Semi-parametric Gaussian Copula Models for Machine Learning

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Dinu Johannes Kaufmann

aus Ufhusen, Luzern, Schweiz

Basel, 2017

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Volker Roth, Dissertationsleiter

Dr. Michael Gutmann, Korreferent

Basel, den 20. Juni 2017

Prof. Dr. Martin Spiess, Dekan

Abstract

The aim of machine learning and statistics is to learn and predict from data. With the introduction of copulas, probabilistic models and algorithms can benefit from the separation of dependency and marginals. The additional flexibility allows to generalise better and increase the prediction accuracy. Based on this observation, this work enlightens different models within the framework of a semi-parametric Gaussian copula model.

The first model we consider is archetypal analysis. We show that the Gaussian copula approximates the dependency structure of the generative model we consider. With copula archetypal analysis, we present a new model, which extends the applicability of the original model. Our second contribution refers to the semi-parametric Gaussian copula extension of principal component analysis. We consider the model in the context of parametric appearance models for facial appearance. We show, that the copula relaxation leads ultimately to a higher specificity and provide a unifying way of combining different data. The third contribution is Bayesian sub-network estimation within the framework of Gaussian graphical models. We show that the Markov blanket of a set of query variables has analytical form and can be efficiently estimated. Our last contribution is the motivation of time-resolved information flows in the context of directed information and Pearlian graphs. We show, how to discover information flows in non-stationary time series and give a convenient estimator.

At the core of these models lies the semi-parametric Gaussian copula model. In this work we show how it allows to relax certain assumptions in the aforementioned models. Ultimately, this leads to non-Gaussian and latent linear models, which better apply to real-world data sets.

Acknowledgments

I am very grateful to Prof. Dr. Volker Roth for accepting me as a PhD student and helping me making this thesis to what it is today. Without his passion for research and his unconditional commitment to machine learning, this thesis would not be possible. I am also very much obliged to Prof. Dr. Thomas Vetter for advising and inspiring me during my time in Basel.

I want to thank all members of the Biomedical Data Analysis Group for the many profound discussions which led to a substantial understanding of science and research. This includes Aleksander Wieczorek, Sonali Parbhoo, Sebastian Keller, Mario Wieser, Damian Murrezan, and my former colleagues David Adametz, Sandhya Prabhakaran, Melanie Rey, Behrouz Tajoddin, Julia Vogt, and Sudhir Raman.

Special thanks goes also to the Graphics and Vision Research Group for all the scientific and informal discussions within the PhD programm as well as in- and outside the office. This includes Bernhard Egger, Sandro Schönborn, Clemens Blumer, Marcel Lüthi, Ghazi Bouabene, Andreas Forster, Thomas Gerig, Adam Kortylewski, Jasenkov Žianov, Andreas Schneider, Tobias Maier, Christoph Jud, Thomas Albrecht, and Christoph Langguth.

Collaborating with different research groups outside the institute of computer science was an enriching experience which let me get an insight in other research areas. This includes Peter Fuhr, Ute Gschwandtner, Martin Hardmeier, Florian Hatz, Stephan Rüegg, Menorca Chaturvedi, Antonia Meyer, Vitali Cozac, Anne Roesch, Henrik Hörster, Habib Bousleiman, and Ronan Zimmermann, Paul Jenö, and Pankaj Shende. Many thanks also to Jürgen Dölz, Monica Bugeanu, Dennis Trönde, Claudiu Tanase, Nenad Stojnic, Filip-Martin Brinkmann, Ivan Giangreco, and Manolis Sifalakis for giving me an enjoyable time in Basel.

I am grateful to my family, Megumi and Hanstoni, René and Natalie, and Christian for supporting me during all my studies. A special mention deserves Noëmi for her devotion, support, and loyalty. My gratitude also goes to her familiy, Mägi, Heiri, Lisbeth, Salome and Pirmin, and Jan and Sara.

Symbols and Notation

	Symbol	Description	Format
Scalar, Vector, and Matrix	x	scalar	1×1
	$\boldsymbol{x}, x^p$	column vector	$p \times 1$
	$\boldsymbol{x}^\top$	row vector	$1 \times p$
	$\mathbf{X}$	matrix	$p \times n$
	$\mathbf{X}^\top$	transpose of matrix $\mathbf{X}$	$p \times n$
	$\mathbf{X}^{-1}$	inverse of matrix $\mathbf{X}$	$n \times p$
	$\boldsymbol{0}_p$	column vector of 0s	$p \times 1$
	$\mathbf{0}_{p \times n}$	matrix of 0s	$p \times n$
	$\boldsymbol{1}_p$	column vector of 1s	$p \times 1$
	$\mathbf{I}_{p \times n}$	matrix of 1s	$p \times n$
Random Variables	X	random variable	1×1
	x	realisation of X	1×1
	$\boldsymbol{X}, X^p$	random vector	$p \times 1$
	$\boldsymbol{x}$	realisation of $\boldsymbol{X}$	$p \times 1$
	$\mathbf{X}$	random matrix	$p \times n$
	$\mathbf{X}$	realisation of $\mathbf{X}$	$p \times n$
	$E[]$	expectation	

	Symbol	Description
Distributions	F	probability distribution
	f, p	probability density
	C	copula
	c	copula density
	$\mathcal{N}$	normal distribution
	$\mathcal{MN}$	matrix normal distribution
	$\mathcal{W}, \mathcal{W}_c$	central Wishart distribution
	$\mathcal{W}_{nc}$	non-central Wishart distribution
	$\mathcal{W}^{-1}$	inverse Wishart distribution
	Γ	Gamma distribution
	$\mathcal{IG}$	inverse Gaussian distribution
	$\mathcal{MGIG}$	matrix generalised inverse Gaussian
Information Theory	H, h	entropy
	I	mutual information
	M	multiinformation
	$D_{KL}(\cdot\|\|\cdot)$	KL-divergence
	$I(X^n \to Y^n)$	directed information
	$I(X^n \leftrightarrow Y^n)$	instantaneous coupling
Graphs	G	graph
	V	set of vertices
	E	set of edges
Sets	$\mathbb{N}$	set of all natural numbers $\{1, 2, 3, \ldots\}$
	$\mathbb{Z}$	set of all integers $\{\ldots, -1, 0, 1, \ldots\}$
	$\mathbb{R}$	set of all real numbers
	$[n]$	the set $\{1, \ldots, n\}$

Contents

1. Introduction

Machine learning and statistics address the problem of learning and prediction from data which comes from a complex system or unknown phenomenon. In a supervised setting, the data is recorded at the input and output of the system and the goal is to understand or learn the systems behaviour as well as to predict its outcome for an unseen input. In an unsupervised setting, where only data from the output is available, the goal is to infer patterns in the data. Learning from data means finding regularities in the data which generalise well for the observed system. Despite the identical goal of statistics and machine learning communities, Breiman et al. (2001) identified two different cultures of how the data is addressed to infer conclusions. On one hand, the statistics community uses (parametric) models to represent the system. The output of the system is generally modelled as a parametrised function of the input and is observed subject to random noise. Learning then corresponds to fitting the parameters of the function. The applied models are generally very well known and this leads to properly understood conclusions. Such an approach may simplify learning substantially, since a model abstracts from the potentially complex system and only represents the mechanisms of interest with a limited set of parameters. However, for complex systems, this modelling approach requires prior knowledge of the data-generating process or imposes assumptions on it. Often, these assumptions are idealised and oversimplifying or do not reproduce correctly the nature of the system. Thus, the machine learning approach keeps the system as a black box, and the primary goal is to predict accurate outputs for unseen inputs. Here, the models are often more complex and interpretation of the results are in general harder.

However, learning and prediction are impossible if there are no regularities in the data. A fundamental role for the discovery of such regularities are dependency concepts, since they allow to understand associations in the data. In this thesis, we will look at a specific model for multivariate distributions, namely at the Gaussian copula model. For any multivariate distribution, a copula is a stochastic function for modelling the dependency between random variables. Moreover, a copula is invariant against the marginal distributions. In this way, a copula abstracts the dependency between the random variables and only describes a pure association pattern. In the sequel, we will use a parametrised copula which is called a Gaussian copula. It is the inherent copula of the multivariate Gaussian distribution. However, being invariant against the marginal distributions, the Gaussian copula can also be used with non-Gaussian data. By this means, the model assumptions are weaker and the applicability of a Gaussian copula model is broadened substantially.

Models from statistics and machine learning have contrary assumptions: Typically, models from statistics are associated with low-dimensions (number of samples is much larger than number of dimensions) and the data is assumed to be generated from a given stochastic model. The goal is then to infer the parameters in the model. In the simplest case, a parametric linear model with Gaussian marginals will meet the requirements. On the other hand, approaches from machine learning treat the data mechanism as unknown and use algorithmic models in order to learn from data. This led to concepts like regularisation, bagging, boosting, neural networks and kernel machines. These concepts are very general in the sense that they are non-parametric, non-linear, and non-Gaussian and also fit well in the high-dimensional setting, where the number of samples are much lower than the number

of dimensions. Between these opposite cultures, a Gaussian copula model might be seen as a model which lies between these extremes and takes a relevant role in statistics as well as machine learning: it is a semi-parametric, latent linear model which fits to non-Gaussian data.

In the light of these considerations, we will extend various established models with the Gaussian copula and see that it improves learning and prediction. This thesis builds around four contributions which all are based on the Gaussian copula model. In particular, we discuss

- copula archetypal analysis (Kaufmann et al., 2015),
- copula eigenfaces: an application of copula principal component analysis to facial appearance (Egger et al., 2016)
- sub-network estimation in a probabilistic graphical model (Kaufmann et al., 2016), and
- causal information flows in time series.

The first two topics are related, since they concern dimensionality reduction algorithms and thus play a central role in the representation of data. Copula archetypal analysis and extends archetypal analysis to a Gaussian copula model. We show that the dependency pattern of the generative model of archetypal analysis can be approximated with a Gaussian copula.

The second topic considers principal component analysis (PCA) in the framework of a Gaussian copula. Here, the extension to the Gaussian copula model is simple since PCA assumes that the data is Gaussian distributed. In the context of facial appearance, we apply the relaxed model to parametric appearance models. The increased flexibility of the Gaussian copula model allows to increase the specificity: the non-Gaussian distributed colour is better captured by the model and a unifying combination with different data modalities like shape is possible.

The third topic focuses on estimating a undirected graphical model. We consider the Bayesian view of discovering a sub-network and focus on estimating the neighbourhood of a set of query variables. We show that the posterior conditionals have analytic form and propose an efficient Gibbs sampler. While this framework is valid for Gaussian distributed data, the Gaussian copula extension provides an elegant way that allows to apply it to non-Gaussian distributed data. We further extend the real world applicability by allowing mixed discrete and continuous non-Gaussian distributed data.

The fourth topic considers causal information flows between time series. We propose estimators which quantify the causal associations between time series in a non-stationary setting. Analogously to directed information and transfer entropy, these estimators are motivated as being defined as an difference between an oberservational and an interventional distribution. We apply the model to electroencephalogram data, and show how non-stationary information flows can be discovered.

Before we delve deeply into those topics, we provide some basics about probability theory, copulas, and information theory: they form the foundation for modelling random phenomenons and quantifying the information which is contained in an actual data sample.

1.1. Probability Theory

Statistics and machine learning use the language of probability theory to model the non-determinism or random phenomenons of the observed system. In the model, each mea-

surement corresponds to a random variable, and interactions between random variables are modelled with specific mechanisms. We start to describe, how the non-determinism of observations is modelled.

The non-determinisms in a model are described with a triple $(\Omega, \mathcal{F}, P)$ which is called a probability space. Thereby, Ω is the sample space which contains all possible outcomes ω of the model. $\mathcal{F}$ is a σ-algebra, meaning that it is a collection of subsets $A \subseteq \Omega$, which satisfy the following properties:

1. the empty set as well as the full set are elements of the σ-algebra, i.e. $\varnothing \in \mathcal{F}$ and $\Omega \in \mathcal{F}$
2. closed under the complement, i.e. if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
3. closed under countable unions, i.e. if $A_i \in \mathcal{F}, i = 1, 2, \ldots$, then $\bigcup_i A_i \in \mathcal{F}$.

Finally, $P : \mathcal{F} \mapsto \mathbb{R}$ is a probability measure that assigns a probability to each subset in $\mathcal{F}$. The properties of the probability measure P are

1. normalisation, i.e. $P(\varnothing) = 0$ and $P(\Omega) = 1$
2. non-negativity, i.e. $P(A) \geq 0, \forall A \in \mathcal{F}$
3. countable additivity, i.e. for n disjoint sets $A_1, \ldots, A_n$, $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$.

A real-valued random variable $X : \Omega \mapsto \mathbb{R}$ is a mapping from the sample space Ω to the real line $\mathbb{R}$, such that $\forall x \in \mathbb{R} : \{\omega | X(\omega) \leq x\} \in \mathcal{F}$. The condition is a mesurability condition which stems from measure theoretic considerations. However, we forgo such considerations since they won't impair our results. Instead, we use random variables as quantities, whose values are described by probability distributions. Thus, a random variable is just a mapping from the sample space to the domain of a probability distribution.

1.1.1. Univariate Distribuitons

Let $X : \Omega \mapsto \mathbb{R}$ be a random variable. A cumulative distribution function (cdf) $F_X(x) : \mathbb{R} \mapsto [0, 1]$ assigns a probability to each value $x \in \mathbb{R}$, such that

$$F_X(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}). \tag{1.1}$$

We distinguish a random variable depending on whether x is allowed to take value on a discrete or a continuous subset of $\mathbb{R}$.

Let $X : \Omega \mapsto S$ be a discrete random variable with $S \subseteq \mathbb{R}$ being a discrete subset of $\mathbb{R}$. The probability mass function (pmf) $f_X(x) : S \mapsto [0, 1]$ assigns a probability to each value in S, such that

$$f_X(x) = P(X = x), \quad \forall x \in S. \tag{1.2}$$

A continuous random variable $X : \Omega \mapsto \mathbb{R}$ exists, if it can be written in terms of a probability density function (pdf) $f_X(x)$, such that

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt \tag{1.3}$$

Note that $f_X(x) = 0$ for any given value $x \in \mathbb{R}$, but $\int_a^b f_X(x)dx = P(a \leq X \leq b) \in [0, 1]$ for $a < b$, and that the derivative $\frac{d}{dx} F_X(x) = f_X(x)$.

Normal Distribution

The normal (or Gaussian) distribution is often used for describing distributions which are not known. A normal distributed random variable X is denoted as

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2}, \tag{1.4}$$

where μ is the mean, and $\sigma^2 > 0$ is the variance. The cdf has no analytical form and is

$$F_X(X) = \Phi\mu, \sigma^2(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} \exp\left(-\frac{t^2}{2}\right) dt. \tag{1.5}$$

Due to its frequent occurence, the cdf of a normal distribution is denoted as Φ_{μ,σ^2}, and the cdf of the standard normal distribution ($\mu = 0$, $\sigma^2 = 1$) is often denoted without subscript as $\Phi(x)$.

The normal distribution has several amenities.

- Central limit theorem: averages of mutually independent random variables with constrained variance converge in distribution to the normal distribution Lyon (2014).
- Many derived expressions have analytic form (Roweis, 1999).
- The normal distribution is the maximum entropy distribution with a specified mean μ and variance σ^2 (Cover and Thomas, 2012).

Inverse Gaussian Distribution

Let $X > 0$ be a random variable which is inverse Gaussian distributed with mean $\mu > 0$, and shape $\lambda > 0$, then

$$X \sim \mathcal{IG}(\mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left(\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right). \tag{1.6}$$

Uniform Distribution

The uniform distribution is used to model a random variable with equiprobable outcomes with a limited support. If X is uniform distributed, then

$$X \sim \mathcal{U}(a, b) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b) \\ 1 & x \geq b \end{cases} \tag{1.7}$$

where a and b define the lower and upper limit of its support. The pdf is

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{1.8}$$

The uniform distribution is the maximum entropy distribution for a random variable under the sole constraint of the distribution's support (Park and Bera, 2009).

Gamma Distribution

Let $X > 0$ be a random variable which is Gamma distributed with shape $\alpha > 0$, and scale $\beta > 0$, then

$$X \sim \Gamma(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \tag{1.9}$$

where $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$ denotes the gamma function.

Probability Integral Transform

Theorem 1 (Probability Integral Transform)**.** *If the random variable X has a continuous cdf $F_X(x)$, then the random variable $Y = F_X(X)$ follows the uniform distribution $\mathcal{U}(0,1)$.*

Proof.

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(F_X(X) \le y) \\ &= P(X \le F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y \end{aligned} \tag{1.10}$$

which is the uniform distribution on the unit interval. □

For a more general proof, see Angus (1994).

1.1.2. Multivariate Distributions

Whenever modelling a system with more than one random variable, it is meaningful to consider their interactions as well. In order to do so, we define multivariate distributions which describe multiple random variables jointly.

Let $\boldsymbol{X} = (X_1, \ldots, X_p)$, $p \ge 2$ be a real-valued random vector. A multivariate cumulative distribution function (multivariate cdf) $F_{\boldsymbol{X}}(\boldsymbol{x}) : \mathbb{R}^p \mapsto [0,1]$ assigns a probability to each vector $\boldsymbol{x} \in \mathbb{R}^p$, such that

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \le x_1, \ldots, X_p \le x_p) \tag{1.11}$$

If the cdf is continuous everywhere, there exists a density

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\partial^p}{\partial x_1 \cdots \partial x_p} F_{\boldsymbol{X}}(\boldsymbol{x}) \Big|_x \tag{1.12}$$

A conditional distribution is the distribution of a random vector, when a subset of random variables are set to fixed values. In a multivariate distribution, where the dependencies between random variables are modelled as well, the fixed values provide some information about the other variables. The conditional distribution takes account of this information and provides an adjusted distribution for the other variables. In the case, where we set $\boldsymbol{Y} = \boldsymbol{y}$, the conditional distribution of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})}{f_{\boldsymbol{Y}}(\boldsymbol{y})} \tag{1.13}$$

whenever $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$.

In contrast to the conditional distribution, where we set $\boldsymbol{Y} = \boldsymbol{y}$ to fixed values, the marginal distribution provides the distribution of $\boldsymbol{X}$ without refering to the values of $\boldsymbol{Y}$. The random variables in $\boldsymbol{Y}$ are called marginalised. The marginal distributions of X is denoted as

$$\begin{aligned} f_X(x) &= \sum_{\boldsymbol{Y}} f_{X,\boldsymbol{Y}}(x, \boldsymbol{y}) \\ f_X(x) &= \int_{\boldsymbol{Y}} f_{X,\boldsymbol{Y}}(x, \boldsymbol{y}), \end{aligned} \tag{1.14}$$

for a discrete and continuous random vector, respectively.

Whenever the variables have no influence on each other, these variable are called independent, and the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be factorised as

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Y}}(\boldsymbol{y}). \tag{1.15}$$

Some random variables become independent when conditioned on other random variables. This concept called conditional independence is beneficial in analysing and describing complex distributions. If $\boldsymbol{X}$ is conditionally independent of $\boldsymbol{Y}$ given $\boldsymbol{Z}$, the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ factorises according to

$$f_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) = f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|\boldsymbol{z}) f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) \tag{1.16}$$

and is denoted as $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$.

In the same way as $f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{y})$ is defined in Eq. 1.13, the conditional distribution $f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{x})$ can be defined. This leads to the following equation which is known as the Bayes' theorem

$$f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{\boldsymbol{Y}}(\boldsymbol{y})}. \tag{1.17}$$

Multivariate Normal Distribution

Let $\boldsymbol{X}$ follow a multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, then

$$\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{1.18}$$

Matrix Normal Distribution

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ follows a Matrix Normal distribution with mean $\mathbf{M} \in \mathbb{R}^{n \times p}$, row covariance $\boldsymbol{\Omega} \in \mathbb{R}^{n \times n}$, and column covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, then

$$\begin{aligned} \mathbf{X} &\sim \mathcal{MN}_{n \times p}(\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}) \\ &= \frac{1}{(2\pi)^{\frac{pn}{2}}} \det(\boldsymbol{\Omega})^{-\frac{p}{2}} \det(\boldsymbol{\Sigma})^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Omega}^{-1}(\mathbf{X} - \mathbf{M})\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{M})^\mathsf{T}\right)\right). \end{aligned} \tag{1.19}$$

(Central) Wishart Distribution

Let $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, and the sample covariance $\mathbf{S} = \mathbf{X}^\mathsf{T}\mathbf{X}$. Then $\mathbf{S} \in \mathbb{R}^{p \times p}$ is Wishart distributed, i.e.

$$\mathbf{S} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n) = \frac{\det(\mathbf{S})^{\frac{n-p-1}{2}}}{2^{\frac{np}{2}} \det(\boldsymbol{\Sigma})^{\frac{n}{2}} \Gamma_p\left(\frac{n}{2}\right)} \exp\left(-\frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})\right), \tag{1.20}$$

where $n > p-1$ are degrees of freedom, $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ is a positive defininte scale matrix, $\Gamma_p = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma\left(\frac{n}{2} + \frac{1-j}{2}\right)$ is the multivariate gamma function, $\Gamma(z) = \int_0^\infty x^{z-1}\exp(-z)dz$ is the Gamma function, and tr is the trace function.

Non-Central Wishart Distribution

Let $\mathbf{X} \sim \mathcal{MN}_{n\times p}(\mathbf{M}, \mathbf{I}_n, \mathbf{\Sigma})$, and the sample covariance $\mathbf{S} = \mathbf{X}^\mathsf{T}\mathbf{X}$. Then $\mathbf{S} \in \mathbb{R}^{p\times p}$ is non-central Wishart distributed, i.e.

$$\begin{aligned}\mathbf{S} &\sim \mathcal{W}_{nc}\left(n, \mathbf{\Sigma}, \mathbf{\Theta}\right) \\ &= \frac{\det\left(\mathbf{S}\right)^{\frac{n-p-1}{2}}}{2^{\frac{np}{2}}\det\left(\mathbf{\Sigma}\right)^{\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)} {}_0F_1\left(\frac{n}{2}; \frac{1}{4}\mathbf{\Theta}\mathbf{\Sigma}^{-1}\mathbf{S}\right)\exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{S} + \mathbf{\Theta}\right)\right),\end{aligned} \tag{1.21}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ is a positive defininte scale matrix, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}\mathbf{M}^\mathsf{T}\mathbf{M}$ is the non-centrality parameter matrix, and ${}_0F_1$ is the hypergeometric function (Bessel function). If $\mathbf{\Theta} = \mathbf{0}$, the non-central Wishart distribution reduces to the central Wishart distribution.

1.1.3. Notation

We use simplified notation to prevent clutter: we overload the symbol f for a pmf and a pdf, since it should be clear from the context, if it refers to a discrete or continuous variable. We also omit the subscript of f_X and F_X, whenever it is clear, to which random variable the probability distribution refers to.

1.2. Copulas

It is always possible to write a multivariate cdf as in 1.11, however, there is only a limited set of distributions for jointly modelling multiple random variables. In many cases, the analytic form of these distributions forces to accept questionable approximations. Assume for example the use of the multivariate normal distribution for modelling a multivariate data set: the actual data distribution may completely mismatch the modelling assumptions due to non-Gaussian marginals and non-linear dependencies. In such situations, considering a copula model can be meaningful. A copula (Nelsen, 2013; Joe, 1997) is a function which links a multivariate joint distribution function to its univariate maginals.

Suppose having p random variables $X_1, \ldots, X_p$ with univariate marginals $F(x_i)$, $i = 1, \ldots, p$ which follow a joint distribution $F(x_1, \ldots, x_p)$. If the random variables are independent, the joint distribution can be written as the product of the marginals

$$F(x_1, \ldots, x_p) = F(x_1)\cdots F(x_p) \tag{1.22}$$

However, if the random variables depend on each other, the joint distribution has not anymore this simple form. In order to account for the interactions, a copula C links the univariate marginals to the joint distribution as follows

$$F(x_1, \ldots, x_p) = C\left(F(x_1), \ldots, F(x_p)\right). \tag{1.23}$$

This expression can be interpreted as follows: by the univariate distributions, each value x_i corresponds to a value $u_i = F(x_i) = P(X_i \leq x_i) \in [0, 1]$ which lies in the unit interval.

In the same way, the joint distribution associates the values $(x_1, \ldots, x_p)$ with a probability $u = F(x_1, \ldots, x_p) = P(X_1 \leq x_1, \ldots, X_p \leq x_p) \in [0,1]$. In this sense, the copula relates the outcomes from the individual random variables to the probability, that these outcomes occur jointly. Thus, the copula defines how the marginal distributions are coupled to a joint distribution.

A different view makes things even more clear. By the probability integral transform, the variables $U_i = F(X_i) \sim \mathcal{U}(0,1)$ follow uniform distributions on the unit interval. The copula is then defined as a multivariate cdf of marginally uniform distributed random variables, i.e.

$$C(U_1, \ldots, U_p) = P(U_1 \leq u_1, \ldots, U_p \leq u_p). \tag{1.24}$$

Definition 1 (Copula)**.** *An d-dimensional copula is a function $C : [0,1]^d \mapsto [0,1]$ with the following properties:*

1. *C is grounded, i.e. for every $\boldsymbol{u} \in [0,1]^d$, $C(\boldsymbol{u}) = 0$ if at least one coordinate of $\boldsymbol{u}$ is 0.*
2. *C has uniform margins, i.e. $C(\boldsymbol{u}) = u_k$ if all coordinates of $\boldsymbol{u} \in [0,1]^d$ are 1 except u_k.*
3. *C is d-increasing, i.e. the C-volume $V_C([\boldsymbol{u}, \boldsymbol{v}]) \geq 0$ for $\boldsymbol{u} \leq \boldsymbol{v}$.*

Here, $B = [\boldsymbol{a}, \boldsymbol{b}] = [a_1, b_1] \times \cdots \times [a_d, b_d]$ is a d-box, the Cartesian product of d intervals. For $\boldsymbol{a} \leq \boldsymbol{b}$, meaning that $a_k \leq b_k, \forall k$, the C-volume of B is given by $V_C(B) = \Delta_{\boldsymbol{a}}^{\boldsymbol{b}} C(\boldsymbol{u}) = \Delta_{a_d}^{b_d} \cdots \Delta_{a_1}^{b_1} C(\boldsymbol{u})$ which is an nth order difference of C on B.

Up to here, it is not yet clear, what form a copula is of and how it links a joint distribution function to its marginals. However, the highly celebrated theorem of Sklar clarifies the ambiguity. The theorem states the existence and uniqueness of a copula C for a joint distribution $F_{X_1,\ldots,X_p}$ with given marginals F_{X_i}, $i = 1, \ldots, p$.

Theorem 2 (Sklar)**.** *Let X and Y be random variables with (marginal) cdfs $F_X(x)$ and $F_Y(y)$, respectively, and joint distribution function $F_{X,Y}(x,y)$. Then, there exists a copula C such that for all $x \in \mathbb{R}$ and $y \in \mathbb{R}$*

$$F_{X,Y}(x,y) = C(F_X(x), F_Y(y)) \tag{1.25}$$

If $F_X(x)$ and $F_Y(y)$ are continuous cdfs, then C is unique. Otherwise, C is uniqueliy determined on $range(F_X) \times range(F_Y)$.

Conversely, if C is a copula and $F_X(x)$ and $F_Y(y)$ are probability distribution functions, then the function $F_{X,Y}(x,y)$ as defined in Eq. 1.25 is a joint distribution function with margins $F_X(x)$ and $F_Y(y)$.

Proof. See (Sklar, 1959), (Nelsen, 2013)[Theorem 2.3.3 and Theorem 2.4.3] □

In this spirit, a copula links a multivariate joint distribution to its univariate marginals. Sklar's theorem can be interpreted in the following ways:

- a copula is a multivariate distribution with uniform marginal distributions, or
- a multivariate distribution is composed of a copula and marginal distributions.

The first interpretation reveals the connection to analysis of dependence between random variables. With this in mind, copulas can be analysed with respect to the dependency structure which may be non-linear, asymmetric, or exhibits tail dependence. This leads to the notion of scale-free measures of dependence.

The second interpretation gives rise to constructions of multivariate distributions. Specifically, analytic forms for multivariate distributions with arbitrary dependency and marginals are possible.

From a machine learning point of view, copula models are interesting because the decoupled modelling of copula and marginals leads to increased flexibility or reduced assumptions in a stochastic model (Elidan, 2013). Keeping in mind the ultimate goal of machine learning, namely to learn and predict form data, the relaxed model assumption can improve prediction substantially. Not alone the modelling side takes advantage from copula models, but also algorithms: recently, variational Bayesian methods derived benefit from the copula framework by preserving structure in the variational model (Tran et al., 2015; Han et al., 2015).

Independence

An important special case of a copula is independence. The copula for independent random variables is

$$C(u_1, \ldots, u_d) = \Pi(u_1, \ldots, u_d) = u_1 \cdots u_d \tag{1.26}$$

and is called the independent or product copula. For random variables, Eq. 1.22, where the joint distribution factorises into the product of its marginals, follows directly.

Fréchet-Hoeffding Copula Bounds

The Fréchet-Hoeffding copula bounds describe limiting cases of copulas in terms of a minimum and a maximum copula. Random variables X and Y with a minimum copula are called comonotonic, in the sense that X is (almost surely) an increasing funciton of Y. For $U = F_X(X)$, and $V = F_Y(Y)$, the comonotonicity also means $P[U = V] = 1$. Random variables with a maximum copula are called countermonotonic in the sense that X is (almost surely) a decreasing function of Y, or that $P[U + V = 1] = 1$.

The Fréchet-Hoeffding lower bound, defining the maximum copula, is

$$W(u_1, \ldots, u_d) = \max(1 - d + \sum_{i=1}^{d} u_i, 0), \tag{1.27}$$

the Fréchet-Hoeffding upper bound, defining the minimum copula, is

$$M(u_1, \ldots, u_d) = \min(u_1, \ldots, u_d), \tag{1.28}$$

and, the Fréchet-Hoeffding copula bounds are

$$W(u_1, \ldots, u_d) \leq C(u_1, \ldots, u_d) \leq M(u_1, \ldots, u_d). \tag{1.29}$$

As a direct consequence of Sklar's theorem, the Fréchet-Hoeffding copula bounds can be written for random variables as follows

$$\max(1 - d + \sum_{i=1}^{d} F(x_i), 0) \leq F(X_1, \ldots, X_d) \leq \min(F(x_1), \ldots, F(x_d)). \tag{1.30}$$

Scale Invariance

An important property of copulas is scale invariance. Random variables under strictly monotone increasing transformations still have the same copula. In other words, a copula

is invariant to strictly monotone increasing transformations. For strictly monotone functions g and h with domains $range(X)$ and $range(Y)$, and transformed random variables $\tilde{F}(x) = P(g(X) \leq x) = P(X \leq g^{-1}(x)) = F(g^{-1}(x))$, and $\tilde{F}(y) = F(h^{-1}(y))$, respectively,

$$\begin{aligned} C_{g(X),h(Y)}(\tilde{F}(x), \tilde{F}(y)) &= P(g(X) \leq x, h(Y) \leq y) \\ &= P(X \leq g^{-1}(x), Y \leq h^{-1}(y)) \\ &= C_{X,Y}(F(g^{-1}(x)), F(h^{-1}(y))) \\ &= C_{X,Y}(\tilde{F}(x), \tilde{F}(y)) \end{aligned} \tag{1.31}$$

Copula Density

If the copula C and the marginals F_i are continuous, then the joint pdf f can be written in terms of the copula density $c(u_1, \ldots, u_d)$ as

$$\begin{aligned} f(x_1, \ldots, x_d) &= \frac{\partial^d}{\partial x_1 \cdots \partial x_d} C\left(F_1(x_1), \ldots F_d(x_d)\right) \\ &= c(u_1, \ldots, u_d) \prod_{j=1}^{d} f_j(x_j) \end{aligned} \tag{1.32}$$

Conditional Density

For random variables $\boldsymbol{X} : \mathbb{R}^p \mapsto [0,1]$ and $\boldsymbol{Y} : \mathbb{R}^q \mapsto [0,1]$, and $U_i = F(X_i)$, $V_j = F(Y_j)$, the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$ is

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{y}) &= \frac{f(\boldsymbol{x}, \boldsymbol{y})}{f(\boldsymbol{y})} \\ &= \frac{c(u_1, \ldots, u_p, v_1, \ldots, v_q) \prod_{i=1}^{p} f(x_i) \prod_{j=1}^{q} f(y_i)}{c(v_1, \ldots, v_q) \prod_{j=1}^{q} f(y_i)} \\ &= c(u_1, \ldots, u_p | v_1, \ldots, v_q) \prod_{i=1}^{p} f(x_i) \\ &= c(\boldsymbol{u}|\boldsymbol{v}) \prod_{i=1}^{p} f(x_i) \end{aligned} \tag{1.33}$$

The conditional copula density has following form

$$\begin{aligned} c(\boldsymbol{u}|\boldsymbol{v}) &= \frac{c(\boldsymbol{u}, \boldsymbol{v})}{c(\boldsymbol{v})} \\ &= \frac{c(\boldsymbol{u}, \boldsymbol{v})}{\int_{\boldsymbol{X}} c(F(x_1), \ldots, F(x_p), \boldsymbol{v}) \prod_{i=1}^{p} f(x_i) d\boldsymbol{X}} \\ &= \frac{c(\boldsymbol{u}, \boldsymbol{v})}{\frac{\partial^q}{\partial \boldsymbol{v}} C(\boldsymbol{1}_p, \boldsymbol{v})}. \end{aligned} \tag{1.34}$$

As elaborated in (Elidan, 2010), the final derivative form of the conditional copula has a more useful form than the integral from. This is because the integral term depends on both, the copula *and* the univariate marginals, and thus is generally difficult to compute. On the other hand, the derivative form has an explicit form for many copula distributions.

1.2.1. Gaussian Copula

The Gaussian copula is defined as

$$C^{\mathcal{N}}_{\mathbf{R}}(\boldsymbol{u}) = \Phi_{\mathbf{R}}\left(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\right). \tag{1.35}$$

where Φ^{-1} is the inverse cdf of the standard normal distribution, and $\Phi_{\mathbf{R}}$ is the joint cdf of a zero-mean multivariate normal distribution parametrised by a correlation matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$. The correlation matrix $\mathbf{R}$ is the only parameter of a Gaussian copula model.

The copula density of the Gaussian copula is

$$\begin{aligned} c^{\mathcal{N}}_{\mathbf{R}}(\boldsymbol{u}) &= \frac{\partial^d}{\partial u_1 \cdots \partial u_d} \Phi_{\mathbf{R}}\left(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\right) \\ &= \frac{\partial^d}{\partial z_1 \cdots \partial z_d} \Phi_{\mathbf{R}}(z_1, \dots, z_d) \prod_{i=1}^{d} \frac{\partial}{\partial u_i} \Phi^{-1}(u_i) \\ &= \frac{\phi_{\mathbf{R}}\left(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\right)}{\prod_{i=1}^{d} \phi(\Phi^{-1}(u_i))} \\ &= \frac{1}{\sqrt{\det \mathbf{R}}} \exp\left(-\frac{1}{2} \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix}^T \left(\mathbf{R}^{-1} - \mathbf{I}\right) \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix} \right), \end{aligned} \tag{1.36}$$

where we used the chain rule and the inverse function derivative theorem, ϕ is the pdf of a standard normal distribution, and $\phi_{\mathbf{R}}$ is the pdf of a zero-mean multivariate normal distribution with correlation $\mathbf{R}$.

Latent Space

By construction, the Gaussian copula model inherently implies a latent space by the transformation

$$Z_i = \Phi^{-1}(U_i) = \Phi^{-1}(F_i(X_i)) \tag{1.37}$$

and the copula is defined as

$$\begin{aligned} C^{\mathcal{N}}_{\mathbf{R}}(\boldsymbol{z}) &= \Phi_{\mathbf{R}}\left(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\right) \\ &= \Phi_{\mathbf{R}}(z_1, \dots, z_d). \end{aligned} \tag{1.38}$$

This is just the cdf of a zero-mean multivariate normal distribution with correlation matrix $\mathbf{R}$, thus

$$\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}_d, \mathbf{R}). \tag{1.39}$$

In practice, when using a Gaussian copula, it is meaningful to always going through calculation in the latent space, since the Gaussian distribution allows to use the many derived expressions in analytic form, see e.g. (Roweis, 1999). Fig. 1.1 gives a graphical summary of the spaces involved in a Gaussian copula.

Accordinlgy, the density has following form.

$$\begin{aligned} c^{\mathcal{N}}_{\mathbf{R}}(\boldsymbol{u}) &= \frac{\phi_{\mathbf{R}}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right)}{\prod_{i=1}^{d} \phi(\Phi^{-1}(u_i))} \\ &= \frac{\phi_{\mathbf{R}}\left(z_1, \ldots, z_d\right)}{\prod_{i=1}^{d} \phi(z_i)} \\ &= \frac{1}{\sqrt{\det \mathbf{R}}} \exp\left(-\frac{1}{2} \begin{pmatrix} z_1 \\ \vdots \\ z_d \end{pmatrix}^T \left(\mathbf{R}^{-1} - \mathbf{I}\right) \begin{pmatrix} z_1 \\ \vdots \\ z_d \end{pmatrix}\right) \end{aligned} \tag{1.40}$$

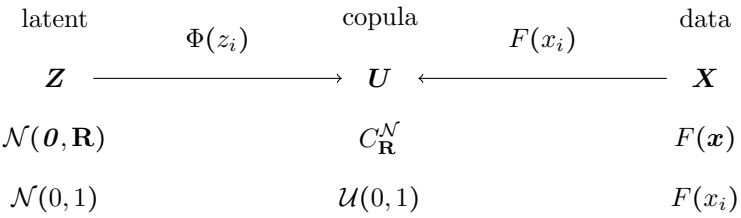


Figure 1.1.: Spaces, random variables, multivariate distributions, marginal distributions (from top to bottom), and mappings (on arrows) of a Gaussian copula model.

Semi-parametric Inference

A semi-parametric model is noteworthy for inference in multivariate distributions when the data is non-Gaussian distributed and contains a moderate amount of outliers. In a semi-parametric copula model, a parametric copula is used for modelling the associations within the data, whereas the non-parametric part is related to the marginal distributions. In this thesis, we use a semi-parametric Gaussian copula model, where we focus on analysing the associations in the data without imposing assumptions on the marginal distributions.

Motivated by the Gaussian distributed latent space in a Gaussian copula model, the approach for analysis of dependence is as follows:

1. Estimate the parameters of the Gaussian copula, namely the latent correlation matrix $\mathbf{R}$.
2. Compute the measure of your choice, using the convenient analytic forms and methods which were devised with the assumption of the multivariate normal distribution.

Using non-parametric marginals has implications on inference in a Gaussian copula model. In particular, the correlation matrix is computed on the ranks of the data only. The correlation between random variables X and Y is the Gaussian rank correlation

$$\rho_G(X,Y) = \frac{\sum_{i=1}^{n} \Phi^{-1}\left(\frac{R(x_i)}{n+1}\right) \Phi^{-1}\left(\frac{R(y_i)}{n+1}\right)}{\sum_{i=1}^{n} \Phi^{-1}\left(\frac{i}{n+1}\right)^2}, \tag{1.41}$$

where $R(x_i)$ and $R(y_i)$ are the ranks of x_i and y_i respectively. Note, the denominator does not depend on the data. An interpretation of this expression is as follows: the probability integral transform is approximated by the ranks which were rescaled to lie between $0 < R(x_i) < 1$, i.e.

$$u(x_i) = F(x_i) \approx \frac{R(x_i)}{n+1}. \tag{1.42}$$

with subsequent transformation to the latent space $z_i = \Phi^{-1}(u(x_i))$. The Gaussian rank correlation is then just the Pearson's correlation in the latent space.

Though, the Gaussian rank correlation is not the only choice for estimating the correlation matrix in a non-parametric way. Depending on the actual data distribution, Kendall's τ_K and Spearman's ρ_S are suitable alternatives (Liu et al., 2012; Xue et al., 2012). Nonetheless, the Gaussian rank correlation has appealing properties (Boudt et al., 2012) which were summarised as

- consistency: Compared to Kendall's τ_K and Spearman's ρ_S, no transformation is needed to obtain consistency for the correlation coefficient of a bivariate normal distribution. This allows for estimating a correlation matrix of a multivariate normal distribution by estimating each element by its bivariate Gaussian rank correlation coefficient.
- positive-definiteness: The resulting correlation matrix is always positive definit.
- complexity: $\mathcal{O}(d^2 n \log n)$ fast to compute, also in high dimensions
- robustness: a breakdown point of 12.4%, showing robustness to small amounts of outliers

In chapters 2 - 5, we investigate the implications of the Gaussian copula under different models.

1.3. Information Theory

Intuitively, information refers to facts, knowledge, and data about a system. Information can be gained by receiving a message or by observing anything. When observing a system, information can be gained by observing the random variables and recording the outcomes. The outcomes of random variables provide information about the configuration of the system. Thus, information reduces the uncertainty about a system.

When quantifying information, a lot of information corresponds to surprise or the unexpected. No information corresponds to what is deterministic or already known. Analogously, an outcome with low probability conveys a lot of information, and an outcome with high probability conveys few information. Thus, information is inverse proportional to probability of an outcome and is directly related to the distribution of a random variable. Information is a quantity that measures the uncertainty in the outcome of a random variable or an experiment to be performed.

Information theory goes back to Shannon (1948), who devised methods to send messages over noisy communication channels, such that the information content of the messages is optimised. This notion of optimal channel capacity is strongly related to statistical dependence. Later, Massey (1990) devised causal quantities which measure the increase of the capacity of a communication channel, when feedback is present. Nonetheless, information theory is not limited to the analysis of communiction channels, but rather is a more general concept (Brillouin, 1962) which is used in different areas such as statistics (Kullback, 1997;

Akaike, 1998), statistical mechanics (Jaynes, 1957) , and quantum computation and quantum information (Nielsen and Chuang, 2002). In the following, we briefly define the most relevant terms and refer to Cover and Thomas (2012) for a comprehensive work of the topic. We put more emphasis on the recent results which are relevant for this thesis.

1.3.1. Entropy

Entropy refers to the amount of uncertainty of a random variable. Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be discrete random variables. The entropy of X is

$$H(X) = -\sum_{x \in \mathcal{X}} f(x) \log f(x) = -E[\log f(x)]. \tag{1.43}$$

The joint entropy is

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x,y) \log f(x,y). \tag{1.44}$$

For a joint distribution $f(X,Y)$, the conditional entropy is

$$H(X|Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x,y) \log f(x|y). \tag{1.45}$$

The chain rule for entropy states that

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, X_2, \ldots, X_{i-1}). \tag{1.46}$$

1.3.2. Entropy Rate

Let $\{X_i\}_{i=1}^n \in \mathcal{X}^n$ be a stochastic process. The entropy rate of the stochastic process $\{X_i\}_{i=1}^n$ is

$$H(\{X_i\}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n). \tag{1.47}$$

One can also show that the entropy of the stochastic process $\{X_i\}_{i=1}^n$ is

$$H(\{X_i\}) = \lim_{n \to \infty} H(X_n | X_{n-1}, \ldots, X_1). \tag{1.48}$$

1.3.3. Differential Entropy

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be continuous random variables. The entropy of X is

$$h(X) = -\int_{\mathcal{X}} f(x) \log f(x) dx. \tag{1.49}$$

The joint differential entropy of a set $X_1, X_2, \ldots, X_n$ is defined as

$$h(X_1, X_2, \ldots, X_n) = -\int_{\mathcal{X}} f(x^n) \log f(x^n) dx^n. \tag{1.50}$$

For a joint distribution $f(X,Y)$, the conditional differential entropy of is defined as

$$h(X|Y) = -\int_{\mathcal{X}} f(x,y) \log f(x|y) dx dy. \tag{1.51}$$

The chain rule for differential entropy states that

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}). \tag{1.52}$$

The relation between entropy and differential entropy is somewhat complicated. There is an analogy in the notation of both measures, since the sum for the support of a discrete random variable is replaced by an integral for a continuous random variable. However, the interpretation is different: whereas the description of a discrete random variable requires maximally $H(X) = \log(n)$ bits, the description of a n-bit quantised continuous random variable requires $h(X) + n$ bits. Thus, a continuous analog of discrete entropy would assign an entropy of ∞ to every infinitely resolved continuous random variable. Note, differential entropy can also become negative, whenever $f(x) > 1$. Hence, attention is required, when comparing entropy and differential entropy, since they do not have the same intuition.

Pleasingly, these laborious circumstances vanish when considering relative entropy measures: the intuition for discrete as well as continuous random variables are equal. Thus, the formal separation of the discrete and continuous case is not anymore necessary. In the following, the integral form is used for both, discrete and continuous random variables.

1.3.4. Relative Entropy

Let f and g be probability density functions. The relative entropy or Kullback-Leibler divergence of f with respect to g is

$$D_{KL}(f(x)||g(x)) = \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} dx. \tag{1.53}$$

The conditional relative entropy is

$$D_{KL}(f(x|y)||g(x|y)) = \int_{\mathcal{Y}} f(y) \int_{\mathcal{X}} f(x|y) \log \frac{f(x|y)}{g(x|y)} dxdy. \tag{1.54}$$

The chain rule for relative entropy is

$$D_{KL}(f(x,y)||g(x,y)) = D_{KL}(f(x)||g(x)) + D(f(y|x)||g(y|x)). \tag{1.55}$$

Mutual Information

The mutual information between X and Y is

$$\begin{aligned} I(X;Y) &= D_{KL}\left(f(x,y)||f(x)f(y)\right) \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy. \end{aligned} \tag{1.56}$$

Conditional mutual information is

$$\begin{aligned} I(X;Y|Z) &= D_{KL}\left(f(x,y|z)||f(x|z)f(y|z)\right) \\ &= \int_{\mathcal{Z}} f(z) \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x,y|z) \log \frac{f(x,y|z)}{f(x|z)f(y|z)} dxdydz. \end{aligned} \tag{1.57}$$

The chain rule for mutual information is

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, X_2, \ldots, X_{i-1}). \tag{1.58}$$

Multivariate Mutual Information

Multivariate mutual information (Jakulin and Bratko, 2003a,b) is a multivariate generalisation of mutual information. The recursive definition is given by

$$I(X_1;\ldots;X_n) = I(X_1;\ldots;X_{n-1}) - I(X_1;\ldots;X_{n-1}|X_n). \tag{1.59}$$

Let $\mathcal{V} = \{X_1,\ldots,X_n\}$, and $\mathcal{T} \subseteq \mathcal{V}$, then, an alternating inclusion-exclusion definition of multivariate mutual information is given by

$$I(\mathcal{V}) = -\sum_{\mathcal{T}\subseteq\mathcal{V}} (-1)^{|\mathcal{V}|-|\mathcal{T}|} H(\mathcal{T}). \tag{1.60}$$

Multiinformation

Multiinformation[1] (Studený and Vejnarová, 1999) quantifies the dependency or redundancy among a set of random variables and is defined as the Kullback-Leibler divergence between the joint distribution $f(x_1,\ldots,x_d)$ and the componentwise independent distribution $f(x_1)\cdots f(x_d)$

$$\begin{aligned} M(X_1,X_2,\ldots,X_d) &= D_{KL}\left(f(x_1,\ldots,x_d)\|f(x_1)\cdots f(x_d)\right) \\ &= \int_{x_1}\cdots\int_{x_d} f(x_1,\ldots,x_d)\log\frac{f(x_1,\ldots,x_d)}{f(x_1)\cdots f(x_d)}dx_1\cdots dx_d. \end{aligned} \tag{1.61}$$

Factorising the logarithm, multiinformation reduces to a differences of entropies

$$M(X_1,X_2,\ldots,X_d) = \sum_{i=1}^{d} H(X_i) - H(X_1,\ldots,X_d). \tag{1.62}$$

A decomposition similar to a chain rule (Slonim et al., 2005) for multiinformation is

$$M(X_1,X_2,\ldots,X_d) = \sum_{i=2}^{d} I(X_{i-1};X_i,\ldots,X_d). \tag{1.63}$$

For any $\mathcal{V} = \{X_1,\ldots,X_n\}$, and $\mathcal{T} \subseteq \mathcal{V}$, it is possible to express multiinformation in terms of multivariate mutual information

$$M(\mathcal{V}) = \sum_{\mathcal{T}\subseteq\mathcal{V}:|\mathcal{T}|\geq 2} (-1)^{|\mathcal{T}|} I(\mathcal{T}), \tag{1.64}$$

as well as multivariate mutual information in terms of multiinformation

$$I(\mathcal{V}) = \sum_{\mathcal{T}\subseteq\mathcal{V}:|\mathcal{T}|\geq 2} (-1)^{|\mathcal{V}|-|\mathcal{T}|} M(\mathcal{T}). \tag{1.65}$$

This decomposition also generalises to sets of variables. For two sets X^p and Y^q, the decomposition is

$$I(X^p;Y^q) = M(X^p,Y^q) - M(X^p) - M(Y^q) \tag{1.66}$$

and provides a useful decomposition of mutual information in terms of multiinformation.

[1] Multiinformation is also known as total correlation (Watanabe, 1960) and multivariate constraint (Garner and Carson, 1960). It should not be confused with multivariate mutual information.

1.3.5. Gaussian Random Variables

Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ be a Gaussian distributed random variable, then, the entropy of X is

$$h(X) = \frac{1}{2} \log\left(2\pi e \sigma^2\right). \tag{1.67}$$

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a multivariate Gaussian distributed random vector, then the entropy of $\boldsymbol{X}$ is

$$h(\boldsymbol{X}) = \frac{1}{2} \log\left((2\pi e)^d \det(\boldsymbol{\Sigma})\right).. \tag{1.68}$$

Let $\mathbf{A}$ be a matrix, then

$$H(\mathbf{A}\boldsymbol{X}) = h(\boldsymbol{X}) + \log|\det(\mathbf{A})|. \tag{1.69}$$

1.3.6. Relation between Copulas and Multiinformation

Since the copula is a probability distribution, the entropy of the copula is well defined and is called copula entropy

$$H_c(\boldsymbol{x}) = -\int_{\boldsymbol{u}} c(\boldsymbol{u}) \log\left(c(\boldsymbol{u})\right) d\boldsymbol{u}. \tag{1.70}$$

The following theorem is a direct application of Sklar's theorem to multiinformation. The theorem states that multiinformation depends on the copula only but not on the marginals of a joint distribution.

Theorem 3 ((Ma and Sun, 2011))**.** *Multiinformation is equivalent to negative copula entropy:*

$$M(X_1, \ldots, X_d) = -H_c(X_1, \ldots, X_d) \tag{1.71}$$

Proof.

$$\begin{aligned} M(X_1, \ldots, X_d) &= \int_{\boldsymbol{X}} f(X_1, \ldots, X_d) \log\left(\frac{f(X_1, \ldots, X_d)}{f(X_1)\cdots f(X_d)}\right) d\boldsymbol{X} \\ &= \int_{\boldsymbol{X}} c(U_1, \ldots, U_d) \prod_{i=1}^{d} f(X_i) \log c(U_1, \ldots, U_d) d\boldsymbol{X} \\ &= \int_{\boldsymbol{U}} c(U_1, \ldots, U_d) \log c(U_1, \ldots, U_d) d\boldsymbol{U} \end{aligned} \tag{1.72}$$

where we changed variables $U_i = F(X_i)$ such that $d\boldsymbol{U} = \prod_{i=1}^{d} f(X_i) d\boldsymbol{X}$. □

For a Gaussian copula, multiinformation is

$$
\begin{aligned}
& M(X_1, \ldots, X_d) \\
&= \int_{\boldsymbol{U}} c(U_1, \ldots, U_d) \log c(U_1, \ldots, U_d) d\boldsymbol{U} \\
&= \int_{\boldsymbol{U}} \frac{\phi_{\mathbf{R}}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right)}{\prod_{i=1}^{d} \phi(\Phi^{-1}(u_i))} \log \left(\frac{\phi_{\mathbf{R}}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right)}{\prod_{i=1}^{d} \phi(\Phi^{-1}(u_i))} \right) d\boldsymbol{U} \\
&= \int_{\boldsymbol{Z}} \phi_{\mathbf{R}}(z_1, \ldots, z_d) \log \left(\frac{\phi_{\mathbf{R}}(z_1, \ldots, z_d)}{\prod_{i=1}^{d} \phi(z_i)} \right) d\boldsymbol{Z} \\
&= \int_{\boldsymbol{Z}} \phi_{\mathbf{R}}(z_1, \ldots, z_d) \log \left(\phi_{\mathbf{R}}(z_1, \ldots, z_d) \right) d\boldsymbol{Z} - \prod_{i=1}^{d} \int_{Z_i} \phi(z_i) \log \phi(z_i) dZ_i \\
&= -\frac{1}{2} \log \left((2\pi e)^d \det(\mathbf{R}) \right) + \frac{d}{2} \log(2\pi e) \\
&= -\frac{1}{2} \log \det(\mathbf{R}),
\end{aligned}
\tag{1.73}
$$

where we changed variables $U_i = \Phi(Z_i)$ such that $d\boldsymbol{U} = \prod_{i=1}^{d} \phi(z_i) d\boldsymbol{Z}$.

1.3.7. Directed Information Theory

Causal conditioning builds the foundation of directed information theory. Assume two sequences of random variables $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, Y_n)$ with temporally aligned indices, such that a random variable is associated for every time point in every time series. Then, the joint distribution of X^n and Y^n can be factorised as follows:

$$
\begin{aligned}
& p(X^n, Y^n) \\
&= \prod_{i=1}^{n} p(X_i, Y_i | Y^{i-1}, X^{i-1}) \\
&= \prod_{i=1}^{n} p(X_i | X^{i-1}, Y^{i-1}) p(Y_i | Y^{i-1}, X^i) \\
&= \prod_{i=1}^{n} p(X_i | X^{i-1}, Y^{i-1}) \frac{p(Y_i, X_i | Y^{i-1}, X^{i-1})}{p(X_i | Y^{i-1}, X^{i-1})} \\
&= \prod_{i=1}^{n} p(X_i | X^{i-1}, Y^{i-1}) \frac{p(Y_i | Y^{i-1}, X^{i-1}) p(X_i | Y^i, X^{i-1})}{p(X_i | Y^{i-1}, X^{i-1})} \\
&= \prod_{i=1}^{n} p(X_i | X^{i-1}, Y^{i-1}) p(Y_i | Y^{i-1}, X^{i-1}) \frac{p(Y_i, X_i | Y^{i-1}, X^{i-1})}{p(Y_i | Y^{i-1}, X^{i-1}) p(X_i | Y^{i-1}, X^{i-1})} \\
&= \prod_{i=1}^{n} p(X_i | X^{i-1}, Y^{i-1}) p(Y_i | Y^{i-1}, X^{i-1}) c(Y_i, X_i | Y^{i-1}, X^{i-1}) \\
&= p(X^n \| Y^{n-1}) p(Y^n \| X^{n-1}) c(X^n, Y^n \| X^{n-1}, Y^{n-1}),
\end{aligned}
\tag{1.74}
$$

where the first and second equation follow from the chain rule. The third equation follows from conditioning. Note that the terms are asymmetric in the second argument of the conditionings. In the communication literature, this was interpreted as that due to

propagation delays in a physical channel, the transmitted symbol X_i is slightly prior to the received symbol Y_i. However, in the following, we will rigorously use the the idea of causal conditioning (Massey, 1990; Kramer, 1998), such that the conditioning of a random variable at time i is only allowed for random variables which are prior to time i. The fourth equation follows from the chain rule, the fifth equation from conditioning. This eliminates the asymmetry and reveals a symmetric instantaneous term instead (Amblard and Michel, 2012). In the sixth equation, the instantaneous term is identified as a conditional copula. The last equation follows by defining groups of the individual terms over time. We use the notion for causal conditioning, i.e. $p(X^n||Y^{n-1}) = \prod_{i=1}^{n} p(X_i|X^{i-1}, Y^{i-1})$, and $c(X^n, Y^n||X^{n-1}, Y^{n-1}) = \prod_{i=1}^{n} c(Y_i, X_i|Y^{i-1}, X^{i-1})$. The first term is associated to feedforward information, the second to feedback, and the third to instantaneous coupling.

Directed Information

Applying the same reasoning to entropy and mutual information the following quantities can be defined: Causal conditional entropy is

$$H(Y^n||X^n) = \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X^i), \tag{1.75}$$

which also follows the causality principle of only conditioning on variables which are prior in time. From the same principle, Massey (1990) generalised the symmetric mutual information to a asymmetric quantity called directed information

$$\begin{aligned} I(X^n \to Y^n) &= D_{KL}\left(p(X^n, Y^n)||p(X^n||Y^{n-1})p(Y^n)\right) \\ &= H(Y^n) - H(Y^n||X^n) \\ &= \sum_{i=1}^{n} I(X^i; Y_i|Y^{i-1}), \end{aligned} \tag{1.76}$$

which is, in case of a system with feedback, a more useful quantity, since it only measures the mutual information in one direction. Instantaneous coupling, on the other hand, measures the remaining contemporaneous mutual information and thus might be seen as a non-causal coupling. It is defined as

$$I(X^n \leftrightarrow Y^n) = \sum_{i=1}^{n} I(X_i; Y_i|X^{i-1}, Y^{i-1}). \tag{1.77}$$

Including instantaneous coupling, Amblard and Michel (2012) showed that mutual information decomposes into the sum of directed informations and instantaneous coupling. This fundamental decomposition is

$$I(X^n; Y^n) = I(X^{n-1} \to Y^n) + I(Y^{n-1} \to X^n) + I(X^n \leftrightarrow Y^n). \tag{1.78}$$

Side Information

Accounting for side information, i.e. a third time series Z^n which acts as a confounder or mediator, the corresponding conditional form of the aforementioned quantities can be defined (Amblard and Michel, 2012).

Causal conditional directed information is

$$\begin{aligned} I(X^n \to Y^n||Z^n) &= H(Y^n||Z^n) - H(Y^n||X^n, Z^n) \\ &= \sum_{i=1}^{n} I(X^i; Y_i|Y^{i-1}, Z^i). \end{aligned} \tag{1.79}$$

Causal conditional instantaneous coupling is

$$I(X^n \leftrightarrow Y^n \| Z^n) = \sum_{i=1}^{n} I(X_i; Y_i | Y^{i-1}, X^{i-1}, Z^i). \tag{1.80}$$

The fundamental decomposition of mutual information in the conditional form is

$$\begin{aligned} &I(X^n; Y^n \| Z^{n-1}) \\ &\quad = I(X^{n-1} \to Y^n \| Z^{n-1}) + I(Y^{n-1} \to X^n \| Z^{n-1}) + I(X^n \leftrightarrow Y^n \| Z^{n-1}). \end{aligned} \tag{1.81}$$

Transfer Entropy

Transfer entropy, as defined by Schreiber (2000); Amblard and Michel (2012), implicitly assumes stationarity and thus is defined as a rate. It explicitly does not contain the instantaneous coupling term

$$\begin{aligned} I_\infty(X^{n-1} \to Y^n) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(X^{i-1}; Y_i | Y^{i-1}) \\ &= \lim_{n\to\infty} I(X^{n-1}; Y_n | Y^{n-1}). \end{aligned} \tag{1.82}$$

Causally conditioned transfer entropy is

$$\begin{aligned} I_\infty(X^{n-1} \to Y^n \| Z^{n-1}) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(X^{i-1}; Y_i | Y^{i-1}, Z^{i-1}) \\ &= \lim_{n\to\infty} I(X^{n-1}; Y_n | Y^{n-1}, Z^{n-1}). \end{aligned} \tag{1.83}$$

Analogously, a directed information rate, which includes instantaneous coupling, is defined as

$$\begin{aligned} I_\infty(X^n \to Y^n) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}) \\ &= \lim_{n\to\infty} I(X^n; Y_n | Y^{n-1}), \end{aligned} \tag{1.84}$$

and the causal conditioned directed information rate is

$$\begin{aligned} I_\infty(X^n \to Y^n \| Z^n) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}, Z^i) \\ &= \lim_{n\to\infty} I(X^n; Y_n | Y^{n-1}, Z^n). \end{aligned} \tag{1.85}$$

Note, the conditioning sets of causally conditioned transfer entropy and causally conditioned directed information rate differ.

2. Copula Archetypal Analysis

In machine learning, many problems are approached by the paradigm of collecting as much evidence as possible. Often the number of observations is limited but it is possible to collect many features from the phenomenon of interest. For example in neurophysiology, the cohorts are limited, but collecting many clinical variables and combining them with high-dimensional time series from magnetic resonance imaging (MRI) as well as electroencephalography (EEG) or magnetoencephalography (MEG) are possible. Also in gene expression analysis, the number of patients is limited, but next generation sequencing allows to sequence the whole human genome with reasonable costs. Nevertheless, finding structure in heterogeneous and high-dimensional data sets can still be challenging and thus, basis transformations to compact representations often facilitate the analysis of data. Archetypal analysis suits this need as it is a data-adaptive technique which represents the data in a lower dimensional manifold. As a special virtue, archetypal analysis represents the data while keeping extremal characteristics of the data set.

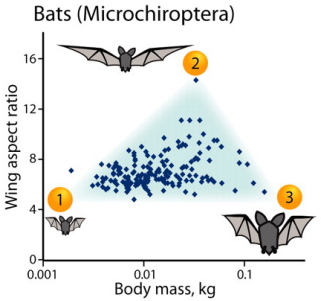
However, combining different data sources is still a difficult task, since different modalities are quantified on different scales. As a workaround, the data is often transformed or normalised to enable a meaningful analysis. Though, finding suitable transformations can be demanding because structure within the data set vanish or emerge depending on the transformations. In this chapter, we will use the copula framework to give a principled way to approach this problem and we will show the benefits which come with a relaxation from a Gaussian to a Gaussian copula model.

Before we deepen into the copula version of archetypal analysis, we will introduce classical archetype analysis. Several algorithms for basis transformations to data-adaptive compact representations are outlined. Subsequently, copula archetypal analysis is presented as providing a unified method for absorbing monotone transformations. Moreover, the Gaussian copula is motivated to be a justified approximation for the probabilistic and generative model we consider. Finally, we highlight additional benefits which come with the copula extension and conclude with an example in computational biology.

2.1. Archetypal Analysis

Archetypal analysis is an unsupervised learning concept in machine learning. Given a data sample in a multi-dimensional space, archetypal analysis tries to find a lower-dimensional manifold which approximates the data by representing it with respect to convex mixtures of itself. By the nature of the problem, the new representation will be in terms of extremal vertices, the so-called archetypes which lie close to the convex hull of the data sample.

An example from biology, presented by Shoval et al. (2012), makes this concept intuitively comprehensible: from a total of 108 species, Norberg and Rayner's study of bat wings (Norberg and Rayner, 1987) identified $K = 3$ archetypes which explain – to some degree – almost all different species. The archetypal bats were found as to outperform all other bats at a single given task, see Fig. 2.1.



	Archetypes	Task
1	Low aspect ratio, small body	Prey: small insects, near vegetation
2	High aspect ratio, medium body	Prey: high flying large insects
3	Low aspect ratio, large body	Prey: animals, near vegetation

Figure 2.1.: Wing aspect ratio of bats versus their body mass (Shoval et al., 2012). Three archetypes were identified; their inferred tasks are listed in the table above. The convex hull is the border of the light blue area. Figure from (Shoval et al., 2012).

Cutler and Breiman (1994) introduced archetypal analysis with the intention that "archetypal analysis represents each individual in a data set as a [convex] mixture of individuals of pure type or archetype." In more precise terms, let the data $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ be in an p-dimensional space, and $\boldsymbol{z}_k \in \mathbb{R}^p$, $k = 1, \ldots, K$, be the archetypes of cardinality K. The data is approximated as a convex combination of the archetypes, i.e.

$$\boldsymbol{x}_i \approx \sum_{k=1}^{K} a_{ik} \boldsymbol{z}_k, \quad \text{s.t.} \quad a_{ik} \geq 0, \quad \text{and} \quad \sum_{k=1}^{K} a_{ik} = 1, \quad i = 1, \ldots, n \tag{2.1}$$

and the archetypes are convex combinations of the data itself, i.e.

$$\boldsymbol{z}_k = \sum_{i=1}^{n} b_{ki} \boldsymbol{x}_i, \quad \text{s.t.} \quad b_{ki} \geq 0, \quad \text{and} \quad \sum_{i=1}^{n} b_{ki} = 1, \quad k = 1, \ldots, K \tag{2.2}$$

where a_{ik} are the convex coefficients of the data with respect to the archetypes and b_{ki} are the coefficients of the archetypes with respect to the data. Eq. 2.1 describes the p-dimensional data $\boldsymbol{x}_i$ in new coordinates $\boldsymbol{a}_i^\top$ with respect to a K-dimensional basis $\boldsymbol{z}_i$, $i = 1, \ldots, K$. Given the data $\boldsymbol{x}_i$, $i = 1, \ldots, n$, the goal of archetypal analysis is to find all coefficients a_{ik}, b_{ki}, and the archetypes $\boldsymbol{z}_k$, $k = 1, \ldots, K$. The corresponding optimisation problem minimises the representation error in terms of residual sum of squares

$$RSS = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \sum_{k=1}^{K} a_{ik} \sum_{j=1}^{n} b_{kj} \boldsymbol{x}_j\|_2^2 \tag{2.3}$$

under the constraints of Eqs. 2.1 and 2.2.

For notational simplicity, we repeat the above optimisation problem in matrix form: the data is represented as the rows of matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and are convex combinations of the archetypes which are the rows of matrix $\mathbf{Z} \in \mathbb{R}^{K \times p}$. Then, the optimisation problem is

$$\begin{aligned} &\underset{\mathbf{A},\mathbf{B}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2 \\ &\text{s.t.} \quad a_{ik} \geq 0, \quad \mathbf{A}\boldsymbol{1}_K = \boldsymbol{1}_n, \quad \text{and} \\ &\qquad b_{ki} \geq 0, \quad \mathbf{B}\boldsymbol{1}_n = \boldsymbol{1}_K, \end{aligned} \tag{2.4}$$

where $\mathbf{A} \in \mathbb{R}^{n \times K}$ and $\mathbf{B} \in \mathbb{R}^{K \times n}$ are row stochastic matrices.

Cutler and Breiman (1994) also introduced an elegant way for inferring the coefficients $\mathbf{A}$ and $\mathbf{B}$. The optimisation problem is convex in $\mathbf{A}$ if we fix $\mathbf{B}$ as well as in $\mathbf{B}$ if we fix $\mathbf{A}$. This leads to an alternating optimisation algorithm which is outlined in Alg. 1. Each iteration consists of two constrained least squares problems which are solved by minimising a penalised version of the non-negative least squares algorithm of Lawson and Hanson (1974).

Algorithm 1 Alternating non-negative least squares algorithm for archetypal analysis.

Require: Data $\mathbf{X}$
Result: Archetypes $\mathbf{Z}$, coefficients $\mathbf{A}$, $\mathbf{B}$
Initialise: $\mathbf{Z}$
1: **while** not converged **do**
2: $\quad \mathbf{A} \leftarrow \operatorname{argmin}_{\mathbf{A}} \|\mathbf{X} - \mathbf{A}\mathbf{Z}\|_F^2 \quad \text{s.t.} \quad a_{ik} \geq 0, \quad \mathbf{A}\boldsymbol{1}_K = \boldsymbol{1}_n$
3: $\quad \mathbf{Z} = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{X}$
4: $\quad \mathbf{B} \leftarrow \operatorname{argmin}_{\mathbf{B}} \|\mathbf{Z} - \mathbf{B}\mathbf{X}\|_F^2 \quad \text{s.t.} \quad b_{kj} \geq 0, \quad \mathbf{B}\boldsymbol{1}_n = \boldsymbol{1}_K$
5: $\quad \mathbf{Z} = \mathbf{B}\mathbf{X}$
6: **end while**

Hitherto, many advances were proposed to accelerate the optimisation problems. Among those, one important idea is to pre-select points on the convex hull by computing the convex hull in two-dimensional projections (e.g. by pairwise PCA projections or random projections), cf. (Thurau et al., 2009; Bauckhage and Thurau, 2009; Kersting et al., 2010; Bauckhage, 2014). Further advances in convex optimisation allowed to accelerate the constrained least squares problems. Mørup and Hansen (2010) proposed a projected gradient descent approach where the non-negativity constraints were satisfied with (costly) back-projections into the feasible set. Preventing the projections was possible by noting that the constraints force the rows of coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ to reside in the standard simplicies, i.e. the update steps are constituted of convex minimisation problems over convex sets. This enabled the use of more elegant techniques. Prabhakaran et al. (2012) proposed the use of monotone increasing forward stagewise regression (MIFSR), a monotone LASSO approach, cf. Hastie et al. (2007). Later on, we will look at an other detail of this approach.

For the same reason, Bauckhage et al. (2015) proposed to use the Frank-Wolfe procedure. In each iteration, this algorithm only solves a linear approximation of the problem, but with the same set of constraints. By construction, the algorithm will automatically satisfy the constraints. Expanding the residual sum of squares

$$\begin{aligned} RSS &= \|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2 \\ &= \operatorname{tr}\left((\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X})^\mathsf{T}(\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X})\right) \\ &= \operatorname{tr}\left(\mathbf{X}^\mathsf{T}\mathbf{X} - 2\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{B}\mathbf{X} + \mathbf{X}^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{B}\mathbf{X}\right) \end{aligned} \tag{2.5}$$

and ignoring the constraints, the linear approximations are given by the gradients with

respect to $\mathbf{A}$ and $\mathbf{B}$ are

$$\begin{aligned}\nabla_{\mathbf{A}} RSS &= -2\left(\mathbf{X}\mathbf{X}^\intercal\mathbf{B}^\intercal - \mathbf{A}\mathbf{B}\mathbf{X}\mathbf{X}^\intercal\mathbf{B}^\intercal\right)\\ &= -2\left(\mathbf{X}\mathbf{Z}^\intercal - \mathbf{A}\mathbf{Z}\mathbf{Z}^\intercal\right)\\ \nabla_{\mathbf{B}} RSS &= -2\left(\mathbf{A}^\intercal\mathbf{X}\mathbf{X}^\intercal - \mathbf{A}^\intercal\mathbf{A}\mathbf{B}\mathbf{X}\mathbf{X}^\intercal\right)\end{aligned} \tag{2.6}$$

The advantage of this approach lies in guarantees for fast achievement of ϵ-approximations of the optimal solution that are provably sparse, cf. e.g. Jaggi (2013). The algorithm is outlined in Alg. 2. Here, $\boldsymbol{e}_j$ is a unit vector with component j having value 1. Note, if $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ are initialised such that they meet the convexity constraints, then they also fulfill the constraints after an update.

Algorithm 2 Frank-Wolfe for archetypal analysis.

Require: Data $\mathbf{X}$
Result: Archetypes $\mathbf{Z}$, coefficients $\mathbf{A}$, $\mathbf{B}$
Initialise: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{Z}$
while not converged **do**
  while not converged **do**
    $\mathbf{G} = \nabla_{\mathbf{A}} RSS = -2\left(\mathbf{X}\mathbf{Z}^\intercal - \mathbf{A}\mathbf{Z}\mathbf{Z}^\intercal\right)$
    for $i \in \{1, \ldots, n\}$ **do**
      $j = \operatorname{argmin}_l G_{il}$
      $\boldsymbol{a}_i \leftarrow \boldsymbol{a}_i + 2/(t+2)(\boldsymbol{e}_j - \boldsymbol{a}_i)$
    end for
  end while
  while not converged **do**
    $\mathbf{G} = \nabla_{\mathbf{B}} RSS = -2\left(\mathbf{A}^\intercal\mathbf{X}\mathbf{X}^\intercal - \mathbf{A}^\intercal\mathbf{A}\mathbf{B}\mathbf{X}\mathbf{X}^\intercal\right)$
    for $j \in \{1, \ldots, K\}$ **do**
      $i = \operatorname{argmin}_l G_{jl}$
      $\boldsymbol{b}_j \leftarrow \boldsymbol{b}_j + 2/(t+2)(\boldsymbol{e}_i - \boldsymbol{b}_j)$
    end for
  end while
  $\mathbf{Z} = \mathbf{B}\mathbf{X}$
end while

The general optimisation problem of archetypal analysis, the trade-off between the representation error (in terms of e.g. RSS) and the complexity of the model (in terms of e.g. number of parameters or compression) is not fully solved in the aforementioned algorithms: all these models work with a predefined number of archetypes K and model selection is often done a posteriori by cross-validation. Prabhakaran et al. (2012) slightly changed the optimisation problem in Eq. 2.4 and introduced an intermediate step in order to determine the number of archetypes. One change concerns the dimensions of matrices $\mathbf{A}$ and $\mathbf{B}$ which are fixed to $\mathbb{R}^{n\times n}$, and $\mathbf{Z} \in \mathbb{R}^{n\times p}$, such that every single observation is a candidate for being an archetype. Model selection is then performed by introducing another constraint in the

optimisation problem, in particular

$$\boldsymbol{z}^{GL} \leftarrow \underset{\boldsymbol{z}^{GL}}{\operatorname{argmin}} \|\boldsymbol{x}^{GL} - \mathcal{A}\boldsymbol{z}^{GL}\|_2^2, \quad \text{s.t.} \quad \sum_{j=1}^{n} \|\boldsymbol{z}_j^{GL}\|_{1,2} \leq \kappa \tag{2.7}$$

where

$$\boldsymbol{x}^{GL} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{0}_n & \cdots & & \boldsymbol{a}_p & \boldsymbol{0}_n & \cdots & \boldsymbol{0}_n \\ \boldsymbol{0}_n & \boldsymbol{a}_1 & \cdots & & \boldsymbol{0}_n & \boldsymbol{a}_p & \cdots & \boldsymbol{0}_n \\ & \ddots & & \cdots & & \ddots & & \\ \boldsymbol{0}_n & \cdots & \boldsymbol{a}_1 & & \boldsymbol{0}_n & \boldsymbol{0}_n & \cdots & \boldsymbol{a}_p \end{pmatrix}, \quad \boldsymbol{z}^{GL} = \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_n \end{pmatrix}. \tag{2.8}$$

are the expanded observation vector, the expanded weighting matrix, and the expanded archetype vector, respectively. Then, the Group-Lasso is used for sparse selection of archetypes and the Bayesian Information Criterion (BIC) for model selection. Solving the Group-Lasso optimisation problem is approached with an active-set algorithm (Roth and Fischer, 2008) which samples the solution path at discrete sets of regularisation parameter κ. Since no additional costs emerge in computing the BIC scores over the entire solution path, the method proves to be computationally efficient. The algorithm is outlined in Alg. 3.

Algorithm 3 Group-Lasso extension for model selection archetypal analysis.

Require: Data $\mathbf{X}$
Result: Archetypes $\mathbf{Z}$, coefficients $\mathbf{A}$, $\mathbf{B}$
Initialise: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{Z}$

1: **while** not converged **do**
2: $\quad \mathbf{A} \leftarrow \operatorname{argmin}_{\mathbf{A}} \|\mathbf{X} - \mathbf{A}\mathbf{Z}\|_F^2 \quad \text{s.t. } a_{ik} \geq 0, \mathbf{A}\boldsymbol{1}_n = \boldsymbol{1}_n \quad \triangleright$ using MIFSR
3: $\quad \boldsymbol{z}_{GL} \leftarrow \operatorname{argmin}_{\boldsymbol{z}^{GL}} \|\boldsymbol{x}^{GL} - \mathcal{A}\boldsymbol{z}^{GL}\|_2^2, \quad \text{s.t. } \sum_{j=1}^{n} \|\boldsymbol{z}_j^{GL}\|_{1,2} \leq \kappa$
4: $\quad \mathbf{B} \leftarrow \operatorname{argmin}_{\mathbf{B}} \|\mathbf{Z} - \mathbf{B}\mathbf{X}\|_F^2 \quad \text{s.t. } b_{kj} \geq 0, \mathbf{B}\boldsymbol{1}_n = \boldsymbol{1}_n \quad \triangleright$ using MIFSR
5: $\quad Z = BX$
6: **end while**

Archetypal analysis extends rather trivially to non-linear kernel models, since the data $\mathbf{X}$ only occurs as Gram matrices $\mathbf{X}\mathbf{X}^\mathsf{T}$ in the optimisation problem, cf. Eq. 2.6, where the inner products $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{x}_j$ can be replaced by any kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. This extensions have been proposed in (Mørup and Hansen, 2010) and (Bauckhage and Manshaei, 2014), efficient methods, where the Gram matrix is approximated by the Nyström method, have been proposed in (Zhao et al., 2016).

The probabilistic interpretation of archetypal analysis (Seth and Eugster, 2013, 2016) may be the closest related work to ours. From a probabilistic viewpoint, classical archetypal analysis can be seen as a linear latent variable model: The n observations are described as convex mixtures of K archetypes arranged as the rows of the matrix $\mathbf{Z}$, thus the mixing components sum to one, i.e. $\sum_{k=1}^{K} a_k = \boldsymbol{1}_K^\mathsf{T}\boldsymbol{a} = 1$. In a probabilistic archetype model we might assume that $\boldsymbol{a}_i \sim \mathrm{Dir}_K(\boldsymbol{\alpha})$, and that the observations $\boldsymbol{x}_i \in \mathbb{R}^p$ scatter around the means $\mathbf{Z}^\mathsf{T}\boldsymbol{a}_i$ according to isotropic Gaussian noise with variance η, such that we arrive at the generative model

$$\begin{aligned} \boldsymbol{a}_i &\sim \mathrm{Dir}_K(\boldsymbol{\alpha}), \quad i = 1, \ldots, n \\ \boldsymbol{x}_i | \mathbf{Z}, \boldsymbol{a}_i &\sim \mathcal{N}_p(\mathbf{Z}^\mathsf{T}\boldsymbol{a}_i, \eta\mathbf{I}_p). \end{aligned} \tag{2.9}$$

Thus, identifying the archetypes can be probabilistically formulated as minimising the negative log-likelihood

$$\sum_{i=1}^{n} (\boldsymbol{x}_i - \mathbf{Z}^\mathsf{T} \boldsymbol{a}_i)^2 = \|\mathbf{X} - \mathbf{AZ}\|_F^2 \tag{2.10}$$

Seth and Eugster (2013) noticed that the model in Eq. 2.9 approximates the convex hull in the parameter space under a Gaussian observation model. This observation can be generalised to other observation models. In this way, they provided efficient optimisation methods for observation models having Bernoulli, multinomial, and Poisson distribution. In Seth and Eugster (2016), the model is further extended to nominal observations and a variational Bayes inference scheme which selects a suitable number of archetypes, at least as for a moderate number of archetypes ($K < 6$).

Several variations of archetypal analysis have been successfully applied to e.g. image collections (Thurau and Bauckhage, 2009; Ebert and Schiele, 2013), document collections (Canhasi and Kononenko, 2014), economic market studies (Li et al., 2003), game strategies (Sifa and Bauckhage, 2013), and audio dictionary learning (Diment and Virtanen, 2015).

2.2. Copula Archetypal Analysis

Remark

The sequel closely follows Kaufmann et al. (2015).

Finding the archetypes is a geometric concept that crucially depends on the representation of the observations in $\mathbb{R}^p$. One major problem in classical archetypal analysis, which we like to address, is its sensitivity to monotone transformations of the coordinate axes: it can make a huge difference if one measures a certain property for example in meters or log(meters). This problem is illustrated in Fig. 2.2: After a transformation of the original data by a strictly monotone increasing transformation, the lower left panel would suggest a total of four archetypes, one located at each corner. Whereas the lower right panel, reconstructed by a semi-parametric copula, identifies approximately the same three archetypes as in the original data.

As long as only Euclidean lengths are concerned, one might argue that the sensitivity to monotone transformations is a problem of somewhat artificial nature, but in high-dimensional real-world applications with features of different types and different domains, the representation problem from above indeed defines an inherent limitation of classical archetypal analysis.

As a means for overcoming this representational problem we introduce a copula based preprocessing step thus making archetypal analysis invariant against all (strictly) monotone increasing transformations: being inherently invariant against such strictly monotone increasing transformations, copula densities prove to be *exactly* the invariance class needed for this task.

Presumably the most elegant solution for the problem of inferring the archetypes would be to complement the model with priors over all (hyper-) parameters and analyse the posterior distribution of the archetypes in a fully Bayesian fashion. In general, we think this would be feasible but this is not the main focus of this work. Instead, we would like to maintain a

probabilistic “flavour”, but we still want to make use of existing highly efficient algorithms

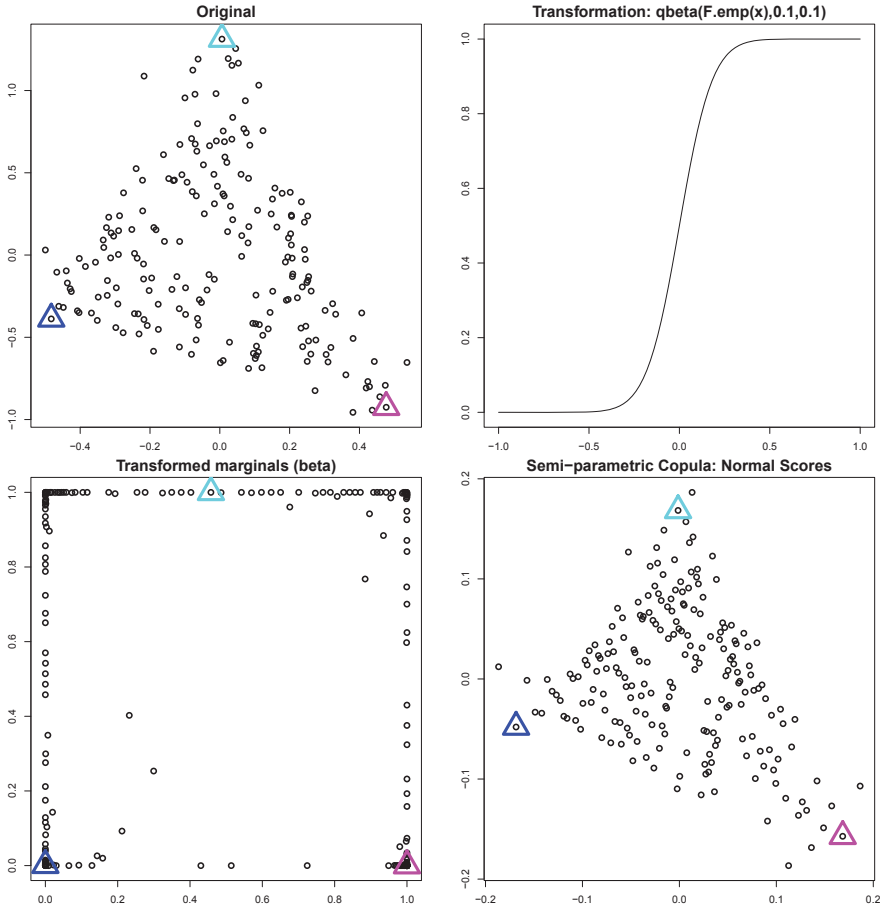


Figure 2.2.: Scale invariance of copula archetypal analysis. Upper left panel: 200 points sampled as (noisy) convex mixtures of 3 archetypes (triangle symbols) in two dimensions. Upper right: Strictly monotone transformation applied to each dimension. Lower left: Transformed data points and location of the original archetypes after transformation. Lower right: Reconstruction of the transformed dataset by copula-PCA.

which we introduced in the previous section.

2.2.1. Model

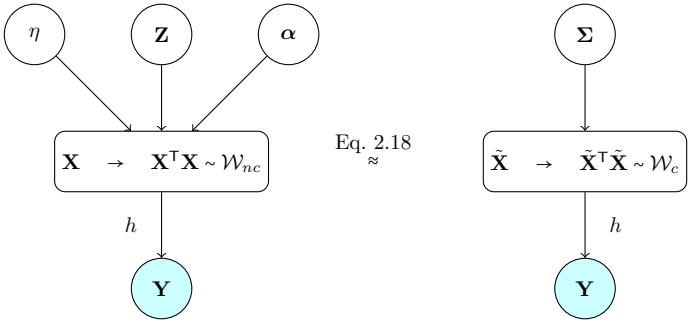


Figure 2.3.: Probabilistic graphical models of Archetypal Analysis (left) and Gaussian Copula (right).

In this section we show how to overcome the representational problem by embedding archetypal analysis in a copula framework (Nelsen, 1999; Joe, 1997). The framework includes a strictly monotone increasing mapping h: $\boldsymbol{y} = h(\boldsymbol{x})$, $\mathbb{R}^p \mapsto \mathbb{R}^p$, thereby treating $\boldsymbol{X}$ as latent variables which are estimated on the observations $\boldsymbol{Y}$, as shown in the probabilistic graphical model in Fig. 2.3. The formulation with latent variables allows to re-use existing algorithms for recovering the archetypes.

Sklar's theorem (Sklar, 1959) allows the decomposition of every continuous multivariate cumulative distribution function (cdf) $F(Y_1, \ldots, Y_p)$ into it's univariate marginals $F_1(Y_1), \ldots, F_p(Y_p)$ and a copula C comprising the dependency pattern only. More precisely, the theorem states the existence and uniqueness of a copula C such that

$$F(Y_1, \ldots, Y_p) = C(U_1, \ldots, U_p), \tag{2.11}$$

where the uniformly distributed $U_j = F_j(Y_j)$ are generated with the probability integral transformation of the univariate marginal cdfs. In the following, we will look for a parametric copula C which suitably represents the dependency structure in the space of U.

2.3. Inference

2.3.1. Special Case: Continuous Observations Without Missing Values.

If all observations $\mathbf{Y}$ are continuous and if there are no missing values, the simplest way of estimating each column $X_{\bullet,j}$, $j = 1, \ldots, p$ is to compute the *normal scores* based on the empirical marginal cdfs F_{emp} and the standard normal inverse cdf: $\hat{U}_{\bullet,j} = F_{\text{emp}}(Y_{\bullet,j}) =$

$\text{ranks}(Y_{\bullet,j})/(n+1)$ is a uniformly distributed random variable, and $X_{\bullet,j} = \Phi^{-1}(\hat{U}_{\bullet,j})$ further transforms the density (element-wise) to standard normal. Given the normal scores, the correlation matrix Σ which fully parametrises the Gaussian copula, is then just the expected covariance of the normal scores.

Using the empirical marginals F_{emp}, corresponds to the non-parametric part in the inference, since only the ranks are used in the transformation. This establishes invariance against arbitrary continuous cdfs F and also against their composition with an arbitrary strict monotone increasing transformation $(F \circ g)(\boldsymbol{y})$. This makes inference invariant against different representations as well as insensitive against outliers. Note that we might have different cdfs and transformations in every component of $\boldsymbol{y}$.

The algorithm, outlined in Alg. 4, now proceeds by estimating the latent $\mathbf{X}$ based on the Gaussian copula model, and then calling an arbitrary function `FindArchetypes(X)` that minimises Eq. 2.10 and returns archetypes $\boldsymbol{Z}$ and mixing coefficients $\mathbf{A}$. We assume that this function implements some classical archetype reconstruction algorithm, together with some mechanism for selecting the number of archetypes. In practice, we use the group-Lasso based algorithm proposed in (Prabhakaran et al., 2012) which uses the Bayesian Information Criterion (BIC)-score for automatically choosing an appropriate number of archetypes.

Algorithm 4 Copula archetypal analysis for continuous observations.

Require: Observations $\mathbf{Y}$
Result: Archetypes $\mathbf{Z}$
1: **for** all dimensions **do**
2: Compute normal scores $X_{\bullet,j} = \Phi^{-1}\left(\frac{\text{ranks}(Y_{\bullet,j})}{n+1}\right)$
3: **end for**
4: $Z \leftarrow$ `FindArchetypes(X)`

2.3.2. General Case: Mixed Data and Missing Values.

In general, however, we allow the observations to be continuous and/or discrete (ordered factors), and we also allow missing values. For discrete observations, the ranks in the empirical cdf contain ties that might be broken in some arbitrary way. However, the resulting cdf-mapping will *not* make the marginal densities uniform, since transformations of discrete random variables do not change the distribution, but affect only the sampling space.

In order to deal with such situations, it has been proposed to use the extended rank likelihood (Hoff, 2007). The elementary observation is that for non-decreasing marginal cdfs, $y_{i,j} < y_{k,j}$ implies $x_{i,j} < z_{k,j}$. For the entire set of observations $\mathbf{Y}$, this generalises such that $\mathbf{X}$ must lie in the set

$$\mathcal{D} = \left\{\mathbf{X} \in \mathbb{R}^{n\times p} : \max(x_{k,j} : y_{k,j} < y_{i,j}) < x_{i,j} < \min(x_{k,j} : y_{i,j} < y_{k,j})\right\} \tag{2.12}$$

This enables us to see the marginal cdfs F_j as nuisance parameters in the likelihood and hence to estimate the correlation matrix $\boldsymbol{\Sigma}$ on $\mathcal{D}$.

Bayesian inference for $\boldsymbol{\Sigma}$ includes an inverse-Wishart prior distribution $p(\boldsymbol{\Sigma}) \sim \mathcal{W}^{-1}(\boldsymbol{\nu}_0, \nu_0\mathbf{V}_0)$, with degrees of freedom v_0 and scale $\mathbf{V}_0$. It can be achieved by constructing a Markov chain having its stationary distribution at $\boldsymbol{\Sigma}$'s posterior distribution $p(\boldsymbol{\Sigma}|\mathbf{X} \in \mathcal{D}) \propto p(\boldsymbol{\Sigma})p(\mathbf{X} \in$

$\mathcal{D}|\mathbf{\Sigma}$). Sampling is done in a Gibbs fashion, alternating between $\mathbf{X}|\mathbf{\Sigma}, \mathbf{Y}$ and $\mathbf{\Sigma}|\mathbf{X}$, as outlined in Alg. 5.

Resampling the latent variable $\mathbf{X}|\mathbf{Y}, \mathbf{\Sigma}$ corresponds to sampling from a truncated normal

$$x_{i,j} \sim \mathcal{N}_{\text{trunc}}\left(\mu_{i,j}, \sigma_j^2, lo, up\right), \tag{2.13}$$

where the lower truncation $lo = \max(x_{i,j} : y_{i,j} < \text{unique}(y_{n,j}, \ldots, y_{n,j})$ and the upper truncation $up = \min(x_{i,j} : y_{i,j} > \text{unique}(y_{n,j}, \ldots, y_{n,j})$ are determined by the set $\mathcal{D}$ in Eq. 2.12. Thereby, the mean $\mu_{i,j} = X_{i,-j}\left(\Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\right)^{\mathsf{T}}$ and the variance $\sigma_{i,j}^2 = \Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}$ are conditioned on the remaining variables.

Resampling the correlation matrix $\mathbf{\Sigma}|\mathbf{X}$ means drawing from the inverse-Wishart, augmented with the data term $\mathbf{X}^{\mathsf{T}}\mathbf{X}$

$$\mathbf{\Sigma} \sim \mathcal{W}^{-1}\left(v_0 + n, v_o\mathbf{V}_o + \mathbf{X}^{\mathsf{T}}\mathbf{X}\right). \tag{2.14}$$

In order to accommodate for missing values $y_{i,j}$, the step in Eq. 2.13 is adjusted to use the unconstrained (i.e. untruncated) normal distribution.

Now, in every Gibbs iteration, we run an existing algorithm for drawing a set archetypes. For every object $\boldsymbol{x}$, we update a score $S(\mathbf{X})$, measuring the average proximity to the closest archetypes. Clustering of the score landscape and, within each cluster, selecting the objects with the highest score, finalises the algorithm. An example is given in Fig. 2.4.

Algorithm 5 Copula archetypal analysis for mixed observations and missing values.

Require: Observations $\mathbf{Y}$
Result: Archetypes $\mathbf{Z}$
Initialise: $(\mathbf{X}, \mathbf{\Sigma})$
for N Gibbs sweeps **do**
  for all observations **do**
    for all dimensions **do**
      conditioned on $(\mathbf{Y}, \mathbf{X})$, compute bounds $\{lo, up\}$
      conditioned on $(\mathbf{\Sigma}, \mathbf{X})$, compute conditional mean $\mu_{i,j}$ and conditional variance $\sigma_{i,j}^2$
      draw $x_{i,j} \sim N_{\text{trunc}}(\mu_{i,j}, \sigma_{i,j}^2, lo, up)$ from truncated normal;
    end for
  end for
  conditioned on $\mathbf{X}$, draw $\mathbf{\Sigma}$ from inverse Wishart
  $A \leftarrow$ `FindArchetypes(X)`
  update average archetypal scores in $S(X)$
end for
find clusters in set $\{x|S(x) > 0\}$
return in every cluster the object with highest score S

2.4. Motivation for Gaussian Copula

A parametric copula C is used in order to define a likelihood $l(\theta; \{\boldsymbol{y}_i\}_{i=1}^n)$ which makes it possible to estimate the latent vectors $\boldsymbol{x}_i$. Subsequently these are used as input for classical archetype reconstruction. A particularly simple choice of a dependency structure is a Gaussian copula model C_Σ which inherently implies a latent space by transforming $\tilde{X}_j = \Phi^{-1}(U_j)$ with the standard normal inverse cdf, i.e. the quantile function. The latent space $\tilde{\boldsymbol{X}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ is jointly normal distributed with zero mean and correlation $\boldsymbol{\Sigma}$. A probabilistic graphical model is given in Fig. 2.3, right panel. Clearly, the latent sample covariance

$$\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}} \sim \mathcal{W}_c(n, \boldsymbol{\Sigma}) \tag{2.15}$$

is central Wishart. In general, Gaussian copulas are very restrictive examples of copulas, in particular if a certain application domain requires proper modelling of tail-dependencies. For the purpose of reconstructing archetypes, however, the Gaussian copula is highly suited, because in the generative archetype model outlined in Eq. 2.9, the dependency structure is indeed approximately Gaussian. To see this, it is useful to rewrite Eq. 2.9 in matrix form:

$$\mathbf{X}|\mathbf{Z}, \mathbf{A} \sim \mathcal{MN}(\mathbf{A}\mathbf{Z}_{K\times p}, \mathbf{I}_n, \eta\mathbf{I}_p), \tag{2.16}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ contains the observations $\boldsymbol{x}_i$ as rows, $\mathbf{A} \in \mathbb{R}^{n\times K}$ contains the mixing components $\boldsymbol{\alpha}$ as rows and $\mathcal{MN}(\mathbf{M}, \boldsymbol{\Omega}, \boldsymbol{\Sigma})$ denotes the matrix normal distribution with mean matrix $\mathbf{M}$, row covariance $\boldsymbol{\Omega}$ and column covariance $\boldsymbol{\Sigma}$. Since in Eq. 2.9, the individual components of $\boldsymbol{x}$ are independent given the means, we might say that the means $\mathbf{M}$ capture the full dependency structure of $\boldsymbol{x}$. This interpretation can be formally expressed by analysing the covariance structure of the observations $\boldsymbol{x}_i$. Since $\mathbf{X}$ follows the matrix normal distribution from Eq. 2.16, it follows that the sample covariance $\mathbf{S} = \mathbf{X}^\mathsf{T}\mathbf{X}$ is non-central Wishart with non-centrality matrix $\eta^{-1}\mathbf{M}^\mathsf{T}\mathbf{M}$, where $\mathbf{M} = \mathbf{A}\mathbf{Z}$, i.e.

$$\mathbf{S}|\mathbf{Z}, \mathbf{A} \sim \mathcal{W}_{nc}\left(n, \eta\mathbf{I}_p, \eta^{-1}\mathbf{M}^\mathsf{T}\mathbf{M}\right). \tag{2.17}$$

Moreover, the non-central Wishart distribution can be approximated with a central Wishart, because the first order moments of the distribution are identical and the second order moments differ in terms of order $\mathcal{O}(n^{-1})$ only. This approximation, derived via moment matching (Steyn and Roux, 1972), allows to write

$$\mathbf{S}|\mathbf{Z}, \mathbf{A} \sim \mathcal{W}_{nc}(n, \eta\mathbf{I}_p, \eta^{-1}\mathbf{M}^\mathsf{T}\mathbf{M}) \approx \mathcal{W}_c(n, \frac{1}{n}\mathbf{M}^\mathsf{T}\mathbf{M} + \eta\mathbf{I}_p) \tag{2.18}$$

Comparing Eq. 2.15 with Eq. 2.18 shows that under the generative archetype model in Eq. 2.9, the covariance structure of the observed vectors $\boldsymbol{x}$ is approximately Gaussian which, in turn, formally justifies the use of a Gaussian copula model for estimating the latent space.

2.5. Demo-Application in Computational Biology

We applied the Copula archetype model to analyse the genetic stress response induced by heat shock in *Saccharomyces cerevisiae* (yeast). Two different information sources are used: (i) time-resolved gene expression measurements of yeast genes under heat shock conditions, i.e. temporal changes in the process of synthesizing gene products under heat stress. (ii) Binding affinity scores for certain stress-related transcription factors. A transcription factor (TF) is a protein that binds to DNA sequences near genes and regulates gene expression. The

first dataset has been published in (Gasch et al., 2000) and can be downloaded from their web supplement, the second one refers to (Harbison et al., 2004) and can be downloaded at `http://fraenkel.mit.edu/Harbison/release_v24/` as p-values for TF binding events. Probe names in this dataset are matched to genes in order to combine the TF data with the gene expression data. The p-values are exponentially transformed to a binding affinity score on $[0, 1]$ such that the upper half of the unit interval is associated with highly significant bindings with $p < 5 \cdot 10^{-3}$. Combination of both datasets leads to a 10-dimensional description of 6105 yeast genes, expression values at 4 different time points and binding affinities to the 6 transcription factors ADR1, GAT1, HSF1, MSN2, SKN7, YAP1.

In the context of archetypal analysis, we look for a few genes that show prominent expression/binding patterns that explain all observed patterns as convex mixtures in the latent copula space. Since roughly 13% of the genes contain missing values in one or more dimensions, we use the Gibbs-sampling strategy in Alg. 5 for inferring archetypal genes. Fig. 2.5 summarises the result of this analysis. Copula archetypal analysis identified 6 archetypal gene clusters that roughly correspond to the following patterns. *Stress response* (genes near the green diamond): these are known heat-stress response genes, they are highly over-expressed and have high binding affinity to the stress-related transcription factor SKN7 which is one of the two major transcriptional stress-response regulators in yeast. *Ribosomal RNA processing* (red): these genes play an essential role in protein synthesis. As expected, they are down-regulated under heat stress, and this regulatory process is mediated by binding to YAP1 which is the second major regulator, cf. (Lee et al., 1999). Two archetypes, depicted by the magenta and blue diamond, represent genes with mainly catalytic function that are regulated by exactly these two different stress response regulons, and two further archetypes (cyan and yellow diamond) have opposite binding affinity to the transcriptional activator ADR1. For further details see Fig. 2.5. Note that our findings nicely corroborate the results in (Shoval et al., 2012), where essentially the same major groups of archetypal genes have been identified under environmental stress conditions but in a different organism. Classical archetypal analysis has severe problems on this dataset: first, genes with missing values have to be removed, and second, several archetypes that have a clear biological interpretation (like the magenta one) could not be found by the classical algorithm, see Fig. 2.6.

2.6. Conclusion

We introduced copula archetypal analysis which wraps classical archetypal analysis into a copula framework. This ensures invariance of archetypal analysis against the class of strictly monotone increasing functions. We think, this is the largest invariance class since it only keeps the rank relation of the data, while the representation of the data can change. Furthermore, we devised the possibility to include mixed data and missing values. This is an important property, since in many real world datasets, mixed data and missing values are very common. Moreover, our algorithm is formulated as a preprocessing step, such that established algorithms can be re-used in order to efficiently recover the archetypes. Lastly, we have demonstrated that our model works well on both simulated data and in a real world application.

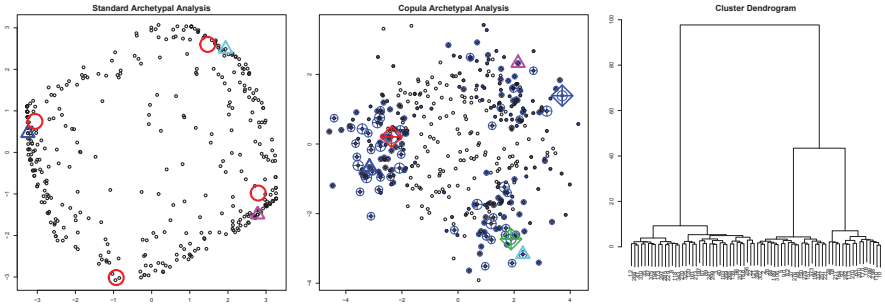


Figure 2.4.: Copula archetypal analysis for general case with discrete data. Left: 400 data points sampled as (noisy) convex mixtures of 3 archetypes in 10 dimensions, monotonically transformed (beta marginal densities) and linearly quantised into 10 levels. Shown is the projection on the first two principal components, the reconstructed archetypes (red circles) and the original archetypes after transformation (triangles). Middle: reconstruction with copula archetypal analysis. The size of the blue circles indicates the archetype score for each data point. Points with a non-zero archetype score are hierarchically clustered. The coloured diamonds show the highest-scoring data point in every cluster found by cutting the dendrogram in the right panel.

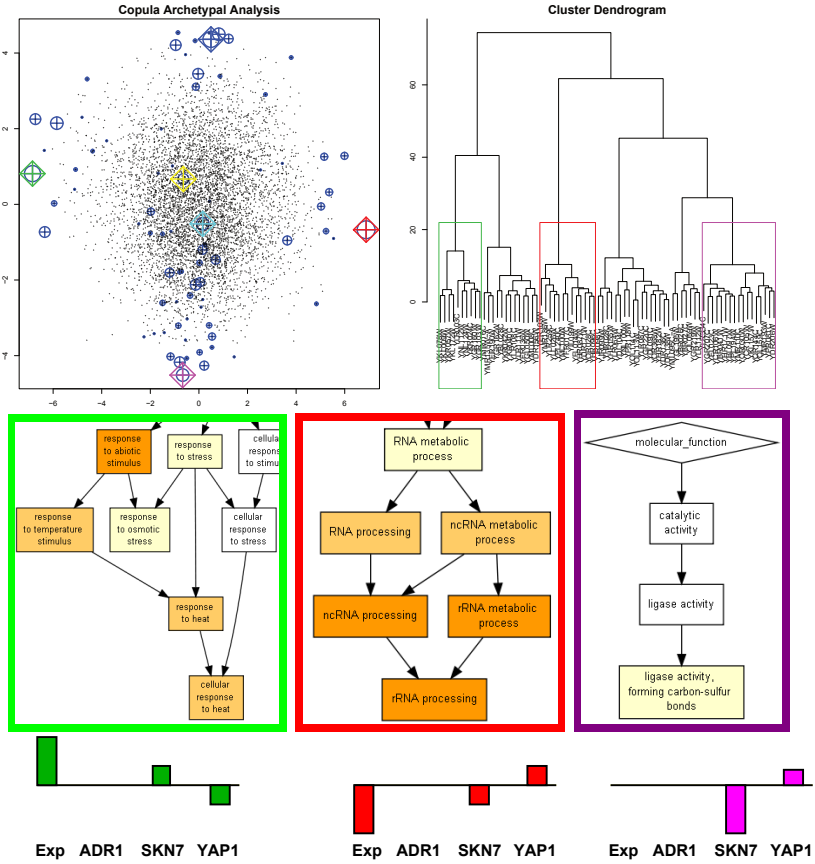


Figure 2.5.: Yeast genes under heat stress, characterised by gene expression values and binding affinity scores to stress-related transcription factors. Top left: PCA-plot of archetype reconstruction with our copula model. Coloured diamonds show genes with highest archetype score in each of the clusters found by cutting the dendrogram in the top right panel (the boxes indicate clusters with significantly enriched gene functions represented by *Gene Ontology* (GO) terms). Middle row: enrichment analysis of genes in the cluster indicated by the green, red, and magenta boxes in the dendrogram, computed with the GOrilla software `http://cbl-gorilla.cs.technion.ac.il/`. Color encodes p-values of enriched GO-term: yellow $= 10^{-3}$ to 10^{-5}, orange $= 10^{-5}$ to 10^{-7}, dark-orange $= 10^{-7}$ to 10^{-9}. Bottom row: archetype-specific gene-expression and binding pattern (schematic).

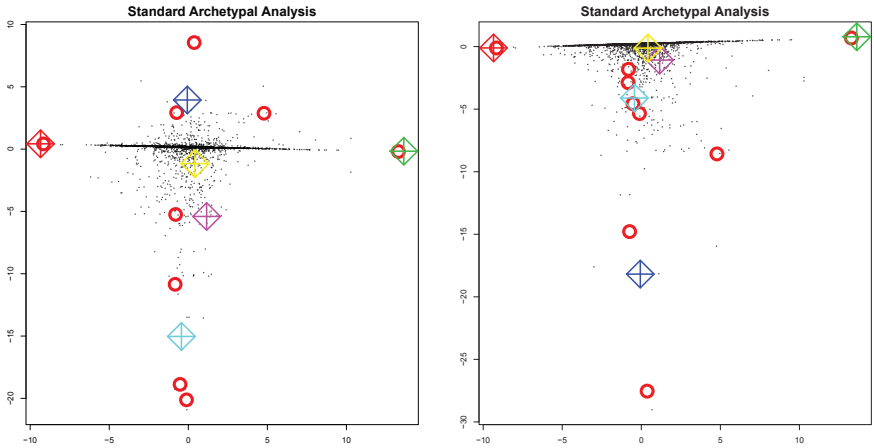


Figure 2.6.: Comparison of standard and copula archetypal analysis. Left: 1st principal component (PC) vs. 2nd one, Right: 1st PC vs. 3rd PC. The red circles indicate the location of the archetypes. For comparison, the coloured diamonds show the archetypal objects identified by copula archetypal analysis. The projection in the right panel reveals that there is no archetype in close proximity of the magenta- and yellow-coloured diamonds.

3. Copula Eigenfaces

In computer vision and graphics, principal component analysis (PCA) is a ubiquitous method for describing dependency and variance in a data set (Jolliffe, 2002). The probabilistic formulation of PCA (Tipping and Bishop, 1999) assumes that the observed data is Gaussian distributed. We show that this assumption is not fulfilled in the context of facial appearance. The model mismatch leads to unnatural artefacts which are severe to human perception. In order to prevent these artefacts, we propose to use a semi-parametric Gaussian copula for modelling the colour distribution of faces. Our extension relaxes the Gaussian assumption of the PCA model and allows us to apply the model to non-Gaussian distributed data.

Remark

The sequel closely follows Egger et al. (2016).

3.1. Introduction

Parametric Appearance Models (PAM) describe objects in an image in terms of pixel intensities. In the context of faces, the Eigenfaces approach (Sirovich and Kirby, 1987; Turk et al., 1991) is a PAM of high renown. The approach uses aligned facial images to analyse and synthesize faces. Since then, many advances were introduced to bring forward facial PAMs. Among those, Active Appearance Models (Cootes et al., 1998) add a shape component which allows to model the shape independently from the appearance. 3D Morphable Models (Blanz and Vetter, 1999) uses a dense registration and extend the shape model to 3D and add camera and illumination parameters. 3D Morphable Models allow handling appearance independently from pose, illumination and shape. Photo-realistic face synthesis methods like Visio-lization (Mohammed et al., 2009) uses PCA as a basis for example-based photo-realistic appearance modelling. All these models are established PAMs that have a common core: the dominant method for learning the parameters of a PAM is principal component analysis (Jolliffe, 2002). PCA is used to describe the variance and dependency in the data.

In general, a PAM which uses PCA is a generative model that is also able to synthesize new random instances. However, the ability to synthesize *natural* random instances is a challenging task, cf. (Mohammed et al., 2009) and in face manipulation (Walker and Vetter, 2009). This is because human perception is very sensitive to unnatural variability in a face. On the other hand, PCA face models are also used as strong priors in probabilistic facial image interpretation algorithms (Schönborn et al., 2013). Hence, such applications put high demands on PCA and require a prior which follows the underlying distribution as closely as possible and are highly specific to faces.

Using PCA to model facial appearance leads to models which synthesize instances which appear unnaturally. This is due to the assumption that the colour intensities or, in other words, the marginals at a pixel are assumed to be Gaussian distributed. We show that this is a severe simplification: the pixel intensities of new samples will follow a joint Gaussian distribution. This approximation is far from the actual observed distribution of the training data and leads to unnatural artefacts in appearance. In order to enhance the specificity of a PCA-based model, an obvious improvement would be the extension to a Gaussian mixture model (Rasmussen, 1999). Here, each colour channel at a pixel is modelled with an (infinite) mixture of Gaussians. However, we skip this step and propose to use a semi-parametric copula model directly. A copula model provides the decomposition of the dependency and the marginal distributions such that the copula contains the dependency structure only. This separate modelling allows us to drop the parametric Gaussian assumption on the colour channels and to replace them with non-parametric empirical distributions. We keep the parametric dependency structure; in particular, we use a Gaussian copula because of its inherent Gaussian latent space. PCA can then be applied in the latent Gaussian space and is used to learn the dependencies of the data independently from the marginal distribution.

A semi-parametric Gaussian copula model provides us with a new flexibility which results in more natural images generated images. Using the model provides us with additional benefits:

1. Learning is robust to outliers and scale-invariant. Moreover, the scale-invariance makes possible a principled way of combining facial appearance and shape data.
2. Implementation in existing pipelines is simple: the additional overhead for using the semi-parametric copula model can be implemented as simple pre- and post-processing steps of PCA. The copula model transforms the non-Gaussian data into a latent space which is Gaussian distributed. PCA is then performed on the transformed data. Generating data is possible by simply reversing this pipeline.
3. The model also allows for changing the colour space. For facial appearance modelling, the HSV colour space is more appropriate than RGB. The HSV colour space is motivated by the separation of the hue and saturation components and brightness value. On the other hand, without adaptation, PCA is not applicable to facial appearance in the HSV colour space because of its sensitivity to differently-scaled colour channels.

In summary, those benefits are general and all PAMs on the basis of a PCA model can easily benefit from these advantages to improve their performance.

In the remainder, we discuss the model and the benefits: The next section explains the copula extension for PCA and presents the theoretical background and practical information for an implementation is provided. Finally, we show qualitatively and quantitatively that the proposed model leads to a facial appearance model which is more specific to faces.

3.2. Methods

3.2.1. PCA for Facial Appearance Modelling

Let $x \in \mathbb{R}^{3n}$ describe a zero-mean vector representing 3 colour channels of an image with n pixels. In an RGB image, the colour channels and the pixels are stacked such that $x = (r_1, g_1, b_1, r_2, b_2, b_3, \ldots, r_n, g_n, b_n)^T$. We assume that the mean of every dimension is already subtracted. The training set of m images is arranged as the data matrix $X \in \mathbb{R}^{3n \times m}$.

PCA (Jolliffe, 2002) aims at diagonalizing the sample covariance $\Sigma = \frac{1}{m} X X^T$, such that

$$\Sigma = \frac{1}{m} U S^2 U^T \tag{3.1}$$

where S is a diagonal matrix and U contains the transformation to the new basis. The columns of matrix U are the eigenvectors of Σ and the corresponding eigenvalues are on the diagonal of S.

PCA is usually computed by a singular value decomposition (SVD). In case of a rank-deficient sample covariance with rank $m < n$ we cannot calculate U^{-1}. Therefore, SVD leads to a compressed representation with a maximum of m dimensions. The eigenvectors in the transformation matrix U are ordered by the magnitude of the corresponding eigenvalues.

When computing PCA, the principal components are guided by the variance as well as the covariance in the data. While the variance captures the scattering of the intensity value of a pixel, the covariance describes which regions contain similar colour. This mingling of factors leads to results which are sensitive to different scales and to outliers in the training set. Regions with large variance and outliers could influence the direction of the resulting principal components in an undesired manner.

We uncouple variance and dependency structure such that PCA is only influenced by the dependency in the data. Our approach for uncoupling is a copula model which provides an analytical decomposition of the aforementioned factors.

3.2.2. Copula Extension

Copulas (Nelsen, 2013; Joe, 1997) allow a detached analysis of the marginals and the dependency pattern for facial appearance models. We consider a relaxation to a semi-parametric Gaussian copula model (Genest et al., 1995; Tsukahara, 2005). We keep the Gaussian copula for describing the dependency pattern, but we allow non-parametric marginals.

Let $x \in \mathbb{R}^{3n}$ describe the same zero-mean vector as used for PCA, representing 3 colour channels of an image with n pixels. Sklar's theorem allows the decomposition of every continuous and multivariate cumulative probability distribution (cdf) into its marginals $F_i(X_i), i = 1, \ldots, 3n$ and a copula C. The copula comprises the dependency structure, such that

$$F(X_1, \cdots, X_{3n}) = C(W_1, \ldots, W_{3n}) \tag{3.2}$$

where $W_i = F_i(X_i)$. W_i are uniformly distributed and generated by the probability integral transformation[1].

For our application, we consider the Gaussian copula because of its inherently implied latent space

$$\tilde{X}_i = \Phi^{-1}(W_i), \quad i = 1, \ldots, 3n \tag{3.3}$$

where Φ is the standard normal cdf. The multivariate latent space is standard normal-distributed and fully parametrized by the sample correlation matrix $\tilde{\Sigma} = \frac{1}{m} \tilde{X} \tilde{X}^T$ only. PCA is then applied on the sample correlation in the latent space $\tilde{X}$. Such a model is analytically analysed in (Han and Liu, 2012) and is called Copula Component Analysis (COCA).

The separation of dependency pattern and marginals provides multiple benefits: First, the Gaussian copula captures the dependency pattern invariant to the variance of the colour

[1] The copula literature uses U instead of W. We changed this convention due to the singular value decomposition which uses $X = USV^T$.

space[2]. Second, whilst PCA is distorted by outliers and is generally inconsistent in high dimensions, the semi-parametric copula extension solves this problem (Han and Liu, 2012). Third, the non-parametric marginals maintain the non-Gaussian nature of the colour distribution. Especially when generating new samples from the trained distribution, the samples do not exceed the colour space of the training set.

3.2.3. Inference

We learn the latent sample correlation matrix $\tilde{\Sigma} = \frac{1}{m}\tilde{X}\tilde{X}^T$ in a semi-parametric fashion using non-parametric marginals and a parametric Gaussian copula. Compared to Han and Liu (2012), we use the Gaussian rank correlation to estimate latent sample correlation matrix. Thus, we compute $\hat{w}_{ij} = \hat{F}_{\text{emp},i}(x_{ij}) = \frac{r_{ij}(x_{ij})}{m+1}$ using empirical marginals $\hat{F}_{\text{emp},i}$, where $r_{ij}(x_{ij})$ is the rank of the data x_{ij} within the set $\{x_{i\bullet}\}$. Then, $\tilde{\Sigma}$ is simply the sample covariance of the normal scores

$$\tilde{x}_{ij} = \Phi^{-1}\left(\frac{r_{ij}(x_{ij})}{m+1}\right), \quad i = 1,\dots,3n, \quad j = 1,\dots,m. \tag{3.4}$$

Equation (3.4) contains the non-parametric part, since $\tilde{\Sigma}$ is computed from the ranks $r_{ij}(x_{ij})$ solely and contains no information about the marginal distribution of the x's. Note, $\tilde{x} \sim \mathcal{N}(0, \tilde{\Sigma})$ is standard normal distributed with correlation matrix $\tilde{\Sigma}$. Subsequently, an eigendecomposition is applied on the latent correlation matrix $\tilde{\Sigma}$.

Algorithm 6 Learning.

Require: Training set $\{X\}$
Result: Projection matrices U, S
1: **for** all dimensions **do**
2: **for** all samples **do**
3: $\tilde{x}_{ij} = \Phi^{-1}\left(\frac{r_{ij}(x_{ij})}{m+1}\right)$
4: **end for**
5: **end for**
6: find $\tilde{U}, \tilde{S}$ such that $\tilde{\Sigma} = \frac{1}{m}\tilde{U}\tilde{S}^2\tilde{U}^T$ (via SVD)

Generating a sample using PCA then simply requires a sample from the model parameters

$$h \sim \mathcal{N}(0, I) \tag{3.5}$$

which is projected to the latent space

$$\tilde{x} = \tilde{U}\frac{\tilde{S}}{\sqrt{m}}h \tag{3.6}$$

and further transformed component-wise to

$$w_i = \Phi(\tilde{x}_i), \quad i = 1,\dots,3n. \tag{3.7}$$

[2]More general, a copula model is invariant against all monotonic transformations of the marginals.

Algorithm 7 Sampling.

Result: Random sample x

1: $h \sim \mathcal{N}(0, I)$
2: $\tilde{x} = \tilde{U} \frac{\tilde{S}}{\sqrt{m}} h$
3: **for** all dimensions i **do**
4: $\quad w_i = \Phi(\tilde{x}_i)$
5: $\quad x_i = \hat{F}_{\text{emp},i}(w_i)$
6: **end for**

Finally, the projection to the colour space requires the empirical marginals

$$x_i = \hat{F}_{\text{emp},i}(w_i), \quad i = 1, \ldots, 3n. \tag{3.8}$$

It is also possible to smoothen the empirical marginals with a kernel k and replace Equation (3.8) by $x_i = k(w_i, X_{i\bullet}), \quad i = 1, \ldots, 3n$. All necessary steps are summarized in Algorithms 6 and 7.

3.2.4. Implementation

The additional effort for using COCA can be implemented as simple pre- and post-processing steps to PCA. Basically the data is mapped into a latent space where it is Gaussian distributed. The mapping is performed in two steps. First, the data is transformed to an uniform distribution by ranking the intensity values. Then it is transformed to a standard normal distribution. On the transformed data, we perform PCA to learn the dependency structure in the data. To generate new instances from the model, all steps have to be reversed. The following listings give an overview of all necessary adaptations to an existing PCA pipeline in MATLAB.

```
% calculate empirical cdf
[empCDFs, indexX] = sort(X, 2);

% transform emp. cdf to uniform
[~, rank] = sort(indexX, 2);
uniformCDFs = rank / (size(rank, 2)+1);

% transform uni. cdf to std. normal cdf
normCDFs = norminv(uniformCDFs',0,1)';

% calculate PCA
[U,S,V] = svd(normCDFs, 'econ');
```

Listing 3.1: Learning

To generate an image from model parameters, the following steps are necessary:

```
% random sample
m = size(normCDFs, 2);
h = random('norm' ,0 ,1 ,m ,1);
sample = U * S / sqrt(m) * h;

% std. normal to uniform
uniformSample = normcdf(sample, 0, 1) * (m - 1) + 1;

% uniform to emp. cdf
empSample = empCDFs(sub2ind(size(empCDFs), 1:size(data, 1), ...
    round(uniformSample')))';
```

Listing 3.2: Sampling

These are the additional steps which have to be performed as pre- and post-processing for the analysis of the data and the synthesis of new random samples. In terms of computing resources we have to consider the following: The empirical marginal distributions F_{emp} are now part of the model and have to be kept in memory. In the learning part, the complexity of sorting the input data is added. In the sampling part, we have to transform the data back by looking up their values in the empirical distribution.

The copula extension comes with low additional effort: it is easy to implement and has only slightly higher computing costs. We encourage the reader to implement these few steps since the increased flexibility in the modeling provides a valuable extension.

3.3. Experiments and Results

For all our experiments, we used the texture of 200 face scans used for building the Basel Face Model (BFM) (Paysan et al., 2009). The scans are in dense correspondence and were captured under an identical illumination setting. We work on texture images and use a resolution of 1024x512 pixels. Our experiments are based on the appearance information only, the last experiment merging the appearance and shape to a combined model. We used the empirical data directly as marginal distribution. The results are rendered with an ambient illumination on the mean face shape of the BFM.

3.3.1. Facial Appearance Distribution

In a first experiment we investigate if the colour intensities in our face data set are Gaussian distributed. We followed the protocol of the Kolmogorov-Smirnov Test (Massey Jr, 1951). We estimate a Gaussian distribution for every colour channel per pixel and compare it to the observed data. The null hypothesis of the test is that the observed data is drawn by the estimated Gaussian distribution. The test measures the maximum distance of the cumulative density function of the estimated Gaussian $\Phi_{\hat{\mu},\hat{\sigma}^2}$ and the empirical marginal distribution F_{emp} of the observed data:

$$d = \sup_x \left\| F_{\text{emp}}(x) - \Phi_{\hat{\mu},\hat{\sigma}^2}(x) \right\| \tag{3.9}$$

Here, $\hat{\mu}$ and $\hat{\sigma}^2$ are maximum-likelihood estimates for the mean and variance of a Gaussian distribution respectively. In Figure 3.1 we visualize the maximal distance value over all colour

channels per point on the surface.

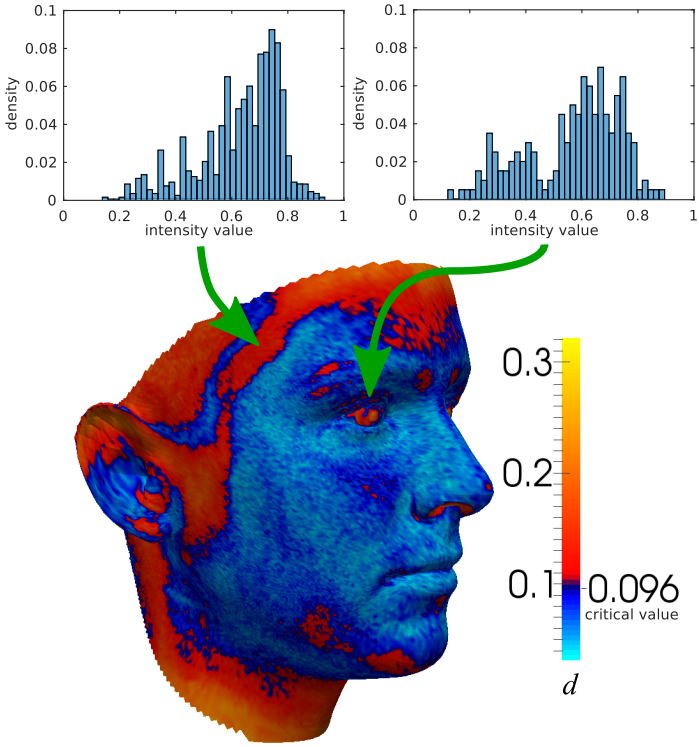


Figure 3.1.: The result of the Kolmogorov-Smirnov Test to compare the empirical marginal distribution of colour values from our 200 face scans with a Gaussian-reference probability distribution. We plot the highest value of the three colour channels per pixel, because the values for the individual colour channels are very similar. We show two exemplary marginal distributions in the eye and temple region. They are not only non-Gaussian but also not similar.

We assume a significance level of $1 - \alpha = 0.05$. The critical value d_α is approximated using the following formula (Lothar Sachs, 2006):

$$d_\alpha = \frac{\sqrt{\ln(\frac{2}{\alpha})}}{\sqrt{2n}} \tag{3.10}$$

With $n = 200$ training samples we get a critical value of 0.096. Non-Gaussian marginal distributions of colour intensities are present in the region of the eyebrows, eyes, chin and

hair, where multi-modal appearance is present. In total for 49% of the pixels over all colour channels, the null hypothesis has to be rejected. In simple monotonic regions, like the cheek, the marginal distributions are close to a Gaussian distribution. In more structured regions like the eye, eyebrow or the temple region, the appearance is highly non-Gaussian. This leads to strong artefacts when modelling facial colour appearance using PCA (see Figure 3.2 and Figure 3.3). Since those more structured regions are fundamental components of a face, it is important to model them properly.

3.3.2. Appearance Modelling

We evaluate our facial appearance model by its capability to synthesize new instances. We measured this capability by comparing the major eigenmodes, random model instances, the sample marginal distributions and the specificity of both models. The specificity is measured qualitatively by visual examples and quantitatively by a model metric.

Model Parameters

The first few principal components store the strongest dependencies. We visualize the first two components by setting their value h_i to $\sigma = 3$ standard deviations and show the result in Figure 3.2. The first parameters of PCA and COCA appear very similar in the variation of the data they model. The second principal component of PCA causes artefacts in the temple region. These artefacts are caused by the linearity of PCA. COCA is a non-linear method and therefore, the artefacts are not present.

Random Samples

The ability to generate new instances is a key feature for generative models. A model which can produce more realistic samples is desirable for various applications. For example, the Visio-lization method to generate high resolution appearances is based on a prototype generated with PCA (Mohammed et al., 2009).

Another field of application for the generative part of models are Analysis-by-Synthesis methods based on Active Appearance Models (AAM) or 3D Morphable Models (3DMM). They can profit from a stronger prior which is more specific to faces and reduces the search space (Schönborn et al., 2013).

Generating a random parameter vector leads to a random face from our PCA or COCA model. We sample h according to Equation (3.5) independently for all 199 parameters and project them via PCA or COCA on the colour space following Equation 3.6. Random samples using COCA contain fewer artefacts and, therefore, appear much more natural (see Figure 3.3). These artefacts are caused by the linearity of PCA. For non-Gaussian distributed marginals, PCA does not only interpolate within the trained colour distribution but also extrapolates to colour intensities not supported by the training data.

The most obvious problem is the limited domain of the colour channels: using PCA, colour channels have to be clamped. The linearity constraint of PCA leads to much brighter or darker colour appearance than those present in the training data in regions which are not Gaussian distributed. In the next experiment, we show that the higher specificity is not only a qualitative result but can also be measured by a model metric.

Few samples of COCA contain artefacts arising from outliers in the training data which appear at the borders of the empirical cdfs. Those artefacts can be removed by slightly

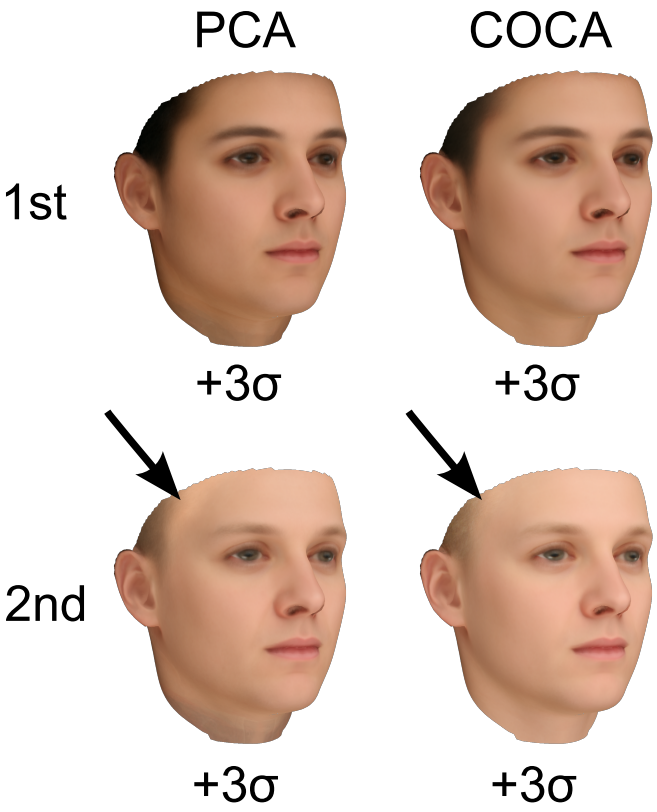


Figure 3.2.: PCA and COCA are compared by visualizing the first two eigenvectors with 3 standard deviations on the mean. The components look very similar, except that the PCA artefacts on the temple (arrows) in the second eigenvector do not appear using COCA.

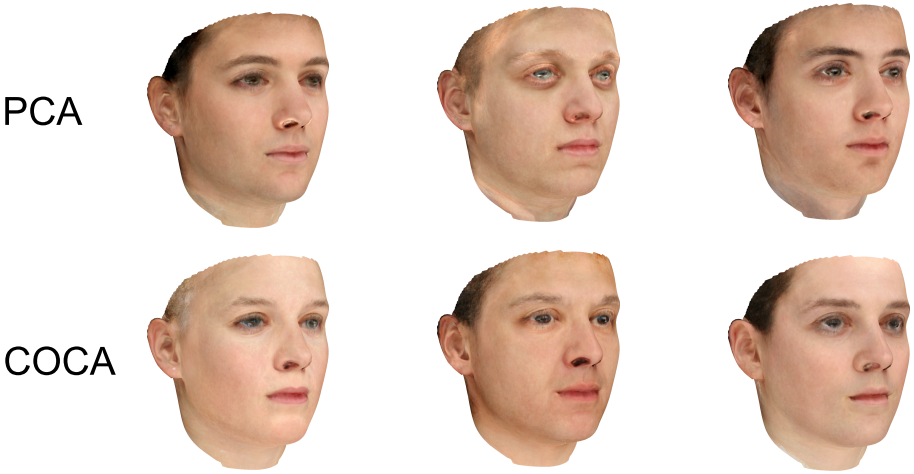
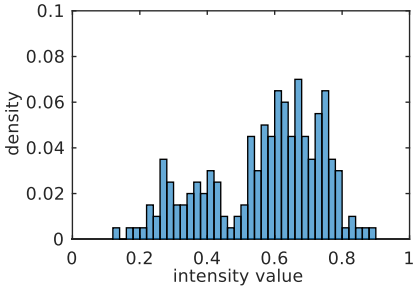


Figure 3.3.: The first and second row show random samples projected by PCA and COCA respectively. Using PCA, we can observe strong artefacts in the regions where the marginal distribution is not Gaussian (see Figure 3.1). The improvement of COCA can be observed in the temple region, on the eyebrows, around the nostrils, the eyelids and at the border of the pupil. We chose representative samples for both methods.

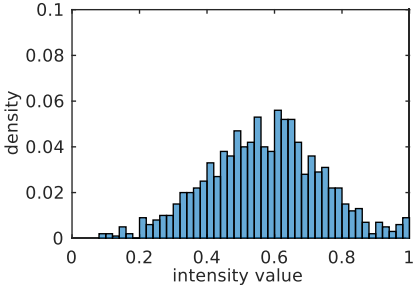
cropping the marginal distributions (removing the outliers) or by applying COCA in the HSV colour space.

3.3.3. Appearance Marginal Distribution

We analyse the marginal distributions of our random faces at a single point at the border between the pupil and the sclera of the eye. In this region the Kolmogorov-Smirnov Test rejected the null hypothesis. We show the empirical intensity distribution of a single colour channel at this point in Figure 3.4a. We compare this distribution to the sample marginal distribution of the PCA model in Figure 3.4b which was generated from 1000 random instances. Whilst COCA respects the empirical distribution, PCA is approximating a Gaussian distribution which is inaccurate in a lot of facial regions.



(a) Empirical marginal distribution



(b) PCA sample marginal distribution

Figure 3.4.: The marginal distribution of the red colour intensity of a single point in the eye region. (a) shows the distribution observed in the training data, (b) shows the distribution of samples drawn from a PCA model. This distribution shows a clear discrepancy to the true marginal distribution.

Specificity and Generalization

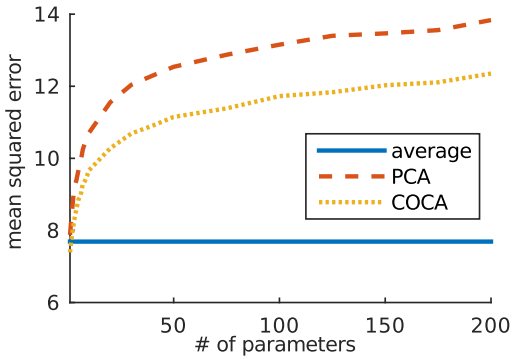


Figure 3.5.: The specificity shows how close generated instances are to instances in the training data. The average distance of 1000 random samples to the training set (mean squared error per pixel and colour channel) is shown. A model is more specific if the distance of the generated samples to the training set is smaller. We observe that COCA is more specific to faces (lower is better).

To measure the quality of the PCA and COCA models, we use model metrics motivated by the shape modelling community (Styner et al., 2003). The first metric is specificity: Instances generated by the model should be similar to instances in the training set. Therefore, we draw 1000 random samples from our model and compare each one to its nearest neighbour in the training data. We measure the distance using the mean absolute error over all pixels and colour channels in the RGB-colour space. The COCA model is more specific to facial appearance (see Figure 3.5). This corresponds to our observation of a more realistic facial appearance (Figure 3.3).

Specificity should always be used in combination with the generalization model metric (Styner et al., 2003). The generalization measures how exactly the model can represent unseen instances. We measure the generalization ability of both models using a test set and use the same distance measure as for specificity. The test data consists of 25 additional face scans not contained in the training data. We observe that both models generalize well to unseen data. PCA generalizes slightly better, see Figure 3.6.

The third model metric is compactness - the ability to use a minimal set of parameters (Styner et al., 2003). The compactness can be measured directly by the number of used parameters. In our experiments, the number of parameters is always the same for both models.

There is always a tradeoff between specificity and generalization. Whilst PCA performs slightly better in generalization, COCA performs better in terms of specificity. The better generalization ability of PCA comes at the price of a lower specificity and clearly visible artefacts.

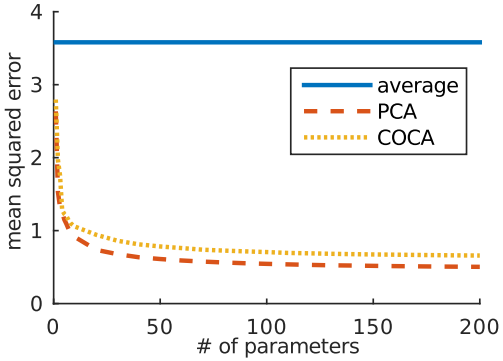


Figure 3.6.: The generalization ability shows how exactly unseen instances can be represented by a model. The lower the error, the better a model generalizes. As a baseline, we present the generalization ability of the average face. We observe that PCA generalizes slightly better (lower is better).

Combined Shape and Colour Model

Colour appearance and shape are modelled independently in AAMs and 3DMMs. Recently, it was demonstrated that facial shape and appearance are correlated (Schumacher and Blanz, 2015) and those correlations were investigated using Canonical Correlation Analysis on separate shape and appearance PCA models.

The main reason to build separate models is a practical one - shape and colour values are not in the same range. Some approaches accommodate this issue by normalization (Edwards et al., 1998). However, this approach is highly sensitive to outliers. Since Copula Component Analysis is scale invariant, we can directly apply it to the unscaled data.

We learned a COCA model combining the colour and shape information (see Figure 3.8 and Figure 3.7). Shape and texture vectors are combined by simply concatenating them. By integrating this additional dependency information, the model becomes more specific (Edwards et al., 1998).

As a future extension, COCA allows us to also integrate attributes like age, weight and size or even social attributes like trustworthiness or social competence directly into the model.

3.4. Conclusions

We showed that the marginal distribution of facial colour is not Gaussian distributed for large parts of the face and that PCA is not able to model facial appearance properly. In a statistical appearance model, this leads to unnatural artefacts which are easily detected by human perception. To avoid such artefacts, we propose to use PCA in a semi-parametric Gaussian copula model which allows to model the marginal colour distribution separately

Figure 3.7.: Random samples projected by a common shape and appearance model using COCA.

from the dependency structure. In this model, the parametric Gaussian copula describes the dependency pattern in the data and the non-parametric marginals relax the restrictive Gaussian requirement of the data distribution.

The separation of marginals and dependency pattern enhances the model flexibility. We showed qualitatively that COCA models facial appearance better than PCA. This finding is also supported by a quantitative evaluation using specificity as a model metric. Moreover, the COCA model enables to add further data to the model: Age, weight, size, and other data like social attributes living on different scales can be incorporated in the model in an unified way. To demonstrate this feature, we showed that the inclusion of shape also increased the specificity of the model.

The computer graphics and vision community is heavily modelling and working with colour intensities. We believe that these intensities are most often not Gaussian distributed and, therefore, our findings can be transferred to a lot of applications. Finally, we again want to encourage the reader to replace PCA with a COCA model, since the additional model flexibility comes with almost no implementation effort.

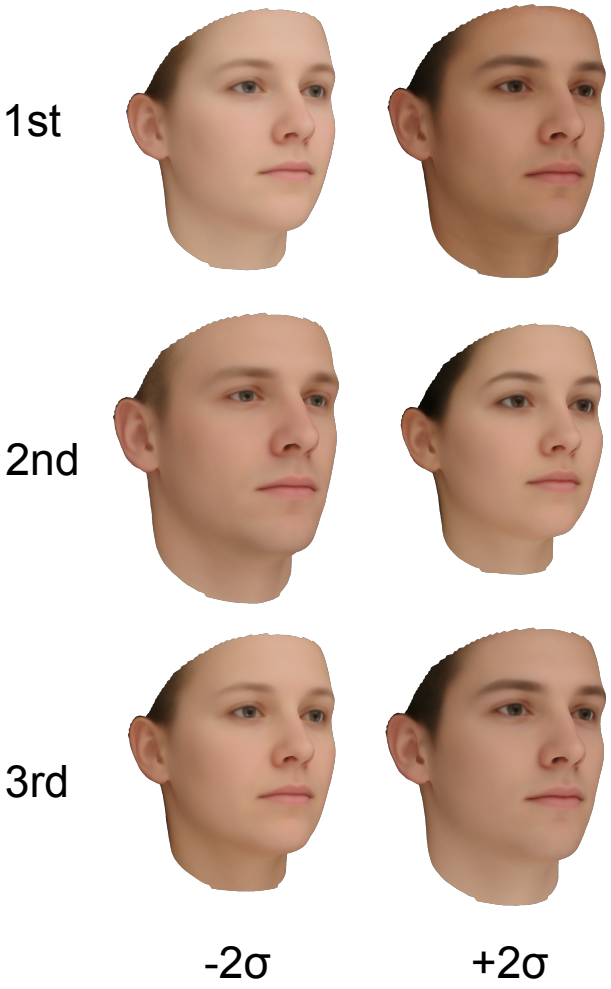


Figure 3.8.: We learned a common shape and appearance model using COCA. We visualize the first eigenvectors with 2 standard deviations which show the strongest dependencies in our training data. Whilst the first parameter is strongly dominated by appearance the later parameters are targeting shape and appearance. Since the model is built from 100 females and 100 males, the first components are strongly connected to sex.

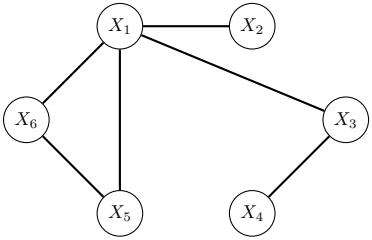
4. Bayesian Markov Blanket Estimation

Probabilistic graphical models (Koller and Friedman, 2009) are a ubiquitous tool for representing dependencies among several random variables. In a typical application, the network of dependencies is unknown and the goal is to identify the dependencies from observations. Modern data analysis scenarios are often confronted with high-dimensional data, where the number of observations n is typically in the same order of magnitude or less than the number of dimensions p. In this setting, identifying the full network can be difficult due to the stability and complexity of estimators, and it is important to introduce sparsity. In other cases, identifying the full network may be undesirable, since the number of variables is that high, such that relevant structures are blurred, and/ or when estimating the whole network is just irrelevant because one is not interested in parts of the network. In such situations it is advisable to reduce the focus on estimating a sub-network. In this part of the thesis, we are looking at undirected networks and focus on a specific sub-network, namely on the Markov blanket of a limited set of query variables. The goal is to identify the Markov blanket, i.e. the nodes among a large set of candidates which are the neighbours of the query variables.

We first give a short introduction into Gaussian graphical models and present the most related algorithms, namely the graphical lasso and its Bayesian interpretation. These algorithms are devised to estimate a full network. Subsequently, the model for estimating the Markov blanket is presented, including an extension to non-Gaussian data using the Gaussian copula. The chapter is then concluded with experiments on artificial and real data.

4.1. Gaussian Graphical Models

Estimating a network of dependencies among a set of objects is a difficult problem in statistics and machine learning, especially in high-dimensional settings or where the observed measurements are noisy. Gaussian graphical models are a ubiquitous tool for representing such relationships in an interpretable way. For a multivariate Gaussian distributed random vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^p$ parametrised with covariance $\boldsymbol{\Sigma}$, the zero pattern of the inverse covariance matrix $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$ encodes the conditional independences between the variables, i.e. if $w_{ij} = 0$, then $X_i \perp\!\!\!\perp X_j | X_{\setminus ij}$, and we say that X_i and X_j are conditionally independent given $X_{\setminus ij}$, where $X_{\setminus ij}$ denotes the set $\{X_k | k \neq i, j\}$. By way of illustration, we relate this correspondence directly to a graph $G = (V, E)$ with vertices $V = \{X_{[p]}\}$ and edge set $E = \{(X_i, X_j) | w_{ij} \neq 0\}$. We illustrate this relation in Fig. 4.1.



$$\mathbf{W} = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & w_{13} & 0 & w_{15} & w_{16} \\ w_{21} & w_{22} & 0 & 0 & 0 & 0 \\ w_{31} & 0 & w_{33} & w_{34} & 0 & 0 \\ 0 & 0 & w_{43} & w_{44} & 0 & 0 \\ w_{51} & 0 & 0 & 0 & w_{55} & w_{56} \\ w_{61} & 0 & 0 & 0 & w_{65} & w_{66} \end{pmatrix} \end{array}$$

Figure 4.1.: Graph $G = (V, E)$ and exemplary inverse covariance matrix.

4.2. Graphical Lasso and its Bayesian Formulation

In a typical application setting, the aim is to estimate the structure of graph G from n realisations of the random vector $\boldsymbol{X} \in \mathbb{R}^p$. The data consists of a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. If we have at $n \geq p$ independent observations and $\boldsymbol{x} \in \mathbb{R}^p$ is normally distributed, it can be shown that the sample covariance $\mathbf{S} = \mathbf{X}^\mathsf{T}\mathbf{X}$ follows a Wishart distribution $\mathbf{S} \sim \mathcal{W}_p(n, \boldsymbol{\Sigma})$, with n degrees of freedom, i.e.

$$\begin{aligned} p(\mathbf{S}|\boldsymbol{\Sigma}) &\propto \det(\boldsymbol{\Sigma})^{-\frac{n}{2}} \det(\mathbf{S})^{\frac{n-p-1}{2}} \exp \operatorname{tr}\left(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) \\ &= \det(\mathbf{W})^{\frac{n}{2}} \det(\mathbf{S})^{\frac{n-p-1}{2}} \exp \operatorname{tr}\left(-\frac{1}{2}\mathbf{W}\mathbf{S}\right). \end{aligned} \tag{4.1}$$

Whenever we are located in a high dimensional setting, i.e. the number of observations n are in the same order of magnitude or even less than the number of dimensions p, the sample covariance gets ill-conditioned and the maximum likelihood estimate of $\mathbf{W}$ becomes error-prone. Consequently, various estimators have been proposed that reduce the number of parameters by imposing sparsity constraints on $\mathbf{W}$. Among these, the popular graphical lasso procedure (Meinshausen and Bühlmann, 2006; d'Aspremont et al., 2008; Friedman et al., 2008; Banerjee et al., 2008; Hastie et al., 2015) computes a point estimate of the graph by minimizing the penalised log-likelihood

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \log\det(\mathbf{W}) - \operatorname{tr}\left(\bar{\mathbf{S}}\mathbf{W}\right) - \rho\|\mathbf{W}\|_1, \tag{4.2}$$

where $\rho \geq 0$ is a shrinkage parameter, and $\|\mathbf{W}\|_1 = \sum_{i \neq j} |\omega_{ij}|$ denotes the L_1 norm of the non-diagonal elements of $\mathbf{W}$. The L_1 norm penalty can be interpreted as a convex relaxation of a L_0 norm penalty, which is related to selection in a graphical model. Here $\bar{\mathbf{S}} = \frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}$ is a scaled sample covariance matrix[1]. Now, Eq. 4.2 is a convex problem and sub-gradient methods can be used to find the solution of the penalised log-likelihood. More precisely, the sub-gradient to Eq. 4.2 is

$$\mathbf{W}^{-1} - \bar{\mathbf{S}} - \rho\boldsymbol{\Psi} = \mathbf{0}, \tag{4.3}$$

where $\boldsymbol{\Psi}$ is a symmetric matrix having diagonal elements equal to zero, and $\psi_{ij} = \operatorname{sgn}(w_{ij})$ if $w_{ij} \neq 0$, and $\psi_{ij} \in [-1, 1]$ if $w_{ij} = 0$. Minimising the likelihood can be done with block-wise

[1] To remain as close as possible to the literature, we introduce the scaled sample covariance $\bar{\mathbf{S}} = \frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}$ with a bar, while the unscaled sample covariance $\mathbf{S} = \mathbf{X}^\mathsf{T}\mathbf{X}$ is denoted without bar.

coordinate descent, where block $k \in \{1, \ldots, p\}$ is established by fixing all but row and column k. For $k = 1$ the partitioning of the matrices is

$$\bar{\mathbf{S}} = \begin{matrix} \\ 1 \\ p-1 \end{matrix} \begin{matrix} \begin{matrix} 1 & p-1 \end{matrix} \\ \begin{pmatrix} \bar{s}_{11} & \bar{\boldsymbol{s}}_{12} \\ \bar{\boldsymbol{s}}_{21} & \bar{\mathbf{S}}_{22} \end{pmatrix} \end{matrix}, \quad \mathbf{W} = \begin{matrix} \\ 1 \\ p-1 \end{matrix} \begin{matrix} \begin{matrix} 1 & p-1 \end{matrix} \\ \begin{pmatrix} w_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{21} & \mathbf{W}_{22} \end{pmatrix} \end{matrix}, \quad \boldsymbol{\Psi} = \begin{matrix} \\ 1 \\ p-1 \end{matrix} \begin{matrix} \begin{matrix} 1 & p-1 \end{matrix} \\ \begin{pmatrix} 0 & \boldsymbol{\psi}_{12} \\ \boldsymbol{\psi}_{21} & \boldsymbol{\Psi}_{22} \end{pmatrix} \end{matrix}, \tag{4.4}$$

and we can use this partitioning without loss of generality for all $k \in \{1, \ldots, p\}$. Using the inverse of the partitioned matrix $\mathbf{W}$, the sub-gradient of equation of Eq. 4.2 for the partitioned problem becomes

$$\hat{\boldsymbol{\Sigma}}_{22}\boldsymbol{\beta} - \bar{\boldsymbol{s}}_{12} + \rho\boldsymbol{\psi}_{12} = \mathbf{0}, \tag{4.5}$$

where $\hat{\boldsymbol{\Sigma}}$ is a current estimate of $\mathbf{W}^{-1}$ [2], and $\boldsymbol{\beta} = -\boldsymbol{w}_{12}/w_{11}$. The algorithm is outlined in Alg. 8.

In summary, the graphical lasso computes a point estimate of the network of dependencies among a set of variables in a high-dimensional setting. More specifically, it efficiently computes a L_1-penalised maximum likelihood estimate of the inverse covariance matrix.

Algorithm 8 Graphical Lasso.

Require: Scaled sample covariance matrix $\bar{\mathbf{S}}$
Result: Inverse covariance $\hat{\mathbf{W}}$
Initialise: $\hat{\mathbf{W}} = \bar{\mathbf{S}}$

```
while not converged do
    for k = 1, ..., p do
        Partition matrix according to Eq. 4.4.
        Solve the subgradient equations Eq. 4.5 using a pathwise coordinate-
    descent algorithm.
        ŵ_12 ← Ŵ_11 β̂
    end for
end while
for k = 1, ..., p do
    Solve ŵ_12 = −β̂ŵ_11, with 1/ŵ = ŵ_11 − ŵ_12^T β̂
end for
```

A Bayesian interpretation of the graphical lasso is presented by Wang et al. (2012). This approach also uses the partitioning of Eq. 4.4 with a block-wise

[2] i.e. $\hat{\boldsymbol{\Sigma}}\mathbf{W} = \mathbf{I}$

but Bayesian inference scheme. The approach also uses the Wishart likelihood from Eq. 4.1 for $\mathbf{W}$ but adds a hierarchical prior

$$p(\mathbf{W}|\mathbf{T}, \lambda) = p(\mathbf{W}|\mathbf{T}, \lambda)p(\mathbf{T}|\lambda), \tag{4.6}$$

where the first part is a double exponential/ Laplace prior on the non-diagonal elements, represented as a scale mixture of Gaussians, multiplied with a product of exponential densities for the diagonal elements

$$p(\mathbf{W}_{i\leq j}|\mathbf{T}, \lambda) \propto \prod_{w_{i\leq j}} \left(\frac{1}{\sqrt{2\pi t_{ij}}} \exp\left(-\frac{w_{ij}^2}{2t_{ij}} \right) \right) \prod_{i=1}^{p} \frac{\lambda}{2} \exp\left(-\frac{\lambda}{2} w_{ii} \right). \tag{4.7}$$

This prior enforces sparsity on the non-diagonal elements of $\mathbf{W}$. Here, $\mathbf{T} = (t_{i\neq j})$ are latent scale parameters. The second part in the hierarchy defines the mixing density for the scale mixtures

$$p(\mathbf{T}|\lambda) \propto \prod_{i<j} \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2} t_{ij} \right), \tag{4.8}$$

which couples the non-diagonal elements such that they meet the positive-definit constraint. This term also ensures that the marginal distribution of individual non-diagonal elements w_{ij} follow the double exponential prior, for details c.f. Wang et al. (2012). Then, the posterior is

$$\begin{aligned} p(\mathbf{W}, \mathbf{T}|\mathbf{X}, \lambda) \propto & \det(\mathbf{W})^{\frac{n}{2}} \exp \operatorname{tr}\left(-\frac{1}{2}\mathbf{S}\mathbf{W} \right) \\ & \times \prod_{i<j} \left(t_{ij}^{-\frac{1}{2}} \exp\left(-\frac{w_{ij}^2}{2t_{ij}} \right) \exp\left(-\frac{\lambda^2}{2} t_{ij} \right) \right) \\ & \times \prod_{i=1}^{p} \exp\left(-\frac{\lambda}{2} w_{ii} \right). \end{aligned} \tag{4.9}$$

Posterior inference involves iterating through the dimensions $k \in \{1, \ldots, p\}$ to estimate the entire network in a block Gibbs sampling manner. For the posterior conditionals, the matrices $\mathbf{W}$ and $\mathbf{S}$ are partitioned into blocks by fixing all but row and column k. Without loss of generality, the partitioning for the first row and column as in Eq. 4.4 is sufficient for the description of the sampler. The posterior conditional for $k = 1$ is

$$\begin{aligned} p(w_{11}, \boldsymbol{w}_{12}|\mathbf{W}_{22}, \mathbf{T}, \lambda) \propto & \left(w_{11} - \boldsymbol{w}_{12}\mathbf{W}_{22}^{-1}\boldsymbol{w}_{12}^{\mathsf{T}} \right)^{\frac{n}{2}} \\ & \times \exp\left(-\frac{1}{2}\boldsymbol{w}_{12}\mathbf{D}_{\mathbf{T}_1}^{-1}\boldsymbol{w}_{12}^{\mathsf{T}} + 2\boldsymbol{s}_{12}^{\mathsf{T}}\boldsymbol{w}_{12} + (s_{11} + \lambda)w_{11} \right), \end{aligned} \tag{4.10}$$

where $\mathbf{D}_{\boldsymbol{t}} = \text{diag}(\boldsymbol{t})$ is a diagonal matrix containing the values of $\boldsymbol{t}_{12}$. The transformation $(w_{11}, \boldsymbol{w}_{12} \mapsto \delta = w_{11} - \boldsymbol{w}_{12}\mathbf{W}_{22}^{-1}\boldsymbol{w}_{12}^{\mathsf{T}}, \boldsymbol{\beta} = \boldsymbol{w}_{12})$ helps in identifying the distributions to draw of. Finally,

$$\begin{aligned} p(\boldsymbol{\beta}, \delta | \mathbf{W}_{22}, \mathbf{T}, \lambda) \propto \delta^{\frac{n}{2}} \exp\left(-\frac{1}{2}(s_{11} + \lambda)\delta\right) \\ \times \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^{\mathsf{T}}\left(\mathbf{D}_{\boldsymbol{t}}^{-1} + (s_{11} + \lambda)\mathbf{W}_{22}^{-1}\right)\boldsymbol{\beta} + 2\boldsymbol{s}_{12}\boldsymbol{\beta}\right)\right) \end{aligned} \tag{4.11}$$

and

$$(\delta, \boldsymbol{\beta} | \mathbf{X}, \mathbf{W}_{22}, \mathbf{T}, \lambda) \sim \Gamma\left(\frac{n}{2} + 1, \frac{s_{11} + \lambda}{2}\right) \mathcal{N}\left(-\mathbf{C}\boldsymbol{s}_{12}^{\mathsf{T}}, \mathbf{C}\right), \tag{4.12}$$

where $\Gamma(a, b)$ is the Gamma distribution with shape parameter a and scale parameter b, and $\mathbf{C} = \left((s_{11} + \lambda)\mathbf{W}_{22}^{-1} + \mathbf{D}_{\boldsymbol{t}}^{-1}\right)^{-1}$. We conclude the description of the Bayesian graphical lasso with the block Gibbs sampler in Alg. 9. For details of the t_{ij}'s, we refer to Wang et al. (2012).

Algorithm 9 Bayesian Graphical Lasso.

Require: Sample covariance matrix $\mathbf{S}$
Result: Inverse covariance $\hat{\mathbf{W}}$
Initialise: $\hat{\mathbf{W}} = \mathbf{S}$
while not converged **do**
  for $j = 1, \ldots, p$ **do**
    Partition matrices $\mathbf{W}$, $\mathbf{S}$, and $\mathbf{T}$ according to Eq. 4.4.
    $\delta \sim \Gamma\left(n/2 + 1, (s_{11} + \lambda)/2\right)$
    $\boldsymbol{\beta} \sim \mathcal{N}\left(-\mathbf{C}\boldsymbol{s}_{12}^{\mathsf{T}}, \mathbf{C}\right)$, where $C = \left((s_{11} + \lambda)\mathbf{W}_{22}^{-1} + \mathbf{D}_{\boldsymbol{t}}^{-1}\right)^{-1}$
    $w_{11} \leftarrow \delta + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{W}_{22}^{-1}\boldsymbol{\beta}$, $\boldsymbol{w}_{12} \leftarrow \boldsymbol{\beta}$
  end for
  for $i < j$ **do**
    $t_{ij}^{-1} \sim \mathcal{IG}(\sqrt{\lambda^2 / w_{ij}^2}, \lambda^2)$
  end for
end while

4.3. Motivation

We now introduce a methodology for instances, where it is more interesting to only estimate a sub-network as opposed to estimating an entire network of

all the associations. Consider the example in gene analysis where the dependency between only a few clinical factors and thousands of genetic markers is required. Here, it is important to limit the focus on the clinical factors and only estimate the *Markov blanket* of the variables we are interested in. This is the set of variables that, when conditioned on, render the variables of interest conditionally independent of the rest of the network. In this context, a question of fundamental importance arises: can we estimate the Markov blanket directly as opposed to estimating the entire network with subsequent pruning to the variables of interest? We thus approach the problem in the spirit of Vapnik (2013): "When solving a given problem, try to avoid solving a more general problem as an intermediate step."

Remark

The sequel closely follows Kaufmann et al. (2016), but adds further details.

In the following, we provide a Bayesian perspective of estimating the Markov blanket of a set of p query variables in an undirected network. The Bayesian view enables the computation of a posterior distribution and thus offers a means of assessing the (un-)certainty of an estimate. This contrasts with the maximum likelihood approach of the graphical lasso which only provides a point estimate of the network. The approach is closely related to the Bayesian graphical lasso (Wang et al., 2012) introduced beforehand in Section 4.2. This approach partitions the matrix $\mathbf{W}$ as shown on the left in Fig. 4.2 and fixes all but one row and column. Posterior inference involves iterating through the individual variables to estimate the entire network. In particular, inference of the $\boldsymbol{w}_{12}$ block relies on estimating both w_{11} and $\mathbf{W}_{22}$. However, the coupling of $\boldsymbol{w}_{12}$ and $\mathbf{W}_{22}$ is a limiting factor that can be avoided in the context of Markov blanket estimation. This is the crucial idea which forms the basis in what follows. The approach follows the partitioning as shown on the right in Fig. 4.2, where we consider $p > 1$ query variables, and estimating the Markov blanket implies to estimate the block $\mathbf{W}_{12}$.

An important observation for the model we present here, is that the Wishart likelihood may be factorised such that the blocks $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ are de-coupled from $\mathbf{W}_{22}$. This result is provided as Lemma 1 in Section 4.4 and is also depicted in the difference in the shading of the blocks in Fig. 4.2. The difference in the shading of $\mathbf{W}$ indicates that estimation of $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ (and hence $\mathbf{W}_{21}$)

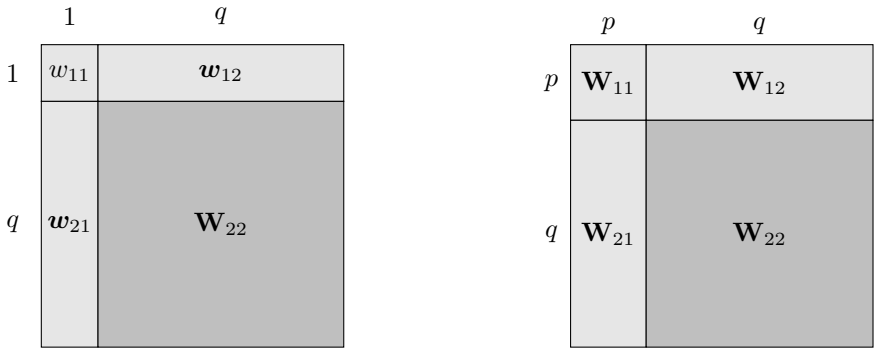


Figure 4.2.: Left: factorisation of the (Bayesian) graphical lasso. Right: the factorisation for the Markov blanket, where the number of query variables is $p > 1$.

is invariant of estimating $\mathbf{W}_{22}$. We show that by combining the factorised likelihood with an appropriate choice of prior, we obtain a posterior distribution that preserves this independence structure. Most importantly, this posterior distribution has an analytic form and can hence be sampled from. We formalise this in Section 4.5 as Theorem 5. A further consequence of this result is formulated in Theorem 6 which demonstrates that sampling from the posterior distribution can be done efficiently. Overall, this means that the Markov blanket of p query nodes, can be estimated efficiently *without explicitly inferring the entire network*. We conclude by summarising the key results in Fig. 4.3.

T1: $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ are conditionally independent of $\mathbf{W}_{22}$ given $\mathbf{S}$

T2: The posterior conditionals $\mathbf{W}_{11}|\mathbf{W}_{12}$ and $\mathbf{W}_{12}|\mathbf{W}_{11}$ have analytic form

T3: Sampling from the posterior costs $O(qp^3)$ per Gibbs sweep

Figure 4.3.: Key results.

The remainder is structured as follows. We begin by exploring the block

factorization of the Wishart likelihood in Section 4.4. We subsequently derive the posterior distribution and construct a Gibbs sampler to efficiently sample from the different blocks in Section 4.5. Section 4.6 describes how Bayesian Markov blanket estimation can be extended to deal with mixed data types with the copula framework. Finally, we demonstrate the practical applicability of the scheme in Section 4.7 with examples of artificial and real data.

4.4. Model

Problem Formulation Let $\mathbf{X} \in \mathbb{R}^{n \times (p+q)}$ be a matrix containing n independent observations of p query and q remaining variables. Assume, that each observation $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$ is drawn from multivariate Gaussian distribution with $p+q$ dimensions with mean $\boldsymbol{0}$ and covariance $\boldsymbol{\Sigma}$. Then, the sample covariance $\mathbf{S} = \mathbf{X}^\mathsf{T}\mathbf{X}$ follows the Wishart distribution $\mathbf{S} \sim \mathcal{W}_{p+q}(n, \boldsymbol{\Sigma})$ with $n > p+q-1$ degrees of freedom. That is,

$$\begin{aligned} p(\mathbf{S}|\boldsymbol{\Sigma}) &\propto \det(\boldsymbol{\Sigma})^{-\frac{n}{2}} \det(\mathbf{S})^{\frac{n-(p+q)-1}{2}} \exp \operatorname{tr}\left(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) \\ &= \det(\mathbf{W})^{\frac{n}{2}} \det(\mathbf{S})^{\frac{n-(p+q)-1}{2}} \exp \operatorname{tr}\left(-\frac{1}{2}\mathbf{W}\mathbf{S}\right), \end{aligned} \tag{4.13}$$

where $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$. We are interested in estimating the Markov blanket of the p query variables with respect to the q remaining variables. Assume that $\mathbf{S}$ and $\mathbf{W}$ are partitioned according to

$$\mathbf{S} = \begin{matrix} \\ p \\ q \end{matrix} \begin{matrix} \begin{matrix} p & q \end{matrix} \\ \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \end{matrix}, \quad \mathbf{W} = \begin{matrix} \\ p \\ q \end{matrix} \begin{matrix} \begin{matrix} p & q \end{matrix} \\ \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix} \end{matrix},$$

where the matrices have been reordered such that the query variables lie in the upper left block. Given $\mathbf{S}$, we would like to infer $\mathbf{W}_{12}$, which is the Markov blanket of the p variables that constitute the block $\mathbf{S}_{11}$. We restrict the problem to the case where $p \ll q$ such that $\mathbf{S}_{11}$ is small, corresponding to the few variables of interest, and $\mathbf{S}_{22}$ is large.

Factorising the Likelihood We begin by showing a block-wise factorisation of the likelihood, which builds the foundation of the model. Let $\mathbf{W}_{22.1} = \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}$ be the Schur complement of the block $\mathbf{W}_{11}$ in $\mathbf{W}$.

Lemma 1. *The likelihood of the covariance matrix factorises in terms of* $\mathbf{W}$ *as follows:*

$$\mathcal{L}_{\mathbf{S}}(\mathbf{W}) \propto \mathcal{L}_1(\mathbf{W}_{11}, \mathbf{W}_{12})\mathcal{L}_2(\mathbf{W}_{22\cdot 1}).$$

The proof of this lemma is analogous to Gupta and Nagar (1999) (Chapter 3, pp. 94–95).

Proof. Factorising the Wishart density according to

$$\begin{aligned}\mathbf{WS} &= \mathbf{W}_{11}\mathbf{S}_{11} + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + \mathbf{W}_{22}\mathbf{S}_{22} \\ &= \mathbf{W}_{11}\mathbf{S}_{11} + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + \\ &\quad + (\mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12} + \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12})\mathbf{S}_{22} \\ &= \mathbf{W}_{11}\mathbf{S}_{11} + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}\mathbf{S}_{22} + \mathbf{W}_{22\cdot 1}\mathbf{S}_{22}\end{aligned} \tag{4.14}$$

and

$$\det(\mathbf{W})^{\frac{n}{2}} = \det(\mathbf{W}_{11})^{\frac{n}{2}} \det(\mathbf{W}_{22\cdot 1})^{\frac{n}{2}}, \tag{4.15}$$

the independence follows from

$$p(\mathbf{W}|\mathbf{S}) = p(\mathbf{W}_{11}, \mathbf{W}_{12}|\mathbf{S})p(\mathbf{W}_{22\cdot 1}|\mathbf{S}), \tag{4.16}$$

where

$$\begin{aligned}p(\mathbf{W}_{11}, \mathbf{W}_{12}|\mathbf{S}) &\propto \det(\mathbf{W}_{11})^{\frac{n}{2}} \\ &\exp \operatorname{tr}\left(-\frac{1}{2}\left(\mathbf{W}_{11}\mathbf{S}_{11} + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}\mathbf{S}_{22}\right)\right)\end{aligned} \tag{4.17}$$

and

$$p(\mathbf{W}_{22\cdot 1}|\mathbf{S}) \propto \det(\mathbf{W}_{22\cdot 1})^{\frac{n}{2}} \exp \operatorname{tr}\left(-\frac{1}{2}\left(\mathbf{W}_{22\cdot 1}\mathbf{S}_{22}\right)\right). \tag{4.18}$$

□

Lemma 1 is a pure functional statement without any statistical reasoning. The factorisation of the likelihood in Lemma 1 then translates to the analogous independence structure in the posterior distribution of $\mathbf{W}$ as shown in Theorem 4.

4.4.1. Prior

The natural conjugate prior to the likelihood is the Wishart distribution. However, in order to ensure sparsity, we also use a double exponential prior as in Wang et al. (2012). Since the focus is on the Markov blanket, we only place the latter on the block $\mathbf{W}_{12}$. This results in a compound prior:

$$\begin{aligned} p(\mathbf{W}, \mathbf{T}) &= \mathcal{W}(p+q+1, \mathbf{I}) p\left(\mathbf{W}_{12} | \mathbf{T}\right) p\left(\mathbf{T} | \lambda\right) \\ &\propto \exp \operatorname{tr}\left(-\tfrac{1}{2} \mathbf{I} \mathbf{W}\right) \prod_{w_{ij} \epsilon \mathbf{W}_{12}} \frac{1}{\sqrt{2 \pi t_{ij}}} \exp\left(-\frac{w_{ij}^2}{2 t_{ij}}\right) \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2} t_{ij}\right), \end{aligned} \tag{4.19}$$

where $\mathbf{T} = \{t_{ij}\}$ are scale parameters introduced by Wang et al. (2012). Analysing the conditional posterior distribution of the t_{ij}^{-1}, we see that

$$t_{ij}^{-1} | w_{ij}, \lambda \sim \mathcal{IG}\left(\sqrt{\lambda^2 / w_{ij}^2}, \lambda^2\right) \tag{4.20}$$

where $\mathcal{IG}$ denotes the inverse-Gaussian distribution and λ is a hyperparameter. Most importantly, the compound prior also possesses the factorisation in terms of $\mathbf{W}$ proved for the likelihood in Lemma 1. This follows from from the element-wise independence of the prior. Multiplying the compound prior introduced in Eq. 4.19 by the likelihood yields the posterior distributions for blocks $\mathbf{W}_{12}$ and $\mathbf{W}_{11}$.

4.4.2. Posterior Distribution

A consequence of the factorisation in Lemma 1 is that the posterior distributions of the blocks $(\mathbf{W}_{11}, \mathbf{W}_{12})$ and $\mathbf{W}_{22 \cdot 1}$ are conditionally independent given $\mathbf{S}$.

Theorem 4. *The posterior distribution of* $(\mathbf{W}_{11}, \mathbf{W}_{12})$ *is conditionally independent of* $\mathbf{W}_{22 \cdot 1}$ *given* $\mathbf{S}$.

Proof. The Likelihood, as shown in Lemma 1, as well as the element-wise independent prior in Eqs. 4.19 and 4.20 factorise according to the blocks $\mathbf{W}_{11}$, $\mathbf{W}_{12}$, and $\mathbf{W}_{22}$. □

Because of the conditional independence proved in Theorem 4, we can infer the Markov blanket $\mathbf{W}_{12}$ without the need of estimating the big block $\mathbf{W}_{22 \cdot 1}$. In the next section, we explicitly derive the posterior distribution and show that it has an analytical form.

4.5. Posterior Inference

We now state the main result. Specifically, we show that the posterior distribution required to estimate the Markov blanket can be expressed in an analytical form. Subsequently, we demonstrate how to efficiently sample from it in Section 4.5.1.

Let the Matrix Generalised Inverse Gaussian (MGIG) distribution (Butler, 1998) be defined by the probability density function with parameter γ:

$$p(\mathbf{M};\gamma,\mathbf{A},\mathbf{B}) \propto \det(\mathbf{M})^{-\gamma-1} \exp \operatorname{tr}\left(-\frac{1}{2}(\mathbf{AM}+\mathbf{BM}^{-1})\right). \tag{4.21}$$

Theorem 5. *The posterior conditionals* $\mathbf{W}_{12}|\mathbf{W}_{11},\mathbf{S},\mathbf{T}$ *and* $\mathbf{W}_{11}|\mathbf{W}_{12},\mathbf{S},\mathbf{T}$ *admit an analytical form:*

(1) *Vectorised rows of* $\mathbf{W}_{12}$ *follow a joint normal distribution*

$$vec(\mathbf{W}_{12}^{\mathsf{T}})|\mathbf{W}_{11},\mathbf{S},\mathbf{T} \sim \mathcal{N}_{pq}\left(-\mathbf{C}vec(\mathbf{S}_{12}^{\mathsf{T}}),\mathbf{C}\right), \tag{4.22}$$

where $\mathbf{C} = \left(\mathbf{W}_{11}^{-1} \otimes (\mathbf{S}_{22}+\mathbf{I}) + \mathbf{D}^{-1}\right)^{-1}$ *is the covariance matrix,* $\mathbf{D} = \operatorname{diag}(\mathbf{D}_1,\ldots,\mathbf{D}_p)$*, and* $\mathbf{D}_i = \operatorname{diag}(T_{i\cdot})$ *are diagonal matrices containing* $T_{i\cdot} = (t_{i1},\ldots,t_{iq})$*.*

(2) $\mathbf{W}_{11}$ *follows the Matrix Generalised Inverse Gaussian (MGIG) distribution:*

$$\mathbf{W}_{11}|\mathbf{W}_{12},\mathbf{S},\mathbf{T} \sim \mathcal{MGIG}_{p\times p}\left(\frac{n}{2}+p, \mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{21}, \mathbf{S}_{11}+\mathbf{I}\right). \tag{4.23}$$

Before we prove Theorem 5, we define the matrix normal distribution and present a Lemma, where we establish the conditional distribution of $\mathbf{W}_{12}|\mathbf{W}_{11},\mathbf{S}$. This Lemma builds as the foundation for the proof of Theorem 5.

Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ follow a matrix normal distribution, that is

$$\begin{aligned}\mathbf{X} &\sim \mathcal{MN}_{p\times n}(\mathbf{M},\mathbf{\Sigma},\mathbf{\Omega}) \\ &\propto \frac{1}{(2\pi)^{pn/2}} \det(\mathbf{\Sigma})^{-\frac{p}{2}} \det(\mathbf{\Omega})^{-\frac{n}{2}} \exp \operatorname{tr}\left(-\frac{1}{2}\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{\Omega}^{-1}(\mathbf{X}-\mathbf{M})^{\mathsf{T}}\right)\end{aligned} \tag{4.24}$$

where mean $\mathbf{M} \in \mathbb{R}^{p\times n}$, column covariance $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$, and row covariance $\mathbf{\Omega} \in \mathbb{R}^{n\times n}$.

Lemma 2. *Let* $\mathbf{S} \sim \mathcal{W}_{p+q}(n, \mathbf{\Sigma})$ *and* $\mathbf{W} \sim \mathcal{W}_{p+q}(p+q+1, \mathbf{I})$ *follow Wishart distributions, then*

$$\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{S} \sim \mathcal{MN}_{p\times q}\left(-\mathbf{W}_{11}\mathbf{S}_{12}(\mathbf{S}_{22}+\mathbf{I})^{-1}, \mathbf{W}_{11}, (\mathbf{S}_{22}+\mathbf{I})^{-1}\right) \tag{4.25}$$

is matrix normal distributed.

Proof. The proof is similar to (Gupta and Nagar, 1999). Factorising the product of the Wishart densities according to

$$\begin{aligned}
&\operatorname{tr}(\mathbf{W}(\mathbf{S}+\mathbf{I}))\\
&= \operatorname{tr}(\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I}) + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + \mathbf{W}_{22}(\mathbf{S}_{22}+\mathbf{I}))\\
&= \operatorname{tr}\Big(\mathbf{W}_{11}\left((\mathbf{S}_{11}+\mathbf{I}) - \mathbf{S}_{12}(\mathbf{S}_{22}+\mathbf{I})^{-1}\mathbf{S}_{21} + \mathbf{S}_{12}(\mathbf{S}_{22}+\mathbf{I})^{-1}\mathbf{S}_{21}\right) + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12}\\
&\quad +(\mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12} + \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12})(\mathbf{S}_{22}+\mathbf{I})\Big)\\
&= \operatorname{tr}\Big(\mathbf{W}_{11}\mathbf{S}_{11\cdot 2} + \mathbf{W}_{11}\mathbf{S}_{12}(\mathbf{S}_{22}+\mathbf{I})^{-1}\mathbf{S}_{21} + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\\
&\qquad +\mathbf{W}_{22\cdot 1}(\mathbf{S}_{22}+\mathbf{I}))\\
&= \operatorname{tr}\Big(\mathbf{W}_{11}\mathbf{S}_{11\cdot 2}\\
&\qquad + \underbrace{(\mathbf{S}_{22}+\mathbf{I})\left(\mathbf{W}_{12} + \mathbf{W}_{11}\mathbf{S}_{12}(\mathbf{S}_{22}+\mathbf{I})^{-1}\right)^{\mathsf{T}}\mathbf{W}_{11}^{-1}\left(\mathbf{W}_{12} + \mathbf{W}_{11}\mathbf{S}_{12}(\mathbf{S}_{22}+\mathbf{I})^{-1}\right)}_{\mathcal{MN}}\\
&\qquad +\mathbf{W}_{22\cdot 1}(\mathbf{S}_{22}+\mathbf{I})\Big)
\end{aligned} \tag{4.26}$$

where $\mathbf{S}_{11\cdot 2} = (\mathbf{S}_{11}+\mathbf{I}) - \mathbf{S}_{12}(\mathbf{S}_{22}+\mathbf{I})^{-1}\mathbf{S}_{21}$ and we changed variables $\mathbf{W}_{22\cdot 1} = \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}$ with Jacobian 1. Factorising the determinants according to

$$\begin{aligned}
\det(\mathbf{W})^{\frac{n}{2}} &= \det(\mathbf{W}_{11})^{\frac{n}{2}}\det(\mathbf{W}_{22\cdot 1})^{\frac{n}{2}}\\
&= \det(\mathbf{W}_{11})^{\frac{n+p}{2}}\underbrace{\det(\mathbf{W}_{11})^{-\frac{p}{2}}}_{\mathcal{MN}}\det(\mathbf{W}_{22\cdot 1})^{\frac{n}{2}}
\end{aligned} \tag{4.27}$$

and

$$\begin{aligned}
\det(\mathbf{S}+\mathbf{I})^{\frac{n-(p+q)-1}{2}} &= \det(\mathbf{S}_{22}+\mathbf{I})^{\frac{n-p-q-1}{2}}\det(\mathbf{S}_{11\cdot 2})^{\frac{n-p-q-1}{2}}\\
&= \det(\mathbf{S}_{22}+\mathbf{I})^{\frac{n-p-2q-1}{2}}\underbrace{\det\left((\mathbf{S}_{22}+\mathbf{I})^{-1}\right)^{-\frac{q}{2}}}_{\mathcal{MN}}\det(\mathbf{S}_{11\cdot 2})^{\frac{n-p-q-1}{2}}
\end{aligned} \tag{4.28}$$

gives the desired result. □

Now, having established the conditional distribution of the Markov blanket, we can prove Theorem 5(1). The idea of the proof is as follows: The posterior

conditionals maintain the conditional independence structure proved for the distribution in Theorem 4, i.e.

$$p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22\cdot1}|\mathbf{S}, \mathbf{T}) = p(\mathbf{W}_{11}, \mathbf{W}_{12}|\mathbf{S}, \mathbf{T})p(\mathbf{W}_{22\cdot1}|\mathbf{S}, \mathbf{T}). \tag{4.29}$$

Derivations of the distributions in Eqs. 4.22 and 4.23 follow from factorising the posterior and rearranging terms. We formalise the relevant calculations in the following proof.

Proof. According to Lemma 2, the likelihood in Eq. 4.13 can be expressed as a matrix normal distribution as in Eq. 4.25. Including the Wishart part of the prior changes the distribution in Eq. 4.25 to

$$\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{S} \sim \mathcal{MN}_{p\times q}\left(-\mathbf{W}_{11}\mathbf{S}_{12}(\mathbf{S}_{22}+\mathbf{I})^{-1}, \mathbf{W}_{11}, (\mathbf{S}_{22}+\mathbf{I})^{-1}\right) \tag{4.30}$$

which is equivalent to

$$vec(\mathbf{W}_{12}^{\mathsf{T}})|\mathbf{W}_{11}, \mathbf{S} \sim \mathcal{N}_{pq}\left(-\left(\mathbf{W}_{11}\otimes(\mathbf{S}_{22}+\mathbf{I})^{-1}\right)vec(\mathbf{S}_{12}^{\mathsf{T}}), \mathbf{W}_{11}\otimes(\mathbf{S}_{22}+\mathbf{I})^{-1}\right) \tag{4.31}$$

where $vec(\mathbf{W}_{12}^{\mathsf{T}})$ are the vectorised rows of matrix $\mathbf{W}_{12}$ and $\otimes$ denotes the Kronecker product. For inclusion of the double exponential prior, it has to be rewritten as

$$\begin{aligned}\prod_{w_{ij}\in\mathbf{W}_{12}}\frac{1}{\sqrt{2\pi t_{ij}}}\exp\left(-\frac{w_{ij}^2}{2t_{ij}}\right) &= \prod_{i=1}^{p}\exp\left(-\frac{1}{2}(\mathbf{W}_{12})_{i,\cdot}^{\mathsf{T}}\mathbf{D}_i^{-1}(\mathbf{W}_{12})_{i,\cdot}\right)\\ &= \exp\left(-\frac{1}{2}vec(\mathbf{W}_{12}^{\mathsf{T}})^{\mathsf{T}}\mathbf{D}^{-1}vec(\mathbf{W}_{12}^{\mathsf{T}})\right),\end{aligned} \tag{4.32}$$

where $(\mathbf{W}_{12})_{i,\cdot}$ denotes the ith row of $\mathbf{W}_{12}$, $\mathbf{D} = \mathrm{diag}\,(\mathbf{D}_1, \ldots, \mathbf{D}_p)$, and $\mathbf{D}_i = \mathrm{diag}\,(T_{i\cdot})$ are diagonal matrices containing $T_{i\cdot} = (t_{i1}, \ldots, t_{iq})$. The result follows from multiplying the double exponential part of the prior in Eq. 4.32 by the expanded density in Eq. 4.31. □

Before proving the second part of Theorem 5, we need another Lemma which characterises the MGIG distribution subjected to a variable transformation.

Lemma 3. *Let* $p(\mathbf{M}) \propto \det(\mathbf{M})^{-\gamma-1}\exp\mathrm{tr}\left(-\frac{1}{2}(\mathbf{AM}+\mathbf{BM}^{-1})\right)$ *be MGIG distributed, then* $\mathbf{W} = \mathbf{M}^{-1}$ *is also MGIG distributed:*

$$p(\mathbf{W}) \propto \det(\mathbf{W})^{\gamma-p}\exp\mathrm{tr}\left(-\frac{1}{2}(\mathbf{AW}^{-1}+\mathbf{BW})\right) \tag{4.33}$$

Proof. Transforming Eq. 4.21 with $\mathbf{W} = \mathbf{X}^{-1}$ and $\mathrm{d}\mathbf{X} = -\mathbf{W}^{-1}\mathrm{d}\mathbf{W}\mathbf{W}^{-1}$, we have the Jacobian $J = \det(\mathbf{W})^{-(p+1)}$, thus

$$\begin{aligned} p(\mathbf{W}) &\propto \det(\mathbf{W})^{-(p+1)} \det(\mathbf{W}^{-1})^{-\gamma-1} \exp \operatorname{tr}\left(-\frac{1}{2}(\mathbf{A}\mathbf{W}^{-1} + \mathbf{B}\mathbf{W})\right) \\ &= \det(\mathbf{W})^{\gamma-p} \exp \operatorname{tr}\left(-\frac{1}{2}(\mathbf{A}\mathbf{W}^{-1} + \mathbf{B}\mathbf{W})\right). \end{aligned} \tag{4.34}$$

□

We are finally ready for proving Theorem 5(2)

Proof. The proof is similar to (Butler, 1998). Factorising the Wishart density according to

$$\begin{aligned} &\operatorname{tr}\Big(\mathbf{W}\mathbf{S}\Big) = \operatorname{tr}\Big(\mathbf{W}_{11}\mathbf{S}_{11} + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + \mathbf{W}_{22}\mathbf{S}_{22}\Big) \\ &= \operatorname{tr}\Big(\mathbf{W}_{11}\mathbf{S}_{11} + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + (\mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12} \\ &\qquad + \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12})\mathbf{S}_{22}\Big) \\ &= \operatorname{tr}\Big(\mathbf{W}_{11}\mathbf{S}_{11} + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + (\mathbf{W}_{22\cdot1} + \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12})\mathbf{S}_{22}\Big) \\ &= \operatorname{tr}\Big(\mathbf{W}_{11}\underbrace{\mathbf{S}_{11}}_{=\mathbf{B}} + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + \underbrace{\mathbf{W}_{12}\mathbf{S}_{22}\mathbf{W}_{21}}_{=\mathbf{A}}\mathbf{W}_{11}^{-1} + \mathbf{W}_{22\cdot1}\mathbf{S}_{22}\Big) \end{aligned} \tag{4.35}$$

and

$$\det(\mathbf{W})^{\frac{n}{2}} = \det(\mathbf{W}_{11})^{\frac{n}{2}} \det(\mathbf{W}_{22\cdot1})^{\frac{n}{2}}, \tag{4.36}$$

where we changed variables $\mathbf{W}_{22\cdot1} = \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}$ with Jacobian 1. Comparing Eq. 4.33 with Eq. 4.35, we can identify $\mathbf{A}$, $\mathbf{B}$, and λ:

$$\frac{n}{2} \stackrel{!}{=} \lambda - p \quad \Rightarrow \quad \lambda = \frac{n}{2} + p \tag{4.37}$$

such that

$$\mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{S} \sim \mathcal{MGIG}_{p\times p}\left(\frac{n}{2} + p, \mathbf{W}_{12}\mathbf{S}_{22}\mathbf{W}_{21}, \mathbf{S}_{11}\right). \tag{4.38}$$

Since the double exponential prior does not affect the distribution of $\mathbf{W}_{11}$, we only have to include the Wishart prior. The result follows immediately. □

Theorem 5 shows that estimation of the Markov blanket of the p query variables only requires sampling from the posterior conditionals of $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ which both have an analytical form while remaining independent of $\mathbf{W}_{22}$. Therefore, the amount of parameters in the Markov blanket that need to be estimated, scales linearly with q. This is an improvement over the Bayesian graphical lasso (Wang et al., 2012) approach, where this number grows quadratically with q. Theorem 5 also provides us with the particular distributions to sample from. Having these distributions, we can construct a Gibbs sampler which alternatively draws from these conditional distributions while averaging out the prior. Next, we demonstrate how this sampling can be done efficiently.

Algorithm 10 Block Gibbs sampling scheme for the posterior.

Require: Sample covariance matrix $\mathbf{S}$
Result: Markov Blanket estimate $\hat{\mathbf{W}}_{12}$
1: **while** not converged **do**
2: $\quad T_{ij}^{-1} \sim \mathcal{IG}\left(\sqrt{\lambda^2/w_{ij}^2}, \lambda^2\right)$
3: $\quad vec(\mathbf{W}_{12}^\mathsf{T})|\mathbf{W}_{11}, \mathbf{S}, \mathbf{T} \sim \mathcal{N}_{pq}\left(-\mathbf{C}vec(\mathbf{S}_{12}^\mathsf{T}), \mathbf{C}\right)$
4: $\quad \mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{S}, \mathbf{T} \sim \mathcal{MGIG}_{p\times p}\left(-\frac{1}{2}(n+p+1), \mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{21}, \mathbf{S}_{11}+\mathbf{I}\right)$
5: **end while**
6: return averaged and thresholded $\mathbf{W}_{12}$

4.5.1. Efficiency of Sampling from the Posterior

The block-wise Gibbs sampling scheme for estimating the Markov blanket is summarised in Algorithm 10. This sampling scheme consists of iterative resampling of $\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{S}, \mathbf{T}$ and of $\mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{S}, \mathbf{T}$, according to their definitions in Theorem 5. The estimate of the Markov blanket $\hat{\mathbf{W}}_{12}$ is subsequently computed based on samples drawn from $\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{S}, \mathbf{T}$ following the burn-in period of the sampler.

The distribution of $\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{S}, \mathbf{T}$ is given by Theorem 5(1). The vectorised rows of $\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{S}, \mathbf{T}$ follow a joint normal distribution. For $\boldsymbol{v} = vec(\mathbf{S}_{12}^\mathsf{T})$, the distribution further simplifies to

$$vec(\mathbf{W}_{12}^\mathsf{T})|\mathbf{W}_{11}, \mathbf{S} \sim \mathcal{N}_{pq}\left(-\mathbf{C}\boldsymbol{v}, \mathbf{C}\right). \tag{4.39}$$

The majority of the computational cost incurred in the method arises from sampling from this joint normal distribution. Eq. 4.39 requires us to compute $\mathbf{C}$ which is of size $pq \times pq$. Note that $\mathbf{C}^{-1}$ cannot be represented as a covariance tensor of a matrix normal distribution. Therefore, naïve inversion using

a standard Cholesky decomposition would cost $\mathcal{O}(p^3q^3)$ operations. The efficient sampling strategy exploits the block structure of this matrix. This is the foundation of Theorem 6.

Theorem 6. *Sampling from the distribution in Theorem 5(1) requires $\mathcal{O}(pq^3)$ operations.*

Proof Sketch. We expand the Kronecker product of matrix $\mathbf{C} \in \mathbb{R}^{pq \times pq}$ which comprises p blocks of size $q \times q$:

$$\mathbf{C} = \begin{pmatrix} u_{11}(\mathbf{S}_{22} + \mathbf{I}) + \mathbf{D}_1^{-1} & u_{12}(\mathbf{S}_{22} + \mathbf{I}) & \cdots \\ u_{21}(\mathbf{S}_{22} + \mathbf{I}) & \ddots & \\ \vdots & \cdots & u_{pp}(\mathbf{S}_{22} + \mathbf{I}) + \mathbf{D}_p^{-1}, \end{pmatrix}^{-1} \tag{4.40}$$

where $\mathbf{U} = \mathbf{W}_{11}^{-1}$ is the inverted upper diagonal block. We observe a regular structure within the blocks in $\mathbf{C}^{-1}$: the matrices $\mathbf{D}_i^{-1}$ are added to the diagonals blocks only, and the non-diagonal blocks only differ by scalar factors u_{ij}. With a block-wise Cholesky factorisation, the inversion requires only pq^3 operations. Since the Cholesky decomposition of the blocks also only differs by a factor, we can store its intermediate result. □

Remark If there are further memory constraints, distributed versions of the Cholesky decomposition should be considered to enhance performance.

Theorem 5(2) states that $\mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{S}, \mathbf{T}$ follows the MGIG distribution. In order to sample from this distribution, we make use of a result by Bernadac (1995). It introduces a representation of an MGIG-distributed random variable as a limit of a random continued fraction of Wishart-distributed random variables. The interested reader should refer to Letac (2000); Bernadac (1995); Koudou et al. (2014) for the details. Drawing samples from the MGIG thus reduces to iterated sampling from the Wishart distribution. In practice, we observe the convergence of the random continued fraction within few iterations. The complexity of sampling from the distribution derived in Theorem 5(2) does not depend on q.

4.5.2. A Note About The Graphical Lasso

A natural question that arises is how the BMB solution presented here compares to existing frequentist techniques, particularly the classical graphical lasso due to Friedman et al. (2008). The BMB uses the same likelihood as the graphical lasso. As a result, comparing both techniques reduces to comparing Bayesian

inference with maximum likelihood inference. Evidently, such a comparison reveals that the BMB provides us with a posterior distribution that expresses the confidence in a solution, while the graphical lasso only returns a point estimate. It should also be noted that BMB and the graphical lasso are virtually identical if a highly peaked prior is used.

4.5.3. Regularisation Parameter ρ and λ

We aim to compare the regularisation parameters ρ from the graphical lasso and λ from the Bayesian graphical lasso and Markov blanket estimate. The posterior distribution whereof the Bayesian graphical lasso and the Bayesian Markov blanket estimate draw is composed of the Wishart likelihood and a double exponential/ Laplace prior. Neglecting all terms not including $\mathbf{W}$, this results in

$$p(\mathbf{W}|\mathbf{S}) \propto \det(\mathbf{W})^{\frac{n}{2}} \exp \operatorname{tr}\left(-\frac{1}{2}\mathbf{W}\mathbf{S}\right) \prod_{i<j} \exp(-\lambda|w_{ij}|) \prod_{i=1}^{p} \exp(-\lambda x)_{x>0}. \tag{4.41}$$

On the other hand, the graphical lasso is composed of the logarithm of Wishart likelihood and a L_1 regulariser on $\mathbf{W}$, whereas the log-likelihood is reduced to terms which include $\mathbf{W}$. This results in

$$\log\det(\mathbf{W}) - \operatorname{tr}\left(\bar{\mathbf{S}}\mathbf{W}\right) - \rho\|\mathbf{W}\|_1. \tag{4.42}$$

To compare Eq. 4.41 with Eq. 4.42 the latter is multiplied by $\frac{n}{2}$ and exponentiated

$$\begin{aligned} &\exp\left(\frac{n}{2}\det(\mathbf{W}) - \frac{n}{2}\operatorname{tr}\left(\bar{\mathbf{S}}\mathbf{W}\right) - \frac{n}{2}\rho\|\mathbf{W}\|_1\right) \\ &= \det(\mathbf{W})^{\frac{n}{2}} \exp\operatorname{tr}\left(\frac{1}{2}\mathbf{S}\mathbf{W}\right) \exp\left(\frac{n}{2}\rho\|\mathbf{W}\|_1\right) \\ &= \det(\mathbf{W})^{\frac{n}{2}} \exp\operatorname{tr}\left(\frac{1}{2}\mathbf{S}\mathbf{W}\right) \exp\left(\frac{n}{2}\rho\sum_{i\leq j}|w_{ij}|\right). \end{aligned} \tag{4.43}$$

Now, comparing Eq. 4.41 with Eq. 4.43, we conclude that

$$\lambda = \frac{n}{2}\rho. \tag{4.44}$$

4.6. Extension with Gaussian Copula

We extend the model for non-Gaussian and mixed continuous/discrete data by embedding it within a copula construction. Copulas describe the depen-

dency in a r-dimensional joint distribution $F(Y_1, \ldots, Y_r)$ and represent an invariance class with respect to the marginal cumulative distribution functions (cdf) F_i. In the model, $r = p + q$. For continuous cdfs, Sklar's theorem (Sklar, 1959) guarantees the existence and uniqueness of a copula C, such that $F(Y_1, \ldots, Y_r) = C(F_1(Y_1), \ldots, F_r(Y_r))$. For discrete cdfs, this leads to an identifiability problem (Genest and Neslehova, 2007), such that established methods on empirical marginals (Liu et al., 2009) cannot be used anymore, but a valid copula can still be constructed (Genest and Neslehova, 2007). For this purpose, we follow the semi-parametric approach by Hoff (2007) and restrict the model to the parametric Gaussian copula, but we do not restrict the data to be Gaussian and treat them in a non-parametric fashion. The Gaussian copula inherently implies latent variables $X_i = \Phi^{-1}(F_i(Y_i))$. The model under consideration is

$$(X_1, \ldots, X_r)^\mathsf{T} \sim \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad Y_i = F_i^{-1}(\Phi(X_i)), \tag{4.45}$$

where F_i^{-1} denotes the ith generalised inverse of continuous or discrete cdfs, $\mathbf{X}$ are the latent variables, and $\mathbf{Y}$ are the observations.

Following Hoff (2007), inference in the latent variables uses the non-decreasing property of discrete cdfs for transforming the observed variables to the latent space. This guarantees that for observations $y_{ik} < y_{il}$ we also have $x_{ik} < x_{il}$, and more generally, $\mathbf{X}$ must lie in the set

$$\mathcal{D} = \{\mathbf{X} \in \mathbb{R}^{r \times n} : \max(x_{ik} : y_{ik} < y_{ij}) < x_{i,j} < \min(x_{ik} : y_{ij} < y_{ik})\}.$$

The data likelihood can then be written as

$$\begin{aligned} p(\mathbf{Y}|\boldsymbol{\Sigma}, F_1, \ldots, F_r) &= p(\mathbf{X} \in \mathcal{D}, \mathbf{Y}|\boldsymbol{\Sigma}, F_1, \ldots, F_r) \\ &= p(\mathbf{X} \in \mathcal{D}|\boldsymbol{\Sigma}) p(\mathbf{Y}|\mathbf{X} \in \mathcal{D}, \boldsymbol{\Sigma}, F_1, \ldots, F_r) \end{aligned}$$

and estimation of $\boldsymbol{\Sigma}$ is performed on maximising the sufficient statistics $p(\mathbf{X} \in \mathcal{D}|\boldsymbol{\Sigma})$ only, thus treating the marginals F_i as nuisance parameters. Bayesian inference for $\boldsymbol{\Sigma}$ is achieved by a Markov chain having stationary distribution at the posterior $p(\boldsymbol{\Sigma}|\mathbf{X} \in \mathcal{D}) \propto p(\boldsymbol{\Sigma}) p(\mathbf{X} \in \mathcal{D}|\boldsymbol{\Sigma})$, where a inverse-Wishart prior $p(\boldsymbol{\Sigma})$ is used. Posterior inference can be achieved with a Gibbs sampler which draws alternately between $\mathbf{X}|\boldsymbol{\Sigma}, \mathbf{Y}$ and $\boldsymbol{\Sigma}|\mathbf{X}$. This sampler extends Alg. 10 with an additional outer loop for inferring the latent variables. The Markov blanket is then iteratively estimated on these variables. The sampling scheme easily accommodates for missing values, when omitting conditioning on the set $\mathcal{D}$.

The presented framework is very useful in practice, since the invariance class of copulas extend the model to non-Gaussian data. With the additional

stochastic transformation to the latent space, we can use discrete variables and allow missing values. In real world applications, it becomes apparent that this is a very valuable extension.

4.7. Experiments

4.7.1. Artificial Data

As a first experiment, we attempt to highlight the differences in inference between the Bayesian Markov blanket (BMB) and Bayesian Graphical Lasso (BGL) procedures. We construct an artificial network with 100 variables, where the interest is confined to only the Markov blanket between $p = 10$ query variables and the $q = 90$ remaining variables. In order to create networks with a "small-world" flavour containing *hubs*, i.e. nodes with very high degree, the connectivity structure of the inverse covariance matrix $\mathbf{W}$ is generated by a beta-binomial model. Edge weights are sampled uniformly from the interval $[0.3, 1]$, and edge signs are randomly flipped. Finally, positive definiteness is guaranteed by adding a suitable constant (related to the smallest eigenvalue) to the diagonal. This process produces a sparse network structure where the majority of edges are connected to only a few single nodes. Note that many real-world networks exhibit such small-world properties. Since we are interested in estimating the Markov blanket, we require $\mathbf{W}_{12}$ to contain a minimum of 15 non-zero elements (= edges) out of a theoretical maximum of $p \times q = 900$. The parameters were chosen to produce sparse, but reasonably challenging network topologies: the full matrix $\mathbf{W}$ of size $(p + q) \times (p + q)$ is composed of around 100 non-zero entries in total (excluding its main diagonal to ensure positive definiteness).

Next, we draw $n = 1\,000$ independent samples from a zero-mean normal distribution with covariance matrix $\mathbf{W}^{-1}$ and compute the sample covariance $\mathbf{S}$. Fig. 4.4 depicts a true Markov blanket and its reconstruction by BGL and BMB using the same sparsity parameter $\lambda = 200$. Both methods were run side-by-side for 700 MCMC samples after an initial burn-in phase of 300 samples. From the sampled networks, a representative network structure is constructed by thresholding based on a 85% credibility interval, which is shown in the plots in Fig. 4.4. We repeat the above procedure to obtain a total of 100 datasets. The quality of reconstructed networks is measured in terms of f-score (harmonic mean of precision and recall) between the true and inferred Markov blanket. When computing precision and recall, inferred edges with edge weights having the wrong sign are counted as missing. Both models share the

same sparsity parameter λ, which in this experiment was selected such that for BMB recall and precision have roughly the same value. The results are depicted as box plots in Fig. 4.5, from which we conclude that there are indeed substantial differences in both models. In particular, BGL has the tendency to introduce many unnecessary edges in comparison to BMB. As a result, BGL achieves high recall and low f-score. Since both methods are based on the same likelihood model and (almost) the same prior, the observed differences can only be attributed to differences in the inference procedure: BGL infers a network by iterating over *all* variables and their neighbourhood systems, whereas BMB only estimates the elements in $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$.

To further study the influence of the different Gibbs sampling strategies, we examine tracer plots and auto-correlations of individual variables in Fig. 4.6. In almost all cases, BGL shows significantly higher auto-correlation and poor convergence. In contrast, Markov chains in the BMB sampler seems to mix much better, typically leading to posteriors with smaller bias and variance. While only one example is shown in the figure, similar results can be seen for basically all variables in the network. Overall, BMB has a significant practical advantage when only the Markov blanket of a network is required since it successfully exploits the Wishart factorization to estimate $\mathbf{W}_{12}$ independently of $\mathbf{W}_{22}$. Further, we experience a substantial decrease in run-time, even for these relatively small networks: computing 1 000 MCMC samples for BMB finished on average after 100 seconds, while BGL typically consumed around 370 seconds. Since BGL requires an additional sampling loop over *all* variables, datasets with large $\mathbf{S}_{22}$ quickly become problematic for BGL. We further explore these differences in the next section for a large real-world application.

4.7.2. Real Data

To demonstrate the practical significance of Markov blanket estimation, we turn to the analysis of *colorectal cancer* which in 2012 ranked among the three most common types of cancer globally (Stewart and Wild, 2014). The data set introduced in Sheffer et al. (2009) is publicly available and contains gene expression measurements from biopsies of $n = 260$ cancer patients. A separate table captures discrete/categorical clinical traits such as sex, age or pathological staging/grading. In this context, one particularly interesting research question is to identify connections between the p (macroscopic) clinical descriptors and the q (molecular) gene expression measurements based on n patients which are treated as realizations. Note that in this setting, a network between only clinical features or only genes is *not* of interest and in part already well explored – instead the analysis specifically targets their *interaction*, hence the Markov

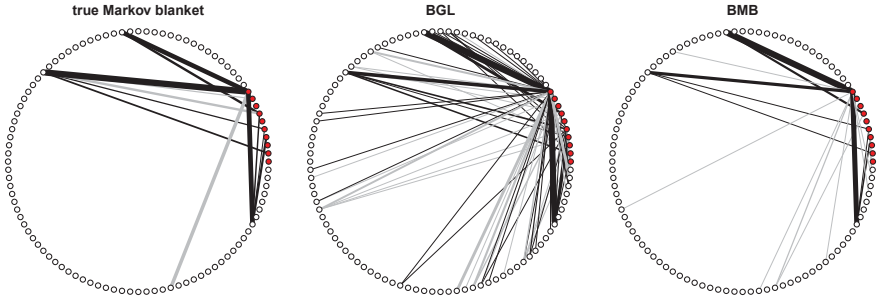


Figure 4.4.: One exemplary Markov blanket ($p = 10$, $q = 90$) and its reconstruction by BGL and BMB. Note that the graphs *only* display edges between p query and q remaining variables. Red nodes represent query variables, white nodes represent all other variables. Black and grey edges correspond to positive and negative edge signs, respectively.

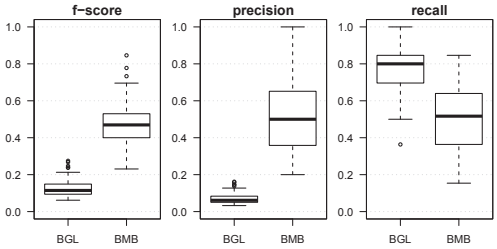


Figure 4.5.: Performance of inferred Markov blankets from 100 datasets.

blanket.

Among the 13 400 genes contained in the dataset, we focus on a specific subset, the so-called "*Pathways in cancer*" as defined in the KEGG database[3]. This particular subset comprises a general class of genes which are known to be involved in various biological processes linked to cancer. For this experiment, we have $q = 312$ candidate genes and $p = 7$ query variables. These are the age and sex of the patient as well as the *TNM* classification, cancer group stage (*GS*) and mutation of the tumour suppressor protein *p53*. Since the observations have mixed continuous/discrete data types with missing values, the

[3]Kyoto Encyclopedia of Genes and Genomes, `http://www.genome.jp/kegg/pathway.html`

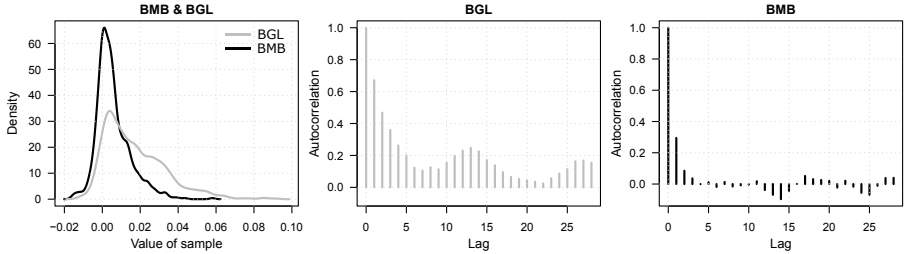


Figure 4.6.: Density and auto-correlation of the Markov chain for a single variable in the Markov blanket. Gray refers to BGL, black to BMB.

Markov blanket estimation is extended by a semi-parametric Gaussian copula framework (Hoff, 2007). Based on this, we calculate 5 000 MCMC samples, consuming just over 2 hours, which finally leads to the Markov blanket in Fig. 4.7. The red nodes represent the p clinical features under consideration, while the white nodes represent the q genes. Here, the result is obtained by defining an 80% credibility interval over the full Markov chain after discarding the first 1 500 burn-in samples.

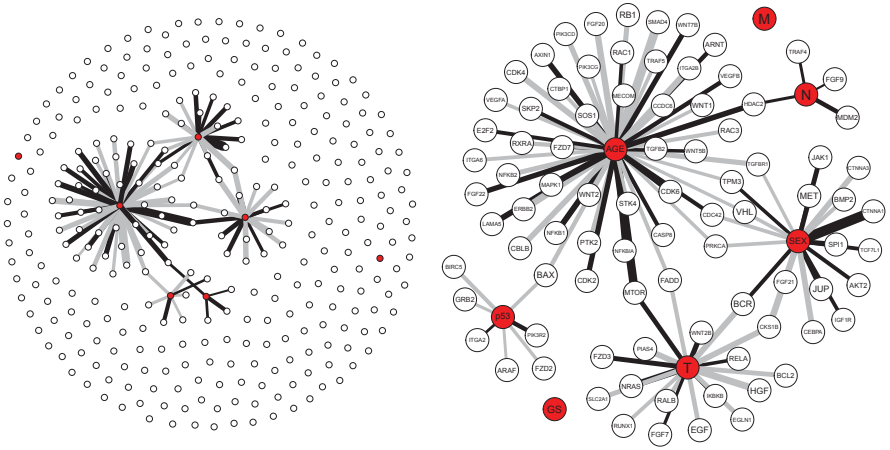


Figure 4.7.: Sparse Markov blanket between $p = 7$ clinical features (red nodes) and $q = 312$ genes in colorectal cancer (Sheffer et al., 2009). Overview of all variables/nodes (left) and enlarged, fully labelled sub-graph (right).

The resulting network structure confirms some well-known properties like the confounding effect of the age and sex variables, both of which (correctly) link to a large number of genes. For example, *FGF21* exhibits significant differences in male and female subjects (Bisgaard et al., 2014), and *CTNNA1* shares connections to survival time in men (Ropponen et al., 1999). Similarly, *mTOR*, the *mechanistic target of rapamycin*, not only represents a key element for cell signaling that triggers a cascade of immune-related pathways, but its function also depends heavily on a subject's age (Johnson et al., 2013). Despite these age- and sex-related observations being non-trivial, they are not of primary interest, which is why the remaining variables carry more practical insights from a clinical point view.

Further, we are able to identify a very interesting network structure around the variable *tumour size* T: almost all direct neighbours control either cell growth (*EGLN1* (Erez et al., 2003), *RELA* (Yu et al., 2004), *HGF* (Renzo et al., 1995; Date et al., 1998) and others) or cell death (*BCL2*, *FADD*). Cancer typically affects the balance between these two fundamental processes and their deregulation eventually leads to tumour development. A second sub-graph concerns variable N, the degree of spread to regional lymph nodes, which is expressed in 4 levels $N0$ to $N3$. Here, all genes in the neighbourhood correspond to the lymphatic system and its direct responses to malignant cell growth, which was confirmed for *FGF9* (Deng et al., 2013), *MDM2* (Leitea et al., 2001; Fridman et al., 2003) and *TRAF4* (Camilleri-Broet et al., 2007) among others. Finally, the following two clinical variables appear to be conditionally independent from genes, yet they may internally depend on other clinical variables (i.e., outside of the Markov blanket): binary M (presence of metastasis in distant organs) and discrete *GS* (group stage of cancer). Interestingly, the latter is only a summary function of T, N and M, hence internal links to the aforementioned variables are very likely.

Despite the study's focus on colorectal cancer and specifics of the intestinal system, the inferred Markov blanket is able to explain rather general properties in accordance with findings in the medical literature. Altogether, this nicely illustrates how the Gaussian copula framework complements the Bayesian Markov blanket estimation – especially pertaining to the clinical domain with mixed observations and missing values.

In contrast to our approach, the high dimensionality of this dataset imposes severe problems for BGL. For BMB, 5 000 Gibbs sweeps could be computed in 2 hours, and MCMC diagnosis did not show any severe convergence problems. For BGL, however, the same number of iterations already took 122 hours (≈ 5 days), and we observed similar (and sometimes severe) mixing problems as

described in the previous section.

4.8. Conclusion

We have presented a Bayesian perspective for estimating the Markov blanket of a set of query nodes in an undirected network. In our experience, it is often the case that we estimate a full network but interpret only part of it. This is especially true in a context where portions of the data are qualitatively different. Here, we would be more interested in establishing the links between these portions, rather than examining the links within the portions themselves. Markov blanket estimation is hence an interesting and relevant sub-problem of network estimation, particularly in high dimensional settings. Existing methods such as the Bayesian graphical lasso iterate through the individual variables to estimate an entire network. While there are several situations in which inference of the entire network is required, there are also cases in which we are only interested in the neighbourhood of a small subset of query variables; for these instances, iterating through all the variables is unnecessary.

In the preceding, we explored the block-wise factorisation of the Wishart likelihood in combination with a suitable choice of prior. The primary contribution in Theorem 5 shows that the resulting posterior distribution of the Markov blanket of a set of query nodes has an analytic form, and is independent of a large portion of the network. The analytic form allows us to explore potentially large neighbourhoods where the Bayesian graphical lasso reaches its limits. We also demonstrated that sampling from the posterior of the Markov blanket is more efficient than the Bayesian graphical lasso. Moreover, we observed fast convergence and superior mixing properties of the Markov chain. We attribute this to the improved flexibility of the sampling strategy.

Including a copula construct in the model further enhances its real world applicability, where mixed data and missing values are prevalent. A particular application in a medical setting is the colorectal example we considered in Section 4.7.2. Using this approach allowed us to make interesting observations about the interactions between various clinical and genetic factors. Such insights could ultimately contribute to a better understanding of the disease.

5. Time-resolved Information Flows

5.1. Introduction

Granger causality (Granger, 1969) is a paradigm to measure causal influence between time series. In an informal way, a time series X^n Granger-causes time series Y^n, if knowing the past of X^n and Y^n helps in predicting the future of X^n compared to only knowing the past of Y^n only. In the context of information theory, directed information (Massey, 1990) and transfer entropy (Schreiber, 2000) quantify the causal influence between time series in the spirit of Granger.

In order to analyse the entire complexity of causal flows between time series, we decompose directed information into its building blocks. This leads to a new notion of time-resolved information flows, which has the capability to represent the associations within the time series in a non-stationary setting. We give a new interpretation of time-resolved information flows within the setting of Pearlian directed acyclic graph (Pearlian DAG) (Dawid, 2010). We show how to estimate approximate information flows with the Gaussian copula and apply the method on electroencephalography (EEG) data of visually evoked potentials (VEP).

It is convenient to use a probabilistic graphical model for representing the dependencies between time series (Eichler, 2012). There are two principled possibilities for representing directed information in graphs. A simple approach is to define a graph $G = (V, E)$, where the vertices $v \in V$ each represent a time series and the edges $e \in E$ the associations between the them. Such graphs, an example is depicted in Fig. 5.1, have appealing properties when the associations refer to transfer entropy (Quinn et al., 2015). However, a causal interpretation is limited to the stationary case. In order to represent the entire complexity of time series, the graph can be unrolled over time so that every time point of every time series is represented by a vertex. Such a graph is depicted in Figs. 5.2 and 5.3. A corresponding interpretation of directed information is presented in (Wieczorek and Roth, 2016), which also covers the non-stationary case.

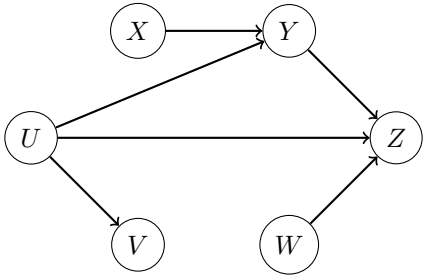


Figure 5.1.: Graphical model for a directed acyclic graph.

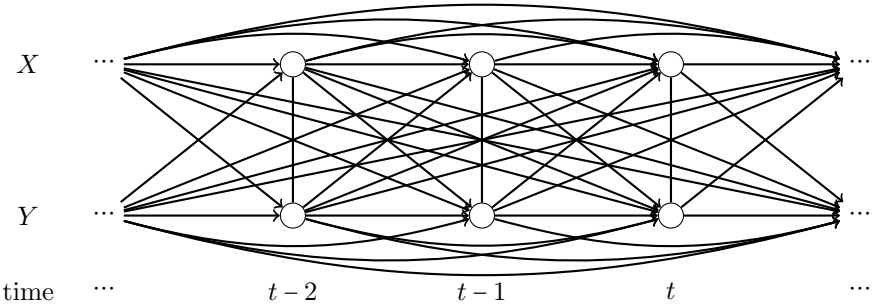


Figure 5.2.: Template for unrolled directed graphical model.

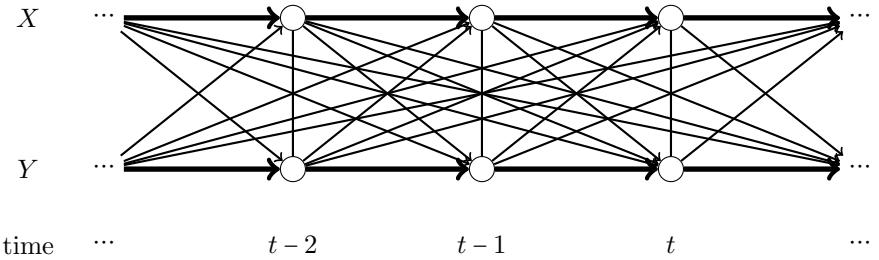


Figure 5.3.: Template for unrolled directed graphical model. Bold arrows represent a Markov order > 1.

5.2. Time-resolved Information Flows

In the following, directed information is factorised into time-resolved information flows. We start with the definition of directed information

$$I\left(X^n \to Y^n\right) = \sum_{i=1}^{n} I\left(X^i; Y_i | Y^{i-1}\right) \tag{5.1}$$

in order to expand it into information flows $I\left(X^i; Y_i | Y^{i-1}\right)$. We then recursively use the chain rule on the information flows

$$\begin{aligned}
&I\left(X^i; Y_i | Y^{i-1}\right) \\
&= I\left(X^{i-1}; Y_i | Y^{i-1}\right) + I\left(X_i; Y_i, | X^{i-1}, Y^{i-1}\right) \\
&= I\left(X^{i-2}; Y_i | Y^{i-1}\right) + I\left(X_{i-1}; Y_i | X^{i-2}, Y^{i-1}\right) + I\left(X_i; Y_i, | X^{i-1}, Y^{i-1}\right) \\
&= I\left(X_1; Y_i | Y^{i-1}\right) + I\left(X_2; Y_i | X_1, Y^{i-1}\right) + \ldots + \\
&\qquad + I\left(X_{i-1}; Y_i | X^{i-2}, Y^{i-1}\right) + I\left(X_i; Y_i, | X^{i-1}, Y^{i-1}\right) \\
&= \sum_{k=1}^{i} I\left(X_k; Y_i | X^{k-1}, Y^{i-1}\right)
\end{aligned} \tag{5.2}$$

and plug this result into the definition in Eq. 5.1

$$I\left(X^n \to Y^n\right) = \sum_{i=1}^{n} \sum_{k=1}^{i} I\left(X_k; Y_i | X^{k-1}, Y^{i-1}\right) \tag{5.3}$$

From this expansion, we identify the time-resolved causally conditioned mutual informations $I\left(X_k; Y_i | X^{k-1}, Y^{i-1}\right)$ as the elementary building blocks for directed information. We call such a building block a *fully time-resolved information flow*, since the information flow can be contributed to a causal flow from time i to k. Having identified such an elementary block, the notion of an information flow can be generalised in order to construct different information flows. In order to do so, note that an unrolled information flow $I\left(X^i; Y_i | Y^{i-1}\right)$ is an asymmetric measure of causality: the asymmetry is not only in the directionality from X to Y, but also from many time points of the past of X to one time point in the presence of Y. Thus, we will call such an information flow *inflow*, an acronym for incoming flow. Now, the motivation is to find its counterpart. In particular, we will define a quantity, which we call *outflow*, short for outgoing flow. An outflow is defined as the information flow from one time point in the presence of X to many time points in the future of Y, in particular for time k, an outflow is defined as

$$I(X_k \to Y^n) = I\left(X_k; Y_k^n | X^{k-1}, Y^{k-1}\right). \tag{5.4}$$

Note, an inflow is a building block for directed information for assessing Granger causality. Analogously, an outflow is an information theoretic equivalent of a building block of Sims causality (Sims, 1972; Saito and Harashima, 1981; Florens and Mouchart, 1982; Kamitake et al., 1984). Sims causality states that X^n does not cause Y^n, if the future of Y^n is conditionally independent of the present of X^n given the past of X^n and the past of Y^n. We summarise the definitions of inflow, instantaneous coupling, fully time-resolved information flow, and outflow graphically in Figs. 5.4 - 5.7, respectively. The shading in the figures corresponds to the respective conditioning sets.

In the following, we present the most important properties of outflows, namely, directed information and transfer entropy can also be expanded in terms of outflow.

Theorem 7 (directed information as outflow)**.** *Let X^n and Y^n be two (possibly non-stationary) time series. Then, directed information $I\left(X^n \to Y^n\right)$ can be expanded in terms of inflow as well as in terms of outflow. Formally,*

$$\sum_{i=1}^{n} I\left(X^i; Y_i | Y^{i-1}\right) = \sum_{k=1}^{n} I\left(X_k; Y_k^n | X^{k-1}, Y^{k-1}\right). \tag{5.5}$$

Proof. We factorise the outflow at time k according to the chain rule for mutual information

$$\begin{aligned}
&I\left(X_k; Y_k^n | X^{k-1}, Y^{k-1}\right) \\
&= I\left(X_k; Y_k | X^{k-1}, Y^{k-1}\right) + I\left(X_k; Y_{k+1}^n | X^{k-1}, Y^{k}\right) \\
&= I\left(X_k; Y_k | X^{k-1}, Y^{k-1}\right) + I\left(X_k; Y_{k+1} | X^{k-1}, Y^{k}\right) + I\left(X_k; Y_{k+2}^n | X^{k-1}, Y^{k+1}\right) \\
&= I\left(X_k; Y_k | X^{k-1}, Y^{k-1}\right) + I\left(X_k; Y_{k+1} | X^{k-1}, Y^{k}\right) + \ldots + \\
&\qquad + I\left(X_k; Y_n | X^{k-1}, Y^{n-1}\right) \\
&= \sum_{\ell=k}^{n} I\left(X_k; Y_\ell | X^{k-1}, Y^{\ell-1}\right)
\end{aligned} \tag{5.6}$$

Summing the outflow over time $k = 1, \ldots, n$

$$\sum_{k=1}^{n} \sum_{\ell=k}^{n} I\left(X_k; Y_\ell | X^{k-1}, Y^{\ell-1}\right) \tag{5.7}$$

and comparing Eq. 5.3 to Eq. 5.7 gives the desired result. □

This result forms an information theoretic equivalent to Chamberlain (1982), who showed a general equivalence of Granger and Sims causality. A similar theorem can also be stated for transfer entropy.

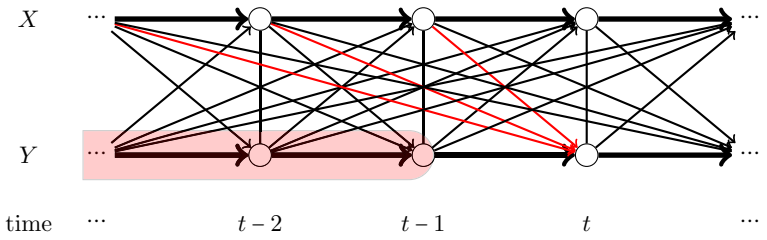


Figure 5.4.: Red arrows: inflow $I(X^{t-1} \to Y_t)$.

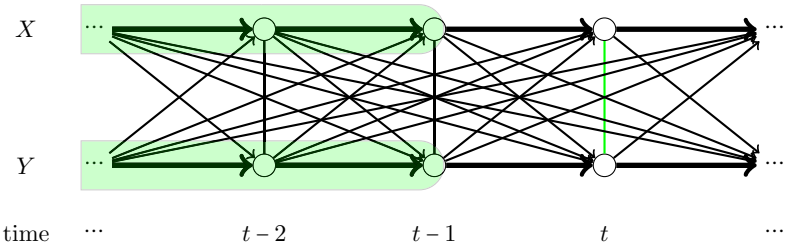


Figure 5.5.: Green edge: instantaneous coupling $I(X_t; Y_t | X^{t-1}, Y^{t-1})$.

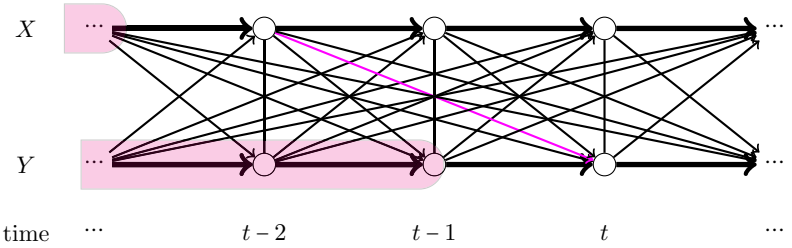


Figure 5.6.: Magenta arrow: fully time-resolved information flow $I(X_{t-2}; Y_t | X^{t-3}, Y^{t-1})$.

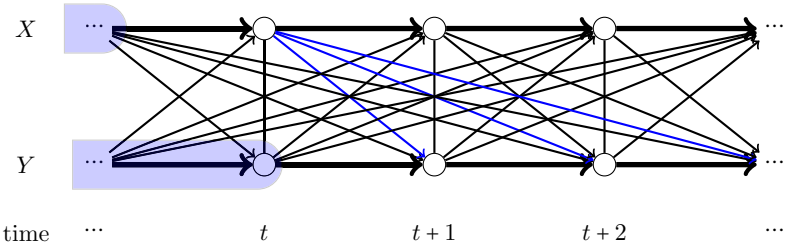


Figure 5.7.: Blue arrows: outflow $I(X_t; Y_{t+1}^n | X^{t-1}, Y^t)$.

Theorem 8 (transfer entropy as outflow)**.** *Let X^n and Y^n be two stationary time series. Then, transfer entropy can be expressed in terms of inflow as well as in terms of outflow. Formally,*

$$\lim_{i\to\infty} I\left(X^{i-1}; Y_i | Y^{i-1}\right) = \lim_{n\to\infty} I\left(X_k; Y_{k+1}^n | X^{k-1}, Y^k\right). \tag{5.8}$$

More precisely, in the limits, the inflow is equal to the outflow.

Proof. In a strict stationary time series, the fully time-resolved information flows do not depend on time k, i.e.

$$I(X_k; Y_{k+\ell} | X^{k-1}, Y^{k+\ell-1}) = I(X_{k+\tau}; Y_{k+\tau+\ell} | X^{k+\tau-1}, Y^{k+\tau+\ell-1}). \tag{5.9}$$

for all $\tau \in \mathbb{Z}$. Comparing Eq. 5.2 to Eq. 5.6 in the limits $\lim_{i\to\infty}$ and $\lim_{n\to\infty}$, respectively, gives the desired result. □

Note, as transfer entropy is defined without instantaneous coupling, we also neglecting the corresponding term. However, it could be included in the theorem as well.

5.2.1. Graphical Interpretation

As mentioned in the beginning of this chapter, a probabilistic graphical model can be useful for analysing information flows. This section aims at clarifying the role of time-resolved information flows in a Pearlian DAG.

As stated in (Massey, 1990) for noisy communication channels, mutual information measures the transmitted information, when no feedback is present. However, whenever feedback is available, directed information is a more useful quantity, since it quantifies the causal information flow in one direction only. Raginsky (2011) showed that this notion of causality actually fits into the framework of interventions by Pearl (2009), thus opening the connections between graphical models and directed information. This led to a more general definition of directed information: for any disjoint sets $S, T \in V$ in a graph $G = (V, E)$, directed information is defined as

$$I(S \to T) = D_{KL}\left(P_{S|T} \| P_{S|do(T)} | P_T\right). \tag{5.10}$$

In particular, directed information can be interpreted as the difference between an observed and an intervened distribution. In words of Pearl, what is called an observed distribution is the pre-interventional distribution $P(S|T)$, and the intervened distribution corresponds to the post-interventional distribution $P(S|do(T))$. Directed information has an interesting interpretation since

it compares those distributions and takes the expectation over the distribution of the intervened variable T. Thus, directed information is the average causal effect (Angrist et al., 1996) in terms of entropy when observing the data in the intervened instead of the observed model. The explicit connection for time series is demonstrated in App. A.3. Note, the KL-divergence emerges naturally in the setting of information theory.

Building on this result, we provide analogous interpretations for time-resolved information flows. In particular, the interpretation of time-resolved information flows as differences between an observed and an intervened distribution is as follows: in a Pearlian DAG, an inflow can be interpreted as

$$I(X^i \to Y_i) = D_{KL}\left(P_{X^i|Y^i} \| P_{X^i|Y^{i-1},do(Y_i)} | P_{Y^i}\right), \tag{5.11}$$

the interpretation of an outflow is

$$I(X_i \to Y^n) = D_{KL}\left(P_{X_i|X^{i-1},Y^n} \| P_{X_i|X^{i-1},Y^{i-1},do(Y_i^n)} | P_{X^{i-1},Y^n}\right), \tag{5.12}$$

and the interpretation of a fully time-resolved information flow is

$$I(X_i \to Y_k) = D_{KL}\left(P_{X_i|X^{i-1},Y^k} \| P_{X_i|X^{i-1},Y^{k-1},do(Y_k)} | P_{X^{i-1},Y^k}\right). \tag{5.13}$$

The derivation of those expressions is given in App. A.3. Note, contrary to standard notation (Cover and Thomas, 2012), we made explicit the averaging over the conditioning sets of the KL-divergences.

5.3. Discovering Information Flows in Non-Stationary Time Series

We aim at discovering time-resolved information flows in non-stationary time series. In our view, outflows and inflows provide a useful measures for discovering information flows in a general setting. What we call a general setting is subject to the following assumptions:

1. the causal information flows are non-stationary
2. the maximal length of a causal flow is limited by the Markov order of the time series

The first assumption describes that the causal flows do not follow any specific structure. The causal flows can be asymmetric and vary over time in length and strength. Thus, arbitrary flows are allowed as long as the network is a

directed acyclic graph. This also includes configurations like hubs: a nodes with a high in-degree represents essentially a temporary information sink with a distributed source, and a node with a high out-degree corresponds generally to a temporary information source with a distributed sink.

The second assumption is mainly a technical assumption. The direct use of in- and outflow is computational impractical since such an estimator considers many variables. Limiting the Markov order as well as maximal length of a causal flow allows using windows of fixed lengths. For a Markov model of fixed ordering, the conditioning set for in- and outflow can be fixed accordingly, such that paths from all potential ancestors are blocked.

The second assumption allows to measure information flows in different segments of the time series, such that a time-resolved analysis of information flows is possible. This essentially enables to recover non-stationary information flows which were described in assumption 1. We define the time-resolved inflow at time i for a time series of Markov order L as

$$I\left(X_{i-L}^{i}; Y_i | Y_{i-L}^{i-1}\right). \tag{5.14}$$

Note, the length of the window is $L+1$. Analogously, we define the time-resolved outflow at time k for a time series of Markov order L as

$$I\left(X_k; Y_k^{k+L} | X_{k-L}^{k-1}, Y_{k-L}^{k-1}\right). \tag{5.15}$$

Note that the length of the window is $2L+1$. For fully time-resolved information flows

$$I\left(X_i; Y_k | X_{i-L}^{i-1}, Y_{k-L}^{k-1}\right). \tag{5.16}$$

These definitions of information flows in windows lead directly to a practical estimator which estimates non-stationary information flows under assumptions 1 and 2. Whereas the last expression is the most general one, the expressions for in- and outflow may have the following interpretation which is equivalent to that of matched filters: in the presence of hub sink and hub sources, they maximise the signal-to-noise ratio, respectively. In the stationary case, the interpretation of time-resolved in- and outflows changes and they can be considered as approximations of transfer entropy which can be recovered for infinitely long windows $L \to \infty$.

5.3.1. A Gaussian Copula Estimator for Information Flows

We aim at computing the time-resolved information flows with a Gaussian copula model. Thus, inflow is decomposed into multiinformations as in A.8

and limit the indices to the window of length L. This gives

$$I(X_{i-L}^{i}; Y_i | Y_{i-L}^{i-1}) = M(X_{i-L}^{i}, Y_{i-L}^{i}) - M(Y_{i-L}^{i}) - M(X_{i-L}^{i}, Y_{i-L}^{i-1}) + M(Y_{i-L}^{i-1}). \tag{5.17}$$

Analogously, outflow is decomposed into multiinformations as in A.12 and the indices are limited to the window of length $2L + 1$

$$\begin{aligned} &I(X_k; Y_k^{k+L} | X_{k-L}^{k-1}, Y_{k-L}^{k-1}) \\ &= M(X_{k-L}^{k}, Y_{k-L}^{k+L}) - M(X_{k-L}^{k-1}, Y_{k-L}^{k+L}) - M(X_{k-L}^{k}, Y_{k-L}^{k-1}) + M(X_{k-L}^{k-1}, Y_{k-L}^{k-1}) \end{aligned} \tag{5.18}$$

Those decompositions enable to use the equivalence between multiinformation and negative copula entropy and the closed-form expression for the entropy of a Gaussian copula, namely

$$M(X_1, \ldots, X_p) = -H\left(c_{\mathbf{R}}^{\mathcal{N}}(U_1, \ldots, U_p)\right) = -\frac{1}{2} \log \det \mathbf{R}. \tag{5.19}$$

What remains is the correlation matrix $\mathbf{R} = \mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ in the latent space. Computing it in a semi-parametric fashion, we use empirical marginals F_{emp} to estimate the copula with normalized ranks

$$U_{\bullet j} = F_{\text{emp}}(Y_{\bullet j}) = \frac{\text{ranks}(\mathrm{Y}_{\bullet \mathrm{j}})}{n+1}, \tag{5.20}$$

and compute subsequently the normal scores

$$Z_{\bullet j} = \Phi^{-1}\left(U_{\bullet j}\right) = \Phi^{-1}\left(\frac{\text{ranks}(Y_{\bullet j})}{n+1}\right). \tag{5.21}$$

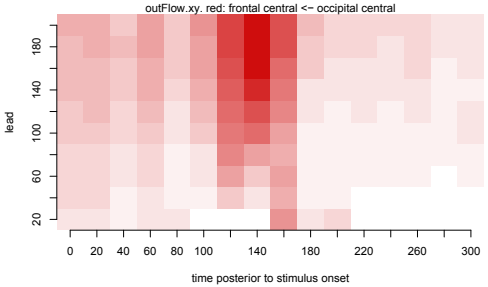
Using the Gaussian copula for estimating time-resolved information flows comes with several benefits. First, the semi-parametric approach, namely using the Gaussian copula with non-parametric marginals, provides a very flexible model. The model does not impose any assumptions on the marginals, but only on the dependency structure. Thus, the estimator is invariant against monotone transformations and is robust against outliers. Second, the decomposition of inflow and outflow into multiinformations and the use of the Gaussian copula allows to use a convenient closed-form expression. Compared to other methods (Hlaváčková-Schindler et al., 2007) based on e.g. binning, k-nearest neighbours, or kernels, the estimator has also amenable properties. First, it only suffers weakly from the curse of dimensionality, second, the computational complexity remains low, and third, there are no parameters to tune.

5.3.2. Application to Visual Evoked Potentials

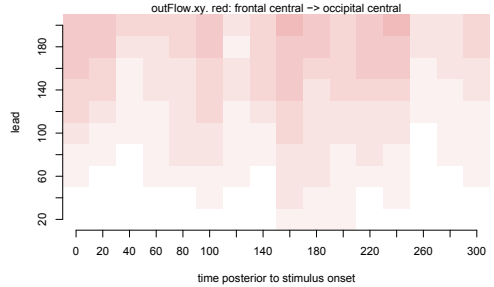
We present a preliminary result for electroencephalography (EEG) recordings of visually evoked potentials (VEP). A healthy control was repeatedly visually stimulated by a colour switching checker board and the electromagnetic response of the brain was recorded with a 256-channels EEG at a sampling rate of 1000 Hz (Hardmeier et al., 2014). The signal is bandpass filtered with passband between 5 and 45 Hz to eliminate high-frequency noise and further down sampled to 50 Hz according to the uncertainty principle/ Gabor limit. Independent component analysis (ICA) is used for localisation of responses in the space of the electrodes. Subsequently, the electrodes in regions with strong response to the stimulus were pooled. Fig. 5.8 shows inflows and outflows computed according to Eq. 5.14 and Eq. 5.15, respectively, for different window sizes L. An information flow is visible from the occipital central to the frontal central region during a period starting at the transition from P100 to N145 and ending at the transition from N145 to P240. Thus, the information flows occur at transition phases between stable topologies.

5.4. Conclusion

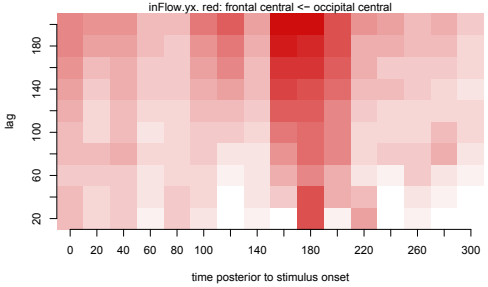
In this chapter, we introduced time-resolved information flows as building blocks of directed information. Based on this notion, we provide the information theoretic interpretation of Granger and Sims causality. In particular, we showed in Thms. 7 and 8 that directed information and transfer entropy can be expressed in terms of inflow as well as in terms of outflow. Those expansions lead to different quantifications of directed information and transfer entropy in terms of differences of observational and interventional distributions in a causal graphical model. Motivated on this interpretation, we defined an estimator for recovering time-resolved information flows and applied it to visually evoked potentials in EEG recordings.



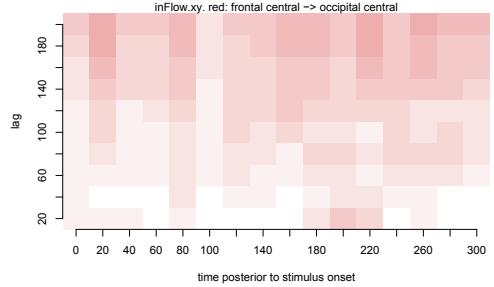
(a) Outflow from occipital central to frontal central.



(b) Outflow from frontal central to occipital central.



(c) Inflow from occipital central to frontal central.



(d) Inflow from frontal central to occipital central.

Figure 5.8.: Preliminary results from visually evoked potentials. (a) and (c): information flow from occipital central to frontal central. (b) and (d): in the opposite direction, no information flow in visible.

6. Conclusion and Outlook

In this thesis, we enlightened the role of copulas in different probabilistic models. The inherent scale-invariance of copula models provides an invariance class which is powerful yet simple and thus extend the applicability of different models to a much larger class of problems. We focused on the Gaussian copula, since its Gaussian distributed latent space allows extending established models to adopt the invariance properties of copulas. Applying a copula model thus enables a deep understanding of associations in data sets and allows for robust generalisation.

6.1. Representation of Data

We studied the Gaussian copula extension in dimensionality reduction algorithms. Within the framework of archetypal analysis, we showed that the dependency structure of the generative model of archetypal analysis is approximately Gaussian. This justifies the use of the Gaussian copula for approximating the the dependency structure of archetypal analysis.

We studied the Gaussian copula extension of principal component analysis in the context of parametric appearance models for faces. Thereby, the eigenfaces approach relaxes the assumption of Gaussian distributed colour marginals. This led to an increased specificity of the model and eliminated artefacts in random generated faces. Moreover, the copula model enabled to combine modalities measured on different scales in an unifying way. Thus, the colour model could be combined with further modalities like shape and other attributes such that the specificity could be further increased.

The copula extension of principal component analysis and archetypal analysis extend the applicability of the models substantially. A further advance could be reached by incorporating neural networks. Combining the invariance class of copulas with (deep) neural networks can lead to models with interesting properties. In particular, the universal approximation theorem (Cybenko, 1989; Hornik et al., 1989) states that a neural network has the capability of approximately computing any continuous function. In particular, a neural network is able to model functions which are not monotonic in the dimensions and

are of deterministic nature. In this way, neural networks are able to extend copula models in a considerable way.

6.2. Networks

In this thesis, we presented a Bayesian perspective for estimating the Markov blanket of a set of query nodes in an undirected network. We showed that in the Bayesian perspective, limiting the focus on the Markov blanket may be advantageous, in particular in the high-dimensional setting, where existing methods suffer from the curse of dimensionality. The extension to the copula network provides us with several benefits. Among those, the ability to include discrete ordered data is of special interest, since it extends the real world applicability substantially.

In the context of the conjecture that causality is a key for bringing forward machine intelligence, the incorporation of copula models in causal structure learning may be beneficial in the sense of providing more robust results than those of Gaussian models. Examples include a structure learning algorithms like a copula PC algorithm for nonparanormal graphical models (Harris and Drton, 2013) with the extension for mixed data (Cui et al., 2016), as well as intervention calculus when the underlying directed acyclic graph is absent (Nandy et al., 2014). However, in the case of hidden variables and/ or cycles/ feedback, there are still open questions.

6.3. Time Series

We analysed directed information and transfer entropy to quantify causal effects in time series. The definition of directed information as a difference between the observational and interventional distribution is an appealing interpretation which can be generalised to time-resolved information flows. This interpretation motivates to define time-resolved estimators for causal flows which we applied preliminarily to EEG recordings.

However, there is always a risk of interpreting statistical dependency as causality and the question arises if it is possible to quantify causal effects from observational data only. This question has several answers with graphical interpretations, e.g. in terms of the (generalised) back-door/ adjustment criterion (Pearl, 2009; Maathuis et al., 2015; Perković et al., 2015) and in terms of direct and indirect or mediated effects (Pearl, 2001, 2014). In the context of time

series and directed information, this question leads to a non-trivial analysis of the underlying graph.

A. Results for Information Theory

A.1. Equivalence of Granger and Sims Causality

We restate (Chamberlain, 1982) in the language of information theory.

We expand directed information in the spirit of Granger causality (Granger, 1969)

$$I(X^n \to Y^n) = \sum_{k=1}^{n} I\left(X^i; Y_i | Y^{i-1}\right) \tag{A.1}$$

and recursively use the chain rule on X^i to factorise the inflows

$$\begin{aligned}
&I\left(X^i; Y_i | Y^{i-1}\right) \\
&= I\left(X^{i-1}; Y_i | Y^{i-1}\right) + I\left(X_i; Y_i, | X^{i-1}, Y^{i-1}\right) \\
&= I\left(X^{i-2}; Y_i | Y^{i-1}\right) + I\left(X_{i-1}; Y_i | X^{i-2}, Y^{i-1}\right) + I\left(X_i; Y_i, | X^{i-1}, Y^{i-1}\right) \\
&= I\left(X_1; Y_i | Y^{i-1}\right) + I\left(X_2; Y_i | X_1, Y^{i-1}\right) + \ldots + I\left(X_{i-1}; Y_i | X^{i-2}, Y^{i-1}\right) + \\
&\quad I\left(X_i; Y_i, | X^{i-1}, Y^{i-1}\right) \\
&= \sum_{k=1}^{i} I\left(X_k; Y_i | X^{k-1}, Y^{i-1}\right)
\end{aligned} \tag{A.2}$$

and plug this result into the definition

$$I\left(X^n \to Y^n\right) = \sum_{i=1}^{n} \sum_{k=1}^{i} I\left(X_k; Y_i | X^{k-1}, Y^{i-1}\right). \tag{A.3}$$

We expand directed information in the spirit of Sims causality (Sims, 1972; Florens and Mouchart, 1982)

$$I(X^n \to Y^n) = \sum_{k=1}^{n} I\left(X_k; Y_k^n | X^{k-1}, Y^{k-1}\right) \tag{A.4}$$

and recursively use the chain rule on Y_k^n to factorise the outflows

$$\begin{aligned}
&I\left(X_k; Y_k^n | X^{k-1}, Y^{k-1}\right) \\
&= I\left(X_k; Y_k | X^{k-1}, Y^{k-1}\right) + I\left(X_k; Y_{k+1}^n | X^{k-1}, Y^k\right) \\
&= I\left(X_k; Y_k | X^{k-1}, Y^{k-1}\right) + I\left(X_k; Y_{k+1} | X^{k-1}, Y^k\right) + I\left(X_k; Y_{k+2}^n | X^{k-1}, Y^{k+1}\right) \\
&= I\left(X_k; Y_k | X^{k-1}, Y^{k-1}\right) + I\left(X_k; Y_{k+1} | X^{k-1}, Y^k\right) + \ldots + I\left(X_k; Y_n | X^{k-1}, Y^{n-1}\right) \\
&= \sum_{\ell=k}^{n} I\left(X_k; Y_\ell | X^{k-1}, Y^{\ell-1}\right)
\end{aligned} \tag{A.5}$$

and plug this result into the definition

$$I(X^n \to Y^n) = \sum_{k=1}^{n} \sum_{\ell=k}^{n} I\left(X_k; Y_\ell | X^{k-1}, Y^{\ell-1}\right). \tag{A.6}$$

Comparing Eq. A.3 to Eq. A.6, we recognise that both expansions contain the same terms.

A.2. Decompositions of Directed Information

Similar to Liu (2012), we decompose directed information to obtain easy to compute expressions.

$$\begin{aligned}
I(X^n \to Y^n) &= \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}) \\
&= \sum_{i=1}^{n} I(X_i; Y_i^n | X^{i-1}, Y^{i-1}) \\
&= \sum_{i=1}^{n} \sum_{k=i}^{n} I\left(X_i; Y_k | X^{i-1}, Y^{k-1}\right)
\end{aligned} \tag{A.7}$$

In the following, we further decompose inflow, outflow, and fully time-resolved information flows.

A.2.1. Decomposition of Inflow

$$
\begin{aligned}
&I(X^i; Y_i|Y^{i-1}) \\
&= H(X^i|Y^{i-1}) - H(X^i|Y^i) \\
&= H(X^i, Y^{i-1}) - H(Y^{i-1}) - H(X^i, Y^i) + H(Y^i) \\
&= H(X^i, Y^{i-1}) - H(Y^{i-1}) - H(X^i, Y^i) + H(Y^i) - H(X^i) + H(X^i) \\
&= I(X^i; Y^i) - I(X^i; Y^{i-1}) \\
&= M(X^i, Y^i) - M(X^i) - M(Y^i) - M(X^i, Y^{i-1}) + M(X^i) + M(Y^{i-1}) \\
&= M(X^i, Y^i) - M(Y^i) - M(X^i, Y^{i-1}) + M(Y^{i-1}).
\end{aligned}
\tag{A.8}
$$

From the third, fifth, and seventh equation, we get the following decompositions of directed information in terms of entropy

$$
\begin{aligned}
I(X^n \to Y^n) &= \sum_{i=1}^{n} \left(H(X^i, Y^{i-1}) - H(Y^{i-1}) - H(X^i, Y^i) + H(Y^i) \right) \\
&= \sum_{i=1}^{n} \left(H(X^i, Y^{i-1}) - H(X^i, Y^i) \right) + H(Y^n),
\end{aligned}
\tag{A.9}
$$

in terms of mutual information

$$
I(X^n \to Y^n) = \sum_{i=1}^{n} \left(I(X^i; Y^i) - I(X^i; Y^{i-1}) \right), \tag{A.10}
$$

in terms of multiinformation

$$
\begin{aligned}
I(X^n \to Y^n) &= \sum_{i=1}^{n} \left(M(X^i, Y^i) - M(Y^i) - M(X^i, Y^{i-1}) + M(Y^{i-1}) \right) \\
&= \sum_{i=1}^{n} \left(M(X^i, Y^i) - M(X^i, Y^{i-1}) \right) - M(Y^n).
\end{aligned}
\tag{A.11}
$$

A.2.2. Decomposition of Outflow

$$
\begin{aligned}
& I(X_i; Y_i^n | X^{i-1}, Y^{i-1}) \\
&= H(X_i | X^{i-1}, Y^{i-1}) - H(X_i | X^{i-1}, Y^n) \\
&= H(X^i, Y^{i-1}) - H(X^{i-1}, Y^{i-1}) - H(X^i, Y^n) + H(X^{i-1}, Y^n) \\
&= I(X_i; X^{i-1}, Y^n) - I(X_i; X^{i-1}, Y^{i-1}) \\
&= M(X^i, Y^n) - M(X^{i-1}, Y^n) - M(X^i, Y^{i-1}) + M(X^{i-1}, Y^{i-1}).
\end{aligned}
\tag{A.12}
$$

From the second, third, and fourth equation, we get the following decompositions of directed information in terms of entropy

$$
\begin{aligned}
I(X^n \to Y^n) &= \sum_{i=1}^{n} \big(H(X^i, Y^{i-1}) - H(X^{i-1}, Y^{i-1}) - H(X^i, Y^n) + H(X^{i-1}, Y^n) \big) \\
&= \sum_{i=1}^{n} \big(H(X^i, Y^{i-1}) - H(X^{i-1}, Y^{i-1}) \big) + H(Y^n) - H(X^n, Y^n) \\
&= \sum_{i=1}^{n} \big(H(X^i, Y^{i-1}) - H(X^i, Y^i) \big) + H(Y^n),
\end{aligned}
\tag{A.13}
$$

in terms of mutual information

$$
I(X^n \to Y^n) = \sum_{i=1}^{n} \big(I(X_i; X^{i-1}, Y^n) - I(X_i; X^{i-1}, Y^{i-1}) \big), \tag{A.14}
$$

in terms of multiinformation

$$
\begin{aligned}
I(X^n \to Y^n) &= \sum_{i=1}^{n} \big(M(X^i, Y^n) - M(X^{i-1}, Y^n) - M(X^i, Y^{i-1}) + M(X^{i-1}, Y^{i-1}) \big) \\
&= \sum_{i=1}^{n} \big(M(X^{i-1}, Y^{i-1}) - M(X^i, Y^{i-1}) \big) - M(Y^n) + M(X^n, Y^n) \\
&= \sum_{i=1}^{n} \big(M(X^i, Y^i) - M(X^i, Y^{i-1}) \big) - M(Y^n).
\end{aligned}
\tag{A.15}
$$

Comparing the decompositions of in- and outflow, the final expressions for entropy and multiinformation are equal.

A.2.3. Decomposition of Fully Time-Resolved Information Flows

$$
\begin{aligned}
&I(X_i;Y_k|X^{i-1},Y^{k-1})\\
&= H(X_i|X^{i-1},Y^{k-1}) - H(X_i|X^{i-1},Y^k)\\
&= H(X^i,Y^{k-1}) - H(X^{i-1},Y^{k-1}) - H(X^i,Y^k) + H(X^{i-1},Y^k)\\
&= I(X_i;X^{i-1},Y^k) - I(X_i;X^{i-1},Y^{k-1})\\
&= M(X^i,Y^k) - M(X^{i-1},Y^k) - M(X^i,Y^{k-1}) + M(X^{i-1},Y^{k-1}).
\end{aligned}
\tag{A.16}
$$

From the second, third, and fourth equation, we get the following decompositions of directed information in terms of entropy

$$
\begin{aligned}
&I(X^n \to Y^n)\\
&= \sum_{i=1}^{n}\sum_{k=i}^{n}\big(H(X^i,Y^{k-1}) - H(X^{i-1},Y^{k-1}) - H(X^i,Y^k) + H(X^{i-1},Y^k)\big)\\
&= \sum_{i=1}^{n}\big(H(X^i,Y^{i-1}) - H(X^i,Y^n) - H(X^{i-1},Y^{i-1}) + H(X^{i-1},Y^n)\big)\\
&= \sum_{i=1}^{n}\big(H(X^i,Y^{i-1}) - H(X^{i-1},Y^{i-1})\big) - H(X^n,Y^n) + H(Y^n)\\
&= \sum_{i=1}^{n}\big(H(X^i,Y^{i-1}) - H(X^i,Y^i)\big) + H(Y^n),
\end{aligned}
\tag{A.17}
$$

in terms of mutual information

$$
I(X^n \to Y^n) = \sum_{i=1}^{n}\sum_{k=i}^{n}\big(I(X_i;X^{i-1},Y^k) - I(X_i;X^{i-1},Y^{k-1})\big),
\tag{A.18}
$$

and in terms of multiinformation

$$
\begin{aligned}
&I(X^n \to Y^n)\\
&= \sum_{i=1}^{n}\sum_{k=i}^{n}\big(M(X^i,Y^k) - M(X^{i-1},Y^k) - M(X^i,Y^{k-1}) + M(X^{i-1},Y^{k-1})\big)\\
&= \sum_{i=1}^{n}\big(-M(X^i,Y^{i-1}) + M(X^i,Y^n) + M(X^{i-1},Y^{k-1}) - M(X^{i-1},Y^n)\big)\\
&= \sum_{i=1}^{n}\big(M(X^{i-1},Y^{i-1}) - M(X^i,Y^{i-1})\big) + M(X^n,Y^n) - M(Y^n)\\
&= \sum_{i=1}^{n}\big(M(X^i,Y^i) - M(X^i,Y^{i-1})\big) - M(Y^n).
\end{aligned}
\tag{A.19}
$$

A.3. Directed Information as the Difference between Observational and Interventional Distribution

Similar to (Wieczorek and Roth, 2016), we adopt the definition of directed information form Raginsky (2011), i.e. the definition of directed information as the KL-divergence between the observational and the interventional distribution. We derive the equivalence from scratch, but we neglect the instantaneous terms. For time series X^n and Y^n of lengths n

$$\begin{aligned} I(X^{n-1} \to Y^n) &= \sum_{i=1}^{n} I(X^{i-1}; Y_i | Y^{i-1}) \quad \text{granger causality} \\ &= \sum_{i=1}^{n} I(X_i; Y_{i+1}^n | X^{i-1}, Y^i) \quad \text{sims causality} \\ &= \sum_{i=1}^{n} H(X_i | X^{i-1}, Y^i) - \sum_{i=1}^{n} H(X_i | X^{i-1}, Y^i, Y_{i+1}^n) \\ &= \sum_{i=1}^{n} H(X_i | X^{i-1}, Y^i) - \sum_{i=1}^{n} H(X_i | X^{i-1}, Y^n) \\ &= H(X^n || Y^n) - H(X^n | Y^n). \end{aligned} \tag{A.20}$$

In the second equation, we used the result from App. A.1, namely, the equivalence between the Granger and Sims expansion of directed information. Note, in the sum, no stationarity assumption required for the equivalence of the expansions. Moreover, this definition is valid for arbitrary disjoint sets of variables X and Y without any specific ordering. We thus omit the time indices.

$$\begin{aligned} I(X \to Y) &= H(X||Y) - H(X|Y) \\ &= \int_{X,Y} P(X,Y) \Big(\log P(X|Y) - \log P(X|do(Y)) \Big) \\ &= E_{P_{X,Y}} \log \frac{P(X|Y)}{P(X|do(Y))} \\ &= D_{KL}(P_{X|Y} \| P_{X|do(Y)} | P_Y), \end{aligned} \tag{A.21}$$

where we used in the second equation, that causal conditional entropy can be written in the notation of Pearl's do-calculus (Pearl, 2009). In the last equation, we used the definition of conditional relative entropy (Cover and Thomas, 2012), where, contrary to standard notation, we made explicit the averaging over the distribution of P_Y.

A.3.1. Interventions for Time-Resolved Information Flows

We interpret causal flows as a KL-divergence between the observational and interventional distribution. We expand the inflow, which is the building block of directed information in the sense of Granger causality.

$$
\begin{aligned}
&I(X^i \to Y_i) \\
&= I(X^i; Y_i | Y^{i-1}) \\
&= H(X^i|Y^{i-1}) - H(X^i|Y^{i-1}, Y_i) \\
&= H(X^i|Y^{i-1}) - H(X^i|Y^i) \\
&= \int_{X^i,Y^i} P(X^i, Y^i) \Big(\log P(X^i|Y^i) - \log P(X_i|Y^{i-1}, do(Y_i)) \Big) \\
&= E_{P_{X^i,Y^i}} \log \frac{P(X^i|Y^i)}{P(X^i|Y^{i-1}, do(Y_i))} \\
&= D_{KL} \left(P_{X^i|Y^i} || P_{X^i|Y^{i-1},do(Y_i)} | P_{Y^i} \right).
\end{aligned}
\tag{A.22}
$$

Equivalently, we expand the outflow, which is the building block of directed information in the sense of Sims causality.

$$
\begin{aligned}
&I(X_i \to Y^n) \\
&= I(X_i; Y_i^n | X^{i-1}, Y^{i-1}) \\
&= H(X_i|X^{i-1}, Y^{i-1}) - H(X_i|X^{i-1}, Y^{i-1}, Y_i^n) \\
&= H(X_i|X^{i-1}, Y^{i-1}) - H(X_i|X^{i-1}, Y^n) \\
&= \int_{X^i,Y^n} P(X^i, Y^n) \Big(\log P(X_i|X^{i-1}, Y^n) - \log P(X_i|X^{i-1}, Y^{i-1}, do(Y_i^n)) \Big) \\
&= E_{P_{X^iY^n}} \log \frac{P(X_i|X^{i-1}, Y^n)}{P(X_i|X^{i-1}, Y^{i-1}, do(Y_i^n))} \\
&= D_{KL} \left(P_{X_i|X^{i-1},Y^n} || P_{X_i|X^{i-1},Y^{i-1},do(Y_i^n)} | P_{X^{i-1},Y^n} \right).
\end{aligned}
\tag{A.23}
$$

We also interpret fully time-resolved information flows as the KL-divergence between an observational and an interventional distribution. We thus expand a fully time-resolved information flow, which is the the fundamental building

block of directed information.

$$
\begin{aligned}
&I(X_i \to Y_k) \\
&= I(X_i; Y_k | X^{i-1}, Y^{k-1}) \\
&= H(X_i | X^{i-1}, Y^{k-1}) - H(X_i | X^{i-1}, Y^{k-1}, Y_k) \\
&= H(X_i | X^{i-1}, Y^{k-1}) - H(X_i | X^{i-1}, Y^{k}) \\
&= \int_{X^i, Y^k} P(X^i, Y^k) \Big(\log P(X_i | X^{i-1}, Y^k) - \log P(X_i | X^{i-1}, Y^{k-1}, do(Y_k)) \Big) \\
&= E_{P_{X^i, Y^k}} \log \frac{P(X_i | X^{i-1}, Y^k)}{P(X_i | X^{i-1}, Y^{k-1}, do(Y_k))} \\
&= D_{KL} \left(P_{X_i | X^{i-1}, Y^k} || P_{X_i | X^{i-1}, Y^{k-1}, do(Y_k)} | P_{X^{i-1}, Y^k} \right).
\end{aligned}
\tag{A.24}
$$

A.3.2. Copula Extension

Using the decompositions in multiinformation, directed information has amenable form in terms copulas (Wieczorek and Roth, 2016). In particular, directed information can be expressed as the difference between an observational and an interventional copula. Using Eq. A.8, namely the decomposition of directed information in terms of multiinformation, directed information can be written as follows

$$
\begin{aligned}
I(X^n \to Y^n) &= \sum_{i=1}^{n} \left(M(X^i, Y^i) - M(Y^i) - M(X^i, Y^{i-1}) + M(Y^{i-1}) \right) \\
&= \sum_{i=1}^{n} \left(M(X^i, Y^i) - M(X^i, Y^{i-1}) \right) - M(Y^n) \\
&= \sum_{i=2}^{n} \left(M(X^{i-1}, Y^{i-1}) - M(X^i, Y^{i-1}) \right) + M(X^n, Y^n) - M(Y^n) \\
&= M(X^n, Y^n) - M(Y^n) - \sum_{i=2}^{n} M(X^i | X^{i-1}, Y^{i-1}) \\
&= M(X^n | Y^n) - M(X^n || Y^n),
\end{aligned}
\tag{A.25}
$$

where we used the conditional multiinformation $M(A|B) = M(A, B) - M(B)$, for any sets of random variables A and B, $|B| \geq 2$, and causally conditioned multiinformation $M(A^n || B^n) = \sum_{i=2}^{n} M(A^i | A^{i-1}, B^{i-1})$. Omitting the indices and using that multiinformation is negative copula entropy, directed information

can be expressed as follows

$$\begin{aligned} I(X \to Y) &= M(X|Y) - M(X||Y) \\ &= \int_{U,V} c(U,V) \log c(U|V) - \int_{U,V} c(U,V) \log c(U|do(V)) \\ &= \int_{U,V} c(U,V) \Big(\log c(U|V) - \log c(U|do(V)) \Big) \\ &= E_{C_{U,V}} \log \frac{c(U|V)}{c(U|do(V))} \\ &= D_{KL}(C_{U|V} || C_{U|do(V)} | C_V), \end{aligned} \tag{A.26}$$

where we used $U = F_X(x)$ and $V = F_Y(y)$. This copula expression is the most concise form of directed information, since it does not depend on the marginal distributions of X and Y but only their copula. Analogously to Raginsky (2011), this is expression is valid for any sets X and Y and not only for time series.

Bibliography

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

Amblard, P.-O. and Michel, O. J. (2012). The relation between granger causality and directed information theory: A review. *Entropy*, 15(1):113–143.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Angus, J. E. (1994). The probability integral transform and related results. *SIAM review*, 36(4):652–654.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Bauckhage, C. (2014). A note on archetypal analysis and the approximation of convex hulls. *arXiv preprint arXiv:1410.0642*.

Bauckhage, C., Kersting, K., Hoppe, F., and Thurau, C. (2015). Archetypal analysis as an autoencoder. In *Workshop New Challenges in Neural Computation 2015*, page 8. Citeseer.

Bauckhage, C. and Manshaei, K. (2014). Kernel archetypal analysis for clustering web search frequency time series. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1544–1549. IEEE.

Bauckhage, C. and Thurau, C. (2009). Making archetypal analysis practical. In Denzler, J., Notni, G., and Se, H., editors, *Pattern Recognition*, volume 5748 of *Lecture Notes in Computer Science*, pages 272–281. Springer Berlin Heidelberg.

Bernadac, E. (1995). Random continued fractions and inverse gaussian distribution on a symmetric cone. *Journal of Theoretical Probability*, 8(2):221–259.

Bisgaard, A., Sørensen, K., Johannsen, T. H., Helge, J. W., Andersson, A.-M., and Juul, A. (2014). Significant Gender Difference in Serum Levels of Fibroblast Growth Factor 21 in Danish Children and Adolescents. *International Journal of Pediatric Endocrinology*, 2014(1):7.

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co.

Boudt, K., Cornelissen, J., and Croux, C. (2012). The gaussian rank correlation estimator: robustness properties. *Statistics and Computing*, 22(2):471–483.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Brillouin, L. (1962). Science and information theory.

Butler, R. W. (1998). Generalized inverse gaussian distributions and their wishart connections. *Scandinavian journal of statistics*, 25(1):69–75.

Camilleri-Broet, S., Cremer, I., Marmey, B., Comperat, E., Viguie, F., Audouin, J., Rio, M.-C., Fridman, W.-H., Sautes-Fridman, C., and Regnier, C. H. (2007). TRAF4 Overexpression is a Common Characteristic of Human Carcinomas. *Oncogene*, 26(1):142–147.

Canhasi, E. and Kononenko, I. (2014). Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. *Expert Systems with Applications*, 41(2):535–543.

Chamberlain, G. (1982). The general equivalence of granger and sims causality. *Econometrica: Journal of the Econometric Society*, pages 569–581.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (1998). Active appearance models. In *Computer VisionECCV98*, pages 484–498. Springer.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Cui, R., Groot, P., and Heskes, T. (2016). Copula pc algorithm for causal discovery from mixed data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 377–392. Springer.

Cutler, A. and Breiman, L. (1994). Archetypal analysis. *Technometrics*, 36(4):338–347.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66.

Date, K., Matsumoto, K., Kuba, K., Shimura, H., Tanaka, M., and Nakamura, T. (1998). Inhibition of Tumor Growth and Invasion by a Four-Kringle Antagonist (HGF/NK4) for Hepatocyte Growth Factor. *Oncogene*, 17(23):3045–3054.

Dawid, A. P. (2010). Beware of the dag! *NIPS Causality: Objectives and Assessment*, 6:59–86.

Deng, M., Tang, H., Lu, X., Liu, M., Lu, X., Gu, Y., Liu, J., and He, Z. (2013). miR-26a suppresses tumor growth and metastasis by targeting FGF9 in gastric cancer. *PloS one*, 8(8):e72662.

Diment, A. and Virtanen, T. (2015). Archetypal analysis for audio dictionary learning. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*, pages 1–5. IEEE.

Ebert, S. and Schiele, B. (2013). Where next in object recognition and how much supervision do we need? In *Advanced Topics in Computer Vision*, pages 35–64. Springer.

Edwards, G. J., Lanitis, A., Taylor, C. J., and Cootes, T. F. (1998). Statistical models of face imagesimproving specificity. *Image and Vision Computing*, 16(3):203–211.

Egger, B., Kaufmann, D., Schönborn, S., Roth, V., and Vetter, T. (2016). Copula eigenfaces. In *11th International Conference on Computer Graphics Theory and Applications (GRAPP)*, volume 1, pages 50–58. Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016).

Eichler, M. (2012). Causal inference in time series analysis. *Causality: statistical perspectives and applications*, pages 327–354.

Elidan, G. (2010). Copula bayesian networks. In *Advances in neural information processing systems*, pages 559–567.

Elidan, G. (2013). Copulas in machine learning. In *Copulae in mathematical and quantitative finance*, pages 39–60. Springer.

Erez, N., Milyavsky, M., Eilam, R., Shats, I., Goldfinger, N., and Rotter, V. (2003). Expression of Prolyl-Hydroxylase-1 (PHD1/EGLN2) suppresses Hypoxia Inducible Factor-1α Activation and Inhibits Tumor Growth. *Cancer Research*, 63(24):8777–8783.

Florens, J.-P. and Mouchart, M. (1982). A note on noncausality. *Econometrica: Journal of the Econometric Society*, pages 583–591.

Fridman, J. S., Hernando, E., Hemann, M. T., de Stanchina, E., Cordon-Cardo, C., and Lowe, S. W. (2003). Tumor Promotion by MDM2 Splice Variants Unable to Bind p53. *Cancer Research*, 63(18):5703–5706.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Garner, W. and Carson, D. (1960). A multivariate solution of the redundancy of printed english. *Psychological reports*.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, 11(12):4241–4257.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Genest, C. and Neslehova, J. (2007). A primer on copulas for count data. *Astin Bulletin*, 37(2):475.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix variate distributions*, volume 104. CRC Press.

Han, F. and Liu, H. (2012). Semiparametric principal component analysis. In *Advances in Neural Information Processing Systems*, pages 171–179.

Han, S., Liao, X., Dunson, D. B., and Carin, L. (2015). Variational gaussian copula inference. *arXiv preprint arXiv:1506.05860*.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104.

Hardmeier, M., Hatz, F., Naegelin, Y., Hight, D., Schindler, C., Kappos, L., Seeck, M., Michel, C. M., and Fuhr, P. (2014). Improved characterization of visual evoked potentials in multiple sclerosis by topographic analysis. *Brain topography*, 27(2):318–327.

Harris, N. and Drton, M. (2013). Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(1):3365–3383.

Hastie, T., Taylor, J., Tibshirani, R., Walther, G., et al. (2007). Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., and Bhattacharya, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46.

Hoff, P. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.*, 1(1):265–283.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435.

Jakulin, A. and Bratko, I. (2003a). Analyzing attribute dependencies. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 229–240. Springer.

Jakulin, A. and Bratko, I. (2003b). Quantifying and visualizing attribute interactions. *arXiv preprint cs/0308002*.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Johnson, S. C., Rabinovitch, P. S., and Kaeberlein, M. (2013). mTOR is a Key Modulator of Ageing and Age-Related Disease. *Nature*, 493(7432):338–345.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Kamitake, T., Harashima, H., and Miyakawa, H. (1984). A time-series analysis method based on the directed transinformation. *Electronics and Communications in Japan (Part I: Communications)*, 67(6):1–9.

Kaufmann, D., Keller, S., and Roth, V. (2015). Copula archetypal analysis. In *German Conference on Pattern Recognition*, pages 117–128. Springer.

Kaufmann, D., Parbhoo, S., Wieczorek, A., Keller, S., Adametz, D., and Roth, V. (2016). Bayesian markov blanket estimation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 333–341.

Kersting, K., Wahabzada, M., Thurau, C., and Bauckhage, C. (2010). Hierarchical convex nmf for clustering massive data. *ACML*, 10:253–268.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Koudou, A. E., Ley, C., et al. (2014). Characterizations of GIG laws: A survey. *Probability Surveys*, 11:161–176.

Kramer, G. (1998). *Directed information for channels with feedback*. PhD thesis, University of Manitoba, Canada.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Lawson, C. L. and Hanson, R. J. (1974). Solving least squares problems. *Prentice-Hall Series in Automatic Computation, Englewood Cliffs: Prentice-Hall, 1974*, 1.

Lee, J., Godon, C., Lagniel, G., Spector, D., Garin, J., Labarre, J., and Toledano, M. B. (1999). Yap1 and skn7 control two specialized oxidative stress response regulons in yeast. *Journal of Biological Chemistry*, 274(23):16040–16046.

Leitea, K. R. M., Franco, M. F., Srougi, M., Nesrallah, L. J., Nesrallah, A., Bevilacqua, R. G., Darini, E., Carvalho, C. M., Meirelles, M. I., Santana, I., and Camara-Lopes, L. H. (2001). Abnormal Expression of MDM2 in Prostate Carcinoma. *Modern Pathology*, 14(5):428–436.

Letac, G. (2000). Symmetric cones as gelfand pairs: probabilistic applications. *Contemporary Mathematics*, 261:109–120.

Li, S., Wang, P., Louviere, J., and Carson, R. (2003). Archetypal analysis: A new way to segment markets based on extreme individuals. In *A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution. Proceedings of the ANZMAC 2003 Conference*, pages 1674–1679.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328.

Liu, Y. (2012). *Directed information for complex network analysis from multivariate time series*. PhD thesis, Michigan State University.

Lothar Sachs, J. H. (2006). *Angewandte Statistik*. Springer Berlin Heidelberg, 7 edition.

Lyon, A. (2014). Why are normal distributions normal? *The British Journal for the Philosophy of Science*, 65(3):621–649.

Ma, J. and Sun, Z. (2011). Mutual information is copula entropy. *Tsinghua Science & Technology*, 16(1):51–54.

Maathuis, M. H., Colombo, D., et al. (2015). A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060–1088.

Massey, J. (1990). Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, pages 303–305. Citeseer.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462.

Mohammed, U., Prince, S. J., and Kautz, J. (2009). Visio-lization: generating novel facial images. *ACM Transactions on Graphics (TOG)*, 28(3):57.

Mørup, M. and Hansen, L. K. (2010). Archetypal analysis for machine learning. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 172–177. IEEE.

Nandy, P., Maathuis, M. H., and Richardson, T. S. (2014). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *arXiv preprint arXiv:1407.2451*.

Nelsen, R. B. (1999). *An introduction to copulas*, volume 139. Springer Science & Business Media.

Nelsen, R. B. (2013). *An introduction to copulas*, volume 139. Springer Science & Business Media.

Nielsen, M. A. and Chuang, I. (2002). Quantum computation and quantum information.

Norberg, U. M. and Rayner, J. M. (1987). Ecological morphology and flight in bats (mammalia; chiroptera): wing adaptations, flight performance, foraging strategy and echolocation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 316(1179):335–427.

Park, S. Y. and Bera, A. K. (2009). Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics*, 150(2):219–230.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. In *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference On*, pages 296–301. IEEE.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, 19(4):459.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2015). A complete generalized adjustment criterion. *arXiv preprint arXiv:1507.01524*.

Prabhakaran, S., Raman, S., Vogt, J., and Roth, V. (2012). Automatic model selection in archetype analysis. In Pinz, A., Pock, T., Bischof, H., and Leberl, F., editors, *Pattern Recognition*, volume 7476 of *Lecture Notes in Computer Science*, pages 458–467. Springer Berlin Heidelberg.

Quinn, C. J., Kiyavash, N., and Coleman, T. P. (2015). Directed information graphs. *IEEE Transactions on Information Theory*, 61(12):6887–6909.

Raginsky, M. (2011). Directed information and pearl's causal calculus. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 958–965. IEEE.

Rasmussen, C. E. (1999). The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560.

Renzo, M. F. D., Olivero, M., Giacomini, A., Porte, H., Chastre, E., Mirossay, L., Nordlinger, B., Bretti, S., Bottardi, S., and Giordano, S. (1995). Overexpression and Amplification of the Met/HGF Receptor Gene During the Progression of Colorectal Cancer. *Clinical Cancer Research*, 1(2):147–154.

Ropponen, K. M. et al. (1999). Reduced Expression of alpha Catenin is Associated with Poor Prognosis in Colorectal Carcinoma. *Journal of Clinical Pathology*, 52(1):10–16.

Roth, V. and Fischer, B. (2008). The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, pages 848–855. ACM.

Roweis, S. (1999). Gaussian identities. *Lectures Notes, University of Toronto.*

Saito, Y. and Harashima, H. (1981). Tracking of information within multichannel EEG record causal analysis in eeg. *Yamaguchi N, Fujisawa K (eds) Recent advances in {EEG} and {EMG} data processing. Elsevier*, pages 133–146.

Schönborn, S., Forster, A., Egger, B., and Vetter, T. (2013). A monte carlo strategy to integrate detection and model-based face analysis. In *Pattern Recognition*, pages 101–110. Springer.

Schreiber, T. (2000). Measuring information transfer. *Physical review letters*, 85(2):461.

Schumacher, M. and Blanz, V. (2015). Exploration of the correlations of attributes and features in faces. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, pages 1–8. IEEE.

Seth, S. and Eugster, M. J. (2013). Probabilistic archetypal analysis. *arXiv preprint arXiv:1312.7604*.

Seth, S. and Eugster, M. J. (2016). Archetypal analysis for nominal observations. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):849–861.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Sheffer, M., Bacolod, M. D., Zuk, O., Giardina, S. F., Pincas, H., Barany, F., Paty, P. B., Gerald, W. L., Notterman, D. A., and Domany, E. (2009). Association of Survival and Disease Progression with Chromosomal Instability: A Genomic Exploration of Colorectal Cancer. In *Proceedings of the National Academy of Sciences*, pages 7131–7136.

Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo, A., Dekel, E., Kavanagh, K., and Alon, U. (2012). Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space. *Science*, 336(6085):1157–1160.

Sifa, R. and Bauckhage, C. (2013). Archetypical motion: Supervised game behavior learning with archetypal analysis. In *Computational Intelligence in Games (CIG), 2013 IEEE Conference on*, pages 1–8. IEEE.

Sims, C. A. (1972). Money, income, and causality. *The American economic review*, 62(4):540–552.

Sirovich, L. and Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *JOSA A*, 4(3):519–524.

Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Slonim, N., Atwal, G. S., Tkacik, G., and Bialek, W. (2005). Estimating mutual information and multi–information in large networks. *arXiv preprint cs/0502017*.

Stewart, B. W. and Wild, C. P. (2014). *World Cancer Report 2014*. IARC Press.

Steyn, H. and Roux, J. (1972). Approximations for the non-central wishart distribution. *South African Statistical Journal*, 6:165–173.

Studenỳ, M. and Vejnarová, J. (1999). The multiinformation function as a tool for measuring stachastic dependence, learning in graphical models.

Styner, M. A., Rajamani, K. T., Nolte, L.-P., Zsemlye, G., Székely, G., Taylor, C. J., and Davies, R. H. (2003). Evaluation of 3d correspondence methods for model building. In *Information processing in medical imaging*, pages 63–75. Springer.

Thurau, C. and Bauckhage, C. (2009). Archetypal images in large photo collections. In *Semantic Computing, 2009. ICSC'09. IEEE International Conference on*, pages 129–136. IEEE.

Thurau, C., Kersting, K., and Bauckhage, C. (2009). Convex non-negative matrix factorization in the wild. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 523–532. IEEE.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Tran, D., Blei, D. M., and Airoldi, E. M. (2015). Variational inference with copula augmentation. *stat*, 1050:10.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3):357–375.

Turk, M., Pentland, A. P., et al. (1991). Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Walker, M. and Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, 9(11):12.

Wang, H. et al. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886.

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82.

Wieczorek, A. and Roth, V. (2016). Causal compression. *arXiv preprint arXiv:1611.00261*.

Xue, L., Zou, H., et al. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571.

Yu, H.-G., Zhong, X., Yang, Y.-N., Luo, H.-S., Yu, J.-P., Meier, J. J., Schrader, H., Bastian, A., Schmidt, W. E., and Schmitz, F. (2004). Increased Expression of Nuclear factor-κB/RelA is Correlated With Tumor Angiogenesis in Human Colorectal Cancer. *International Journal of Colorectal Disease*, 19(1):18–22.

Zhao, C., Zhao, G., and Jia, X. (2016). Hyperspectral image unmixing based on fast kernel archetypal analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Dinu Johannes Kaufmann

University of Basel
Department of Mathematics and Computer Science
Spiegelgasse 1, 4051 Basel, Switzerland

Geissensteinring 48
6005 Luzern
Switzerland

Personal information

Date of birth	August 25, 1985
Place of birth	Arlesheim, Switzerland
Citizenship	Swiss
E-mail	dinu.kaufmann@{unibas.ch,gmail.com}

Education

09/12 – 06/17 Ph.D at the University of Basel, Switzerland
Department of Mathematics and Computer Science
Biomedical Data Analysis Group

Ph.D thesis: "Semi-parametric Gaussian Copula Models for Machine Learning"
Advisor: Prof. Dr. Volker Roth

09/09 – 03/12 MSc in Electrical Engineering at ETH Zürich, Switzerland
Departement for Information Technology and Electrical Engineering

Master thesis at the Institute for Signal- und Informationprocessing in collaboration with Siemens Building Technologies, Zug: "Model-based Fire Detection with Gas Sensors"
Advisors: Prof. Dr. Hans-Andrea Loeliger, Dr. Christoph Reller, and Lukas Bruderer

09/05 – 02/09 BSc in Elecrical Engineering
ETH Zürich, Switzerland

Publications

Kaufmann, D., Keller, S., and Roth, V. (2015). Copula archetypal analysis. In *German Conference on Pattern Recognition*, pages 117128. Springer.

Kaufmann, D., Parbhoo, S., Wieczorek, A., Keller, S., Adametz, D., and Roth, V. (2016). Bayesian markov blanket estimation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 333341.

Egger, B., Kaufmann, D., Schnborn, S., Roth, V., and Vetter, T. (2016). Copula eigenfaces. In *11th International Conference on Computer Graphics Theory and Applications (GRAPP)*, volume 1, pages 5058. Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016).

June 20, 2017, Dinu Johannes Kaufmann