value cannot be undermatched anymore.

Using the above described selection by AIC to yield our end model. In case of **fir** trees we get the following:

$$\begin{aligned}\widetilde{r}^{fir}_{tij} \quad &= \beta_0 + \beta_1 \times \widetilde{summerprec}_t + \beta_2 \times \widetilde{summertemp}_t + \beta_3 \times \widetilde{summerprec\_1}_t + \\ &\beta_4 \times \widetilde{summerpress\_1}_t + \beta_5 \times \widetilde{summertemp\_1}_t + \beta_6 \times \widetilde{summerdry\_1}_t + \\ &\beta_7 \times \widetilde{winterprec}_t + \beta_8 \times \widetilde{winterhum}_t + \beta_9 \times \widetilde{yearlyprec}_t + \\ &\underbrace{\beta_{10} \times \widetilde{yearlypress}_t + \beta_{11} \times \widetilde{yearlytemp}_t + \beta_{12} \times \widetilde{yeardry}_t +}_{\text{fixed}} \\ &\underbrace{u_{0j} + u_{1j} \times s(time_t) + \epsilon_{tij}}_{\text{random}} .\end{aligned} \tag{7.16}$$

The following model results in case of the (log) **spruce** tree data:

$$\begin{aligned}\widetilde{r}^{spruce}_{tij} \quad &= \beta_0 + \beta_1 \times \widetilde{summerprec}_t + \beta_2 \times \widetilde{summerhum}_t + \beta_3 \times \widetilde{summertemp}_t + \\ &\beta_4 \times \widetilde{summerdry}_t + \beta_5 \times \widetilde{summertemp\_1}_t + \beta_6 \times \widetilde{summerdry\_1}_t + \\ &\beta_7 \times \widetilde{winterprec}_t + \beta_8 \times \widetilde{winterhum}_t + \\ &\beta_9 \times \widetilde{winterpress}_t + \beta_{10} \times \widetilde{winterdry}_t + \beta_{11} \times \widetilde{yearlyprec}_t + \\ &\underbrace{\beta_{12} \times \widetilde{yearlyhum}_t + \beta_{13} \times \widetilde{yearlypress}_t + \beta_{14} \times \widetilde{yeartemp}_t +}_{\text{fixed}} \\ &\underbrace{u_{0j} + u_{1j} \times s(time_t) + \epsilon_{tij}}_{\text{random}}\end{aligned} \tag{7.17}$$

From 20 fixed effect parameters in the full model we end with 12 and 14 parameters respectively at this point. A comparison between both tree species, the values and single significances of the fixed effects as well as the estimated parameter values in case of the random effects will be now presented in the next Chapter.

Chapter 8

Results

The estimates of all parameters of the selected models (7.16) and (7.17) together with the p-values of corresponding t-tests (defined in Section 6.3.3) for the fixed parameters are presented in Table 8.1 for both tree species. Due to Section 6.3, the fixed effect parameters (i.e. quasi β_0,...,β_{14}) are estimated by the method of maximum likelihood (ML) while the random effect parameters (i.e. $\sigma_{residual}$, $\sigma_{int:tree}$, $\sigma_{time:tree}$ and $\sigma_{int,time:tree}$) are estimated by the restricted maximum likelihood (REML). Note that in Table 8.1 an estimate of the correlation $\rho_{int,time:tree}$ is presented instead of covariance $\sigma_{int,time:tree}$. Table 8.2 summarizes the relationship of the (standardized) covariates whether they are positively or negatively influencing our (detrended and standardized) tree ring responses.

8.1 Fixed effects

In case of both tree species, the intercept parameters β_0 do not seem to be significant and thus could be excluded from the models in principle. Generally we detect slightly more negative influences of the (standardized) covariates on the corresponding responses than positive ones. In case of the selected spruce tree model we have two more fixed effects included and the response variability is described by more winter and summer meteorological quantities as in the fir tree case. But in return, the fir tree ring widths measured in a specific year seem to be more influenced by meteorological occurrences in summer one year before in contrast to spruce rings. While we have a fixed effect by the longest yearly dry period ($\widetilde{yeardry}$) included in our fir tree model, we detect effects of the longest summer and winter dry period respectively instead, implemented in the spruce tree model which are not modeled in case of fir trees. Further, spruce tree rings also seem to be more influenced by humidity measurements as in the fir trees case.

An interesting finding is given by the positive influence of the yearly precipitation and temperature mean while the corresponding summer means are modeled to have a negative relationship to the reponse of both tree species. The common influence of summer and yearly mean of temperature on the ring width is illustrated in Figure 8.1. One observes that we have modeled that the warmer the summer (and hence the the warmer the summer mean of temperature) of a specific year the smaller the corresponding tree ring width will be (provided that all other covariates stay constant).

Parameter of	Fir trees		Spruce trees	
	Estimate	p-value of t-test	Estimate	p-value of t-test
Intercept (β_0)	0.00	0.90	0.00	1.00
$\widetilde{summerprec}_t$	-0.16	< 0.001	-0.17	< 0.001
$\widetilde{summerhum}_t$	/	/	0.12	0.08
$\widetilde{summerpress}_t$	/	/	/	/
$\widetilde{summertemp}_t$	-0.16	< 0.001	-0.24	< 0.001
$\widetilde{summerdry}_t$	/	/	-0.27	< 0.001
$\widetilde{summerprec\_1}_t$	-0.09	0.01	/	/
$\widetilde{summerhum\_1}_t$	/	/	/	/
$\widetilde{summerpress\_1}_t$	-0.12	< 0.001	/	/
$\widetilde{summertemp\_1}_t$	0.06	0.05	-0.12	< 0.001
$\widetilde{summerdry\_1}_t$	-0.09	< 0.001	0.06	0.05
$\widetilde{winterprec}_t$	0.06	0.07	-0.13	< 0.001
$\widetilde{winterhum}_t$	-0.09	< 0.001	-0.17	< 0.001
$\widetilde{winterpress}_t$	/	/	-0.11	0.01
$\widetilde{wintertemp}_t$	/	/	/	/
$\widetilde{winterdry}_t$	/	/	0.09	0.01
$\widetilde{yearlyprec}_t$	0.08	0.06	0.23	< 0.001
$\widetilde{yearlyhum}_t$	/	/	-0.24	< 0.001
$\widetilde{yearlypress}_t$	-0.09	0.01	0.17	< 0.001
$\widetilde{yearlytemp}_t$	0.21	< 0.001	0.11	0.03
$\widetilde{yeardry}_t$	0.18	< 0.001	/	/
$\sigma_{residual}$	0.79		0.84	
$\sigma_{int:tree}$	0.00		0.00	
$\sigma_{time:tree}$	0.44		1.39	
$\rho_{int,time:tree}$	-0.22		0.00	
AIC (ML)	2570.33		2795.88	
AIC (REML)	2638.08		2869.73	
BIC (ML)	2654.83		2890.94	
BIC (REML)	2722.37		2964.53	
log-lik (ML)	-1268.16		-1378.94	
log-lik (REML)	-1302.04		-1415.87	

Table 8.1: Estimations of the fixed parameters of the covariates (done by maximum likelihood (ML)) together with p-values of the corresponding t-tests as well as the estimations of the random effects (done by restricted maximum likelihood (REML)) for both tree species models. Note that $\rho_{int,time:tree}$ denotes the corresponding correlation. In addition, the AIC, BIC and log-likelihood values of the different methods are also listed.

Variable	Summer mean		Summer mean one year before		Winter mean		Yearly mean	
	Fir	Spruce	Fir	Spruce	Fir	Spruce	Fir	Spruce
Precipitation	−	−	−		+	−	+	+
Humidity		+			−	−		−
Air pressure			−			−	−	+
Temperature	−	−	+	−			+	+
Longest dry period		−	−	+		+	+	

Table 8.2: Summary for both tree species cases whether a seasonal mean has modeled positive or negative influence on the response variable.

Now often a warmer summer induces also a higher yearly mean of temperature, so the tree ring width will not be as small as when we would have only a high temperature mean in summer but a small overall yearly temperature mean in the same year (provided that all other covariates stay constant). However, the year ring width would still be higher (provided that all other covariates stay constant) when we have a high yearly mean of temperature and a lower yearly summer mean of temperature in the same year (thus the seasons except summer must provide warm temperatures). Note that in case of fir trees the yearly mean of temperature has a higher fraction (0.21) on the ring width than in case of spruce trees (0.11) while at the same time the mean of temperature in summer has a lower fraction (-0.16) on fir tree rings than it has on the ones of spruce trees (-0.24). Thus the plane in Figure 8.1 is less steep in case of fir trees than in case of spruce trees. This result corresponds to several studies made so far, like e.g. by Dittmar and Elling [1999], where one observed generally a positive influence of temperature on the year ring width

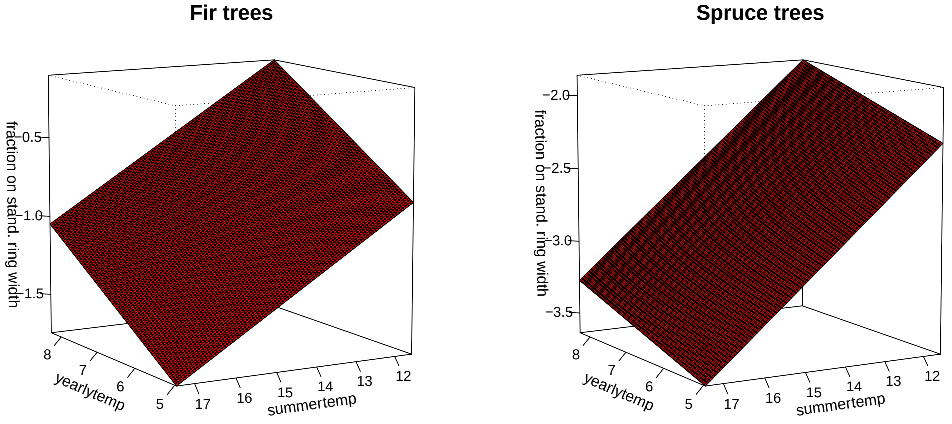


Figure 8.1: Modeled common influence of the yearly and summer temperature mean on the (standardized) tree ring width for both species.

of spruce trees (but it can easily transfer also to fir trees) but, however in contrast one also detected a negative dependence between the year rings and summer temperatures in the region of the alps. In addition at middle-high altitudes as we have in our case (trees located at about 900m above sea level) there was observed a ”change point” from negative to positive dependence between tree ring widths and temperatures in summer in the region from 700 to 1000 meters above sea level. Thus the dependence among responses and summer temperatures is not explict manifested in that region of altitudes and can differ from position to postion. Note that the above description of the phenomenon we have modeled in case of yearly and summer temperature means can equally transfer to the relationship of yearly and summer mean of total precipitation to the ring width responses in both tree species cases.

8.2 Random effects

The assumption of independent and identically normal distributed residuals is underlined in both tree species cases by looking at the diagnostic plots in Figure 8.2. The raw residuals plotted against the observation numbers do not present any systematic pattern and

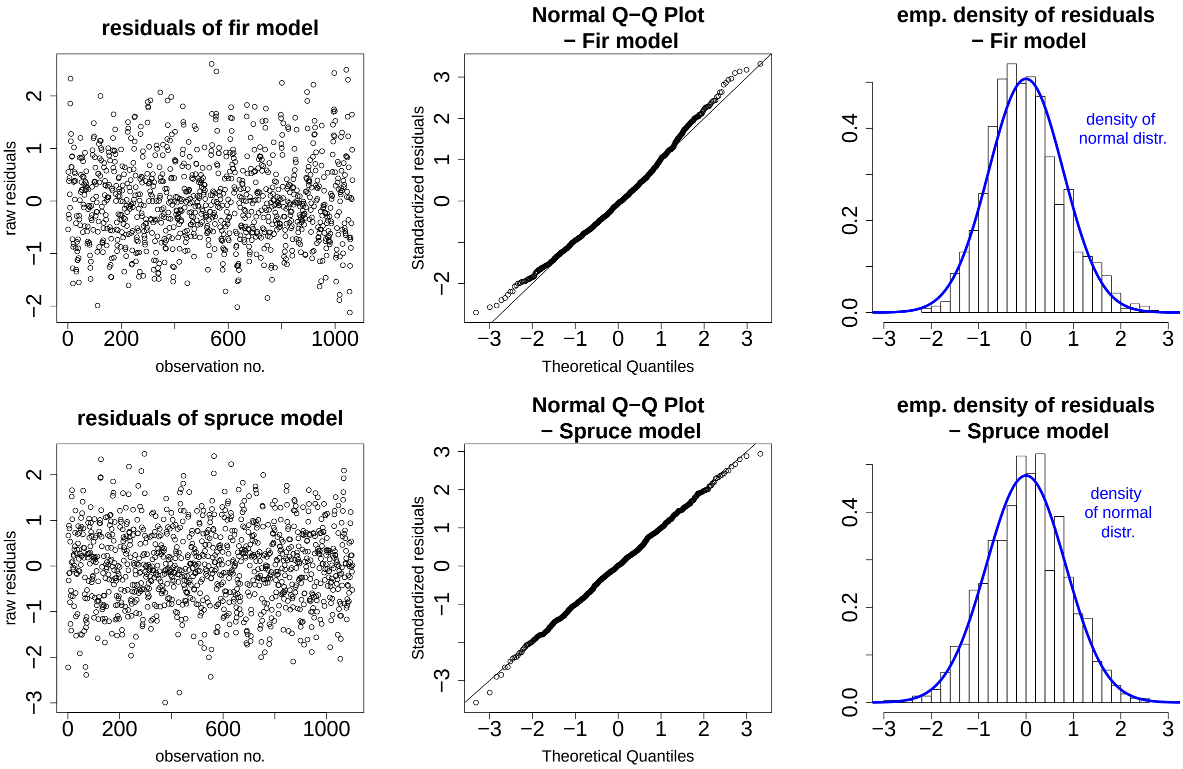


Figure 8.2: From both tree species models we plot each the corresponding residuals against the observation numbers (left panel), a normal Q-Q plot of the residuals (middle panel) and a comparison between the empirical density of the residuals and the theoretical normal distribution (right panel).

the Q-Q plots follow a straight line. Thus it assesses the goodness of our fits in a positive way.

The special feature of linear mixed models is the flexible way of implementing random effects. In our cases the standard deviations of the random effect associated with the time slope a tree number are relatively high (0.44 and 1.39), in case of spruce trees even explicit higher than in case of fir trees. In contrast, the standard deviations of the random effects associated with the intercept for every tree are very small (i.e. < 0.001 for both tree species) but we still modeled a negative correlation among both random effects in case of fir trees (-0.22). In case of spruce trees, this correlation corresponds to zero. Additionally, the estimates of the residual's standard deviations in both tree species cases are similar, namely 0.79 and 0.84.

With these results we can now calculate the modeled tree specific variances for all considered years t. Due to the random part in the selected models (7.16) and (7.17) we get the variance of our detrended tree ring widths for both tree species at time t by

$$\begin{aligned} Var(\widetilde{r}_{tij}) &= Var(u_{0j}) + s(time_t)^2 Var(u_{1j}) + Var(\epsilon_{tij}) + 2s(time_t)Cov(u_{0j}, u_{1j}) \\ &= \sigma^2_{int:tree} + s(time_t)^2\sigma^2_{time:tree} + \sigma^2_{residual} + 2s(time_t)\sigma_{int,time:tree}, \end{aligned} \tag{8.1}$$

for all $i = 1, 2$ and $j = 1, ..., 10$. The modeled variance over the 55 years is illustrated in Figure 8.3 in comparison with the empirical ones from the detrended data of both tree species. Generally the variances of detrended spruce year rings seem to be higher than in the fir tree case. Nevertheless we detect a significant decrease in modeled and empirical variances of both species, observed by simple linear regression of the data against time. The results indicate a decrease of modeled variance by -0.001 p.a. (-0.05 in 55 years) in the fir case and a steeper decrease by -0.003 p.a. (-0.18 in 55 years) in the case of spruce.

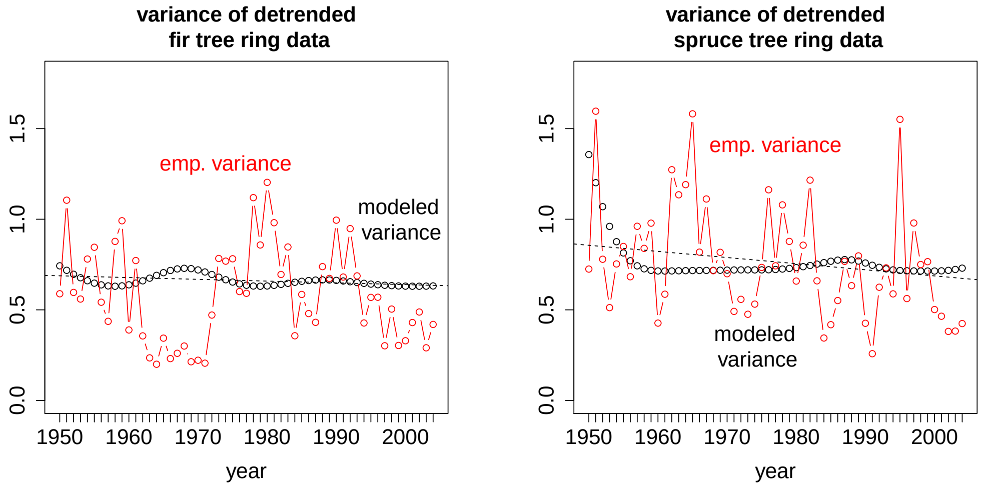


Figure 8.3: Modeled variances of the detrended tree ring widths data for the fir trees (left panel) and spruce trees (right panel) in comparison with the empirical ones of each year. The dashed lines correspond to significantly decreasing simple linear regressions of the detrended ring data against time.

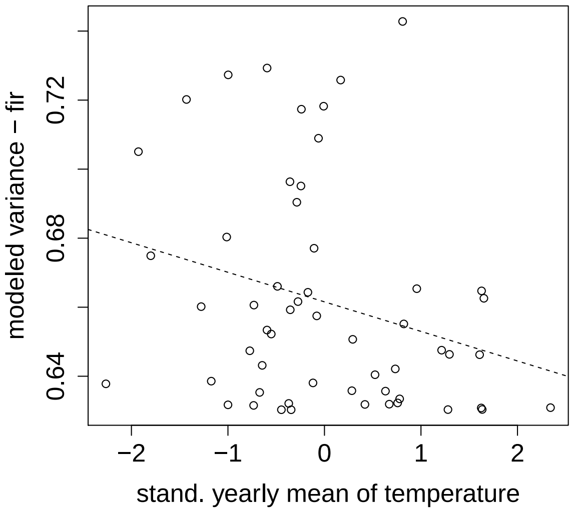


Figure 8.4: Modeled variances of detrended fir tree ring data plotted against the corresponding standardized yearly means of temperature. The dashed line corresponds to a simple linear regression of the modeled variances against the yearly temperature means.

It is consistent with the general studies of tree ring data since year rings are naturally constrained to the tree breadth and thus its variances are also constrained.
Using these outcomings, one could now compare the influences of our covariables contained in the models on the modeled variance. However we cannot detect any strong significant relationships here when we study corresponding plots in both tree species cases. The only observation that we made is a slight significant decrease of modeled variances (measured by a simple linear regression) when the yearly mean of temperature gets higher in case of fir trees. It is presented in Figure 8.4. Together with our model outcome that higher yearly means of temperature induce rather larger tree rings, the significant decrease of the corresponding fir (detrended) tree ring variance exhibit that these large widths will then not vary as much.

As last point we want to remark that we could also calculate the variance of the raw tree ring data y_{tij} for every tree j a drilling direction j and point in time t by using the ratio (7.1). Thus, we get

$$Var(y_{tij}) = s_{tij}^2 Var(treering_{tij}) = s_{tij}^2 Var(\widetilde{r}_{tij}). \qquad (8.2)$$

But in that case it would not make sense to examine the relationship of the raw data variances with our yearly covariates since we would not know exactly which variance developments are caused by climate or biological occurrences and which are caused by non-climate reasons such as time or age trends.

Part III

Summary and Outlook

Chapter 9

Summary and Outlook - Part I

In the first part of the thesis we studied R-vine models for three different periods (i.e. 1955-1959, 1980-1984 and 2005-2009) to model the dependence structures among six meteorological variables during these time spans in Hohenpeissenberg. In order to do this, we first had to model the marginal distributional behaviors of the variables using different regression models. Autoregression and seasonal effects were captured by linear regressions in case of daily mean, minimum and maximum temperature as well as in case of daily mean air pressure. While the residuals in the models of daily mean and maximum temperature follow a normal distribution in the periods 1955-1959 and 1980-1984, we detected slightly skewed normal distributed residuals in the models for both variables in the last period 2005-2009. However, daily minimum air temperature and daily mean air pressure were modeled with skew t distributed error terms in all considered periods. We further noticed significant increases in the modeled means of all temperature variables over the whole time period 1950-2009. The modeled expected values of daily temperature variables increase on average by about 1.4°C in 60 years. Additionally, the modeled variances indicate no significant trend. However, in order to classify these results, one has to consider the overall temperature developments over the last and next centuries. In case of modeling the marginal behavior of daily mean humidity we attained better fits by including the "Fön"-explaining variable of wind direction into our beta regression models. Nevertheless, maybe for a future work, the fits could be improved by using a kind of weighted beta regression. The modeling of daily total precipitation, performed by a two step method with binomial and gamma regressions, offers more rain amount on summer days coming from convective precipitations which occur in the foothills region of the Alps. All in all, our marginal models fit relatively well for all six variables.

To connect the non-continuous (at zero) variable of daily total precipitation to our R-vine models, we first modeled the dependence structure of the five variables without precipitation by a five-dimensional R-vine in the classical way and then connected the variable of positive precipitation amount to the established 5-dimensional vine to get a six-dimensional R-vine copula specification on rain days. Thus, the dependence structure among the variables without precipitation should not change when it rains which is underlined by our empirical data.

The modeled R-vines show dependence structures among the variables as generally expected such as strong dependencies among the temperature variables and negative

dependence between humidity and temperature in all considered periods. Air pressure is modeled to be slightly negative dependendent on temperatures and air pressure as well as slightly positive dependendent on humdity. However, our models offer a further feature, namely to investigate also the tail dependencies among the variables. Based on the log-likelihood values of the different models and the selected pair copulas in the R-vines, we detect a considerable change in the dependencies among the temperature variables from the second (1980-1984) to the last period (2005-2009). Our vine for the last period becomes more "Gaussian" in contrast to the previous ones, in detail, we detect a change from pair copulas with asymmetric tail dependence in our first two R-vines (corresponding to periods 1955-1959 and 1980-1984) to ones with symmetric tail dependence among the temperature variables in the last period. In our case it means that large deviations downwards from the modeled expectation of daily mean temperature strongly depend on large downward deviations from the expected daily maximum temperature in the last period. In the previous two periods we modeled instead that large downward deviations from the expected daily mean temperature rather depended on large downward deviations from the expected daily minimum temperature. So this tail dependence becomes more variable in the last period.
Thus, in a next step one should evaluate R-vine models for all periods, i.e. between the considered ones, to get more information how this detected change in the dependencies among the variables developed over the whole time span and to classify how fundamental it is. Maybe therefore one could also use a more dynamic division of the modeled periods based on the points in time when changes in the dependencies occur.

However, in our case, simulated pairwise Kendall's taus from the models correspond to their empirical counterparts as an indicator for the goodness of our fits. In addition we have simulated the probabilites of different scenarios (compared to the empirical ones) which exihibts a nice feature of R-vine models.

This work could motivate to built R-vine models including further meteorological or other variables, maybe also a spatial extension, to implement them into established forecast and weather derivative pricing models and methods which are dependent on information about the dependence structures among climate variables (as we have mentioned in the introduction).

Chapter 10

Summary and Outlook - Part II

In the second part of the thesis we modeled the (linear) relationship between yearly tree ring widths and climate variables by linear mixed models for two tree species (fir and spruce trees) from the region of Hohenpeissenberg. First we had to detrend the raw data series by using smoothing splines in order to remove any non-climate variance over time (in our case 55 years). The detrended data exhibit different variances per tree so that our approach of using linear mixed models to model the year ring data seems reasonable. However, we still detected an overall time trend (overall time effect) in the detrended data of all considered trees, especially in the case of fir trees, which we modeled by an additional generalized additive model (GAM). The residuals were then modeled by linear mixed models including meteorological quantities as fixed effects for both tree species. Note that in case of spruce trees we attained a better model fit by using log transformed data. The meteorological quantities such as mean temperature, mean humidity, mean air pressure, total precipiation and longest dry period were calculated as means over different seasons which are meaningful in matters of tree growth. We then selected appropriate random and fixed effects. Tree specific random effects were included in the model for both species as well as a number of fixed effects. Thereby, interestingly, we detected positive effects of yearly temperature and precipitation means on tree ring widths but negative effects of summer temperature and precipitation means on ring widths in cases of both tree species. However this result is convenient to previous studies where one observed generally a positive influence of temperature on the year ring width, but in contrast one also detected a negative dependence between the year rings and summer temperatures in the region of the Alps. Further, the modeled variances of the detrended data for both tree species decrease significantly over the whole time span. However, we could not observe a clear relationship between these modeled variances and the meteorological covariates. Only for fir trees in case of yearly mean temperature a slightly negative dependence is detectable. Finally, the assumption of i.i.d. normal distributed residuals seems to be fulfilled in the models of both tree species.

Our here presented methodology should serve as a further approach to extract quantitative and qualitative information about the limiting factors of tree growth. Therefore one could evaluate further data sets by LMMs for comparison and/or modify the LMMs by adding further covariates as well as by adding different covariance structures of the residuals into the models to explain the variability of tree ring widths.

Appendix A

Plots of bivariate copula families

Here we present contour and scatter plots of the bivariate copula families for standard normal margins which we have discussed in Section 2.2.3. We are using the relationships shown in Table 2.4 to calculate the copula parameters for different choices of Kendall's τ. Note that only the Gaussian, t, Frank as well as the rotated Clayton, Gumbel and Joe copulas exhibit negative dependence. For the t copula we choose four and eight degrees of freedom respectively for illustration.

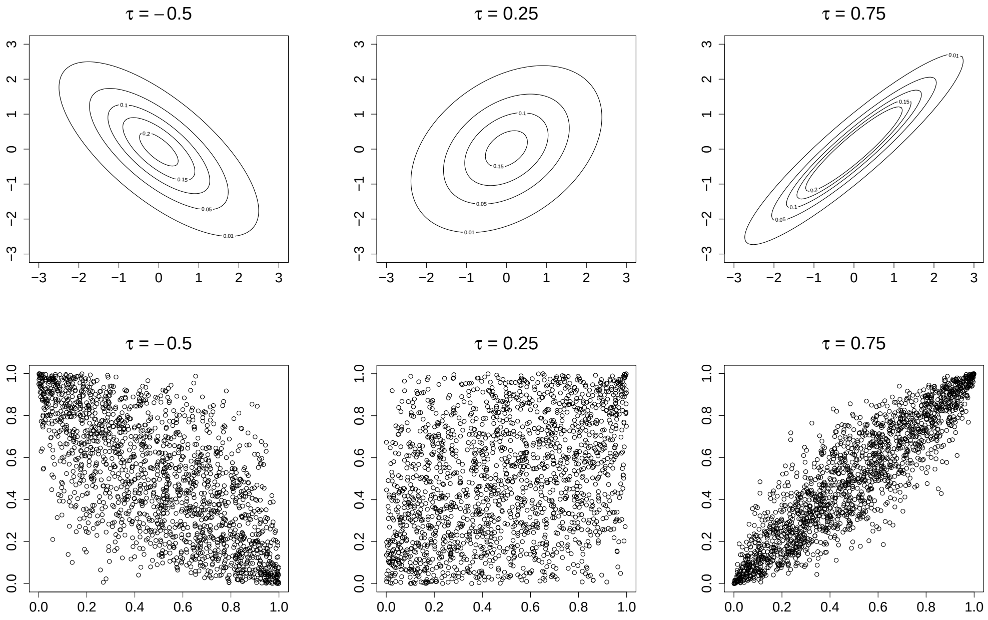


Figure A.1: Contour and scatter plots of the Gaussian copula for three choices of Kendall's τ.

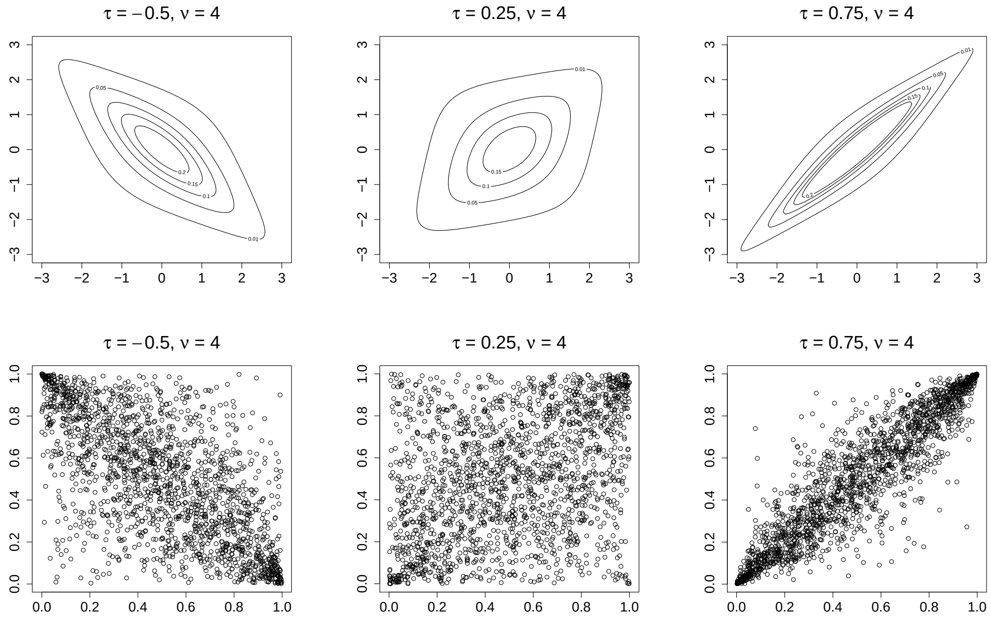


Figure A.2: Contour and scatter plots of the t copula with four degrees of freedom for three choices of Kendall's τ.

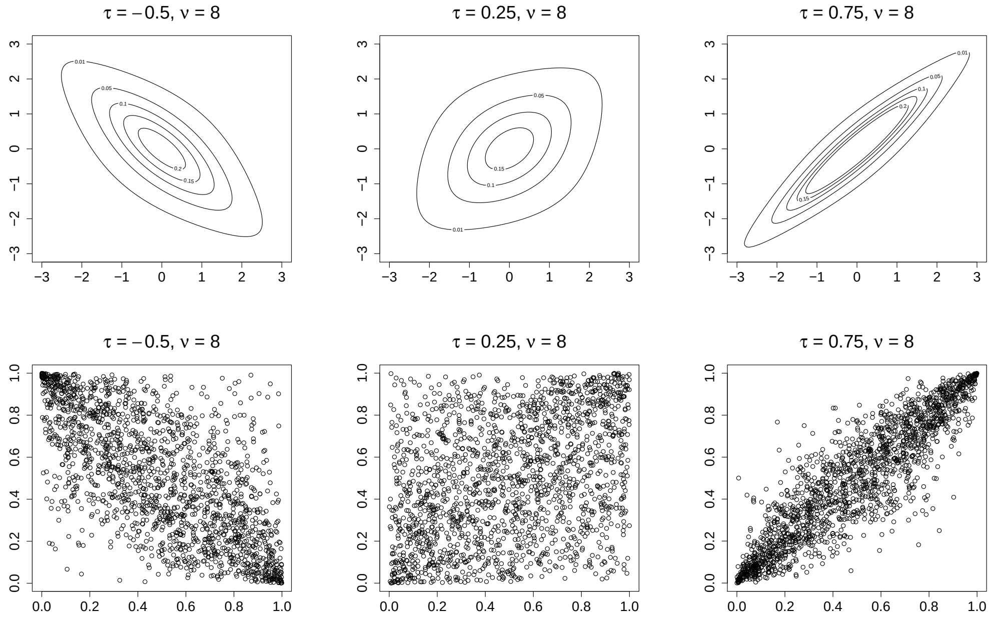


Figure A.3: Contour and scatter plots of the t copula with eight degrees of freedom for three choices of Kendall's τ.

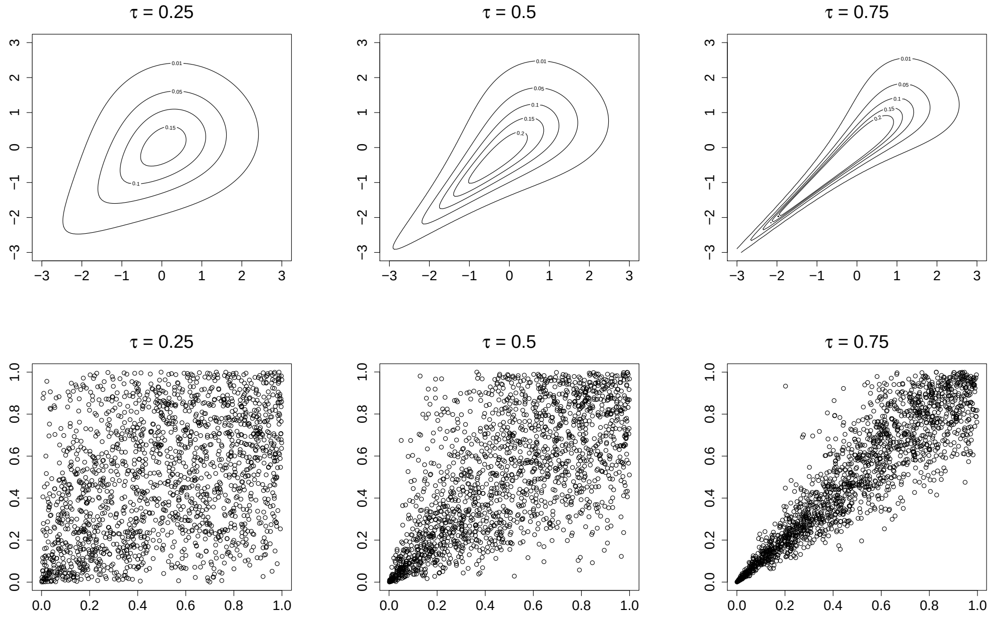


Figure A.4: Contour and scatter plots of the Clayton copula for three choices of Kendall's τ.

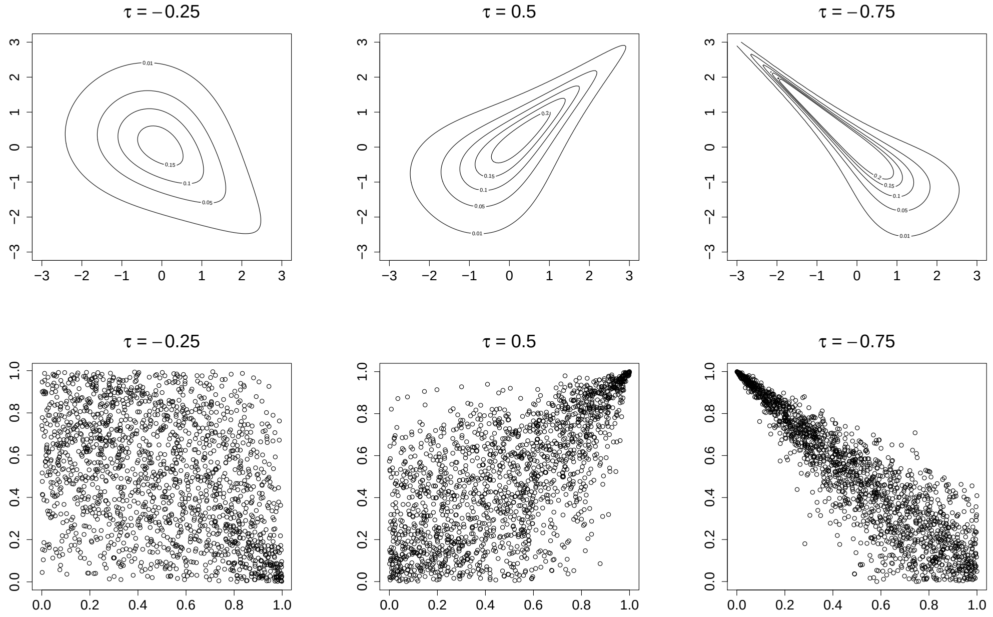


Figure A.5: Contour and scatter plots of rotated Clayton copulas for three choices of Kendall's τ. The left panel corresponds to a rotated Clayton copula by 90 degrees, the middle panel presents a rotated (survival) Clayton copula by 180 degrees and the right panel shows a rotated Clayton copula by 270 degrees.

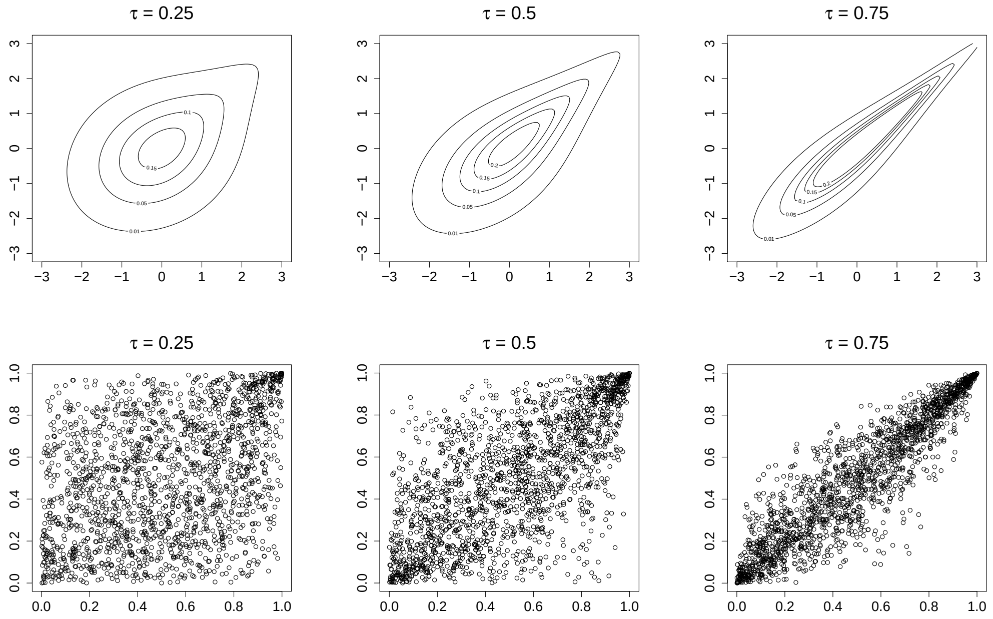


Figure A.6: Contour and scatter plots of the Gumbel copula for three choices of Kendall's τ.

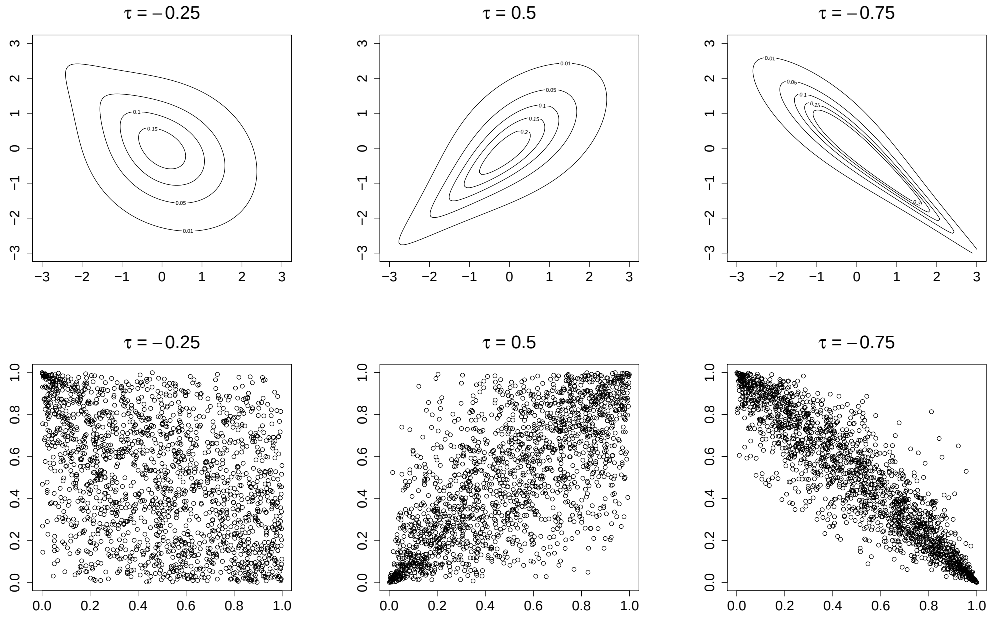


Figure A.7: Contour and scatter plots of rotated Gumbel copulas for three choices of Kendall's τ. The left panel corresponds to a rotated Gumbel copula by 90 degrees, the middle panel presents a rotated (survival) Gumbel copula by 180 degrees and the right panel shows a rotated Gumbel copula by 270 degrees.

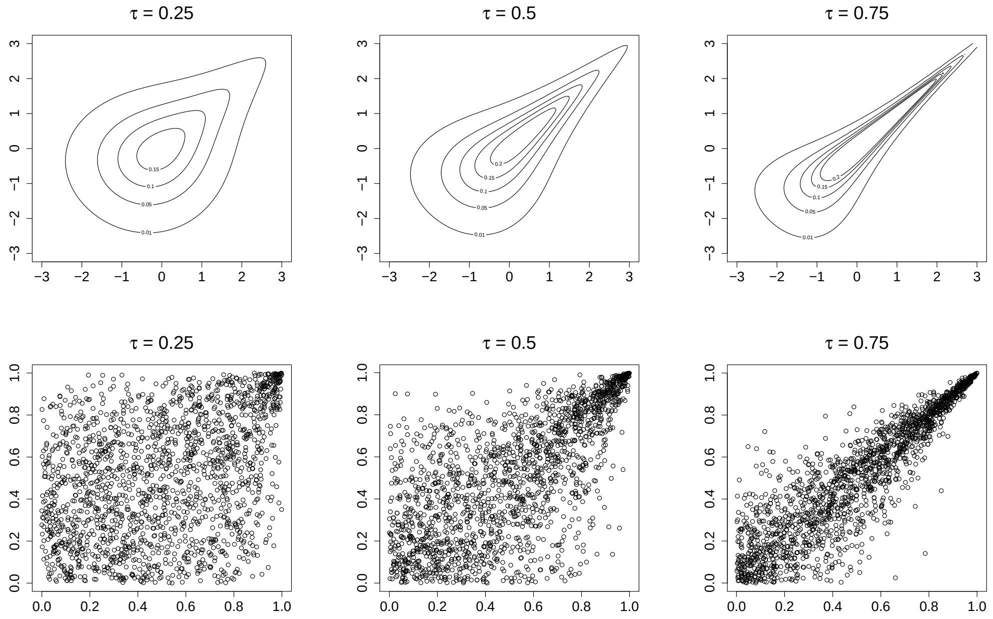


Figure A.8: Contour and scatter plots of the Joe copula for three choices of Kendall's τ.

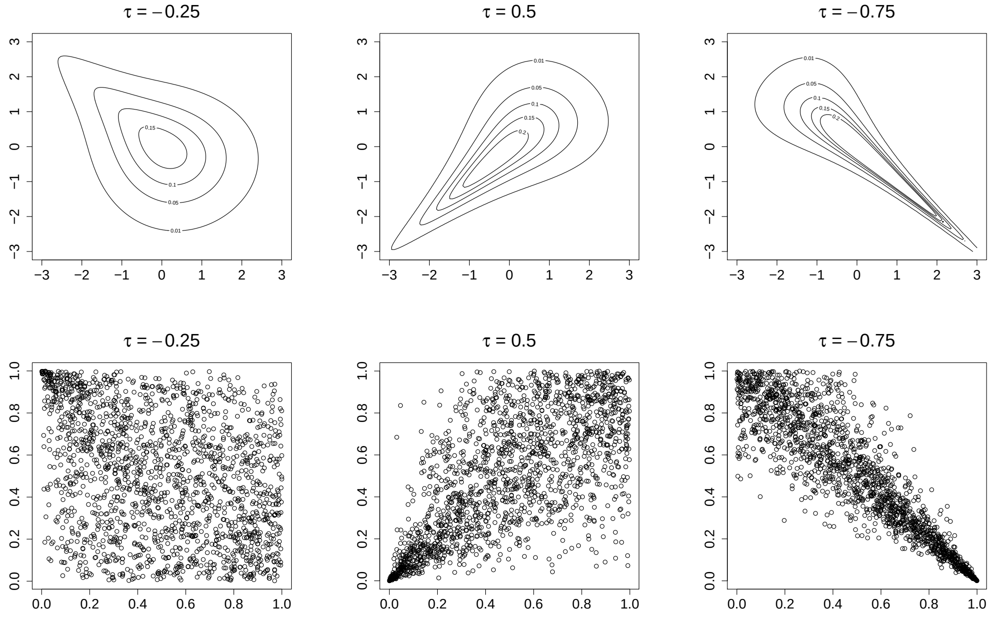


Figure A.9: Contour and scatter plots of rotated Joe copulas for three choices of Kendall's τ. The left panel corresponds to a rotated Joe copula by 90 degrees, the middle panel presents a rotated (survival) Joe copula by 180 degrees and the right panel shows a rotated Joe copula by 270 degrees.

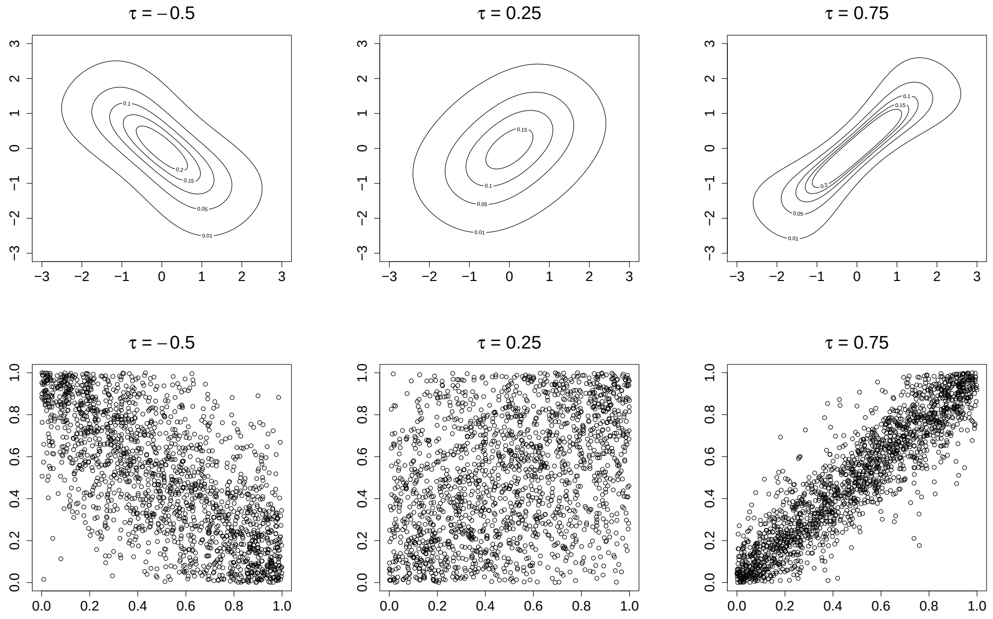


Figure A.10: Contour and scatter plots of the Frank copula for three choices of Kendall's τ.

Appendix B

Method to calculate daily mean wind direction

We have data of wind direction measured every hour (24 measurements per day) and now we want to calculate a daily mean of these wind directions. But how to do it?
The hourly wind direction is measured in degrees from 0 to 360 (0°=North, 45°=East, 180°=South, 270°=West, 360°=North, etc.
Due to the break from 360° to 0° we have to apply trigonometrical functions to be able to calculate a daily mean. Furthermore we have to weight the measurements, because a strong north wind in the first half of a day and a weak south wind in the second half should not result in a daily mean of west wind.
Therefore we use the following procedure explained by an example:

Example: Mean of two measurements:

Measurement 1: 315° (North-west) at wind speed 5,
Measurement 2: 45° (North-east) at wind speed 10.

Calculate weighted cosine:
$C_1 = 5 \times \cos\left(\frac{315\pi}{180}\right)$,
$C_2 = 10 \times \cos\left(\frac{45\pi}{180}\right)$.

Calculate weighted sine:
$S_1 = 5 \times \sin\left(\frac{315\pi}{180}\right)$,
$S_2 = 10 \times \sin\left(\frac{45\pi}{180}\right)$.

Build the means:
$\bar{c} = \frac{(C_1+C_2)}{2}$,
$\bar{s} = \frac{(S_1+S_2)}{2}$.

Since it holds, that $\tan(x) = \frac{\sin(x)}{\cos(x)}$ (see Figure B.1), we get

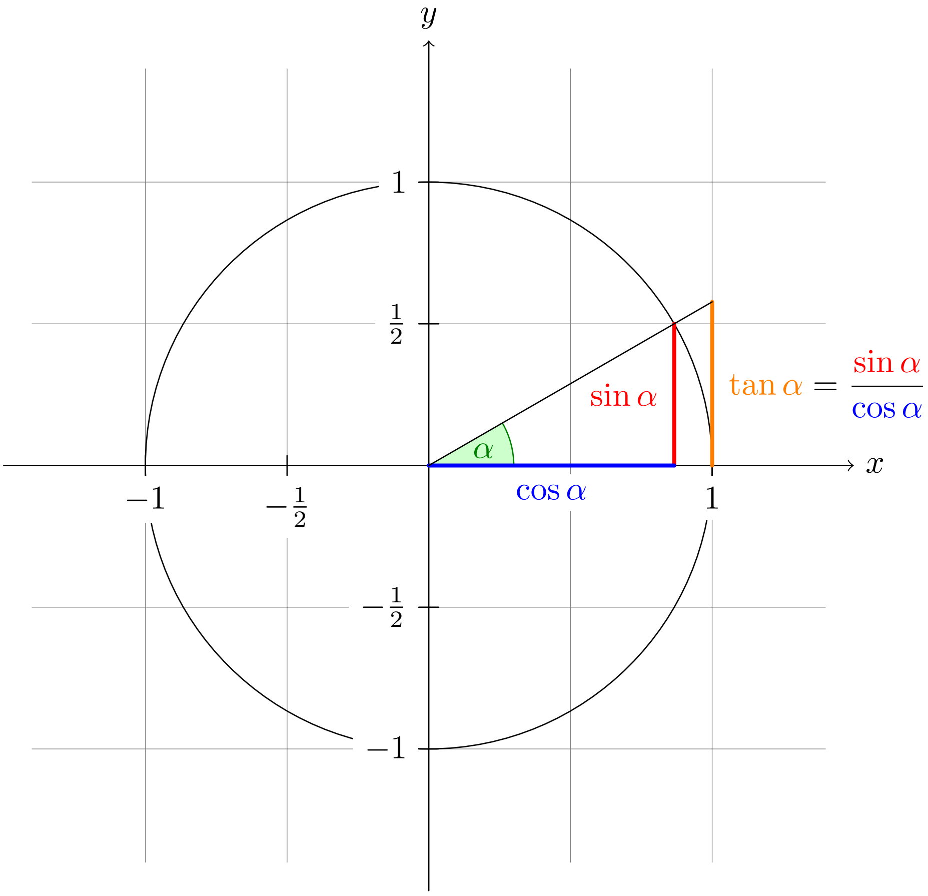


Figure B.1: Connection between sin, cos and tan on the unit circle

$\overline{winddirection} = \arctan(\frac{\bar{s}}{\bar{c}}) \times \frac{180}{\pi}$.

Here $\overline{winddirection} = 18.4°$.

Since we have 24 hourly measurements, we get the daily mean of wind direction $\big($i.e. $\overline{winddirection}\big)$ in general by:

1. Build weighted means over cosine and sine (here, $windspeed_t$ denotes the measured wind speed and $winddirection_t$ the measured wind direction at time t):

$$\bar{c} = \frac{1}{24}\sum_{t=1}^{24}\left(windspeed_t \times \cos\left(winddirection_t \times \frac{\pi}{180}\right)\right),$$

$$\bar{s} = \frac{1}{24}\sum_{t=1}^{24}\left(windspeed_t \times \sin\left(winddirection_t \times \frac{\pi}{180}\right)\right).$$

2. Get daily mean of wind direction by calculating arcus tangent:

$$\overline{winddirection} = \arctan\left(\frac{\bar{s}}{\bar{c}}\right) \times \frac{180}{\pi}.$$

Appendix C

Plots of tree ring widths against covariables

To complete the data analysis in Section 7.1 in Chapter 7 we here present plots of the detrended tree ring widths (as described in (7.1) of both tree species against our twenty covariables, namely the different seasonal means. We start with the plots of the year rings of fir trees.

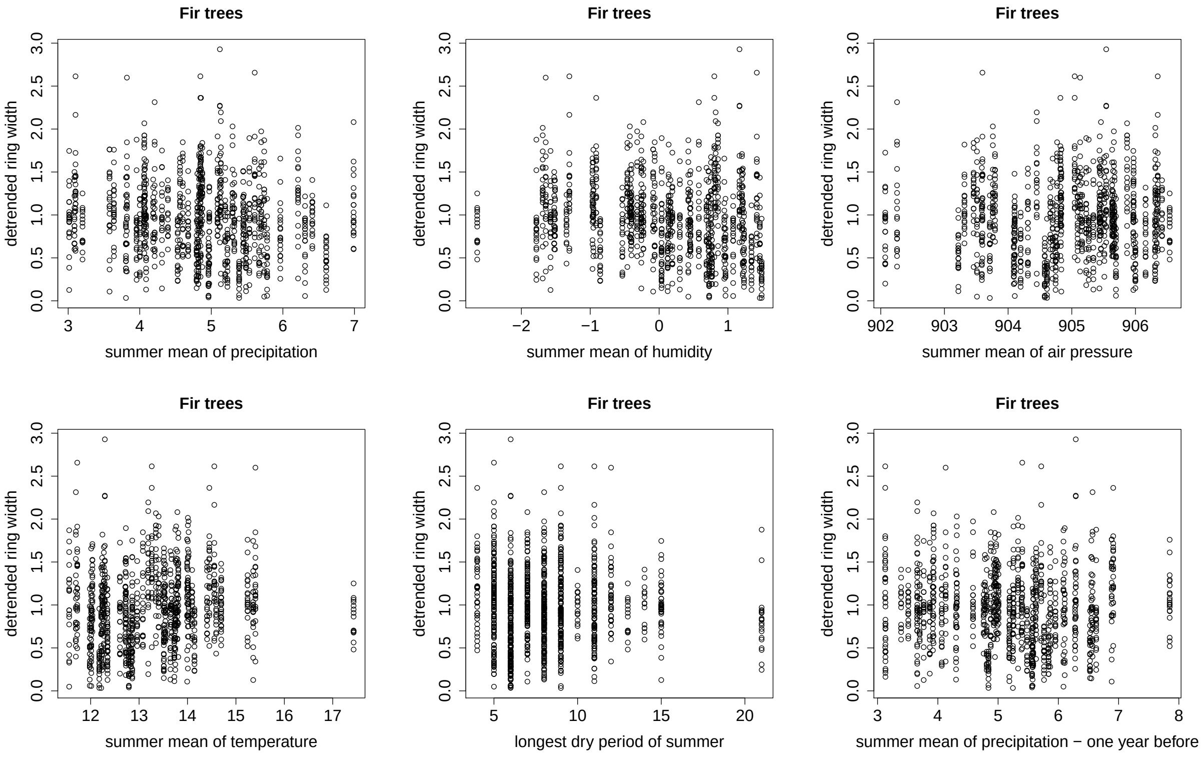


Figure C.1: Plots of detrended fir tree ring widths against the seasonal means - Part I.

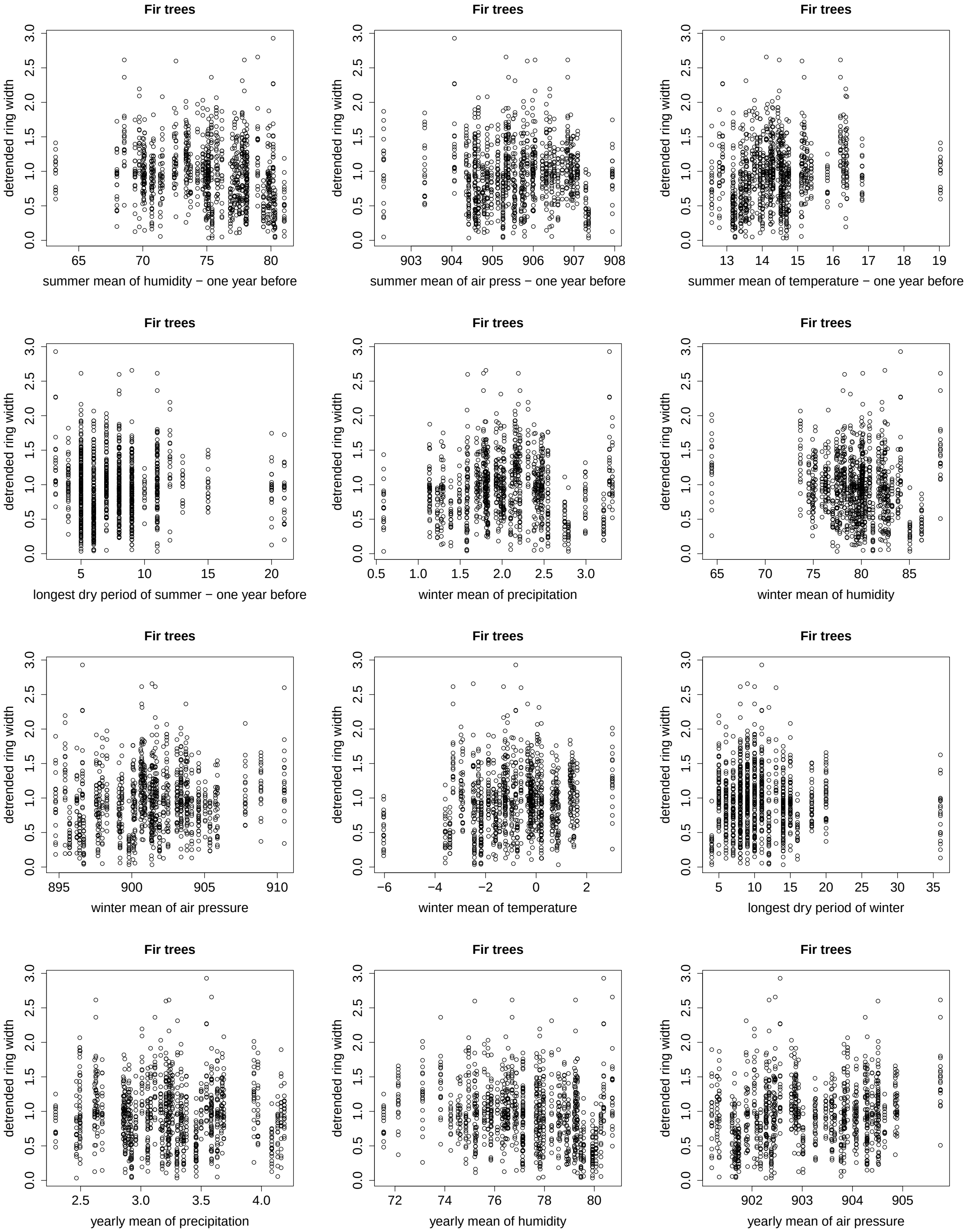


Figure C.2: Plots of detrended fir tree ring widths against the seasonal means - Part II.

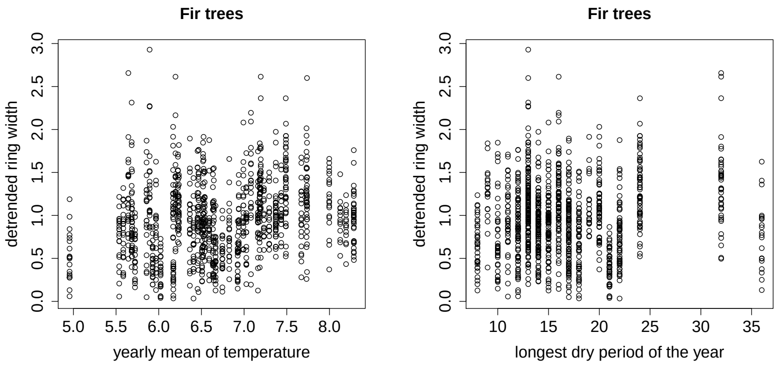


Figure C.3: Plots of detrended fir tree ring widths against the seasonal means - Part III.

We continue with the tree rings widths of spruce trees:

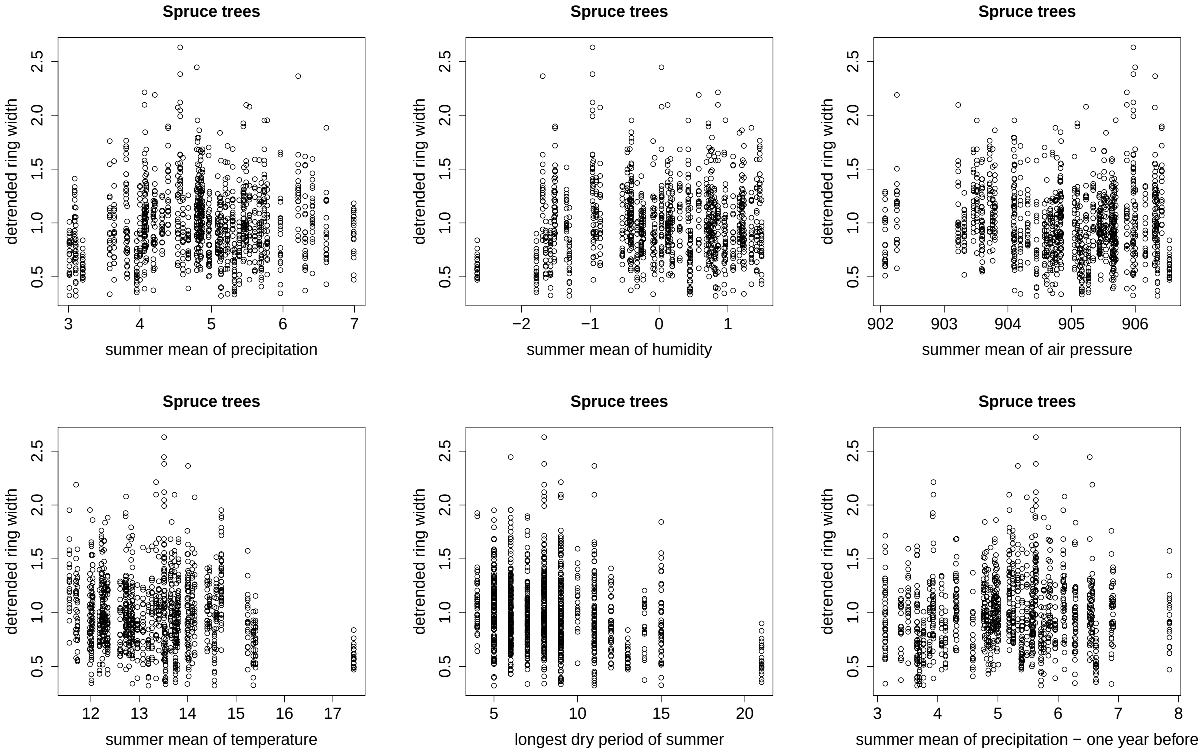


Figure C.4: Plots of detrended spruce ring widths against the seasonal means - Part I.

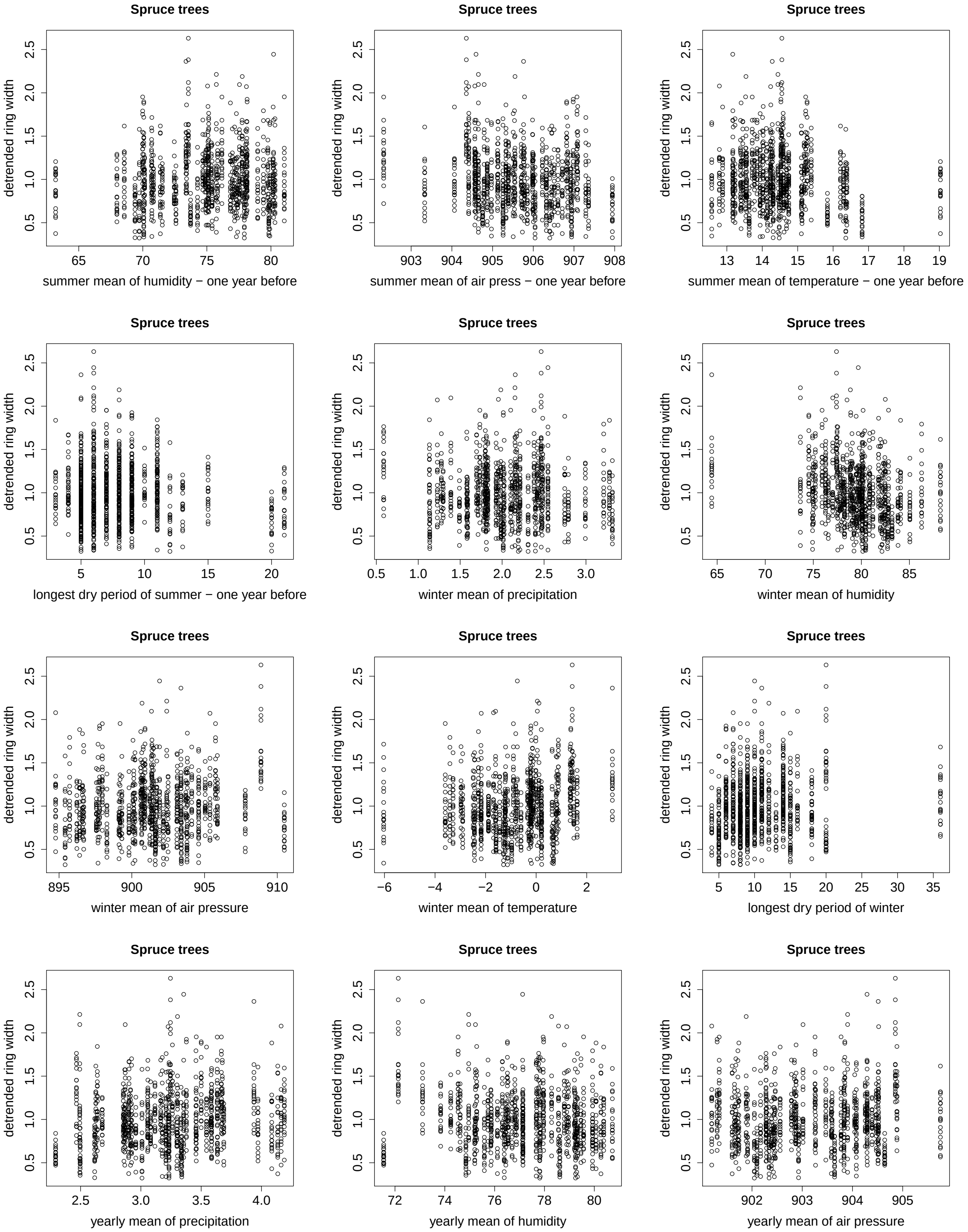


Figure C.5: Plots of detrended spruce ring widths against the seasonal means - Part II.

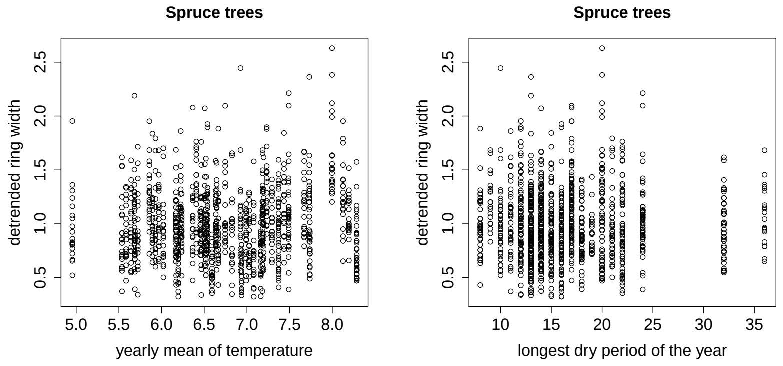


Figure C.6: Plots of detrended spruce ring widths against the seasonal means - Part III.

Bibliography

K. Aas, C. Czado, G. Frigessi, and H. Baaken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44:182–198, 2009.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, editor, *Proceedings of the Second International Symposium of Information Theory*, pages 267–281. Budapest: Akademiai Kiado, 1973.

Z. Anastasiadou and B. López-Cabrera. Statistical Modelling of Temperature Risk. 2012. SFB 649 Discussion Paper 2012-029, Humboldt-Universität zu Berlin.

A. Azzalini. A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985.

A. Azzalini. *R package* `sn`*: The skew-normal and skew-t distributions (version 0.4-17).* Università di Padova, Italia, 2011. URL `http://azzalini.stat.unipd.it/SN`.

A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society, Series B*, 61(3):579–602, 1999.

A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society, Series B*, 65:367–389, 2003.

A. Azzalini and A. Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83:715–726, 1996.

T. Bedford and R. M. Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32:245–268, 2001.

T. Bedford and R. M. Cooke. Vines: A New Graphical Model for Dependent Random Variables. *The Annals of Statistics*, 30:1031–1068, 2002.

V. J. Berrocal, A. E. Raftery, and T. Gneiting. Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *The Annals of Applied Statistics*, 2 (4):1170–1193, 2008.

F. Bertrand and M. Maumy-Bertrand. *Initiation à la Statistique avec R, R package version 1.0.4.* Dunod, Paris, 2012. URL `http://www-irma.u-strasbg.fr/ fbertran/`.

R. Bokusheva. Measuring the dependence structure between yield and weather variables. MPRA Paper 22786, University Library of Munich, Germany, April 2010. URL `http://ideas.repec.org/p/pra/mprapa/22786.html`.

A.W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations.* Oxford University Press, Oxford, 1997.

E. Brechmann. Truncated and simplified regular vines and their applications, 2010. Diploma thesis, Technische Universität München.

E. C. Brechmann and C. Czado. Risk Management with High-Dimensional Vine Copulas: An Analysis of the Euro Stoxx 50. 2012. Submitted for publication.

T. A. Buishand and A. M. G. Klein Tank. Regression model for generating time series of daily precipitation amounts for climate change impact studies. *Stochastic Hydrology and Hydraulic*, 10:87–106, 1996.

A. G. Bunn, M. Korpela, F. Biondi, F. Qeadan, and C. Zang. *dplR: Dendrochronology Program Library in R. R package version 1.5.3*, 2012. URL `http://CRAN.R-project.org/package=dplR`.

S. D. Campbell and F. X. Diebold. Weather Forecasting for Weather Derivatives. *Journal of the American Statistical Association*, 100(469):6–16, 2005.

M. Cao and J. Wei. Pricing Weather Derivative: an Equilibrium Approach. 1999. Department of Economics, Queen's University, Kingston, Ontario, Working Paper.

E. R. Cook and K. Peters. The smoothing spline: A new approach to standardizing forest interior tree-ring width series for dendroclimatic studies. *Tree-ring Bulletin*, 41:45–53, 1981.

E. R. Cook and K. Peters. Calculating unbiased tree-ring indices for the study of climatic and environmental change. *The Holocene*, 7:361–370, 1997.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms.* The MIT Press, Cambridge, Massachusetts, 2009.

F. Cribari-Neto and A. Zeileis. Beta Regression in R. *Journal of Statistical Software*, 34 (2):1–24, 2010. URL `http://www.jstatsoft.org/v34/i02/`.

C. Czado. Pair–copula constructions of multivariate copulas. In W. Hard P. Jaworski, F. Durante and T. Rychlik, editors, *Copula Theory and Its Applications.* Berlin: Springer, 2010.

C. Czado. The world of vines. Presentation, 2012. Presented at workshop: Statistical Methods and Models 2012, February 7th, 2012, Garching, available online at `http://www-m4.ma.tum.de/fileadmin/w00bdb/www/veranstaltungen/vine_world.pdf`, visited on July 20th 2012.

C. Czado and T. Schmidt. *Mathematische Statistik*. Springer-Verlag Berlin, Heidelberg, 2011.

S. Demarta and A. J. McNeil. The t Copula and Related Copulas. *International Statistical Review*, 73:111–129, 2005.

Deutsche Wetterdienst DWD. Meteorological Observatory Hohenpeissenberg, Southern Germany. Website, 2012. Available online at http://www.dwd.de/bvbw/appmanager/bvbw/dwdwwwDesktop?_nfpb=true&_windowLabel=dwdwww_main_book&T15400416501146137673615gsbDocumentPath=Content%2FOeffentlichkeit%2FKU%2Fallgemeines%2Fklimaueberwachung%2Fteaser__klimaueberwachung__mohp.html&switchLang=en&_pageLabel=_dwdwww_klima_umwelt_ueberwachung, visited on August 13th 2012.

R. Diestel. *Graph Theory*. Springer-Verlag, Heidelberg, 2010.

J. Dißmann. Statistical inference for regular vines and application., 2010. Diploma thesis, Technische Universität München.

J. Dißmann, E. C. Brechmann, C. Czado, and D. Kurowicka. Selecting and Estimating Regular Vine Copulae and Application to Financial Returns. 2011. To appear in Computational Statistics & Data Analysis.

C. Dittmar and W. Elling. Radial growth of Norway spruce and European beech in relation to weather and altitude. *Forstw. Cbl.*, 118:251–270, 1999.

V. Erhardt and C. Czado. Modeling dependent yearly claim totals including zero claims in private health insurance. *Scandinavian Actuarial Journal*, 2012(2):106–129, 2012.

L. Fahrmeir, T. Kneib, and S. Lang. *Regression: Modelle, Methoden und Anwendungen*. Springer, Berlin, 2007.

S. Ferrari and F. Cribari-Neto. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.

R. W. Freund and R. H. W. Hoppe. *Stoer/Bulirsch: Numerische Mathematik 1*. Springer, Berlin, 2007.

D. Friedrichs, U. Büntgen, D. Frank, J. Esper, B. Neuwirth, and J. Loffler. Complex climate controls on 20th century oak growth in Central-West Germany. *Tree Physiology*, 29:39–51, 2008.

H. C. Fritts. *Tree Rings and Climate*. Elsevier, New York, 1976.

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

H.-O. Georgii. *Stochastik - Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Walter de Gruyter GmbH & Co. KG, Berlin, 2007.

L. A. Gil-Alana. Statistical Modeling of the Temperatures in the Northern Hemisphere Using Fractional Integration Techniques. *Journal of Climate*, 18:5357–5368, 2005.

A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, Cambridge, UK, 1989.

T. Hastie and R. Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3): 297–318, 1986.

Trevor Hastie. *gam: Generalized Additive Models*, 2011. URL `http://CRAN.R-project.org/package=gam`. R package version 1.06.2.

J. Herzog and G. Müller-Westermeier. Homogenization of Various Climatological Parameters in the German Weather Service. In *Proceedings of the First Seminar for Homogenization of Surface Climatological Data*, pages 101–112. Budapest, Hungary, 6–12 October 1996, 1996.

I. Hobæk Haff, K. Aas, and A. Frigessi. On the simplified pair-copula construction – Simply useful or too simplistic? *Journal of Multivariate Analysis*, 101:1296–1310, 2010.

H. Hult and F. Lindskog. Multivariate Extremes, Aggregation and Dependence in Elliptical Distributions. *Advances in Applied Probability*, 34:587–608, 2002.

H. Joe. Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In B. Schweizer L. Rüschendorf and M. D. Taylor, editors, *Distributions with fixed marginals and related topics*, pages 120–141. Hayward, CA: Institute of Mathematical Statistics, 1996.

H. Joe. *Multivariate Models and Dependence Concepts.* Chapman & Hall, London, 1997.

S.-C. Kao and R. S. Govindaraju. A copula-based joint deficit index for droughts. *Journal of Hydrology*, 380:121–134, 2010.

H.-M. Kim and B. K. Mallick. A Bayesian prediction using the skew Gaussian distribution. *Journal of Statistical Planning and Inference*, 120:85–101, 2004.

W. Kleiber, A. E. Raftery, J. Baars, T. Gneiting, C. F. Mass, and E. Grimit. Locally Calibrated Probabilistic Temperature Forecasting Using Geostatistical Model Averaging and Local Bayesian Model Averaging. *Monthly Weather Review*, 139(8):2630–2649, 2011.

D. Kurowicka and R. Cooke. *Uncertainty Analysis with High Dimensional Dependence Modelling.* John Wiley & Sons Ltd, 2006.

Dorota Kurowicka and Harry Joe, editors. *Dependence Modeling: Vine Copula Handbook.* World Scientific Publishing Co. Pte. Ltd., 2010.

S. Leal, T. M. Melvin, M. Grabner, R. Wimmer, and K. R. Briffa. Tree-ring growth variability in the Austrian Alps: the influence of site, altitude, tree species and climate. *Boreas*, 36:426–440, 2008.

M. A. Little, P. E. McSharry, and J. W. Taylor. Generalized Linear Models for Site-Specific Density Forecasting of U.K. Daily Rainfall. *Monthly Weather Review*, 137: 1029–1045, 2009.

G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65:297–303, 1978.

H. Manner. Estimation and model selection of copulas wit an application to exchange rates, 2007. METEOR research memorandum (RM) 07/056, Maastricht University.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

C. E. McCulloch and S. R. Searle. *Generalized, Linear and Mixed Models*. John Wiley & Sons Inc., New York, 2001.

T. C. Mills. Modelling Current Temperature Trends. *Journal of Data Science*, 7:89–97, 2009.

A. Möller, A. Lenkoski, and T. L. Thorarinsdottir. Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 2012. doi: 10.1002/qj.2009.

G. Müller-Westermeier. Das Klima in Deutschland. In *Klimastatusbericht 2001*, pages 9–11. DWD (Deutscher Wetterdienst), 2001.

O. Morales-Nápoles. Bayesian belief nets and vines in aviation safety and other applications, 2008. Ph. D. thesis, Technische Universiteit Delft.

O. Morales-Nápoles. About the number of regular vines on n nodes. Presentation, 2010. Slides available online at `http://risk2.ewi.tudelft.nl/Work2007files/Oswaldo_number_vines.ppt`, visited on July 23th 2012.

R. B. Nelsen. *An Introduction to Copulas (2nd edition)*. Springer, Berlin, 2006.

J. Owen and R. Rabinovitch. On the Class of Elliptical Distributions and their Applications to the Theory of Portfolio Choice. *The Journal of Finance*, 38:745–752, 1983.

J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Development Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2012. R package version 3.1-103.

G. Quinn and M. Keough. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, UK, 2002.

C. H. Reinsch. Smoothing by Spline Functions. *Numerische Mathematik*, 10:177–183, 1967.

U. Schepsmeier. Maximum likelihood estimation of C-vine pair-copula constructions based on bivariate copulas from different families, 2010. Diploma thesis, Technische Universität München.

U. Schepsmeier, J. Stoeber, and E. C. Brechmann. *VineCopula: Statistical inference of vine copulas*, 2012. URL `http://CRAN.R-project.org/package=VineCopula`. R package version 1.0.

A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

J. M. Sloughter, A. E. Raftery, T. Gneiting, and C. Fraley. Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, 135:3209–3220, 2007.

R. D. Stern and R. Coe. A Model Fitting Analysis of Daily Rainfall Data. *Journal of the Royal Statistical Society. Series A (General)*, 147(1):1–34, 1984.

A. C. Turlapaty, N. H. Younan, and V. G. Anantharaj. Precipitation Data Merging using General Linear Regression. In *IGARSS (3)'09*, pages 259–262, 2009.

K. L.P. Vasconcellos and F. Cribari-Neto. Improved maximum likelihood estimation in a new class of beta regression models. *Brazilian Journal of Probability and Statistics*, 19: 13–31, 2005.

H. Visser and J. Molenaar. Trend Estimation and Regression Analysis in Climatological Time Series: An Application of Structural Time Series Models and the Kalman Filter. *Journal of Climate*, 8:969–979, 1995.

Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989.

B. T. West, K. T. Welch, and A. T. Galecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall/CRC, Boca Raton, 2007.

D. S. Wilks. Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210:178–191, 1998.

A. Y. M. Yao. A statistical model for the surface relative humidity. *Journal of Applied Meteorology*, 13:17–21, 1974.

C. S. Zang. *Growth reactions of temperate forest trees to summer drought - a multispecies tree-ring network approach*. Dissertation, Technische Universität München, 2010. URL `http://d-nb.info/1010952110/34`.

X. Zheng, R. E. Basher, and C. S. Thompson. Trend Detection in Regional–Mean Temperature Series: Maximum, Minimum, Mean, Diurnal Range, and SST. *Journal of Climate*, 10:317–326, 1997.