Copula models in Machine Learning

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Mélanie Rey

aus Montana, Schweiz

Basel, 2015

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Volker Roth, Universität Basel, Dissertationsleiter
Prof. Dr. Gal Elidan, the Hebrew University, Korreferent

Basel, den 22.04.2014

Prof. Dr. Jörg Schibler, Dekan

Namensnennung-Keine kommerzielle Nutzung-Keine Bearbeitung 3.0 Schweiz
(CC BY-NC-ND 3.0 CH)

Sie dürfen: **Teilen** — den Inhalt kopieren, verbreiten und zugänglich machen

Unter den folgenden Bedingungen:

Namensnennung — Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen.

Keine kommerzielle Nutzung — Sie dürfen diesen Inhalt nicht für kommerzielle Zwecke nutzen.

Keine Bearbeitung erlaubt — Sie dürfen diesen Inhalt nicht bearbeiten, abwandeln oder in anderer Weise verändern.

Wobei gilt:

- **Verzichtserklärung —** Jede der vorgenannten Bedingungen kann **aufgehoben** werden, sofern Sie die ausdrückliche Einwilligung des Rechteinhabers dazu erhalten.
- **Public Domain (gemeinfreie oder nicht-schützbare Inhalte) —** Soweit das Werk, der Inhalt oder irgendein Teil davon zur Public Domain der jeweiligen Rechtsordnung gehört, wird dieser Status von der Lizenz in keiner Weise berührt.
- **Sonstige Rechte —** Die Lizenz hat keinerlei Einfluss auf die folgenden Rechte:
    - Die Rechte, die jedermann wegen der Schranken des Urheberrechts oder aufgrund gesetzlicher Erlaubnisse zustehen (in einigen Ländern als grundsätzliche Doktrin des **fair use** bekannt);
    - Die **Persönlichkeitsrechte** des Urhebers;
    - Rechte anderer Personen, entweder am Lizenzgegenstand selber oder bezüglich seiner Verwendung, zum Beispiel für **Werbung** oder Privatsphärenschutz.
- **Hinweis —** Bei jeder Nutzung oder Verbreitung müssen Sie anderen alle Lizenzbedingungen mitteilen, die für diesen Inhalt gelten. Am einfachsten ist es, an entsprechender Stelle einen Link auf diese Seite einzubinden.

Quelle: http://creativecommons.org/licenses/by-nc-nd/3.0/ch/ Datum: 12.11.2013

Abstract

The introduction of *copulas*, which allow separating the dependence structure of a multivariate distribution from its marginal behaviour, was a major advance in dependence modelling. Copulas brought new theoretical insights to the concept of dependence and enabled the construction of a variety of new multivariate distributions. Despite their popularity in statistics and financial modelling, copulas have remained largely unknown in the machine learning community until recently. This thesis investigates the use of copula models, in particular Gaussian copulas, for solving various machine learning problems and makes contributions in the domains of dependence detection between datasets, compression based on side information, and variable selection.

Our first contribution is the introduction of a copula mixture model to perform dependency-seeking clustering for co-occurring samples from different data sources. The model takes advantage of the great flexibility offered by the copula framework to extend mixtures of Canonical Correlation Analyzers to multivariate data with arbitrary continuous marginal densities. We formulate our model as a non-parametric Bayesian mixture and provide an efficient Markov Chain Monte Carlo inference algorithm for it. Experiments on real and synthetic data demonstrate that the increased flexibility of the copula mixture significantly improves the quality of the clustering and the interpretability of the results.

The second contribution is a reformulation of the information bottleneck (IB) problem in terms of a copula, using the equivalence between mutual information and negative copula entropy. Focusing on the Gaussian copula, we extend the analytical IB solution available for the multivariate Gaussian case to meta-Gaussian distributions which retain a Gaussian dependence structure but allow arbitrary marginal densities. The resulting approach extends the range of applicability of IB to non-Gaussian continuous data and is less sensitive to outliers than the original IB formulation.

Our third and final contribution is the development of a novel sparse compression technique based on the information bottleneck (IB) principle, which takes into account side information. We achieve this by introducing a sparse variant of IB that compresses the data by preserving the information in only a few selected input dimensions. By assuming a Gaussian copula we can capture arbitrary non-Gaussian marginals, continuous or discrete. We use our model to select a subset of biomarkers relevant to the evolution of malignant melanoma and show that our sparse selection provides reliable predictors.

Contents

List of Figures

List of Tables

Acknowledgements

My thesis advisor Prof. Dr. Volker Roth played a crucial role in the development of the work presented here which would not have materialised without his insightful ideas and academic guidance. I would like to thank him for his constant support and the many theoretical discussions by which he gave ideas and problems a little more life.

I am very grateful to my co-examiner Prof. Dr. Gal Elidan for reviewing my thesis and feel honored by his interest in my work.

Collaborating with Prof. Dr. Niko Beerenwinkel, Armin Töpfer (ETH Zürich, Basel), Dr. Karin J. Metzner, Dr. Huldrych Günthard, Francesca Di Giallonardo (University Hospital, Zürich) and Dr. Osvaldo Zagordi (University of Zürich, Zürich) was a privilege and an enriching experience.

I would like to thank my colleagues Sandhya Prabhakaran, David Adametz, Behrouz Tajoddin, Dinu Kaufmann, Julia Vogt and Sudhir Raman for their support and friendship. A special thought goes to Sandhya for the countless moments of fun, the many interesting discussions and the enduring support.

Un remerciement tout particulier à ma famille. Merci à mes parents, Angela et Macdonald, pour leur soutien parfois héroïque et pour m'avoir transmis le désir de comprendre. Merci à mes grands-parents, Jeannette et Guy, ainsi qu'à ma tante Mauricia qui m'ont également soutenue durant toutes mes études.

Finally, I would like to thank Andriy for sharing my life across the Channel.

Chapter 1

Introduction

Analysis of dependence is a central task in statistics and many well-known problems revolve around this concept. While the connection to dependence is readily apparent for contingency tables and independence tests, other techniques such as regression and variable selection can also be seen as essentially dependence questions. Furthermore, since any multivariate model involves a dependence structure, the task of specifying, or estimating, this structure is at the heart of high-dimensional modelling. The introduction of *copulas*, which use random vectors with uniform marginals to separate the dependence structure of a multivariate distribution from its marginal behaviour, was a major advance in dependence modelling. Copulas brought new theoretical insights to the concept of dependence and enabled the construction of a variety of new multivariate distributions. Although copula models have been very popular in statistics and financial modelling (Genest et al., 2009), until recently they have remained largely unknown in the machine learning community. As stated in Elidan (2013), this is especially surprising considering the central role probabilistic graphical models, which also focus on high-dimensional dependency structures, play in the field of machine learning. Naturally, an important direction of research on copulas in machine learning deals with constructing and estimating multivariate distributions or graphical models. The first work to bridge the gap between the two fields, Kirshner (2007), introduced tree-averaged copula densities. In other key developments, Liu et al. (2009) estimated high-dimensional sparse networks using a Gaussian copula and Elidan (2010) introduced a more general approach to graphical models using local copulas. There are many natural connections and potential synergies between copulas and various machine learning techniques, also going beyond the construction of multivariate models.

This thesis focuses on applying copulas to three different machine learning problems, showing how the additional flexibility inherent to copula models can improve the existing solutions. We first consider the problem of detecting potential dependencies between two datasets of co-occurring samples. Going beyond methods that assume global linear dependence, such as Canonical Correlation Analysis, we extend the Bayesian non-parametric dependency-seeking clustering method introduced in Klami and Kaski (2008). We show that by using a Gaussian copula we can avoid the model mismatch problems, which can undermine the reliability of dependence detection, while retaining a highly efficient inference. The second problem we consider is data compression with relevance information solved in the Information Bottleneck (IB) framework (Tishby et al., 1999). Although an analytical solution to the IB compression problem is available in the special case of jointly Gaussian variables, no such solution is known for the general case. The resulting optimisation problem for discrete data, for example, involves an iterative algorithm with no guaranties of global convergence. Using a model based on a Gaussian copula, we extend the analytical solution to continuous meta-Gaussian variables (variables with a Gaussian copula and arbitrary marginals), thus substantially increasing the IB's domain of applicability, and establish strong connections between the IB problem and copulas, showing that the problem depends only on the copula of the

data considered. Turning our attention to the problem of variable selection, we introduce a modified version of the IB which, due to the introduction of a new constraint, is able to identify the most informative dimensions in the data. In order to be able to perform variable selection for the numerous mixed (continous-discrete) datasets, such as those arising in the medical and biological fields, we further generalise our previous meta-Gaussian IB to discrete variables.

This thesis is organised as followed. Chapter 2 introduces the mathematical concepts required for the subsequent developments and provides a basic summary of probabilistic models, mixture models, Bayesian inference, Markov Chain Monte Carlo, and statistical dependence. Chapter 3 is dedicated to copulas, introducing the main definitions and some major results. The first main contribution of this thesis, a new copula mixture model for dependency-seeking clustering, is presented in Chapter 4. Before presenting our second innovation, Meta-Gaussian information bottleneck (MGIB) in Chapter 6, we introduce the Information Bottleneck problem in Chapter 5. Chapter 7 is dedicated to a new variable selection method based on the IB principle, for which we also extend MGIB to mixed data.

Chapter 2

Background

This chapter provides a summary of the main concepts, theory and results needed for subsequent developments. Most of the background notions required come from probability theory and statistics but some algorithmic aspects will also be covered. We follow the notation conventions of probability theory and will use some measure theoretical concepts when needed. While this formalism requires some preliminaries, it provides a level of precision which, we hope, facilitates understanding. It also constitutes a unifying framework for continuous and discrete variables, which will be useful in Chapter 7. Moreover, using some formal definitions from probability theory enables a proper description of stochastic processes such as the Dirichlet process which will play an important role in Chapter 4.

2.1 Probability spaces and random variables

Probability spaces and random variables. We denote a *probability space* by $(\Omega, \mathscr{F}, \mathbb{P})$ and a *measurable space* by $(E, \mathscr{E})$, where $\mathscr{F}, \mathscr{E}$ are *σ-algebras* on Ω, E respectively, and $\mathbb{P}$ is a *probability measure* (pm) on $\mathscr{F}$. Random variables (rv) will be written in capital letters: X, Y, whereas observations will be denoted by small letters: x, y. A random variable X is a measurable map from the probability space to a measurable space $(E_x, \mathscr{E}_x)$, where E_x is called the *sample space*. In other words, a random variable is an $(\mathscr{F}, \mathscr{E}_x)$-measurable map $X : \Omega \to E_x$. The *distribution* of X is the probability measure on $(E_x, \mathscr{E}_x)$ implicitly defined by the probability measure $\mathbb{P}$:

$$\mu_X(A) = \mathbb{P} \circ X^{-1}(A) = \mathbb{P}\{X \in A\}, \quad \forall A \in \mathscr{E}_x. \tag{2.1}$$

The sample space E_x can take various forms. For univariate random variables common choices are $\mathbb{R}$, $\mathbb{N}$ or subsets of them. Multivariate random variables will typically take values in product spaces of the form $\mathbb{R}^d$ or $\mathbb{N}^d$, where d represents the number of dimensions which can be infinite [1].

σ-algebra and product spaces. When the sample space is a topological space, a natural choice of σ-algebra is the *Borel σ-algebra* which is the smallest σ-algebra generated by the open sets. Measurable sets are then called *Borel sets*. We denote the smallest *σ-algebra* generated by a collection of sets $U_i, i \in J$, for an index set J, by $\sigma(U_i)$. For product spaces another natural choice of σ-algebra is the *product σ-algebra* defined as the smallest σ-algebra generated by products of

[1] In the literature, a random variable is sometimes defined as a measurable function taking values in $\mathbb{R}$ or $\bar{\mathbb{R}}$. Since the thesis focuses on dependence we use another convention which directly extends the notion of random variable to multi-dimensional sample spaces.

measurable sets (measurable w.r.t. to their respective σ-algebras). As an example, we briefly consider the case of $E_x = \mathbb{R}^d, d \leq \infty$, equipped with the Euclidean metric. The product σ-algebra on $\mathbb{R}^d$ is $\sigma\left(\prod_{i=1}^d B_i\right)$, where $B_i \in \mathcal{B}(\mathbb{R})$ are Borel sets in $\mathbb{R}$, whereas the Borel σ-algebra on $\mathbb{R}^d$ is $\mathcal{B}(\mathbb{R}^d) = \sigma(U_i)$, where U_i are the open sets in $\mathbb{R}^d$. In the case of $\mathbb{R}^d$ the product and the Borel σ-algebras are equal but this does not hold anymore for uncountable products spaces.

Cumulative distribution function. We give in (2.1) the distribution of a random variable. Under particular conditions there exists a simpler method to characterise a random variable using a function instead of a measure: the *cumulative distribution function*. If we consider of a random variable with sample space $E_x = \mathbb{R}^d$, $d < \infty$, and $\mathscr{E}_x = \mathcal{B}(\mathbb{R}^d)$, it is sufficient to specify the value of $\mu_X(A)$ for all sets of the form $A = \prod_{i=1}^d (-\infty, x_i],\ x_i \in \mathbb{R}$. The underlying principle justifying this simplification is that these sets form a *π-system* generating $\mathcal{B}(\mathbb{R}^d)$, see Çinlar (2011). As a consequence, the distribution of X can equivalently be defined using a function $F_X : \mathbb{R}^d \to [0, 1]$:

$$F_X(x) = \mu_X\left(\prod_i (-\infty, x_i]\right) = \mathbb{P}\{X_1 \leq x_1, \ldots, X_d \leq x_d\},\ \ \forall x = (x_1, \ldots, x_d), \tag{2.2}$$

and F_X is called the cumulative distribution function of X.

Density and probability mass function. Another convenient method for characterising a random variable is to use a density function. Let X be a random variable with sample space $E_x = \mathbb{R}^d$, $d < \infty$, and $\mathscr{E}_x = \mathcal{B}(\mathbb{R}^d)$. We further assume that μ_X is a σ-finite distribution (i.e. E_X is a countable union of measurable sets with finite measure). If μ_X is absolutely continuous w.r.t the Lebesgue measure λ on $\mathbb{R}^d$, the *Radon-Nikodym theorem* ensures the existence of a positive measurable function $f : \mathbb{R}^d \to \mathbb{R}_+$ such that

$$\mu_X(A) = \int_A f \mathrm{d}\lambda,\ \ \forall A \in \mathcal{B}(\mathbb{R}^d), \tag{2.3}$$

f is then called the *density* of X. If X takes values in a finite or countably infinite subset of $\mathbb{R}^d$ with probability one, it is called a *discrete random variable*. Discrete variables clearly cannot be absolutely continuous w.r.t λ, however an analogue to density functions exists for the discrete case. Instead of specifying a probability measure we can use a *probability mass function* $f : \mathbb{R}^d \to [0, 1]$ defined by

$$f_X(x) = \mu_X(\{x\}). \tag{2.4}$$

From which it naturally follows that

$$\mu_X(A) = \sum_{a \in A} \mu_X(\{a\}) = \sum_{a \in A} f_X(a) = \int_A f_X(a) \mathrm{d}m(a), \tag{2.5}$$

where m is the counting measure on $\mathbb{R}^d$.

Marginal distributions. Consider a multivariate random variable $X = (X_j)_{j \in J}$, having an arbitrary number of dimensions indexed by a set J (possibly uncountable). The univariate distributions of the different dimensions of X, called the *marginal distributions* μ_{X_j}, are the measures on $(E_{x_j}, \mathscr{E}_{x_j})$ defined by

$$\mu_{X_j}(A) = \mathbb{P}\{X_j \in A\} = \mu_X(E_{x_1} \times \cdots \times A \times \cdots \times E_{x_d}),\ \ \forall A \in \mathscr{E}_{x_j}. \tag{2.6}$$

By extension and depending on the context, the term marginal distribution can also denote the joint distribution of any finite subset of dimensions.

Stochastic processes. Infinite-dimensional random variables are often considered in the context of stochastic processes. A *stochastic process* is an indexed family of univariate random variables on the same sample space.

Definition 2.1 (Stochastic process)**.** *A stochastic process with state space* $(E, \mathscr{E})$ *and index set* T *is a collection* $\{X_t, t \in T\}$ *of random variables on* $(E, \mathscr{E})$.

We write (X_t) for simplicity. The set T can be finite, infinite countable or uncountable. The index t is often interpreted as time when $T = \mathbb{N}$ or $T = \mathbb{R}_+$. We denote the *product space* by $E^T = \prod_{t \in T} E$ which is the set of all functions $f : T \to E$. We can look at a stochastic process from different perspectives. We can envisage it as a collection of univariate rv, as in the above definition, but we can also consider it as a random function: for every fixed event $\omega \in \Omega$ a stochastic process constitutes a function from the index set to the state space. The functions

$$\begin{aligned} p_\omega : T &\to E, \\ t &\mapsto X_t(\omega), \end{aligned}$$

are elements of the product space E^T called the *paths* of (X_t). A stochastic process randomly selects one path which is called the realisation or the observed stochastic process. A third view on stochastic processes, closely related to the random function view, is to consider it as a (possibly infinite) dimensional random variable on a product space:

$$\begin{aligned} X : \Omega &\to E^T, \\ \omega &\mapsto p_\omega. \end{aligned}$$

Since a random variable is a measurable map, for the above specification to be complete we still have to specify a σ-algebra on E^T. We are interested in stochastic processes measurable with respect to the Borel σ-algebra $\mathcal{B}(E^T)$ on the product space. The distribution $\mu_X = \mathbb{P} \circ X^{-1}$ of the stochastic process is then a measure on $(E^T, \mathcal{B}(E^T))$. We call μ_X the *probablility law* of the stochastic process. When the index set is countable, $\mathcal{B}(E^T)$ is also the σ-algebra generated by the product sets

$$\textstyle\prod_{t \in T} A_t = \{a \in E^T | a_t \in A_t,\ A_t \in \mathscr{E},\ \forall t \in T\},$$

for which $|\{t \in T | A_t \neq E\}| < \infty$. In this case, *Kolmogorov extension theorem* [2] ensures that the distribution of X on $\mathcal{B}(E^T)$ exists and is uniquely determined by the values of its final dimensional projections

$$\mathbb{P}\{X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n\},\ \forall n \in \mathbb{N},\ A_i \in \mathscr{E},$$

subject to the condition that these final dimensional projections are consistent, see Çinlar (2011) for more details.

Well-known examples of stochastic processes include Markov chains and Gaussian processes (of which the Wiener process is a special case). In Bayesian statistics, the Dirichlet process has a special role since it defines a measure over distributions and can thereby define priors. Consider a measurable space $(H, \mathscr{H})$. The probability law of a *Dirichlet process* G is a measure on the space of all probability measures on H, i.e. μ_G is a measure on the set

$$M = \{\mu : \mathscr{H} \to [0, 1] |\ \mu \text{ is a probability measure}\} \subset \mathbb{R}^{\mathscr{H}}.$$

More precisely, a Dirichlet process is specified by two parameters: the base measure G_0 on $(H, \mathscr{H})$ and a real valued parameter α. The characteristics of a Dirichlet process are given in the following definition.

Definition 2.2. *A Dirichlet process with state space* $(E, \mathscr{E}) = ([0, 1], \mathcal{B}([0, 1]))$*, base measure* G_0 *and concentration* $\alpha > 0$ *is the stochastic process* G *with index set* $T = \mathscr{H}$ *defined by*

$$(G_{t_1}, \ldots, G_{t_n}) \sim \mathrm{Dir}(\alpha G_0(t_1), \ldots, \alpha G_0(t_n)),\ \forall n \in \mathbb{N},\ \forall \mathcal{P}_{\mathscr{H}}, \tag{2.7}$$

[2] In the general case of uncountable T, the theorem ensures the existence and unicity of the disctribution on the product σ-algebra $\mathscr{E}^T \subseteq \mathcal{B}(E^T)$.

where $t_i \in \mathcal{P}_{\mathscr{H}}$, G_0 is a measure on $(H, \mathscr{H})$, and $\mathcal{P}_{\mathscr{H}}$ denotes a measurable partition of H [3]*. We use the following notation*

$$G \sim \mathrm{DP}(\alpha, G_0). \tag{2.8}$$

The set T in the above definition differs from the intuitive idea of an index set since $T = \mathscr{H}$ is a σ-algebra, and, as a consequence, each component of G is indexed by a set. In (2.7) the indices $t_1, \ldots, t_n$ are restricted to a partition $\mathcal{P}_{\mathscr{H}}$ instead of the entire σ-algebra $\mathscr{H}$, this simplification is induced by the special form of the Dirichlet distribution which draws finite discrete probability distributions. Each component $G_t, t \in \mathscr{H}$ is a rv with values is the real unit interval. However, since we impose that any finite collection follows a Dirichlet distribution, a group of n marginal variables $G_{t_1}, \ldots, G_{t_n}$ takes values in the $(n-1)$-dimensional simplex. The paths

$$\begin{aligned} p_\omega : T = \mathscr{H} &\to E = [0,1] \\ A &\mapsto G_A(\omega) \end{aligned}$$

of a Dirichlet process are special functions, they are almost surely [4] probability measures on $(H, \mathscr{H})$. Since $G : \Omega \to E^T = [0,1]^{\mathscr{H}}$ is a random variable taking values in the space of all functions from $\mathscr{H}$ to the $[0,1]$ interval, μ_G actually is a distribution on the space of probability measures M.

Conditional distributions To close this section on the basics of probability theory we provide below a short summary on conditional distributions, these being essential to Bayesian statistics. We first need to introduce the concept of *conditional expectation* of a random variable. Consider a positive random variable X with state space $(E, \mathscr{E})$ and let $\mathscr{A}$ be a sub-σ-algebra of $\mathscr{F}$. The conditional expectation of X given $\mathscr{A}$, denoted by $\mathrm{E}_{\mathscr{A}}(X)$, is any random variable $\bar{X}$ such that

1. $\bar{X}$ is $\mathscr{A}$-measurable.
2. $\mathrm{E}(VX) = \mathrm{E}(V\bar{X})$, for every $\mathscr{A}$-measurable positive random variable V.

This definition can be extended for general random variables (not necessarily positive) by decomposing the variable in a positive and a negative part. The intuitive interpretation of conditional expectations is that $\mathrm{E}_{\mathscr{A}}(X)$ is the random variable which best estimates X based on the information provided by $\mathscr{A}$. It can be shown that two random variables fulfilling the above requirements are almost surely equal, we can therefore speak of *the* conditional expectation when referring to such a variable. Conditional distributions are built using conditional expectations. We first introduce the *conditional probability* of a set $H \in \mathscr{F}$ given $\mathscr{A}$ which is the random variable defined by

$$\mathbb{P}_{\mathscr{A}}(H) = \mathrm{E}_{\mathscr{A}}(I_H). \tag{2.9}$$

The *conditional distribution* of X given $\mathscr{A}$ is any *transition probability kernel* from $(\Omega, \mathscr{A})$ into $(E, \mathscr{E})$

$$K : \Omega \times \mathscr{E} \to [0,1] \tag{2.10}$$

$$(\omega, B) \mapsto K_\omega(B), \tag{2.11}$$

such that

$$\mathbb{P}_{\mathscr{A}}(X \in B) = K(B). \tag{2.12}$$

We also recall that, by definition of a transition probability kernel, $K_\omega(.)$ defines a measure on $\mathscr{E}$ for every fixed ω, and $K.(B)$ is a random variable measurable w.r.t. the σ-algebras $\mathscr{A}, \mathscr{E}$ for every fixed $B \in \mathscr{E}$. We can finally define the conditional distribution of X given another random

[3] $\mathcal{P}_{\mathscr{H}} = \{H_1, \ldots, H_n\}$ such that $H_i \in \mathscr{H}$, $\cup_{i=1}^n H_i = H$, $H_i \cap H_j = \emptyset$ for $i \neq j$.

[4] i.e. the statement holds up to nullsets

variable Y as the conditional distribution of X given the σ-algebra generated by Y: $\sigma(Y)$. Denote by μ the joint distribution of (X, Y), the conditional probability of $Y \in B | X \in A$ is given by the stochastic kernel K fulfilling

$$\mu(A \times B) = \int_A \mu_X(\mathrm{d}x) K(x, B), \tag{2.13}$$

and the conditional distribution of $Y|X$ is the kernel L defined by

$$L_\omega(B) = K(X(\omega), B). \tag{2.14}$$

For every realisation $x = X(\omega)$ of X, L_ω defines a measure on $\mathscr{E}$ denoted by $\mu_{Y|X}(.)$ for simplicity. We further use the notation $f_{Y|X}$ for the corresponding density of probability mass function.

2.2 Statistical Models

Statistics can roughly be described as the science (some authors say *art*) of drawing conclusions from data. More precisely, we can identify two goals of statistical modelling: *prediction*, we want to be able to predict some quantities associated to the data, and *understanding*, we want to extract information about the data generation process. A necessary assumption is that the data generation process contains some *randomness*, meaning that the data could have differed slightly from the particular set we analyse. This slight difference consist in the *error* caused by pure randomness in the process or by measurements uncertainties. Assuming some randomness is not in contradiction with perfectly determined systems since the random part could be constituted only of measurements imprecisions or could represent imperfect knowledge about the generation process. As described in length in Breiman (2001) they are two main types of statistical models: *data modelling* assumes the data was produced following a stochastic law which we try to identify, *algorithmic modelling* directly tries to answer a data related question (e.g. which variables are more important) using an algorithmic approach. Since algorithmic modelling does not assume any precise distribution for the data, it offers more flexibility and its standard methods can be applied to a large variety of data sets. However, this also implies that traditional statistical analysis methods including tests, confidence intervals and asymptotic results, are not applicable. Another often encountered criticism towards algorithmic techniques is their lack of interpretability. Whereas very good predictions can be achieved, the results are often not readily interpretable and do not provide the same insight as a fitted probability model. On the other hand, probabilistic modelling also raises some concerns. While the vast majority of work in statistics has been dedicated to probabilistic models, lack of fit issues arising from assuming an inadequate or too restrictive distribution often are still underestimated. This thesis will focuses on probabilistic data models, trying to address some lack of fit problems by enlarging the panel of distributions available to solve various tasks. We start by formally introducing parametric and non-parametric data models, both from the frequentist and the Bayesian point of view.

Definition 2.3 (Parametric model)**.** *A parametric model is a family of the form* $\{\mu_X^\theta | \theta \in \Theta\}$*, where the distributions are indexed by* θ *and* Θ *is a subset of* $\mathbb{R}^d$*,* $d \in \mathbb{N}/\{0\}$*.*

A parametric model is a family of distributions where each element is uniquely specified by the value of θ, called *the parameter* of the family. This parameter can have an arbitrary large number of dimensions with the restriction that its dimensionality d must be fixed. In the context of parametric models the task of statistical estimation is to choose within the family the distribution best suited to the data, where "best suited" still needs to be precisely defined. Using a parameter space with many dimensions, the class of available distributions can become very large, however the flexibility of probabilistic models can be further increased by relaxing the assumption of a parameter space with fixed dimensionality. Despite being inherently parametric, such models are called *non-parametric models*. Their distinctive characteristic is that the dimension d of

the parameter space remains unbounded during the estimation procedure. More precisely, this means that the number of parameters is allowed to vary with the sample size: as we see new observations we are allowed to reconsider the number of parameters needed to adequately describe our sample. When the number of observations tends to infinity this number can potentially become arbitrarily large. Non-parametric models are therefore also called *infinite dimensional models* [5]. A classic example of non-parametric model is Parzen density estimator (Parzen, 1962) which is parameterized by a global bandwidth and one location parameter per observation.

We will in this thesis adopt a Bayesian approach to statistical modelling and we therefore present a brief introduction to Bayesian statistics, highlighting its relationship to the frequentist point view. Bayesian inference is a probabilistic data analysis technique which conducts inference conditional on the observed data. The (possibly multivariate and infinite dimensional) parameter θ of the probabilistic model assumed is considered as random, in a sense we will make precise below, and inference consists in determining its *posterior distribution* conditioned on the data. Assuming the probability model is correct, the posterior distribution contains all available information about the parameter. Point estimates of parameters, confidence intervals and hypothesis testing can therefore be constructed from the posterior. The distribution of the parameters before any data has been seen is called the *prior distribution*. It does not describe a potential variability of θ, which is typically assumed to be a fixed unknown variable, the randomness of the parameter rather reflects our uncertainty about its precise value. The term probability is this context is used in the sense of a measure of uncertainty. To be more precise we should add here that this uncertainty measure is conditional under particular conditions which includes assumptions made (e.g. a particular probability model assumed) and eventual additional prior information available. In particular situations, prior information concerning the parameters might be available, e.g. from expert knowledge, in other cases one wants to introduce as little information as possible prior to data observation and a most *uninformative* prior distribution is sought. The latter perspective is treated in the framework of *objective Bayes*. In particular, *reference priors* which are designed to have a minimal effect on the posterior inference, are based on information theoretical principles and include *Maximum entropy* or *Jeffrey's prior* as special cases. A general introduction to Bayesian statistics is provided in Bernardo (2011) and details on reference priors can be found in Berger et al. (2009). We summarise below a few facts highlighting how Bayesian statistics relates to the traditional frequentist analysis.

1. In the frequentist perspective randomness expresses the fact that repeated experiments will not provide identical observations. Bayesians also consider that randomness can express uncertainty about the state of the word.

2. The parameter θ is considered as a fixed unknown both in the frequentist and Bayesian perspectives, however Bayesians consider that θ can be viewed as random in the sense of probability as measure of uncertainty.

3. Frequentist statistics is concerned with the efficiency of a statistical procedure on repeated similar problems [6]. Inference takes into account the particular data sets at hand along with other realisations which were not observed but could potentially be observed, averaging is performed over the possible data X. On the other hand, Bayesian statistics makes inference conditioned on the observed data, averaging is performed over the parameter of interest.

We give below the definition of a Bayesian parametric model, as explained above the parameter θ is a random variable, we thus need to equip the parameter state space Θ with a σ-algebra, denoted by $\mathscr{A}$.

[5] Terminology should be considered with care since the term *non-parametric* has multiple meanings in the literature.

[6] In Bayarri and Berger (2004) the precise meaning of the frequentist principle is explained in more details, in particular it is emphasized that accuracy must be considered for the same task applied to multiple problems.

Definition 2.4 (Bayesian parametric model)**.** *A Bayesian parametric model is a family of conditional distributions* $\{\mu_{X|\theta}|\ \theta \in M_\Theta\}$*, where* M_Θ *is the set of all* $(\mathscr{F}, \mathscr{A})$*-measurable functions from* Ω *to* Θ*.*

Definition 2.4 is based on conditional distributions but inference is simplified if we consider a model based on conditional densities. When there exists a measure ν such that $\mu_{X|\theta} \ll \nu$ for every member $\mu_{X|\theta}$ of our parametric model, $\ll$ denoting absolute continuity, we can directly work with a Bayesian family of densities. *A Bayesian parametric density model* is a family of conditional densities $\{f_{X|\theta}|\ \theta \in M_\Theta\}$, where the densities are all defined w.r.t. to the same measure ν. Recalling that we are interested in the posterior distribution of the parameter given the data, the next question arising is how can we actually calculate this distribution. Bayes theorem provides a formula for the posterior distribution of Bayesian parametric density models. This result shows that when the parametric family is dominated, i.e. there exists a measure ν with respect to which every family member is absolutely continuous, the posterior distribution is absolutely continuous w.r.t. to the prior.

Theorem 2.1 (Bayes theorem)**.** *Consider a Bayesian parametric family density model* $\{f_{X|\theta}|\ \theta \in M_\Theta\}$ *with prior measure* μ_θ *on the parameter sample space* Θ*. The posterior distribution* $\mu_{\theta|X}$ *is absolutely continuous* w.r.t. μ_θ *and has Radon-Nikodym derivative:*

$$\frac{\mathrm{d}\mu_{\theta|X=x}}{\mathrm{d}\mu_\theta}(\theta_0) = \frac{f_{X|\theta=\theta_0}(x)}{\int f_{X|\theta=\theta_0}(x)\mathrm{d}\mu_\theta(\theta_0)}, \tag{2.15}$$

for every x *which is not in the set* $N = \left\{y \in E|\ \int f_{X|\theta=\theta_0}(y)\mu_\theta(\mathrm{d}\theta_0) \in \{0, \infty\}\right\}$*.*

It can be shown that observations actually occur in N with probability zero, and thus equation 2.15 holds for every data we actually observe. As for traditional models we can also define *non-parametric Bayesian models*. Handling a potentially increasing number of dimensions for the parameter seems difficult given the fact that a prior probability on the parameter space needs to be defined. Since the prior distribution need to be defined and fixed [7] prior to conducting inference, the prior will typically be defined on an infinite-dimensional space, leaving therefore enough potential degrees of freedom to explain new observations. In Orbanz (2008), the adaptivity of such models is stressed, also pointing out that their main characteristic is not the infinite dimensionality but their ability to explain any fixed number of observations. Explaining any finite sample will effectively require only a finite subset of dimensions, and the infinite-dimensionality is needed to potentially explain any fixed, but arbitrarily large, sample.

2.2.1 Mixture model

When standard distributions do not show a structure rich enough to explain the data, mixture models offer more flexible alternatives. To obtain more complex models mixture models combine two distributions by conditioning and integration. An intuitive representation is given by a two-staged sampling procedure: draw the value of a first random variable, then sample from a second distribution which depends on the first obtained value. Let X, Z be random variables taking values in the measurable spaces $(E_x, \mathscr{E}_x), (E_z, \mathscr{E}_z)$. We will work with the conditional distribution of X given Z, which requires to impose some conditions on the chosen spaces. A sufficient condition to insure the existence of conditional distributions is to consider Polish spaces equipped with Borel σ-algebras.

Consider the family of the conditional distributions of X given Z:

$$\{\mu_{X|Z}|Z \in M_{E_z}\},$$

[7] The data can also be used to determine a suitable prior but we do not include this first step when speaking of inference.

where M_{E_z} is the space of ($\mathscr{F}$-$\mathscr{E}_z$)-measurable functions. If the family is dominated i.e. if there exists a measure ν on E_x such that $\mu_{X|Z} \ll \nu$ for every member $\mu_{X|Z}$ in the family, then the family of distributions can be represented as a family of densities:

$$\{f_{X|Z}|Z \in M_{E_z}\}. \tag{2.16}$$

By integrating $f_{X|Z=z}$ over z using μ_Z, the distribution of Z, we recover a density function depending on x only:

$$f_X(x) = f_{X|\mu_Z}(x) = \int_{E_z} f_{X|Z=z}(x)\,\mathrm{d}\mu_Z(z). \tag{2.17}$$

The notation $f_{X|\mu_Z}$ in (2.17) emphasizes that the density of X depends on μ_Z, to be more precise f_X is parametrized by μ_Z. The formulation (2.16) looks similar to a Bayesian model but, in the case of a mixture model, the variable Z is not the parameter of interest and is integrated out, the parameter of the model being μ_Z. Previous remarks are summarised in the following definition.

Definition 2.5 (Mixture model)**.** *A mixture model is a family of the form*

$$\left\{\int_{E_z} f_{X|Z=z}(x|z)\,\mathrm{d}\mu_Z(z) \mid \mu_Z \text{ is a pm on } E_z\right\}.$$

In a *Bayesian mixture model* the parameter, which is here μ_Z, is a random variable. Thus to obtain the Bayesian version of the model described in 2.5 one has to define a prior distribution for μ_Z, i.e. a prior distribution over distributions.

Finite mixture

When Z is a discrete random variable taking only a finite number of different values, the obtained model, called a *finite mixture*, has a simplified formulation and becomes easier to interpret. In the finite case, we assume that Z has a probability mass function f_Z which puts mass $p_1, \ldots, p_K \in \mathbb{R}_{+\setminus 0}, \sum_k p_k = 1$, on a finite number of points $\theta_1, \ldots, \theta_K \in E_z$:

$$f_Z(z; p, \theta) = \sum_{k=1}^{K} p_k \delta_{\theta_k}(z) \tag{2.18}$$

where $p = (p_1, \ldots, p_K)$ and $\theta = (\theta_1, \ldots, \theta_K)$. Combining (2.17) and (2.18) we find (by exchanging the sum with the integral and then integrating) that the density of X can be rewritten as a weighted sum of K densities:

$$f_X(x; p, \theta) = \sum_{k=1}^{K} p_k f_{X|\theta_k}(x) \tag{2.19}$$

From equation (2.19) we can see that each observation of f_X is generated from one of the K densities $f_{X|\theta_1}, \ldots, f_{X|\theta_K}$ with probabilites $p_1, \ldots, p_K$. Introducing the latent variables of the observations' class assignments $C_i, i \in \{1, \ldots, n\}$ defined by $C_i = k$ if x_i is an observation generated from the density $f_{X|\theta_k}$, we can generate observations x_i from f_X in a two-staged process:

1. Draw a class assignment c from $C_i|p \sim \text{mult}(p_1, \ldots, p_K, 1)$.
2. Draw x_i from $X_i|\{C_i = c\} \sim f_{X|\theta_c}$.

A Bayesian finite mixture requires only to specify prior distributions for θ and p, avoiding the more complicated task of defining a prior for μ_Z. The prior for θ depends on the particular model

considered. A standard prior for p is the conjugate to the multinomial distribution, the Dirichlet distribution denoted $\mathrm{Dir}(\alpha)$:

$$f(p_1, \ldots, p_{K-1}; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} p_i^{\alpha_i - 1}; \quad p_K = 1 - \sum_{i=1}^{K-1} p_i, \tag{2.20}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K_{+\setminus 0}$ and the normalising constant is the Beta function expressed using the Gamma function:

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$$

Example: finite Gaussian mixture A *finite Gaussian mixture* is a mixture model having Gaussian conditional densities $f_{X|\theta_k}$:

$$f_{X|\theta_k}(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right) \tag{2.21}$$

where $\theta_k = (\mu_k, \Sigma_k)$ and assuming X has dimension d. To fully specify this example from a Bayesian perspective we need to provide prior distributions for the parameters p and θ. The parameters will have different posterior distributions depending on the class but all have the same prior distribution and share common hyperparameters (the parameters of the prior distribution). We give an example of conjugate finite Gaussian mixture model which is fully characterise by the following set of equations:

1. The class probabilities are *a priori* Dirichlet distributed

$$(p_1, \ldots, p_K)|\alpha \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right), \tag{2.22}$$

   with hyperparameter $\alpha \in \mathbb{R}_{+\setminus 0}$.

2. The parameters $\theta_1, \ldots, \theta_K$ are *a priori* independent with the following densities

$$\Sigma_k|\Delta, \nu \overset{\text{iid}}{\sim} \mathrm{IW}(\Delta, \nu) \tag{2.23}$$

$$\mu_k|\Sigma_k, m, \kappa \overset{\text{ind}}{\sim} \mathcal{N}\left(m, \kappa^{-1}\Sigma_k\right), \quad k = 1, \ldots, K, \tag{2.24}$$

   where the hyperparameters are $\nu > 0, \Delta \in \mathbb{R}^{d \times d}$ and $\kappa > 0, m \in \mathbb{R}$.

3. The latent variables are *a priori* iid multinomial

$$C_i|p \overset{\text{iid}}{\sim} \mathrm{mult}(p_1, \ldots, p_k, 1), \;\; i = 1, \ldots, n.$$

4. The components of the mixture $f_{X|\theta_k}$ are normal densities as in equation (2.21)

$$X_i|C_i, \theta \overset{\text{ind}}{\sim} \mathcal{N}_d\left(\mu_{C_i}, \Sigma_{C_i}\right), \;\; i = 1, \ldots, n.$$

5. Finally the multivariate density f_X is as in equation (2.19)

$$f_X(x) = \sum_{k=1}^{K} p_k f_{X|\theta_k}(x|\theta_k).$$

Infinite mixture

Infinite mixture models encompass the cases where we do not restrict Z to a finite number of values. The density f_X cannot be reduced to equation (2.19) and its general form must be retained. In a Bayesian framework, this implies that μ_Z, which is considered as a random variable, needs a prior distribution and we face the problem of defining a distribution over the infinite dimensional space of distributions on E_z:

$$M = \{\mu_Z : \mathscr{E}_z \to [0,1] \mid \mu_Z \text{ is a probability measure}\} \tag{2.25}$$

Keeping the issue of its construction for later, assume that μ_M is a measure on M. We can then integrate over M using μ_M and the unconditional density takes the following form:

$$f_X(x) = \int_M \int_{E_z} f_{X|Z=z}(x) \mathrm{d}\mu_Z(z) \mathrm{d}\mu_M. \tag{2.26}$$

Equation (2.26) provides the unconditional form of a *Bayesian infinite mixture model*, where our parameter μ_Z is a random measure with prior distribution μ_M. A closer look at the space M reveals that it has a product structure, a feature which we can take advantage of to construct the measure μ_M. Indeed, an element of M being a function from $\mathscr{E}_z$ to a real interval, M actually is a set of functions and can be written using the product form $M = \mathbb{R}^{\mathscr{E}_Z}$. This, in particular, implies that random variables taking values in M also have a product structure indexed by $\mathscr{E}_Z$.

Dirichlet process mixture

If we could construct μ_M such that its samples are almost surely discrete distributions, i.e. every μ_Z sampled from μ_M is a discrete distribution with probability one, the density f_X would take the form of a sum, similar to equation (2.19), but infinite instead of limited to K components. In the case of a finite E_z a solution is known, the Dirichlet distribution constitutes a distribution over finite discrete distributions (i.e. discrete distributions on a finite set). The idea underlying the *Dirichlet process* Ferguson (1973) is to construct an extension to the infinite case, going from the distribution to a stochastic process. The Dirichlet process was already introduced in Definition 2.2, where the desired stochastic process was constructed by specifying its finite dimensional margins, and, as previously mentioned, draws of a Dirichlet process are a.s. discrete distributions (Blackwell, 1973).

We recall that a Dirichlet process $G : \Omega \to M$ is parametrized by a base measure $G_0 \in M$ and a concentration parameter by $\alpha \in \mathbb{R}_{+\setminus 0}$, and admits the following product representation:

$$G = \prod_{A \in \mathscr{E}_z} G(A).$$

The stochastic process G can be seen as a function from the product space $\Omega \times \mathscr{E}_z$ to the real line. For every fixed measurable set A, $G(A) : \Omega \to [0,1]$ is a random variable, and for every fixed event ω, $G(\omega) : \mathscr{E}_z \to [0,1]$ is a measure on the parameter space $(E_z, \mathscr{E}_z)$. This later property reveals that G can also be interpreted as a random measure on the parameter space. A classic way to construct a stochastic process, i.e. by specifying all finite margins and then make use of Kolmogorov extension theorem (Kolmogorov, 1950) to obtain the infinite dimensional extension, would require here to specify the value of the process for every Borel set $A \in \mathscr{E}_z$. However, as mentioned in Section 2.1, in the particular case of the Dirichlet process we can actually restrict ourself to the set of measurable finite partitions of E_z

$$\mathcal{P}_{E_z} = \{(H_j)_{j \in J} \mid J \text{ is finite},\ H_j \in \mathscr{E}_Z\ \forall j,\ \cup_{j \in J} H_j = E_Z,\ H_i \cap H_j = \varnothing \text{ for } i \neq j\}, \tag{2.27}$$

and the Dirichlet process G is characterised by

$$\prod_{i=1,\dots,m} G(A_i) \sim \mathrm{Dir}\left(\alpha G_0(A_1), \dots, \alpha G_0(A_m)\right), \quad \forall m \in \mathbb{N}, \forall (A_i)_{i=1,\dots,m} \in \mathcal{P}_{E_z}. \tag{2.28}$$

Equation (2.28) is often written with the following form:

$$(G(A_1), \ldots, G(A_m)) \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_m)). \tag{2.29}$$

The base measure G_0 can be seen as the mean of the Dirichlet process since $E[G(A)] = G_0(A), \forall A \in \mathscr{E}_z$. The concentration parameter α plays the role of an inverse variance, larger values corresponding to more concentrated mass around G_0. However, draws from G do not become arbitrarily close to G_0 as α increases. Indeed, if G_0 is a continuous measure draws from G, which are discrete measures, always are singular with G_0. Existence and uniqueness of the Dirichlet process can be proven by different methods, besides using Kolmogorov extension theorem [8] as mentioned previously, a stick-breaking construction is introduced in Sethuraman (1994) and a proof based on Pólya urn schemes is given in Blackwell and MacQueen (1973). Since this later scheme provides a helpful representation to perform Gibbs sampling (refer Section 2.3.3) we detail it below. A *Pólya sequence* with parameters G_0, α is a sequence of random variables $Z_i, i \geq 1$ such that for every $A \in \mathscr{E}_z$

$$\mathbb{P}(Z_1 \in A) = \frac{G_0(A)}{G_0(E_z)} \tag{2.30}$$

$$\mathbb{P}(Z_{n+1} \in A | Z_1, \ldots, Z_n) = \frac{G_n(A)}{G_n(E_z)}, \text{ with } G_n = \alpha G_0 + \sum_{i=1}^{n} \delta_{Z_i}. \tag{2.31}$$

Define a new measure G_n^* by $G_n^*(A) = G_n(A)/G_n(E_z)$, Blackwell and MacQueen (1973) showed that G_n^* converges with probability one to a discrete measure G_∞ which is distributed according to a Dirichlet process

$$G_\infty \sim \mathrm{DP}(\alpha, G_0). \tag{2.32}$$

Assuming $G_0(E_z) = 1$, equation (2.31) becomes

$$\mathbb{P}(Z_{n+1} \in A | Z_1, \ldots, Z_n) = \frac{\alpha}{\alpha + n} G_0(A) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{Z_i}(A), \tag{2.33}$$

and we can write the predictive distribution of $Z_{n+1}|Z_n$ as

$$Z_{n+1} | Z_1, \ldots, Z_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{Z_i}. \tag{2.34}$$

Equation (2.34) gave rise to the famous urn scheme metaphor of the Dirichlet process and can be interpreted the following way: imagine that each value in the space E_z is a unique colour, and the observations from Z are obtained by drawing a coloured ball and assigning its colour to be the observed value of Z. Imagine also that an urn is kept to collect previously seen balls. At the beginning of the sampling process the urn is empty and we draw a colour from the distribution G_0. In addition, we paint a ball with that observed colour and drop it in the urn. In the subsequent steps, we either:

a) With probability $\frac{n}{\alpha+n}$, we draw a previously seen colour by taking (at random) a ball from the urn, replace the ball in the urn, paint a new ball with this colour and drop it in the urn as well.

b) With probability $\frac{\alpha}{\alpha+n}$, sample a new colour from G_0, paint a new ball with this colour and drop it in the urn.

[8] Kolmogorov's theorem give the existence and unicity on the product σ-algebra $\bigotimes_{A \in \mathscr{E}_z} \mathcal{B}(\mathbb{R})$ but we actually are interested in the Borel σ-algebra on M, $\mathcal{B}(M)$.

2.3 Markov Chain Monte Carlo sampling

2.3.1 Monte Carlo Methods

Different, more or less restrictive, definitions of Monte Carlo (MC) methods exist in the literature. MC can roughly be described as a class of algorithms based on random sampling but the following description, which largely agrees with the definition found in Anderson (1999), provides a more precise definition.

> Monte Carlo methods are algorithms approximating an expectation by the sample mean of a function of simulated random variables.

The above definition might seem restrictive but many problems, such as computing probabilities, integrals or discrete sums, can be reformulated as the computation of an expectation. The key idea in MC methods is to approximate the expectation to be computed by the mean of a large number of samples from the distribution. Consider a random variable $X : \Omega \to \mathbb{R}^d, d < \infty$, with density, in the continuous case, or probability mass function, in the discrete case, denoted by f_X. The expected value of a function of X

$$\mathrm{E}(g(X)) = \int g(x) f_X(x) \mathrm{d}\mu(x),$$

where μ is either the Lebesgue or the counting measure on $\mathbb{R}^d$, can be approximated by the *Monte Carlo estimate*

$$\hat{g}(x) = \frac{1}{n} \sum_{i=1}^{n} g(x_i), \tag{2.35}$$

where $x_i, i = 1, \ldots, n$ are samples from the distribution of X. The key part in the above process is sampling from μ_X, which is obviously not always possible or easy to achieve, and there exist different MC methods tackling the problem, the most famous being rejection sampling and importance sampling. These techniques, however, scale badly with dimensionality and we need more complex algorithms to handle higher dimensional problems. *Markov Chain Monte Carlo* (MCMC) methods provide powerful alternative solutions by building a Markov Chain which admits our distribution of interest μ_X as stationary distribution. Before developing on MCMC, we briefly present some introductory notions on Markov Chains.

2.3.2 Markov Chains

Markov chains are Markov processes in discrete time, however, to simplify the exposition, we directly define the concept required for MCMC: homogeneous Markov chains.

Definition 2.6 (Homogeneous Markov Chain)**.** *A homogeneous Markov chain on* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ *with transition kernel* K *is a discrete time stochastic process* $(X_t)_{t\in\mathbb{N}}$ *with values in* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ *such that*

$$\mathbb{P}\{X_{t+h} \in A | \sigma(X_t; t \geq 0)\} = K^h(X_t, A), \quad \forall A \in \mathcal{B}(\mathbb{R}^d), \tag{2.36}$$

where K^h *is the* h*-th iterate* [9] *of* K *and* $t \in \mathbb{N}, h \in \mathbb{N} \setminus \{0\}$.

The distribution of X_0 is called the *initial distribution*. The transition kernel is the mechanism controlling the evolution of the stochastic process over time by prescribing, as its name indicates,

[9] We recall that the product of two kernels is: $(KQ)(x, A) = \int K(x, \mathrm{d}y) Q(y, A)$.

the transition probabilities from one state to the next. The main characteristic of a Markov chain, expressed in equation (2.36), is that the probability of a future state does not depend on the entire history of the process but only on the current state. In Definition 2.6, the chain is called *homogeneous* because the kernel K is constant over time. Under certain regularity conditions, which we do not detail here [10], Markov chains stabilise over time and adopt an equilibrium distribution, defined below in 2.38.

- A distribution π on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ satisfying:

$$\pi(A) = \int K(x, A)\pi(\mathrm{d}x), \quad \forall A \in \mathcal{B}(\mathbb{R}^d), \forall x \in \mathbb{R}^d, \tag{2.37}$$

  is a *stationary distribution* of the chain.

- A stationary distribution π on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ satisfying:

$$\pi(A) = \lim_{h \to \infty} K^h(x, A), \quad \forall A \in \mathcal{B}(\mathbb{R}^d), \tag{2.38}$$

  for π-almost all x, is an *equilibrium distribution* of the chain.

2.3.3 Markov Chain Monte Carlo

Most of the traditional theory on Markov chains assumes a known transition kernel and investigates the properties of the resulting stochastic process. MCMC approaches the problem from another angle, a desired equilibrium distribution π is first chosen to be the target distribution, i.e. the distribution from which we want samples. The aim of MCMC is then to construct a stochastic kernel, and thereby a Markov chain admitting the target distribution as equilibrium. The empirical distribution of samples generated by the chain will approximate the distribution of interest (sometimes up to a proportionality constant). Two different methods leading to such Markov chains will be used in later chapters and are shortly introduced below: *Metropolis-Hastings* (Metropolis et al., 1953), (Hastings, 1970) and *Gibbs sampling* (Gelfand and Smith, 1990). To construct the desired transition kernel K we assume that π has a density w.r.t. some σ-finite measure μ (e.g. the Lebesque measure) and K has the form

$$K(x, \mathrm{d}y) = k(x, y)\mu(\mathrm{d}y) + r(x)\delta_x(\mathrm{d}y), \tag{2.39}$$

for some real-valued function k with $k(x, x) = 0$ and where $r(x) = 1 - \int k(x, y)\mu(\mathrm{d}y)$. Equation (2.39) can be interpreted as a decomposition of K in one part controlling the probability to leave state x and another part steering the probability of staying in state x. It can be shown that if k satisfies the *detailed balance* condition

$$\pi(x)k(x, y) = \pi(y)k(y, x), \tag{2.40}$$

then π is stationary for the chain defined by K, and under further conditions K^n converges to π when $n \to \infty$. The problem has thus been reduced to the choice of the function k. The detailed balance condition (2.40) can be interpreted as a requirement of reversibility or symmetry in the chain. While collecting approximated samples of π using MCMC techniques, convergence issues should be accounted for. In particular, the first b-th samples should be discarded, where b is an integer to determine, since the chain needs a certain time to converge. This first period before convergence is called the burn-in period. One should finally also bear in mind that samples emerging from a Markov chain are by definition not independent.

[10] An ergodic Markov chain, i.e. an irreducible, aperiodic and positive recurrent chain, will adopt an equilibrium distribution.

Metropolis-Hastings sampling

Assume there exists a transition kernel $Q(x, \mathrm{d}y) = q(x, y)\mu(\mathrm{d}y)$ from which we can sample. In general, q will not satisfy condition (2.40) and the main idea of Metropolis-Hasting (MH) is to correct q to obtain a new kernel which does. This correcting factor is defined as

$$\alpha(x, y) = \begin{cases} \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right), & \text{if } \pi(x)q(x, y) > 0 \\ 1, & \text{if } \pi(x)q(x, y) = 0 \end{cases}. \tag{2.41}$$

The transition kernel of the MH chain finally is

$$K_{\mathrm{MH}}(x, \mathrm{d}y) = q(x, y)\alpha(x, y)\mu(\mathrm{d}y) + r(x)\delta_x(\mathrm{d}y), \tag{2.42}$$

where $r(x) = 1 - \int q(x, y)\alpha(x, y)\mu(\mathrm{d}y)$. Algorithm 1 exposes the MH sampling process. Further introductory details on MH sampling can be found in Chib and Greenberg (1995).

Algorithm 1 Metropolis-Hastings Sampling

Input: Target distribution π, initial state x_0, transition kernel Q.
Results: Approximated samples from π.
for $t = 1, \ldots, n$ **do**
  Draw y from Q;
  Set $x_{t+1} = \begin{cases} y, & \text{with probability } \alpha(x_t, y) \\ x_t, & \text{with probability } 1 - \alpha(x_t, y) \end{cases}$;
end for
Return the samples $x_b, \ldots, x_n$, where b represents the burn-in period.

Gibbs sampling

Assume $X = (X_1, \ldots, X_d)$ is distributed according to π, let f be a function of X and define the following transition kernel

$$K(x, A) = \mathbb{P}\{X \in A | f(X) = x\}. \tag{2.43}$$

The kernel K defined in (2.43) admits π as stationary distribution. The idea underlying *Gibbs sampling* is to sample in turn from the full conditionals by setting $f_i(X) = X^{(i)}$ and defining the conditional kernels $K_i(x, A) = \mathbb{P}\{X \in A | X^{(i)} = x^{(i)}\}$. The Gibbs transition kernel is finally defined by $K = K_1 \ldots K_d$ and still admits π as stationary distribution. The sampling procedure is given in Algorithm 2.3.3 and more details can be found in Gelfand (2000).

Algorithm 2 Gibbs Sampling

Input: Target distribution π, initial state x_0.
Results: Approximated samples from π.
for $t = 1, \ldots, n$ **do**
  Sample $X_{t+1,1} | X_{t,2}, \ldots, X_{t,d}$
  for $j = 2, \ldots, d-1$ **do**
    Sample $X_{t+1,j} | X_{t+1,1}, \ldots, X_{t+1,j-1}, X_{t,j+1}, \ldots, X_{t,d}$
  end for
  Sample $X_{t+1,d} | X_{t+1,1}, \ldots, X_{t+1,d-1}$
end for
Return the samples $x_b, \ldots, x_n$, where b represents the burn-in period.

Gibbs sampling for Dirichlet process mixture

We denote again the latent variable of the observations assignments by C_i, $i = 1, \ldots, n$. Define $C^{(i)} = \{C_1, \ldots, C_n\}/\{C_i\}$ and $Z^{(i)}$ analogously. The conditional posterior distributions of the parameters are:

$$Z_i | Z^{(i)}, C^{(i)} \sim \frac{\alpha}{\alpha + n - 1} G_0 + \frac{1}{\alpha + n - 1} \sum_{j=1}^{k^{(i)}} n_j^{(i)}, \tag{2.44}$$

where $k^{(i)}$ is the number of mixture components when we do not consider the observation x_i and $n_j^{(i)}$ is the number of observations in component j. The posterior distribution of the parameter Z_i for observation i given the data X, the remaining parameters $Z^{(i)}$ and class assignments $C^{(i)}$ is

$$Z_i | X, Z^{(i)}, C^{(i)} \sim q_{i,0} G_{i,0} + \sum_{j=1}^{k^{(i)}} q_{i,j}, \tag{2.45}$$

where

$$q_{i,j} = \begin{cases} c\, \alpha\, h_i\left(x_i\right) & \text{for } j = 0 \\ c\, n_j^{(i)}\, f_{X_i | Z_j}\left(x_i\right) & \text{else} \end{cases}, \tag{2.46}$$

and

1. $G_{i,0}$ is the posterior distribution of Z knowing only the observation x_i.
2. $h_i\left(x_i\right)$ is the marginal density of X_i evaluated at x_i obtained by integrated over the parameter Z_i using the prior measure G_0.

$$h_i\left(x_i\right) = \int f_{X_i | Z = z}\left(x_i\right) \mathrm{d}G_0\left(z\right), \tag{2.47}$$

3. c is a normalisation constant to ensure that $\sum_{j=1}^{k^{(i)}} q_{i,j} = 1$, $i = 1, \ldots, n$.

From equation (2.45) we can see that the posterior distribution of the assignments are:

$$\mathbb{P}\left\{C_i = j | X, Z^{(i)}, C^{(i)}\right\} = q_{i,j}. \tag{2.48}$$

2.4 Dependence and measures of it

2.4.1 Introduction

Copula is primarily a concept of dependence since it aims at distinguishing joint and marginal behaviours of a random variable. In order to later clarify the relationship between copula and dependence we provide in this section an introduction to *stochastic dependence*. As we will see, dependence of random variables is often difficult to measure even though it seems very intuitive. We therefore start by defining a simpler concept: the absence of dependence or independence.

Stochastic independence is defined for a collection of different objects, e.g. events, random variables or stochastic processes, but always relates to the definition of independent σ-algebras.

Definition 2.7 (Independency)**.** *A collection of σ-algebras $\mathscr{F}_1, \ldots, \mathscr{F}_d \subseteq \mathscr{F}$ is an* independency *if*

$$\mathrm{E}(X_1 \ldots X_d) = \mathrm{E}(X_1) \ldots \mathrm{E}(X_d), \tag{2.49}$$

for all positive random variables $X_1, \ldots, X_d$ measurable w.r.t.*, respectively, $\mathscr{F}_1, \ldots, \mathscr{F}_d$. We call the members of an independency* independent*. An infinite collection is an independency if every finite subset is.*

In the above definition the expectation of the different variables' product can be decomposed as the product of the marginal expectations, meaning that the multivariate quantity $\mathrm{E}(X_1 \ldots X_d)$ actually depends solely on univariate quantities $\mathrm{E}(X_1) \ldots \mathrm{E}(X_d)$, and this must hold for all random variables which are "relevant" for the considered σ-algebras. This concept can be naturally extended to random variables by considering the σ-algebras they generate.

Definition 2.8 (Independent random variables)**.** *A collection $X = \{X_t, t \in T\}$ of random variables with index set T is independent if $\{\sigma(X_t), t \in T\}$ is an independency. The random variables $X_t, t \in T$ are then called independent.*

Definition 2.8 is a natural extension of Definition 2.7 but might not seem as intuitive as the following characterisation, a proof of which can be found in Çinlar (2011).

Proposition 2.1. *The random variable $X = (X_1, \ldots, X_d)$ has independent dimensions $X_1, \ldots, X_d$ if and only if the joint distribution is the product of the marginal distribution:*

$$\mu_X = \mu_{X_1} \times \cdots \times \mu_{X_d}. \tag{2.50}$$

Independence also affects conditional distributions, the conditional distribution of X given Y is equal to the marginal distribution of X when both variables are independent :

$$\mu(A \times B) = \int_A \mu_X(\mathrm{d}x) K(x, B) = \int_A \mu_X(\mathrm{d}x) \mu_Y(B) = \mu_X(A) \mu_Y(B). \tag{2.51}$$

In summary, independent variables live in a product space and are determined by the product measure on that space. This means that potential information about the realisation of one variable is irrelevant to our knowledge about the other variables' outcomes. When variables are not independent, complete or partial knowledge about one variable can affect our predictions concerning the remaining variables, this is expressed by a deviation of the conditional distribution from the unconditional distribution. Stochastic dependence occurs whenever random variables are not independent, however dependence is a complex concept for which there exists various definitions corresponding to different types of dependence. We will not provide an exhaustive exposition of dependence related concepts, as this research area is too vast to be completely reviewed here, but will instead present selected examples. Dependence can be characterised in different ways. First, we can verify if random variables fulfill certain *dependence properties*, i.e. conditions designed to express dependence in a particular way. Amongst the most prominent dependence properties counts *linear dependence* which envisage dependence as a linear relationship between variables, further examples are given by 2.2 and 2.3 below. A second possibility consists in defining when a multivariate random variable is more dependent than another, thereby creating a *stochastic ordering* of all random variables w.r.t. dependence. We will not cover stochastic ordering and point to Joe (1997) for an introduction. Finally, a third approach consists in defining a function of the random variables which captures and summarises their dependence, such functions are called *dependence measures*. Note that some dependence properties can be defined using dependence measure, e.g. the strength of linear dependence between two univariate random variables is measured by the linear correlation coefficient defined below and variables are called *linearly dependent* if their coefficient is strictly larger than zero.

Example 2.1 (Linear dependence)**.** The linear correlation between two univariate rv measures their linear dependence:

$$\rho_l(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1)\text{var}(X_2)}}.$$

Example 2.2 (Positive orthant dependence)**.** An example of a broader scoped dependence property is *positive orthant dependence*. A multivariate random variable X is *positive upper orthant dependent* if

$$P(X_j > a_j, j = 1, \ldots, d) \geq \prod_{j=1}^{d} P(X_j > a_j), \quad \forall a_j \in \mathbb{R}, \tag{2.52}$$

and is *positive lower orthant dependent* if

$$P(X_j \leq a_j, j = 1, \ldots, d) \geq \prod_{j=1}^{d} P(X_j \leq a_j), \quad \forall a_j \in \mathbb{R}. \tag{2.53}$$

X is called *positive orthant dependent* when it satisfies both (2.52) and (2.53). The intuitive idea behind this dependence property is that dimensions are more likely to take large (respectively small) values together than independently.

Example 2.3 (Tail dependence)**.** This dependence property is concerned with the probability of joint extreme events, *tail dependence* appears when rare events have a tendency to appear jointly in different dimensions. More precisely, tail dependence is measured by the *tail dependence coefficients* which are defined for the 2-dimensional case as

$$\lambda_L = \lim_{u \searrow 0} \mathbb{P}\left(X_2 \leq F_2^{-1}(u) | X_1 \leq F_1^{-1}(u)\right), \tag{2.54}$$

for the *lower tail dependence coefficient* and as

$$\lambda_U = \lim_{u \nearrow 0} \mathbb{P}\left(X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)\right), \tag{2.55}$$

for the *upper tail dependence coefficient*. When a tail coefficient is larger than zero the variable is said to have *asymptotic dependence* and this dependence becomes stronger when the coefficient's value approaches one.

2.4.2 The axiomatic approach

This sections aims at further clarifying the idea of dependence, bearing in mind that copula are, at least in the case of continuous variables, the right tool to capture dependence. The few examples given above already express a noticeable variety of dependency concepts. This variety naturally raises the question of the choice of a dependence measure. Influential work on this subject has been done by Rényi, in Rényi (1959) he introduces a list of desired properties which a good dependence measure should satisfy, the *Rényi postulates*, and examine different existing measures in the light of this set of axioms. Of the measures considered, which include the linear correlation coefficient, only Gebelein's maximal correlation (Gebelein, 1941) fulfills all requirements. Whereas Rényi originally formulated his axioms for bivariate variables only, we present below the postulates extended to the d-dimensional case. Much work on dependence was first done for the bivariate case, and if some properties are straightforward to adapt for higher dimensions, others are more difficult to generalise which sometimes leads to different multivariate versions. Of the postulates presented below, axioms (A), (B), (C), (D), (F) are trivial extensions of the bivariate case whereas axioms (E) and (G) are less straightforward extensions [11].

(A) $\delta(X)$ is defined for any random variable $X = (X_1, \ldots, X_d)$ such that, with probability one, X_j is not a constant, $\forall j = 1, \ldots, d$, .

[11] The original formulation of axiom (E) for the bivariate case is: $\delta(X) = 1$ if either $X_1 = g(X_2)$ or $X_2 = g(X_1)$ for some real valued, measurable function g.

(B) For any permutation of the dimensions $\sigma = (j_1, \ldots, j_d)$ we have

$$\delta(X_1, \ldots, X_d) = \delta(X_{j_1}, \ldots, X_{j_d}). \tag{2.56}$$

(C) $0 \leq \delta(X) \leq 1$.

(D) $\delta(X) = 0$ if and only if $X_1, \ldots, X_d$ are independent.

(E) $\delta(X) = 1$ if there exists an index i and a real valued, measurable function g such that $X_i = g(X^{(i)})$ with probability one, where $X^{(i)}$ denotes the variable X with dimension i removed.

(F) For every injective transformation T of X, $T(X) = (T_1(X_1), \ldots, T_d(X_d))$, δ remains invariant

$$\delta(T(X)) = \delta(X). \tag{2.57}$$

(G) If (X_1, X_2) is bivariate Gaussian, then we obtain the absolute value of the linear correlation $\delta(X) = |\rho_l(X_1, X_2)|$.

Axiom (A) excludes measures based on the variance which might not be defined. Axiom (B) expresses the intuitive requirement that a dependence measure should be invariant w.r.t. the order of the dimensions, and, in the particular bivariate case, should be symmetric. Axiom (F) adds another desired property of invariance, namely invariance to injective transformations of the margins. Whereas axiom (D), which provides an unambiguous characterisation of independence, is widely adopted, axiom (E) implicitly defines perfect dependence as the case where one dimension is almost surely a measurable function of the others and some authors suggested modified versions of it. In Bell (1962), Bell compares Gebelein's maximal correlation S with two possible normalisations of Shannon's mutual information (see also Section 2.5.2) and, on the basis on his analysis, proposes some revisions of Rényi's postulates. Arguing that S might take the value one too often, he first suggests to replace the "if" statement in postulate (E) by an equivalence relationship and proposes two alternative (non equivalent) versions of axiom (E). He then also suggests using a less restrictive version of axiom (G). We give below the multivariate generalisations of his revised postulates.

(E2a) $\delta(X) = 1$ if and only if there exists an index i and a real valued, measurable function g such that $X_i = g(X^{(i)})$, with probability one.

(E2b) $\delta(X) = 1$ if and only if every variable can be expressed as a function of the others, i.e. for every i there exists some real valued, measurable function g_i such that $X_i = g_i(X^{(i)})$, with probability one.

(G2) If (X_1, X_2) is bivariate Gaussian, then $\delta(X)$ is a strictly increasing function of the linear correlation's absolute value $|\rho_l(X_1, X_2)|$.

The work in Bell (1962) focuses on strictly positive probability spaces, i.e. in which every measurable set of probability zero is the empty set. In particular, his analysis does not encompass continuous variables and he leaves open the question of a possible extension to arbitrary random variables. This problem is later considered in Joe (1989) where several normalisations of Shannon's mutual information for continuous and discrete variables are introduced. Dependence measures based on mutual information will be treated in Section 2.5.2. The case of continuous variables is also developed in Micheas and Zografos (2006). A multivariate extension of Rényi's postulates is proposed where some adaptations are made to suit the continuous case. In particular a less strict formulation of axiom (C) is adopted, axioms (E) is adapted accordingly, and axiom (G) is replaced by axiom (G2) to maintain coherence between the postulates:

(C3) $0 \leq \delta(X) \leq \gamma$ where $\gamma \in [0, \infty]$.

(E3) $\delta(X) = \gamma$ if there exists an index i and a real valued, measurable function g such that $X_i = g(X^{(i)})$ with probability one.

In axiom (C3), γ is allowed to be infinite, this relaxation was introduced to account for the fact that the continuous versions of some discrete measures might take infinite values. Note that axioms (C) and (C3) implicitly assume that no distinction is made between positive and negative dependence since all values are restricted to the positive axis. Further adjustments to Réniy's postulates have been made in Schweizer and Wolff (1981) where it is argued that some of Rényi's conditions are too strong. Axiom (G2) is again adopted in place of (G) and weak convergence condition is added, but maybe the most interesting deviations from the original set of postulates concern axioms (E) and (F). A multivariate version of these modified axioms is given by

(E4) $\delta(X) = 1$ if and only if for some j there exist strictly monotone functions g_i such that $X_i = g_i(X_j), j \neq i$, with probability one.

(F4) For every strictly increasing transformation $T(X) = (T_1(X_1), \ldots, T_d(X_d))$, δ remains invariant $\delta(T(X)) = \delta(X)$.

Postulate (E4) is similar to (E2b) with the important difference that the function g_i must be strictly increasing. Strictly increasing functions are again introduced in (F4) in place of the injective functions in the original formulation. Since every strictly increasing function is also injective, this new axiom constitutes a less strong condition on T. Whereas the original set of axioms considers dependence as a type of association along some measurable function, the postulates introduced in Schweizer and Wolff (1981) consider only association along measurable strictly monotone functions. Measures fullfiling these axioms are therefore also called *measures of monotone dependence*. A link can at this point be drawn to copulas: as explained in Chapter 3, copulas are invariant to strictly increasing transformations.

2.4.3 Measures of dependence

We introduce classical measures of dependence which we broadly classify in three groups: measures based on linear correlation, concordance measures and distance measures. Some measures are included for completeness are will not be discussed in details.

Linear correlation based. Already mentioned in example 2.1, the linear correlation coefficient was amongst the first dependence measures to be introduced and remains one of the most widely used.

Definition 2.9 (Linear correlation)**.** *The linear correlation coefficient between two univariate rv is defined as*

$$\rho_l(X_1, X_2) = \frac{\operatorname{cov}(X_1, X_2)}{\sqrt{\operatorname{var}(X_1)\operatorname{var}(X_2)}} = \frac{\mathrm{E}\left((X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2)\right)}{\sqrt{\mathrm{E}\left((X_1 - \mathrm{E}X_1)^2\right)\mathrm{E}\left((X_1 - \mathrm{E}X_1)^2\right)}}.$$

Definition 2.10 (Maximal correlation)**.** *The maximal correlation is defined for bivariate variables by*

$$S(X_1, X_2) = \sup_{f,g} \rho_l(f(X_1), g(X_2)), \tag{2.58}$$

where the supremum is taken over all measurable functions for which ρ_l is defined.

If we additionally require that f, g in Definition 2.10 fulfill: $\mathrm{E}[f(X_1)] = \mathrm{E}[g(X_2)] = 0$ and $\mathrm{E}[f(X_1)^2] = \mathrm{E}[g(X_2)^2] = 1$, we obtain Gebelein's maximal correlation.

Concordance measures. Concordance in a multivariate random variable can be intuitively understood as a type of "agreement" between the different dimensions i.e. we expect them to be small together or large together. Interestingly, the first attempts to formalise this concept (Consonni and Scarsini, 1982) restricted the definition to random variables in the same Fréchet class, meaning variable having same margins, before it was generalised to any continuous variables using copula in Scarsini (1984). As for the general case of dependence measures, concordance measure have first been studied for bivariate variables. Consider two pairs of continuous bivariate variables $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$ with cdfs F_x and F_y, the pair X is *more concordant* than Y if

$$C_x(u, v) \geq C_y(u, v), \quad \forall u, v \in [0, 1], \tag{2.59}$$

where C_x and C_y are the copulas of X and Y, respectively. Concordance as introduced in equation (2.59) defines a partial stochastic ordering and Scarsini (1984) considers the task of measuring concordance as the construction of a total ordering compatible with (2.59). We present concordance measures as a particular type of dependence measures, however, measures of concordance do not in general satisfy all postulates mentioned in the previous section. Concordance considers dependence in a particular way: dependence is seen as monotone association and takes into account the type of monotonicity (increasing or decreasing). Concordance measures are therefore signed measures, traditionally taking values in $[-1, 1]$. A set of postulate analogue to Rényi's axioms but specially tailored for concordance was proposed in Scarsini (1984). We do not detail them here but it is interesting to note that invariance to strictly increasing transformations (which is not included in the set of axioms) follows as a natural consequence. Concordance as defined in (2.59) is well-defined for every continuous variable but cannot be used in the presence of discrete margins since the unicity of the copula does not hold anymore. This later case is also treated in Scarsini (1984) where a more involved definition of concordance is introduced for discrete variables. The work on concordance mentioned above was limited to the bivariate case and extensions were proposed in Taylor (2007) and Dolati and Úbeda Flores (2006). Even though the definition of concordance we give in equation (2.59) is based on copulas, the most well-known concordance measures, in particular Spearman's rho Spearman (1904) and Kendall's tau Kendall (1938), were originally defined independently and without resorting to copula. It appeared only later that these measures could effectively be reformulated using copulas.

Definition 2.11 (Spearman's rho)**.** *For two univariate rv X_1, X_2 with marginal distribution functions F_1, F_2 Spearman's rho is given by:*

$$\rho_S(X_1, X_2) = \rho_l(F_1(X_1), F_2(X_2))$$

Definition 2.12 (Kendall's tau)**.** *Consider two univariate rv X_1, X_2 and independent copies $\tilde{X}_1, \tilde{X}_2$. Kendall's rank correlation is:*

$$\tau(X_1, X_2) = \mathbb{P}\left((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right) - \mathbb{P}\left((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0\right).$$

Definition 2.13 (Blomqvist's beta)**.** *Denote by $\tilde{x}_1$ and $\tilde{x}_2$ the median of two continuous univariate rv X_1, X_2. Blomqvist beta coefficient is:*

$$\beta(X_1, X_2) = \mathrm{P}\left((X_1 - \tilde{x}_1)(X_2 - \tilde{x}_2) > 0\right) - \mathrm{P}\left((X_1 - \tilde{x}_1)(X_2 - \tilde{x}_2) < 0\right).$$

Distance measures. An intuitive method of measuring dependence is to consider the distance between the variable of interest X and the random variable X_0 having the same margins but independent dimensions. Consider $X = (X_1, \ldots, X_d)$ distributed according to μ with marginal distributions $\mu_1, \ldots, \mu_d$, the variable X_0 is then distributed according to the product measure $\mu_0 = \mu_1 \otimes \cdots \otimes \mu_d$. The crucial point is then the choice of a suitable metric for the space of measures. A classical choice of distance between two measures μ and ν has the form

$$d(\mu, \nu) = \sup_{g \in \mathcal{D}} \left| \int g \mathrm{d}\mu - \int g \mathrm{d}\nu \right|, \tag{2.60}$$

where $\mathcal{D}$ is a set of measurable functions fulfilling particular properties. Every choice for $\mathcal{D}$ then leads to a new distance, e.g. choosing the set of Lipschitz functions with Lipschitz constant 1 we obtain the *Wasserstein distance*. If proper distance measures can be used, we will see in Section 2.5.2 that one of the most prominent measure of dependence, the *Kullback-Leibler* divergence, does not satisfy all properties of a distance but has an interesting interpretation in terms of Information Theory.

2.5 Information Theory

2.5.1 Introduction and entropy

Information theory is centered around the questions of describing and transmitting information. Describing information involves coding, compressing with potential loss; transmitting information involves communication channels and interpretation of the received signal. C.E. Shannon described the fundamental problem of information transmission, or communication, as the task of "reproducing at one point either exactly or approximately a message selected at another point". Important aspects of that description are the facts that the reproduction might be not be exact, meaning that loss of information occurred, and that the message is selected from a set of possible messages, implying that a good communication system should consider all potential messages rather than a particular instance. One aspect which is however not apparent in that description is that the source of information is assumed to be stochastic. Information theory first elaborates the theory of information transmission, thereby describing its fundamental properties and the achievable performances. It then is also concerned with providing coding schemes which would ideally approach the best theoretical performance. Information theory mostly originated from work conducted at the AT&T Bell Laboratories with first papers in the 1920s (Nyquist, 1928), (Hartley, 1928) and the seminal paper by C.E. Shannon (Shannon, 1948) where the basis of the field were set. Far from being limited to the electrical engineering perspective on communication, information theory has strong connexions and applications to many other fields like Physics, Economics, Computer Science and Statistics. We cover in this section only the information-theoretical concepts required for subsequent chapters, which constitute a very little part of the field and suggest Cover and Thomas (1991) for a broad introduction or Gray (1990) for an in-depth coverage.

A fundamental information theoretical concept is the notion of *entropy*. Entropy measures the degree of uncertainty inherent to a stochastic distribution, i.e. how uncertain the outcome from that distribution is. Entropy was first defined for distributions on a finite set, we thus naturally start with the definition for discrete variables.

Definition 2.14 (Discrete Entropy)**.** *Consider a discrete random variable $X : \Omega \to E$ with probability mass function $f : E \to [0, 1]$. The* entropy *of X is:*

$$H(X) = -\sum_{x \in E} f(x) \log f(x) = -\mathrm{E}_f[\log f(X)], \tag{2.61}$$

where we use the convention $0 \log(0) = 0$.

Note that in the above definition the state space E can be a product space, e.g. a subset of $\mathbb{R}^d$, and the case of multivariate random variable is thereby covered. When entropy is defined using the natural logarithm as in (2.61), it is said to be measured in *nats*. Another standard usage defines H with the base 2 logarithm, and it is then measured in *bits*. Both versions slightly differ in interpretation but remain the same in essence. Definition 2.14 might at first seem obscure but a look at some properties of H reveals why it is a sensible measure of uncertainty.

1. H is a continuous function of $f(x)$.
2. If all outcomes $x \in E$ are equally likely, i.e. $f(x) = cst$, then H is an increasing function of $\text{card}(E)$.
3. $H(X) = 0$ if and only if $f(x^*) = 1$ for some $x^* \in E$ and $f(x) = 0$ for $x \in E \setminus \{x^*\}$, i.e. the entropy is null only when the outcome is certain.
4. For fixed $n = \text{card}(E)$, the maximum value of H is attained when $f(x) = \frac{1}{n}, \forall x$ and is $H(X) = \log(n)$, i.e. the entropy is maximal when all events are equiprobable.

The first two properties above were part of a set of three desirable characteristics Shannon required for his definition of entropy (we do not detail here the more technical third characteristic which prescribes the behaviour of H in the case of successive sampling steps). These three requirements are actually sufficient to uniquely determine the functional form of H up to a multiplicative constant, thereby also providing a further justification for its particular form. The entropy does not depend on the particular values taken by X but uniquely on their probabilities, it as such is invariant to permutations in E and can be reformulated in terms of partitions, see Gray (1990).

Entropy for continuous variables can be defined similarly. We first rewrite the discrete entropy (2.61) as

$$H(X) = -\int_E f(x) \log f(x) \mathrm{d}m(x), \tag{2.62}$$

where m is the counting measure. This new expression for the entropy leads to the following extension for continuous variables.

Definition 2.15 (Differential entropy)**.** *Consider a continuous random variable $X : \Omega \to E$ with distribution μ and density f* w.r.t. *the Lebesgue measure λ. The entropy of X is*

$$H(X) = -\int_E f(x) \log f(x) \mathrm{d}\lambda(x) = -\int_E \log(\frac{d\mu}{d\lambda}) \mathrm{d}\mu = -\mathrm{E}_\mu \left[\log(\frac{d\mu}{d\lambda}) \right]. \tag{2.63}$$

Even if both definitions are very similar, differential entropy is not considered as a strict generalisation of discrete entropy because some interesting properties of the discrete case do not carry over to the continuous extension. Whereas $0 \leq H(X) \leq \log(\text{card}(E))$ for a discrete X, the differential entropy can be negative or infinite and loses its interpretation as *uncertainty measure*. To gain a little more insight on differential entropy, example 2.4 considers the case of a uniform random variable.

Example 2.4 (Entropy of the uniform distribution)**.** *For $X \sim \text{Unif}[a, b]$ we have:*

$$H(X) = -\int_{[a,b]} \frac{1}{b-a} \log\left(\frac{1}{b-a}\right) \mathrm{d}x = \log(b-a). \tag{2.64}$$

As mentioned above, differential entropy is not a perfect generalisation of the discrete case. The following example illustrates the relationship between discrete and differential entropy. Consider a real random variable X with density f. X can be approximated by a serie of discrete variables $X_n, n \in \mathbb{N}$. We first partition the real line in a collection of intervals of length $\frac{1}{2^n}$:

$$\Delta_{n,k} = \Big[\frac{k}{2^n}; \frac{k+1}{2^n}\Big). \tag{2.65}$$

X_n is then defined as

$$X_n = \frac{k}{2^n}, \quad \text{where} \quad k \in \mathbb{Z} \quad \text{is such that} \quad X \in \Delta_{n,k}. \tag{2.66}$$

We can compute the entropy of X_n:

$$H(X) = -\sum_{i\in\mathbb{Z}} \mathbb{P}(X=i)\log\mathbb{P}(X=i) = -\sum_{k\in\mathbb{Z}} \mathbb{P}(X_n\in\Delta_{n,k})\log\mathbb{P}(X_n\in\Delta_{n,k}). \tag{2.67}$$

We have that $\mathbb{P}(X_n \in \Delta_{n,k}) \xrightarrow{\text{n}\to\infty} 0$ and $H(X_n) \xrightarrow{\text{n}\to\infty} \infty$, meaning that the entropy of the discrete approximations X_n does not converge to the differential entropy of X. Under further assumptions on the density f we can show, see Ihara (1993), that

$$\lim_{n\to\infty} H(X_n) - n\log(2) = H(X). \tag{2.68}$$

Differential entropy is not an absolute but a relative quantity, (2.68) shows that we need to take the partition size into account to find $H(X)$. Another illustration of that fact is that H is not invariant to changes of coordinates $X' = g(X)$. Another measure turns out to be better suited to continuous distributions, this measure, called *relative entropy*, do not evaluate the uncertainty of a distribution but provides a similarity measure between the distribution of interest and a reference distribution. Relative entropy actually is the opposite of the *Kullback-Leibler divergence* which was introduced in Kullback and Leibler (1951) as a "distance" between probability measures. Before introducing the Kullback-Leibler divergence we define the *conditional entropy* which will also provide a connection to the (discrete or differential) entropy.

Definition 2.16 (Conditional entropy)**.** *Let X, Y be random variables taking values in, respectively, E_x and E_y. We assume that the variables are both discrete or both continuous, and f will denote the joint probability mass function or density according to the case. The Y margin of f will be denoted by f_y. The conditional entropy of X given Y is*

$$H(X|Y) = -\int_{E_x\times E_y} f(x,y)\log\frac{f(x,y)}{f_y(y)}\mathrm{d}\nu(x,y), \tag{2.69}$$

where ν is the counting measure for discrete random variables and the Lebesgue measure for continuous random variables.

Expression (2.69) can be interpreted as the expectation w.r.t. to the joint distribution of the log conditional $f_{x|y}$. We conclude this section on entropy by the following list of properties valid for discrete and differential entropy.

1. $H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$,
2. $H(X|Y) \leq H(X)$, with equality if and only if X is independent of Y,
3. $H(X,Y) \leq H(X) + H(Y)$, with equality if and only if X is independent of Y.

The first property can be interpreted as the result of a two step process: first fix Y (or X), then choose $X|Y$ (or $Y|X$), the total entropy finally is the sum of the entropies in each step. The second property expresses that knowledge reduces entropy unless that knowledge is irrelevant. The third property, which is obtained by combining the two preceding ones, shows that potential dependency between X and Y reduces the entropy of their joint distribution.

2.5.2 Mutual Information

In this section we introduce the concept of *mutual information* between two random variables, also called *negative relative entropy*, which turns out to be a special case of *Kullback-Leibler divergence*

(KL divergence). The Kullback-Leibler divergence is itself a special member of the family of *f-divergences* and, as its name indicates, provides a measure of the dissimilarity between probability measures.

We consider the task of discriminating between two probability measures. In Kullback and Leibler (1951), the "distance" between two measures is described using a statistical approach: how difficult would it be to discriminate between between them based on observations ? Consider two probability measures μ_1, μ_2 on $(E, \mathscr{E})$ such that μ_1, μ_2 are absolutely continuous w.r.t. to each other, i.e. $\mu_1 \ll \mu_2$ and $\mu_2 \ll \mu_1$. Assume that there exists a measure λ such that $\mu_1 \ll \lambda$, $\mu_2 \ll \lambda$ and assume further that there exist functions $f_i, i = 1, 2$ fulfilling $\mu_i(A) = \int_A f_i(x)\mathrm{d}\lambda(x)$, $A \in \mathscr{F}$. If λ is the Lebesgue measure then $f_i, i = 1, 2$ are the traditional density functions, and if λ is the counting measure then $f_i, i = 1, 2$ are probability mass functions. Denote by H_i the hypothesis that an observation x originates from the population with measure μ_i. The information contained in x to discriminate between H_1 and H_2 is defined as (Kullback and Leibler, 1951):

$$\log\left(\frac{f_1(x)}{f_2(x)}\right).$$

This pointwise definition leads to the KL divergence which is defined as the mean information to discriminate between H_1 and H_2 per observation from μ_1.

Definition 2.17 (Kullback-Leibler divergence)**.** *The Kullback-Leibler divergence between the two measures μ_1, μ_2 is:*

$$D_{kl}(\mu_1 \parallel \mu_2) = \int \log\frac{f_1(x)}{f_2(x)}\mathrm{d}\mu_1(x) = \int \log\frac{f_1(x)}{f_2(x)} f_1(x)\mathrm{d}\lambda(x). \tag{2.70}$$

We define the KL divergence between f_1 and f_2 similarly. For rv X_1, X_2 we simply use the corresponding distributions: $D_{kl}(X_1 \parallel X_2) = D_{kl}(\mu_1 \parallel \mu_2)$.

Definition 2.17 assumes that μ_1 and μ_2 are absolutely continuous w.r.t. each other. If μ_1 and μ_2 are singular ($\mu_1 \perp \mu_2$) meaning that $\exists A \in \mathscr{E}$ with $\mu_i(A) = 0$ and $\mu_j(A) > 0, j \neq i$, we can then perfectly discriminate between the measures and we extend Definition 2.17 by setting $D_{kl}(\mu_1 \parallel \mu_2) = \infty$. Definition 2.17 applies to continuous and discrete variables: if λ is the Lebesgue measure then f_1, f_2 are the density functions, and if λ is the counting measure then f_1, f_2 are probability mass functions. The KL divergence is not a proper distance between distributions since it is not symmetric and does not satisfy the triangle inequality, it however has interesting properties. In particular, the following proposition justifies its use as a divergence measure.

Proposition 2.2 (Properties of KL divergence)**.** *Let $\nu, \mu_1, \ldots, \mu_n$ be measures satisfying the requirements of Definition 2.17*

1. *$D_{kl}(\mu_1 \parallel \mu_2) \geq 0$ with equality if and only if $f_1 = f_2$ up to λ-null sets.*
2. *If $\lim_{n\to\infty} D_{kl}(\nu \parallel \mu_n) \to 0$ then $\lim_{n\to\infty} ||\nu - \mu_n||_{TV} = 0$, where*

$$||\nu - \mu_n||_{TV} = \sup_{\mathcal{P}(E)} \sum_i |\nu(A_i) - \mu_n(A_i)|,$$

denotes the total variation, the supremum being taken over the finite partitions of E.

We can recognise in the first property given in Proposition 2.2 the first of the three axioms needed to define a *distance*, and the second property means that convergence in KL divergence implies convergence in total variation.

As mentioned in Section 2.4.3, a dependence measure can be obtained by measuring the "distance" or divergence between the distribution of interest and the product measure having the same margins. In the case of two univariate variables, using the KL divergence leads to the definition of *mutual information*.

Definition 2.18 (Mutual information, univariate case)**.** *Consider the random variable* $Z = (X, Y)$*. The mutual information between* X *and* Y *is defined as*

$$I(X;Y) = D_{kl}(Z \parallel Z_0), \tag{2.71}$$

where Z_0 *is bivariate with independent margins* X *and* Y*.*

Mutual information is also an information theoretical quantity and can be interpreted as a reduction in entropy. The following properties can be directly obtained from the definitions:

1. $I(X;Y) = H(X) + H(Y) - H(XL, Y)$.
2. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

Mutual information is, like similarity measures, inherently a bivariate concept: it provides a measure of the information *between* two random variables X and Y. However, X and Y can perfectly be multivariate, in which case we simply need to be careful in the choice of the reference variable Z_0. Denote by μ_X, respectively, μ_Y the distributions of X and Y, and by μ the joint distribution of (X, Y).

Definition 2.19 (Mutual information)**.** *The mutual information between* X *and* Y *(or equivalently between* μ_X *and* μ_Y*) is defined as*

$$I(X;Y) = D_{kl}(\mu \parallel \mu_X \otimes \mu_Y), \tag{2.72}$$

where $\mu_X \otimes \mu_Y$ *is the product measure of* μ_X *and* μ_Y*.*

Another closely related quantity is the *multi-information* of one multivariate random variable which is a measure of the dependency contained in its joint distribution.

Definition 2.20 (Multi-information)**.** *The multi-information of a joint distribution* μ *with margins* $\mu_1, \ldots, \mu_d$ *is defined as*

$$I(\mu) = D_{kl}(\mu \parallel \mu_0), \tag{2.73}$$

where μ_0 *is the product measure* $\mu_1 \otimes \cdots \otimes \mu_d$*. The multi-information of a rv* X *with joint distribution* μ *is defined as* $I(X) = I(\mu)$*.*

The following proposition clarifies the relation between mutual and multi-information.

Proposition 2.3 (Mutual and multi-information)**.** *Consider* X *and* Y *as in Definition 2.19. The mutual information between* X *and* Y *can be rewritten as*

$$I(X;Y) = I(X,Y) - I(X) - I(Y),$$

where $I(X, Y)$ *denotes the multi-information of the vector* (X, Y)*.*

Proof. We denote by f_{XY} the joint density function (or probability mass function in the case of discrete variables), f_X and f_Y denote the density of, respectively, X and Y, f_0 denotes the corresponding density products.

$$\begin{aligned}
I(X,Y) - I(X) - I(Y) &= \mathrm{E}_\mu \log\left(\frac{f_{xy}(X,Y)}{f_0(X,Y)}\right) - \mathrm{E}_{\mu_X} \log\left(\frac{f_x(X)}{f_0(X)}\right) - \mathrm{E}_{\mu_Y} \log\left(\frac{f_y(Y)}{f_0(Y)}\right) \\
&= \mathrm{E}_\mu \log\left(\frac{f_{xy}(X,Y)}{f_0(X,Y)}\right) - \mathrm{E}_\mu \log\left(\frac{f_x(X)}{f_0(X)}\right) - \mathrm{E}_\mu \log\left(\frac{f_y(Y)}{f_0(Y)}\right) \\
&= \mathrm{E}_\mu \log\left(\frac{f_{xy}(X,Y)}{f_0(X,Y)} \frac{f_0(X)}{f_x(X)} \frac{f_0(Y)}{f_y(Y)}\right) \\
&= \mathrm{E}_\mu \log\left(\frac{f_{xy}(X,Y)}{f_x(X) f_y(Y)}\right) = I(X;Y).
\end{aligned}$$

□

Chapter 3

Copulas

3.1 Introduction

A historical perspective. Copulas have received a lot of attention in recent years. The rapid development of research on copulas, starting in the late nineties, was mainly driven by applications in the domain of finance where non-Gaussian multivariate models are sought e.g. for log-returns. An informative study of the advent of copulas in this area is given in Genest et al. (2009). For a comprehensive introduction to copula methods in finance see Cherubini et al. (2004), a broader introduction to quantitative methods in the field can be found in McNeil et al. (2005). The drastic increase of copula models' popularity has also been a source of criticism (Mikosch, 2006), but they remain the tool of choice for many applications. Another interesting opinion on copulas' wide usage was given in Embrechts (2009).

The story of copulas, however, started long before these recent developments. Already in 1951, M.R. Fréchet considered the problem of finding the set of 2-dimensional distributions compatible with given univariate margins (Fréchet, 1951), this set is now called the *Fréchet class*. Work on the same subject was also conducted by R. Féron in Féron (1956) which studies the three-dimensional case. The term *copula* was first introduced by A. Sklar (Sklar, 1959) who formulated a theorem fundamental for the theory of copulas, today known as *Sklar's theorem*. The genealogy of his work on probability theory and his research collaborations are related in Sklar (2009).

A logical perspective. In Sklar (2009), A. Sklar relates his first serious engagement with probability theory, which happened in the context of number theory, and how, when starting to work under Karl Menger's direction, he decided to revisit probability theory in depth "from the ground up". He explains how his re-development of probability ran into difficulties when it came to multivariate random variables. Indeed, if variables with independent dimensions are easily constructed in the product space, introducing dependency poses a much more complex problem. It is striking that, whereas independence admits a concise definition, a lot of work has been conducted on defining dependence measures. We will come back to the link between copula and dependence measures in Section 3.4 but first elaborate on the task of constructing multivariate random variables. Assume that the problem of constructing various univariate rv is satisfyingly solved and that we have at hand a sufficient variety of univariate distributions represented by their cumulative distribution functions $F_1, \ldots, F_d$. How can a multivariate rv with fixed known margins be constructed ? Are all types of dependence possible or do we encounter restrictions ? As shown in Fréchet (1951) and Hoeffding (1940), restrictions already appear in the 2-dimensional case. Denote by $\Pi(F_1, \ldots, F_d)$ the class of all d-dimensional rv having the prescribed margins,

also called the Fréchet class. In the bivariate case, Π admits a lower and an upper bound, both having a closed form as a function of F_1, F_2. More precisely, there exist cdfs F_m and F_M such that for all $F \in \Pi$ we have

$$F_m(x_1, x_2) \leq F(x_1, x_2) \leq F_M(x_1, x_2), \forall x_1, x_2 \in \mathbb{R}, \tag{3.1}$$

moreover the bounds are given by

$$F_m(x_1, x_2) := \max(0, F_1(x_1) + F_2(x_2) - 1) \quad \text{and} \quad F_M(x_1, x_2) := \min(F_1(x_1), F_2(x_2)). \tag{3.2}$$

Both inequalities in (3.1) remain valid for the d-dimensional case with the difference that F_m is not guaranteed to be a cdf anymore, and thus in general $F_m \notin \Pi$. Once the bounds are known, the next natural step is to try to characterise Π more precisely using a parameterized form for every element in the set. Convex combinations of F_m and F_M do not cover the entire set Π since the independent distribution $F_1(x_1)F_2(x_2)$ does not admit this form, and more complex representations are sought. A complete characterisation of the d-dimensional distributions in Π can be obtained using the so-called *probability transformation*, which transforms a univariate rv to a uniformly distributed one, and *copulas*, which are marginally uniform multivariate distributions. The following proposition recalls the useful *quantile and probability transformations*.

Proposition 3.1. (Quantile and Probability transformations) *Consider a univariate cdf F with generalised inverse $F^{\leftarrow}$, $F^{\leftarrow}(y) = \inf\{x \in \mathbb{R} | F(x) \geq y\}$. The following relations hold:*

1. *If $U \sim \text{Unif}(0, 1)$ then $F^{\leftarrow}(U) \sim F$.*
2. *If $X \sim F$ and F is continuous then $F(X) \sim \text{Unif}(0, 1)$.*

Proof. Both transformations are obtained using basic properties of cumulative distribution functions as follows.

1. $\forall x \in \mathbb{R}$, $\mathbb{P}(F^{\leftarrow}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$, where in the first equality we used that F is right-continuous (holding for every distribution function).
2. $\forall u \in [0, 1]$, $\mathbb{P}(F(X) \leq u) = \mathbb{P}(F^{\leftarrow} \circ F(X) \leq F^{\leftarrow}(u)) = \mathbb{P}(X \leq F^{\leftarrow}(u)) = F \circ F^{\leftarrow}(u) = u$. The first equality holds because $F^{\leftarrow}$ is strictly increasing for a continuous F. The second equality holds for every cdf F and the last equality is true for continuous F.

□

Using Proposition 3.1, any multivariate rv with given margins can be constructed in a two stages process. First, the margins are (independently from each other) mapped to uniform random variables using the probability transformation 3.1 (2). Second, the newly obtained uniform variables are combined to form a multivariate rv using a multivariate cdf with uniform margins, this cdf is then the copula of the constructed rv.

We give bellow two equivalent definitions of copulas, starting with the most concise version, and present the basic results in the field.

Definition 3.1 (Copula)**.** *A d-dimensional copula is a cumulative distribution function $C : [0, 1]^d \to [0, 1]$ with standard uniform marginal distributions i.e. $C_j \sim \text{Unif}(0, 1), \forall j$.*

An equivalent but self-contained definition is:

Definition 3.2 (Copula)**.** *A d-dimensional copula is a function $C : [0, 1]^d \to [0, 1]$ such that:*

1. *$C(u_1, \ldots, u_d)$ is an increasing function in each component u_i.*
2. *$C(u_1, \ldots, u_d) = u_i$ if $u_j = 1, \forall j \neq i$ and $u_i \in [0, 1]$.*
3. *For all $(a_1, \ldots, a_d), (b_1, \ldots, b_d) \in [0, 1]^d$ with $a_i \leq b_i, \forall i$ we have*

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots+i_d} C(u_{1,i_1}, \ldots, u_{d,i_d}) \geq 0,$$

where $u_{j,1} = a_j$ and $u_{j,2} = b_j$.

The first condition must hold for any cdf. The second ensures that the margins are uniform and the third that the mass of any d-dimensional rectangle $\times_{i=1}^{d}[a_i, b_i]$ is non-negative.

The connection between copulas and multivariate random variables is formalised in *Sklar's theorem* (Sklar, 1959) which states that any d-dimensional rv adopts a copula representation, and, conversely, that any construction using copulas leads to a well-defined rv. The theorem provides an equivalence relationship between copulas and multivariate rv with given margins. As we will see later, it indeed is an equivalence (in the mathematical sense of the term) for continuous variables, however the introduction of discrete margins breaks the perfect correspondence.

Theorem 3.1. (Sklar)**.**

1. *Let F be a joint cumulative distribution function with margins $F_1, \ldots, F_d$. Then there exists a copula $C : [0, 1]^d \to [0, 1]$ such that*

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)), \forall x_j \in [-\infty, +\infty]. \tag{3.3}$$

   Moreover, if the margins are continuous, then that copula is unique. Otherwise, the copula is uniquely determined on $F_1(\bar{\mathbb{R}}) \times \cdots \times F_d(\bar{\mathbb{R}})$.

2. *Conversely, if C is a copula and $F_1, \ldots, F_d$ are univariate cdfs, then F defined as in* (3.3) *is a multivariate cdf with margins $F_1, \ldots, F_d$ and copula C.*

Proof. For simplicity we give here a proof for the continuous case only. A proof of the general case was given in Schweizer and Sklar (1983).

1. (a) Existence of C:
   Consider a multivariate random variable X with distribution function F and continuous margins $F_1, \ldots, F_d$. For any cdf F the following holds:

$$F(x_1, \ldots, x_d) = \mathbb{P}(F_1(X_1) \leq F_1(x_1), \ldots, F_d(X_d) \leq F_d(x_d)).$$

   Using the probability transformation 3.1 (2), we have that $F_1(X_1), \ldots, F_d(X_d)$ are uniformly distributed and thus the multivariate distribution of $(F_1(X_1), \ldots, F_d(X_d))$ is a copula, which we denote by C, and we finally obtain:

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

   (b) Unicity :
   If we evaluate expression (3.3) at $x_i = F_i^{\leftarrow}(u_i)$, for $0 \leq u_i \leq 1, i = 1, \ldots, d$, we obtain:

$$F(F_1^{\leftarrow}(u_1), \ldots, F_d^{\leftarrow}(u_d)) = F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) = C(u_1, \ldots, u_d),$$

   which shows unicity. In the last equality we used that $F_i \circ F_i^{\leftarrow}(u_i) = u_i$ since F_i is continuous.

2. Assume that C is a copula and that $F_1, \dots, F_d$ are univariate cdfs. Consider a random vector $U = (U_1, \dots, U_d)$ having cumulative distribution function C. Since C is a copula it is clear that $U_i \sim \text{Unif}(0,1), i = 1, \dots, d$. Define $X = (F_1^{\leftarrow}(U_1), \dots, F_d^{\leftarrow}(U_d))$. We can then compute the distribution of X:

$$\begin{aligned}
\mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d) &= \mathbb{P}(F_1^{\leftarrow}(U_1) \leq x_1, \dots, F_d^{\leftarrow}(U_d) \leq x_d) \\
&= \mathbb{P}(U_1 \leq F_1(x_1), \dots, U_d \leq F_d(x_d)), \text{since } F_i \text{ is right-continuous,} \\
&= C(F_1(x_1), \dots, F_d(x_d)), \text{ by definition of } U. \qquad (3.4)
\end{aligned}$$

Equation (3.4) shows that X has copula C. From Proposition 3.1 we know that $X_i \sim F_i, \forall i$, which concludes the proof.

□

The unicity of the copula for continuous distributions showed in Theorem 3.1 permits the following definition and characterisation.

Definition 3.3 (Copula of F, copula of X)**.** *If F is a continuous cdf, then the unique copula C given by Theorem 3.1 is called the copula of F. For a rv X with continuous cdf F we define the copula of X analogously.*

Corollary 3.1. *The copula of a continuous cdf F with margins $F_1, \dots, F_d$ can be written as*

$$C(u_1, \dots, u_d) = F(F_1^{\leftarrow}(u_1), \dots, F_d^{\leftarrow}(u_d)).$$

Sklar's theorem ensures that combining univariate cdfs using a copula always leads to a valid distribution with the required margins. Moreover, at least for the continuous case where copula unicity is ensured, it provides a theoretical justification for the original idea of using the probability transformation to construct multivariate rvs by asserting that any multivariate distribution indeed adopts such a representation.

The copula construction of multivariate rvs can also be represented as a latent variables model where the quantile transformation is applied on a set of marginally uniform hidden variables. The following generative model defines a rv X with margins $F_1, \dots, F_d$ which still has copula C if the margins are continuous.

Definition 3.4 (Copula latent variables model)**.** *For a d-dimensional copula C and univariate cdf $F_1, \dots, F_d$ we define:*

$$(U_1, \dots, U_d) \sim C \qquad (3.5)$$

$$X_i = F_i^{-1}(U_i), \ i = 1, \dots, d. \qquad (3.6)$$

The random vector resulting from Definition (3.4) has margins $F_1, \dots, F_d$ for any choice of the marginal cdfs, in particular, this construction remains valid when discrete margins are used. However, in the presence of discrete margins, several choices of copulas could lead to the same distribution for X. We will come back to issues arising with discrete margins in Section 3.6.

3.2 Standard Copulas

Sklar's theorem makes explicit the relationship between a multivariate distribution and its margins. It becomes clear that arbitrary cdfs $F_1, \dots, F_d$ can be combined to form a multivariate distribution, and that copulas can be used to create new multivariate models with the desired

margins and dependence properties. On the other hand copulas can be extracted from existing multivariate distributions to later be used for combining other margins. Such copulas obtained from multivariate distributions are called *implicit copulas*. Implicit copulas extracted from a continuous distribution have the form given in Corollary 3.1. Unfortunately but not surprisingly, the variety of interesting implicit copulas is limited, especially in higher dimensions. The most prominent examples of implicit copula are found in the class of *elliptical copulas* which includes the Gaussian and the Student-t copulas. Elliptical copulas are copulas extracted from *elliptical distributions* which we define below.

Definition 3.5 (Elliptical distribution)**.** *A d-dimensional* rv X *has an elliptical distribution with parameters μ, Σ and ϕ, written $X \sim E_d(\mu, \Sigma, \phi)$, if the characteristic function of $X - \mu$ is a function ϕ of the quadratic term $t^T \Sigma t$ only:*

$$\varphi_{X-\mu}(t) = \phi(t^T \Sigma t), \tag{3.7}$$

where $\mu \in \mathbb{R}^d$, Σ is a $d \times d$ symmetric positive definite matrix and $\phi : \mathbb{R} \to \mathbb{R}$.

We briefly recall that the characteristic function φ of a d-dimensional rv X is defined as

$$\varphi_X : \mathbb{R}^d \to \mathbb{R}, \quad \varphi_X(t) = \mathrm{E}\left[\exp(it^T X)\right],$$

where i is the imaginary unit. The function ϕ is must fulfill certain requirements we do not detail here. We list below a few properties of elliptical distributions, for a complete introduction see Cambanis et al. (1981).

1. In one dimension, elliptical distributions coincide with symmetric distributions.
2. Elliptical distributions are radially symmetric.
3. For a given elliptical rv X the representation (3.7) is not unique. To obtain a unique formulation we can additionally require that ϕ and Σ satisfy $\mathrm{cov}(X) = \Sigma$.
4. If a density exists it has the form $|\Sigma|^{-\frac{1}{2}} g\left((X-\mu)^T \Sigma^{-1} (X-\mu)\right)$ for some non-negative real-valued function g.
5. The marginal distributions also are elliptical with the same function ϕ.

The most famous member of the elliptical family is the d-dimensional Gaussian distribution $\mathcal{N}_d(\mu, \Sigma)$ which can be obtained by setting $\phi(u) = \exp(-u/2)$ and has the following density function:

$$f(x; \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right). \tag{3.8}$$

A Gaussian variable having zero mean and identity covariance matrix is called *standard normal* or *standard Gaussian* distributed. Another interesting example of elliptical distribution is the *Student-t* distribution.

Example 3.1 (Student-t distribution)**.** *The d-dimensional Student-t density with parameters μ, Σ and degrees of freedom ν is*

$$f(x; \Sigma, \mu, \nu) = c(d, \nu) |\Sigma|^{-\frac{1}{2}} \left(1 + \frac{1}{\nu}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)^{-\frac{\nu+d}{2}}, \tag{3.9}$$

where the constant value is

$$c(d, \nu) = (\nu\pi)^{-\frac{d}{2}} \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})}.$$

It has mean μ and covariance matrix $\frac{\nu}{\nu-1}\Sigma$.

The *Student-t copula* and the *Gaussian copula* are then naturally the copulas extracted from the corresponding distributions and can be represented in the form given by Corollary 3.1. Since the Gaussian copula plays a prominent role in the subsequent chapters we will discuss it in more details in Section 3.5.

Definition 3.6 (Gaussian copula)**.** *The copula of a multivariate Gaussian rv $X \sim \mathcal{N}_d(\mu, \Sigma)$ is called Gaussian copula with parameter matrix P, where P is the correlation matrix of X. We denote a Gaussian copula by C_P^G or simply C_P when the context is unambiguous.*

Knowing the simple form of a Gaussian density it might seem surprising that the Gaussian copula (which is a cumulative distribution function) does not have a simple closed form but can be expressed as multivariate integral. Note that C_P^G depends on the correlations of X but not on the mean or variance which are properties of the margins. This means that the class of all Gaussian variables having correlation matrix P share the same copula C_P^G, regardless of their respective mean and variance. This invariance w.r.t. the mean and variance is a consequence of one important property of copulas, the invariance of copulas to strictly increasing transformations (see Section 3.3). It can easily be seen that standardising the margins can be achieved by applying the successive transformations $\Phi_{\mu,\Sigma}$ and Φ^{-1}, which are both strictly increasing. Similarly, the Student-t copula depends only on the correlation matrix corresponding to Σ and the degrees of freedom ν. We denote the correlation matrix corresponding to a covariance matrix Σ as $\mathcal{P}(\Sigma)$. The Gaussian and Student copulas are amongst the few interesting implicit copulas which can be used with large dimension d. The majority of copulas used in higher dimensions are "constructed" rather than derived since the variety of "natural" (meaning not created using a copula) multivariate cdfs available is limited. We present below some fundamental copulas. More information about the principal types of copula can be found in Joe (1997) or Nelsen (1999).

Independence between dimensions is represented by the independence copula.

Definition 3.7 (Independence copula)**.** *The independence copula is defined by $C^\pi(u_1, \ldots, u_d) = \prod_{j=1}^d u_j$.*

Due to the product structure of C^π, distributions with copula C^π clearly have independent dimensions. In the continuous case, there is an equivalence relationship between independence copula and independent margins: a continuous rv has independent dimensions if and only if it has copula C^π. In the discrete case, the independence copula still implies independent margins but the reverse does not hold, see also Section 3.6. At the other end of the dependency spectrum, perfect positive dependence is obtained with the *comonotonicity copula*.

Definition 3.8 (Comonotonicity copula)**.** *The comonotonicity copula is defined by $C^M(u_1, \ldots, u_d) = \min\{u_1, \ldots, u_d\}$.*

In Definition 3.8 we can recognise the Fréchet upper bound from (3.1). In the axiomatic approach to dependence presented in Section 2.4.2, perfect dependence was defined in axiom (E4) as occurring when all dimensions are strictly increasing functions of one of them. It can be shown (see McNeil et al. (2005)) that a perfect positive dependent rv in the sense of axiom (E4) with continuous margins must have the comonotonicity copula. The copula C^M is also the multivariate cdf of the random variable $(U, \ldots, U)$, where $U \sim \text{Unif}(0, 1)$. As previously mentioned the Fréchet lower bound, representing perfect negative dependence is a copula only in two dimensions. In this case, it is also the distribution of $(U, 1 - U)$, where $U \sim \text{Unif}(0, 1)$.

If significant work has been conducted on the properties of known copulas, the search for new copulas is also a prolific area of research. There is a large diversity in copulas, and far from aiming at a thorough review, we briefly mention a few more names. Three famous examples in the parametric familiy of *Archimedean copula* are *Gumbel*, *Clayton* and *Frank* copulas. Whereas the Frank copula is radially symmetric with no tail dependence, the other two are asymmetric

copulas with tail dependence, the Gumbel copula exhibiting upper tail dependence and the Clayton copula lower tail dependence. Archimedean copulas have a simple parametric form which can be extended to any number of dimensions, however additional conditions are required to ensure that the resulting form indeed is a copula (McNeil and Nešlehová, 2009). Recently, *paired-copula models* were introduced as a general method to obtain high-dimensional copulas by combining bivariate copulas. In particular, *regular vines models* provide very flexible structures in high dimensions for which inference remains feasible (Bedford and Cooke, 2001).

3.3 Further properties of copulas

We present in this section some important properties of copulas. A fundamental characteristic of copulas is their invariance to strictly increasing transformations of continuous margins. This invariance property was already required in the axiomatic approach to dependence (axiom (F4) in Section 2.4.2), and any dependence measure based on the copula will then naturally satisfy this condition. Since strictly increasing functions are order preserving, this invariance also implies that any measure based on the ranks will depend on the copula only.

Proposition 3.2. *Consider a rv $X = (X_1, \dots, X_d)$ with continuous margins and copula C. If $T_1, \dots, T_d : \mathbb{R} \to \mathbb{R}$ are strictly increasing functions then $(T_1(X_1), \dots, T_d(X_d))$ also has copula C.*

Proposition 3.2 already appeared in Schweizer and Wolff (1981). A thoroughly detailed proof of it can be found in Embrechts and Hofert (2013), we give below a proof closer to the formulation found in McNeil et al. (2005), which first requires the following result.

Lemma 3.1. *Consider a real-valued rv X and an increasing function $T : \mathbb{R} \to \mathbb{R}$, then $\{X \leq x\} \subset \{T(X) \leq T(x)\}$ and: $\mathbb{P}(T(X) \leq T(x)) = \mathbb{P}(X \leq x) + \mathbb{P}(T(X) = T(x), X > x)$. A proof can be found in McNeil et al. (2005).*

Proof of Proposition 3.2. First, we show that the distribution of $T_i(X_i), i = 1, \dots, d$, is given by $\tilde{F}_i$ which we define by $\tilde{F}_i(y) = F_i \circ T_i^{\leftarrow}(y)$.

$$\begin{aligned}
\tilde{F}_i(y) &= \mathbb{P}\left(X_i \leq T_i^{\leftarrow}(y)\right), \text{ by definition,} \\
&= \mathbb{P}\left(T_i^{\leftarrow} \circ T_i(X_i) \leq T_i^{\leftarrow}(y)\right), \text{ since } T_i \text{ is strictly increasing,} \\
&= \mathbb{P}\left(T_i(X_i) \leq y\right) + \mathbb{P}\left(X_i = T_i^{\leftarrow}(y), T_i(X_i) > y\right), \text{ by Lemma 3.1,} \\
&= \mathbb{P}\left(T_i(X_i) \leq y\right), \text{ since } F_i \text{ is continuous.}
\end{aligned}$$

For $u_1, \dots, u_d \in [0, 1]$ we can write:

$$\begin{aligned}
C\left(u_1, \dots, u_d\right) &= \mathbb{P}\left(F_1(X_1) \leq u_1, \dots, F_d(X_d) \leq u_d\right), \text{ by definiton of } C, \\
&= \mathbb{P}\left(\tilde{F}_1 \circ T_1(X_1) \leq u_1, \dots, \tilde{F}_d \circ T_d(X_d) \leq u_d\right),
\end{aligned}$$

since $\tilde{F}_i \circ T_i(x) = F_i \circ T_i^{\leftarrow} \circ T_i(x) = F_i(x)$. Since $\tilde{F}_i$ is the distribution of $T_i(X_i)$ the above equality shows that C is also the copula of $(T_1(X_1), \dots, T_d(X_d))$. □

In the presence of discrete margins, we need to impose an additional condition on the functions $T_1, \dots, T_d$ to ensure that $(T_1(X_1), \dots, T_d(X_d))$ still admits C as one of its copulas: these functions must be continuous (Embrechts and Hofert, 2013). Beside being continuous multivariate functions in the interval $[0, 1]$, copulas have the advantage of being Lipschitz with constant 1: for all $u = (u_1, \dots, u_d)$ and $v = (v_1, \dots, v_d)$ we have

$$|C(u) - C(v)| \leq \sum_{j=1}^{d} |u_j - v_j|. \tag{3.10}$$

Copulas are not necessarily differentiable, however when the d-th order partial derivatives exist we define the *copula density* which simply is the density function corresponding to the cdf C.

Definition 3.9 (Copula density)**.** *If a copula C has a density function $c(u_1, \dots, u_d) = \frac{\partial C(u_1,\dots,u_d)}{\partial u_1 \dots \partial u_d}$ we call c the copula density of C.*

The copula density provides a very useful formula for the density of a continuous rv which can be expressed as a product the copula density and the marginal densities.

Proposition 3.3. *If F with copula C and margins $F_1, \dots, F_d$ is absolutely continuous wrt the Lebesgue measure on $\mathbb{R}^d$, then its density is given by:*

$$f(x_1, \dots, x_d) = c\left(F_1(x_1), \dots, F_d(x_d)\right) \prod_{j=1}^{d} f_j(x_j), \tag{3.11}$$

where $c(u_1, \dots, u_d) = \frac{\partial C(u_1,\dots,u_d)}{\partial u_1 \dots \partial u_d}$ is the copula density of C.

Proof.

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

□

In cases where c has a simple closed form we can obtain an analytical expression for f using (3.11). This is true for the multivariate normal case as we will see in Section 3.5.

3.4 Measures of dependence revisited

There is a fundamental link between copulas and measures of dependence. As pointed out in Schweizer and Wolff (1981), any property of a continuous joint distribution which is invariant to strictly increasing transformations depends only on its copula. It is therefore natural to define dependence measures which are functions of the copula only. Such measures also have the practical advantage that their estimation requires to estimate the copula only instead of the whole distribution. Interestingly, widely used dependence measures which had been defined without any references to copulas can be reformulate as functions of the copula only. This is the case of Spearman's rho and Kendall's tau as stated in the following proposition.

Proposition 3.4. *Consider two continuous univariate rv X_1, X_2 having a joint cdf F and copula C. Spearman's and Kendall's rank correlations can then be expressed as:*

$$\begin{aligned} \tau(X_1, X_2) &= 4 \int_0^1 \int_0^1 C(u_1, u_2) \mathrm{d}C(u_1, u_2) - 1, \\ \rho_S(X_1, X_2) &= 12 \int_0^1 \int_0^1 \left(C(u_1, u_2) - u_1 u_2\right) \mathrm{d}u_1 \mathrm{d}u_2. \end{aligned}$$

Proof. We give a proof for τ only, the case of ρ_S being solved analogously. Starting with Definition 2.12 we successively obtain:

$$\begin{aligned}
\tau(X_1, X_2) &= \mathbb{P}\left((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right) - \mathbb{P}\left((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0\right) \\
&= \mathbb{P}\left((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right) - \left\{1 - \mathbb{P}\left((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0\right)\right\} \\
&= 4\mathbb{P}\left(X_1 - \tilde{X}_1 > 0, X_2 - \tilde{X}_2 > 0\right) - 1 \\
&= 4\int_{\mathbb{R}^2} F(x)\mathrm{d}F(x) - 1 = 4\int_{[0,1]^2} C(u_1, u_d)\mathrm{d}C(u_1, u_2) - 1.
\end{aligned}$$

In the last equality we used the probability transformation to substitute $u_i = F_i(x_i), i = 1, 2$. □

It should not be too surprising that τ and ρ_S depend only on the copula since it is known that both can be reformulated as functions of the observations' ranks and ranks are preserved by strictly increasing increasing transformations. Proposition 3.4 also leads to an interesting interpretation of Spearman's rho which appears to be the integral of the difference between the copula of interest $C(u_1, u_2)$ and the independence copula $C^\pi(u_1, u_2) = u_1 u_2$. Spearman's rho can therefore also be considered as a divergence measure between C and C^π (see also Section 2.4.3). Schweizer and Wolff (1981) studies different variants of divergences between $C(u_1, u_2)$ and C^π, amongst which are the L^1, L^2 and L^∞ distances, and provides interesting comparisons and examples. We do not introduce the multivariate generalisations of τ and ρ_S, which are not unique and have a slightly more complicated form, for more information on the subject see Joe (1990).

3.5 Gaussian copula models

Copula models can be interpreted as latent variables models for which the hidden variables have uniform margins on $[0, 1]^d$. The Gaussian copula model can be interpreted as a Gaussian latent variables model: starting with a multivariate Gaussian variable we transform the margins to obtain a rv with a Gaussian copula but new margins, as formalised in Definition 3.10 and illustrated in Figure 3.1.

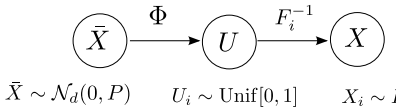

Figure 3.1: Starting with a jointly Gaussian rv $\bar{X} \sim \mathcal{N}_d(0, P)$, we can apply several transformations to obtain the desired margins. If we want to construct a rv with marginal cdfs $F_i, i = 1, \ldots, d$, we can first use the probability transformation Φ to obtain a marginally uniform rv U, and then use the quantile transformations F_i^{-1} to achieve the desired margins. When $F_i, i = 1, \ldots, d$ are continuous $F_i^{-1}, i = 1, \ldots, d$ are strictly increasing functions and, since Φ also is strictly increasing, the resulting rv X has the same Gaussian copula as $\bar{X}$.

Definition 3.10 (Gaussian copula model)**.** *Let P be a correlation matrix and $F_i, i = 1, \ldots, d$ be univariate, continuous or discrete, cdfs. A Gaussian copula model is constituted of the following elements:*

$$\bar{X} \sim \mathcal{N}_d(0, P) \tag{3.12}$$

$$X_i := F_i^{-1}(\Phi(\bar{X}_i)), \quad i = 1, \ldots, d \tag{3.13}$$

In the above definition, we could have chosen for the latent variable any other Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ such that $\mathcal{P}(\Sigma) = P$ without changing the distribution of X, meaning that the latent variables are only identifiable if we fix their mean and variance. Gaussian copula models enable to construct multivariate distributions combining a Gaussian dependence structure and arbitrary margins. The model defined in 3.10 remains valid when discrete margins are used but requires particular care in estimation and interpretation, we will discuss inference in presence of discrete margins in Chapter 7. When all margins are continuous, any variable with a Gaussian copula can be represented using the *normal scores* as latent variables as explained in the following proposition.

Proposition 3.5. *Consider a random vector X with continuous margins $F_1, \dots, F_d$ and Gaussian copula C_P. If the density f of X has a connected support then the vector of the normal scores is jointly Gaussian:*

$$\tilde{X} := \left(\Phi^{-1}\left(F_1(X_1)\right), \dots, \Phi^{-1}\left(F_d(X_d)\right)\right) \sim \mathcal{N}_d(0, I).$$

Proof. We first show that $\tilde{X}$ has copula C_P. Since the univariate Gaussian cdf Φ is strictly increasing, Φ^{-1} is also strictly increasing. If f has a connect support then F is a strictly increasing function $F : supp(f) \to [0, 1]$. Using Proposition 3.2 it immediately follows that $\tilde{X}$ has copula C_P. From Proposition 3.1 it is clear that $\tilde{X}$ has standard Gaussian margins. The proposition follows then directly from the unicity of the copula representation for continuous variables in Theorem 3.1. □

We mentioned in Section 3.2 that the Gaussian copula does not have a convenient analytical form, however it has the advantage of having a copula density of a similar form to the Gaussian density.

Proposition 3.6. *The copula density of a Gaussian copula C_P is given by:*

$$c_P(u_1, \dots, u_d) = c_P(\Phi(\tilde{x}_1), \dots, \Phi(\tilde{x}_d)) = |P|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\tilde{x}^T(P^{-1} - I)\tilde{x}\right\}, \tag{3.14}$$

where $\tilde{x}_j = \Phi^{-1}(u_j)$ for $u_j \in [0, 1]$, $j = 1, \dots, d$.

Proof. The d-dimensional Gaussian density with mean $(\mu_1, \dots, \mu_d)$, correlation matrix P and variance $(\sigma_1^2, \dots, \sigma_d^2)$ can be written as

$$f(x) = (2\pi)^{-\frac{d}{2}}|P|^{-\frac{1}{2}} \frac{1}{\prod_{j=1}^d \sigma_j} \exp\left\{-\frac{1}{2}\tilde{x}^T P^{-1}\tilde{x}\right\}, \tag{3.15}$$

where $x = (x_1, \dots, x_d)$ and $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_d)$ with $\tilde{x}_j = \frac{x_j - \mu_j}{\sigma_j}$. Equation (3.15) can then be rewriten as

$$f(x) = |P|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\tilde{x}^T(P^{-1} - I)\tilde{x}\right\} \prod_{j=1}^d \frac{1}{\sigma_j}\phi(\tilde{x}_j), \tag{3.16}$$

where ϕ denotes the univariate standard normal density $\mathcal{N}(0, 1)$. Combining equations (3.16) and (3.11) we obtain the assertion. □

Finally, another very useful property of Gaussian copulas is that independent dimensions can, as for a jointly Gaussian distribution, be identified by looking at at correlation matrix.

Proposition 3.7 (Gaussian copula independence structure)**.** *If a continuous random variable $X = (X_1, \dots, X_d)$ has a Gaussian copula C_P, then a zero entry in P implies independence between the corresponding dimensions.*

Proof. Since uncorrelated dimensions in a jointly Gaussian distribution implies independence, Proposition 3.3 shows that the Gaussian copula density c_P must factorise accordingly. This implies (using Proposition 3.3 again) that the density of X also factorises accordingly, which by definition means independence.

□

3.6 Copula for discrete marginals

Copulas appear as a natural method to model dependence for continuous distributions: given a random variable X with distribution F, there exists a unique copula C such that $F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$, and C admits the representation $C(u_1, \ldots, u_d) = F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right)$. The copula of F can be obtained from the joint distribution using the margins but it does not depend on them and it appears as a "margin-free" representation of F. Moreover, as we have seen, many dependence properties, like the rank correlations or the tail dependence parameter, depend on the copula only. The situation is however far less straightforward when discrete margins are involved. A complete and accessible account on the problems encountered is given in Genest and Nešlehová (2007), we also mention the interesting Nešlehová (2007). In the following we try to clarify the main issues and challenges of copula modeling in the discrete case.

1. *Lack of uniqueness in Sklar's representation.*
   If F is multivariate with discrete margins, Sklar's Theorem gives the existence of a corresponding copula but garanties unicity only on the range of the margins:

   $$\text{Range}\{F_1, \ldots, F_d\} := F_1(\bar{\mathbb{R}}) \times \cdots \times F_d(\bar{\mathbb{R}}).$$

   This lack of unicity implies that there exists a set $\mathcal{C}_F$ of copulas such that:

   $$F(x_1, \ldots, x_d) = C_F(F_1(x_1), \ldots, F_d(x_d)), \ \forall C_F \in \mathcal{C}_F. \tag{3.17}$$

   Denote by $\mathcal{A}_F$ the set of functions $A : [0,1]^d \to [0,1]$ for which equation (3.17) holds i.e. $F(x_1, \ldots, x_d) = A(F_1(x_1), \ldots, F_d(x_d))$. It is clear that $\mathcal{C}_F \subset \mathcal{A}_F$, but both sets are not equal. In particular, the function defined by

   $$B(u_1, \ldots, u_d) = F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right), \ u_i \in [0,1],$$

   which gives the copula of F in the continuous case is a member of $\mathcal{A}_F$ but is not a copula. B is not even a distribution function since F_i^{-1} is not right-continuous when F_i is discrete. We can also easily see that the distribution of $(F_1(X_1), \ldots, F_d(X_d))$, which is the copula of F is the continuous case, is not a copula anymore since its margins are not uniformly distributed. The natural question to ask next is: how different are members of $\mathcal{C}_F$? The answer is not completely encouraging. Carley (2002) derived lower C_F^- and upper C_F^+ bounds for $\mathcal{C}_F$. In general the set $\mathcal{C}_F$ is not small and copula based measures of dependence like Kendall's τ or Spearman's ρ_S can take significantly different values for C_F^- and C_F^+.

2. *The dependence structure is not characterized by the copula alone.*
   As mentioned above, the values of τ and ρ_S do not agree on the spectrum of $\mathcal{C}_F$, although these quantities are uniquely defined for the joint distribution F (see Definition 2.12 and 2.11). This fact indicates that the dependence structure is not completely characterized by the copula when margins are discrete. For example, mutual independence is not fully characterized by the independence copula $\Pi(u_1, u_2) = u_1 u_2$. Consider a joint distribution constructed using discrete margins F_1, F_2 and copula C. If $C = \Pi$ then X_1 and X_2 are independent, but independence between X_1 and X_2 does not imply that $C = \Pi$.

3. *Consequences for the latent variable representation.*
When using a copula model we implicitly assume that $X = (X_1, \dots, X_d)$ with margins $F_1, \dots, F_d$ is a transformation of a multivariate random vector $U = (U_1, \dots, U_d)$ with copula C_U and uniform margins, the transformation being given by $X_i = F_i^{-1}(U_i)$. When X has continuous margins the relationship between X and U is one-to-one, however in the discrete case X can be generated using different latent vectors U. The copula of the latent vector C_U uniquely determines the copula of X but the converse is not true.

3.7 Copula with conditional distributions

Conditional distributions are important building blocks for high-dimensional models like Bayesian networks (BN) and play an important role in recently developed models such as Copula Bayesian networks (CBN) (Elidan, 2010) or Vines (Bedford and Cooke, 2001). We therefore present a few results showing how copulas behave with conditional distributions.

We first mention that Sklar' s Theorem can also be applied to conditional distributions, meaning that if F denotes the conditional cdf of X given Y, then there exists a *conditional copula* C_y such that

$$F(x|y) = C_y(F_1(x_1|y), \dots, F_d(x_d|y) \mid y). \tag{3.18}$$

We use the notation C_y to emphasize that the copula in equation (3.18) is a conditional distribution on Y. When F is absolutely continuous its density has the form

$$f(x|y) = c_y(F_1(x_1|y), \dots, F_d(x_d|y) \mid y) \prod_{j=1}^{d} f_j(x_j|y), \tag{3.19}$$

where c_y is the copula conditional density and $f_j(x_j|y)$ are the conditional marginal densities. The density represented in (3.19) remains impracticable for modelling since it involves only conditional densities on the right-hand side. Elidan (2010) proposes another parametization of a conditional density $f_{X|Y}$ which combines the unconditional marginal densities with a copula quotient.

Proposition 3.8 (Copula parametrization of the conditional density)**.** *Consider the rv X and Y with conditional density $f_{X|Y}$. Denote by f_{X_i} the margins of X and by f_{Y_i} the margins of Y, which we assume to be strictly positive. For simplicity we write $F(x_i)$ instead of $F_{X_i}(x_i)$ in the notation of the copulas. There exists a copula density function $c(F(x_1), \dots, F(x_p), F(y_1), \dots, F(y_q))$ such that:*

$$f_{X|Y}(x|y) = \frac{c(F(x_1), \dots, F(x_p), F(y_1), \dots, F(y_q))}{c_m(F(y_1), \dots, F(y_q))} \prod_{j=1}^{p} f_{X_j}(x_j), \tag{3.20}$$

where c_m is the marginal copula density of c with the components $F(x_1), \dots, F(x_p)$ integrated out. The converse is also true, starting with a copula density for (X, Y) the right hand side of (3.20) *give the conditional density $f_{X|Y}$.*

Proof. Using Sklar's theorem we know that there exists copula densities c_{XY} and c_Y such that

$$f_{(X,Y)}(x, y) = c_{XY}(F(x_1), \dots, F(x_p), F(y_1), \dots, F(y_q)) \prod_{j=1}^{p} f_{X_j}(x_j) \prod_{k=1}^{q} f_{Y_q}(y_q),$$

$$f_Y(y) = c_Y(F(y_1), \dots, F(y_q)) \prod_{k=1}^{q} f_{Y_q}(y_q).$$

Since the marginal densities are strictly positive, we can then rewrite the conditional density as

$$f_{X|Y}(x|y) = \frac{f_{(X,Y)}(x, y)}{f_Y(y)} = \frac{c_{XY}(F(x_1), \dots, F(x_p), F(y_1), \dots, F(y_q))}{c_Y(F(y_1), \dots, F(y_q))} \prod_{j=1}^{p} f_{X_j}(x_j)$$

Finally since f_Y is a marginal density of $f_{(X,Y)}$ we have that

$$c_Y(u_1^y, \ldots, u_d^y) = \int c_{XY}(u_1^x, \ldots, u_p^x, u_1^y, \ldots, u_q^y) \mathrm{d}u_1^x \ldots \mathrm{d}u_p^x, \quad u_i^x, u_j^y \in [0,1] \forall i, j, \tag{3.21}$$

which can also be rewritten as

$$c_Y(u_1^y, \ldots, u_d^y) = \frac{\partial^q C_{XY}(1, \ldots, 1, u_1^y, \ldots, u_q^y)}{\partial u_1^y \ldots \partial u_q^y}. \tag{3.22}$$

□

This new copula representation of a conditional density is central to the construction of Copula Bayesian networks since it can be used to introduce copulas into directed graph models as shown by following result.

Proposition 3.9 (Graph Decomposition)**.** *Consider a direct acyclic graph $\mathcal{G}$ and a rv X with copula density $c(F_1(x_1), \ldots, F_p(x_p))$ and strictly positive margins $f_1(x_1), \ldots, f_p(x_p)$. If $f(x)$ the joint density of X decomposes according to $\mathcal{G}$ then the copula density $c(F_1(x_1), \ldots, F_p(x_p))$ also decomposes according to $\mathcal{G}$.*

$$c(F_1(x_1), \ldots, F_p(x_p)) = \prod_{i=1}^{p} \frac{c_{i,pa}(F_i(x_i), F(pa_{i_1}), \ldots, F(pa_{i_k}))}{c_{pa}(F(pa_{i_1}), \ldots, F(pa_{i_k}))}, \tag{3.23}$$

where $pa_i = \{pa_{i_1}, \ldots, pa_{i_k}\}$ denotes the parents of X_i and we write $F(pa_{ij})$ instead of $F_{pa_{ij}}(pa_{ij})$ for simplicity. $c_{i,pa}$ and c_{pa} denote the copula of $(X_i, pa_{i_1}, \ldots, pa_{i_k})$ and $(pa_{i_1}, \ldots, pa_{i_k})$, respectively.

Proof. Using the strict positivity of the marginal densities and the graph decomposition we obtain:

$$c(F_1(x_1), \ldots, F_p(x_p)) = \frac{f(x)}{\prod_{i=1}^{p} f_i(x_i)} = \frac{\prod_{i=1}^{p} f_{X_i|pa_i}(x_i|pa_i)}{\prod_{i=1}^{p} f_i(x_i)}, \tag{3.24}$$

Using Proposition 3.8 then gives

$$c(F_1(x_1), \ldots, F_p(x_p)) = \frac{1}{\prod_{j=1}^{p} f_j(x_j)} \prod_{i=1}^{p} \frac{c_{i,pa}(F_i(x_i), F(pa_{i_1}), \ldots, F(pa_{i_k}))}{c_{pa}(F(pa_{i_1}), \ldots, F(pa_{i_k}))} f_i(x_i), \tag{3.25}$$

which completes the proof. □

Chapter 4

Copula Mixture Model

4.1 Introduction

This chapter presents a new model for detecting dependencies which demonstrates how the flexibility offered by copula models can help in solving model-mismatch issues occurring with too restrictive modelling. We use copulas to significantly increase the scope of existing models while retaining efficient inference in a Bayesian setup. The general problem we consider is of detecting potential dependencies between two datasets of co-occuring observations. When different types of measurements concerning a same underlying phenomenon are available, often appearing in the form of co-occurring samples, combining them is more informative than analysing them separately. First, if we assume that these different measurements, also referred to as the different views, are generated by several data sources with independent noise, analysing them jointly can increase the signal to noise ratio. Second, only a combined analysis can take into consideration the dependencies existing between the different types of measurements. As pointed out in Klami and Kaski (2007), possible dependencies between the views often contain some of the most relevant information about the data. Dependency modelling captures what is common between the views, i.e. the shared underlying signal, and in many applications where several experiments are designed to measure the same object this shared aspect is the focus of interest.

The task of detecting dependencies has traditionally been solved by Canonical Correlation Analysis (CCA). This method can however detect only global linear dependency. When the data express not only one global dependency but different local dependencies, a mixture formulation is more adequate. Fern et al. (2005) introduces a mixture of local CCA model which groups pairs of points expressing together a particular linear dependency between the two views. This model is adapted to cases where the data express several different local correlations, but it still focuses exclusively on linear dependencies since it assumes that within each cluster the two views are linearly correlated.

Dependency-seeking clustering goes one step further in the generalisation process by assuming that the views become independent when conditioned on the cluster structure. The aim is to perform clustering in the joint space of the multiple views, while focussing explicitly on inter-view dependencies [1]. In the case of two views, the objective is then to group the co-occurring pairs of datapoints according to their inter-view dependency pattern such that when the cluster assignments are known these views become independent. As a consequence, the group structure now has a semantic interpretation in terms of dependency with the partition capturing the dependencies.

[1] The term inter-view dependencies refers to the dependence structure between the different views, whereas intra-view dependencies refers to the dependence structure between the different dimensions of one view.

The starting point of existing dependency-seeking methods is the probabilistic interpretation of CCA given in Bach and Jordan (2005) which provides the mathematical formalism on which dependency-seeking techniques are first based. In Klami and Kaski (2007) a Dirichlet prior Gaussian mixture for dependency-seeking clustering is introduced. However, as pointed out in Klami et al. (2010), when the data are not normally distributed, this method can suffer from a severe model mismatch problem. On application to non-normally distributed data these models have to increase the number of clusters to achieve a reasonable fit. Additional clusters are used to compensate for the inadequate Gaussian assumption. The components of these mixtures will not only be used to reflect differences in dependence structures but will also be used to approximate a non-Gaussian distribution. As a result some points expressing a similar inter-view dependence can be assigned to different groups and the interpretation of the clusters in terms of dependencies is lost. Moreover, the model needs to find a compromise between the cluster homogeneity and the approximation of a non-Gaussian mixture, so that non-homogenous clusters might emerge. Figure 4.1 illustrates how several Gaussian components can be used to approximate a beta density. An exponential family dependency-seeking method is proposed in Klami et al. (2010) to overcome this problem. This model can however be too restrictive when the views are multidimensional. Although the 1-dimensional exponential family covers many interesting distributions, only a few of them have convenient multivariate forms. In particular their dependence structure between dimensions is often very restrictive. Another restriction of that model is that all the dimensions in all the views must have the same univariate distribution whereas in practice different data sources are likely to produce differently distributed data.

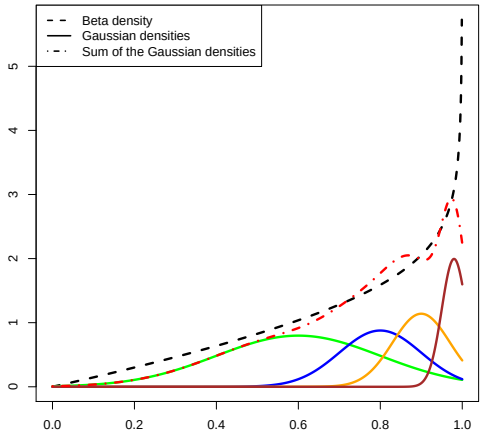

Figure 4.1: Gaussian components approximating a beta density.

To overcome these limitations we take advantage of the copulas framework to build a dependency-seeking clustering method suitable for data with any type of continuous densities. We use Gaussian copulas to construct Dirichlet prior mixtures of multivariate distributions with arbitrary continuous margins, the only restriction being that a density must exist. The model combines the adaptability of Bayesian non-parametric mixtures with the flexibility of copula-based distributions. Our approach focusses on Gaussian copulas for two main reasons. Firstly, their parametrisation using a correlation matrix covers many different dependence patterns ranging from independence to comonotonicity (perfect dependence). Secondly, the model can be reformulated using multivariate Gaussian latent variables which enables efficient MCMC inference.

4.2 Dependency-seeking clustering

Consider a p-dimensional random variable X and a q-dimensional rv Y which constitute two different sources of information about an object of interest. For example, several corporal measurements of a patient and the levels of different drugs administrated can serve as two sources of information about a medical treatment. We assume that X and Y have co-occurring samples $\left(x_{(1p),1}, \ldots, x_{(1p),n}\right)$ and $\left(y_{(1q),1}, \ldots, y_{(1q),n}\right)$ with $x_{(1p),i} \in \mathbb{R}^p$ and $y_{(1q),i} \in \mathbb{R}^q$, $i = 1, \ldots, n$. The probabilistic interpretation of CCA given by Bach and Jordan (2005) uses the following latent variable formulation:

$$\begin{aligned} Z &\sim \mathcal{N}_d\left(0, I_d\right), \\ (X, Y)\,|Z &\sim \mathcal{N}_{p+q}\left(WZ + \mu, \Psi\right), \end{aligned}$$

where $\mu = (\mu_x, \mu_y) \in \mathbb{R}^{p+q}$, $W = \begin{pmatrix} W_x \\ W_y \end{pmatrix} \in \mathbb{R}^{(p+q)\times d}$, $1 \leq d \leq \min(p, q)$ and the covariance matrix Ψ has a block diagonal form:

$$\Psi = \begin{pmatrix} \Psi_x & 0 \\ 0 & \Psi_y \end{pmatrix}. \tag{4.1}$$

They showed that the maximum likelihood estimate of W is connected to the canonical directions and correlations:

$$\hat{W}_x = \tilde{\Sigma}_x U_x M_x, \quad \hat{W}_y = \tilde{\Sigma}_y U_y M_y,$$

where $\tilde{\Sigma}_x$, $\tilde{\Sigma}_y$ are the sample covariance matrices, and U_x and U_y are the first d canonical directions. M_x and M_y are matrices such that $M_x M_y^T = P_d$ where P_d is the diagonal matrix containing the first d canonical correlations. Based on the above formulation, the following dependency-seeking clustering model is derived in Klami and Kaski (2008) :

$$Z \sim \text{Mult}\left(\theta\right), \tag{4.2}$$

$$(X, Y)\,|Z \sim \mathcal{N}_{p+q}\left(\mu_z, \Psi_z\right), \tag{4.3}$$

where Ψ_z has a block structure as in (4.1):

$$\Psi_z = \begin{pmatrix} \Psi_{zx} & 0 \\ 0 & \Psi_{zy} \end{pmatrix}, \tag{4.4}$$

and μ_z is a mean vector depending on Z. The latent variable Z now represents the clustering assignment. A key property of this model is the block diagonal structure of the covariance matrix Ψ_z. This special form implies that given the cluster assignment the two views are independent, thereby enforcing the cluster structure to capture all the dependencies, see Figure 4.2. This model however explicitly makes a conditional Gaussian assumption and can perform badly when data within a cluster are non-normally distributed as mentioned in Section 4.1. To relax this normality assumption, we present a dependency-seeking clustering model constructed using Gaussian copulas which can be applied to almost any type of continuous data.

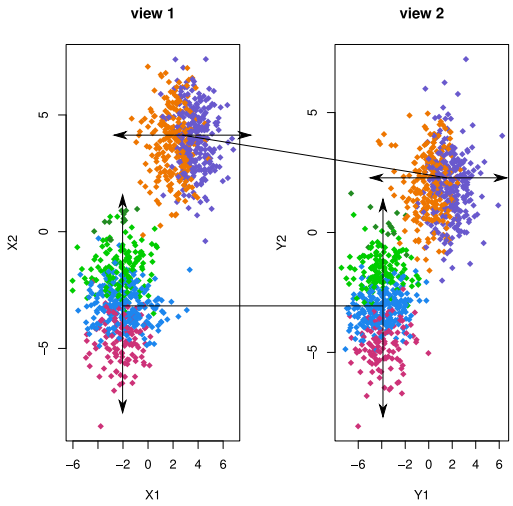

Figure 4.2: Simulated example of Gaussian data expressing a non-uniform dependence pattern between views 1 and 2. The points in the two bottom clouds have an inter-view dependence between their respective second components ($\text{corr}(X2, Y2) = 0.8$), whereas the points in the two upper clouds show inter-view dependence between their respective first dimensions ($\text{corr}(X1, Y1) = 0.6$). The result of dependency-seeking clustering is represented by the colour-coded clusters. The upper clouds are divided along the horizontal axis which is the dimension showing inter-view dependence, and the bottom clouds are divided along the vertical axis. The stronger vertical inter-view dependence in the bottom clouds causes a finer split in three coloured clusters, whereas the slightly weaker horizontal dependence between the upper groups causes a further split in two coloured clusters.

4.3 Multi-view clustering with meta-Gaussian distributions

4.3.1 Model specification

Our model is based on a Gaussian copula C_P parametrized by a correlation matrix P. Using Sklar's theorem 3.1 with C_P, we can construct multivariate distributions with arbitrary margins and a Gaussian dependence structure. These distributions, called meta-Gaussian distributions, provide a natural way to extend models based on a multivariate normality assumption. To avoid the various issues occurring with copula modelling for discrete margins we restrict our model to the continuous case. As the latent variable representation of the Gaussian copula model 3.10 precisely expresses, when using a Gaussian copula we do not attempt to directly model the correlation of the original variables, but instead we first apply the transformation $\Phi^{-1}(F_j(\,.\,))$ to every margin to obtain normally distributed variables $\Phi^{-1}(F_j(X_j)) \sim \mathcal{N}_1(0, 1)$ and then use P to describe their correlation. A determining advantage of using a Gaussian copula is that zero values in P encode independence between the corresponding marginal variables. Therefore, if P has a block diagonal structure as in (4.1), the conditional independence of $X|Z$ and $Y|Z$, which was a key property of equation (4.3), will be preserved in a meta-Gaussian model. Moreover, as mentioned in Section 3.5, absolutely continuous distributions constructed using a Gaussian copula have a density of the convenient form:

$$f(x) = |P|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\tilde{x}^T(P^{-1} - I)\tilde{x}\right\} \prod_{j=1}^{d} f_j(x_j), \tag{4.5}$$

where $\tilde{x}_j = \Phi^{-1}\left(F_j(x_j)\right)$, $x = (x_1, \ldots, x_d)$, $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_d)$. We denote this density by $\mathcal{M}(\theta, P)$, where θ is the vector containing all parameters of the marginal distributions.

Consider the two rv $X = (X_1, \ldots, X_p)$ and $Y = (Y_1, \ldots, Y_q)$. We assume their joint distribution is a Dirichlet prior mixture (DPM) given by:

$$f_{(X,Y)}(x, y) = \int\int f_{(X,Y)|\theta,P}(x, y)\mathrm{d}\mu_{\theta,P}\mathrm{d}\mu_G(\lambda, G_0),$$

where μ_G is the distribution of a Dirichlet process (Ferguson, 1973) with base distribution G_0 and concentration parameter λ. The novelty here is the choice of $f_{(X,Y)|\theta,P}$. We model the marginal distributions and the dependence structure separately to allow for more freedom:

1. The margins can be arbitrary continuous distributions (providing the corresponding density exists):

$$X_j|\theta = X_j|\theta_j^x \sim F_{X_j|\theta},\ j = 1, \ldots, p,$$
$$Y_j|\theta = Y_j|\theta_j^y \sim F_{Y_j|\theta},\ j = 1, \ldots, q,$$

   where $\theta = \left(\theta_1^x, \ldots, \theta_p^x, \theta_1^y, \ldots, \theta_q^y\right)$. Note here that $F_{X_j|\theta}$ can be different types of distributions for the multiple dimensions j.

2. The dependence structure is then specified by a Gaussian copula C_P with correlation matrix P having a block diagonal structure as in (4.1).

3. Finally the constructed multivariate distribution will have the form:

$$F_{(X,Y)|\theta,P}(x, y) = C_P\left(F_{X_1|\theta}(x_1), \ldots, F_{Y_q|\theta}(x_q)\right). \tag{4.6}$$

4.3.2 Bayesian inference

Separating the modelling task between specification of the margins and specification of the dependence structure simplifies the choice of the prior distributions. If we assume *a priori* independence for θ and P we can specify prior distributions for the margins and separately choose a prior for the parameters of the copula C_P. We specify independent prior distributions for the blocks P_x and P_y, where $P = \begin{pmatrix} P_x & 0 \\ 0 & P_y \end{pmatrix}$. For P_x and P_y we choose the marginally uniform prior given in Barnard et al. (2000). This prior is a multivariate distribution on the space of correlation matrices with uniform margins, i.e. P_{ij} is a uniform variable for $i \neq j$, and is connected to the inverse-Wishart distribution: if a covariance matrix $\Psi \in \mathbb{R}^{d\times d}$ is standard inverse-Wishart distributed with parameter I_d and $d+1$ degrees of freedom, then the corresponding correlation matrix R follows the marginally uniform prior distribution. The density of R is explicitly given by

$$f(R|d+1) \propto |R|^{\frac{d(d-1)}{2}-1}\left(\prod_{i=1}^{d}|R_{(i)}|\right)^{-\frac{(d+1)}{2}}, \tag{4.7}$$

where $|R|$ is the determinant of R and $R_{(i)}$ is the ith principal sub-matrix of R.

Inference can be done using MCMC sampling methods for Dirichlet process mixture models. We use a sampling scheme for models with non-conjugate prior given in Neal (2011). The method, detailed in Algorithm 3, is composed of three steps: a modified Metropolis-Hastings step, partial Gibbs sampling updates and an update of the parameters θ, P. In the third step we need to update the parameters of every cluster according to their posterior distribution. Since we cannot sample directly from this conditional posterior we developed a sampling scheme similar to the algorithm

proposed in Hoff (2007). The main idea is to overparametrize the model by introducing a normally distributed latent vector $(\tilde{X}, \tilde{Y})$. The variables in the complete model are then given by:

$$\begin{aligned}(\tilde{X}, \tilde{Y}) &\sim \mathcal{N}_{p+q}(0, \Sigma),\\ (X, Y) &\sim \mathcal{M}(\theta, P),\\ G &\sim \mathrm{DP}(\lambda, G_0),\\ (\theta, P) &\sim G,\end{aligned}$$

where Σ is a covariance matrix with corresponding correlation matrix P and DP denotes a Dirichlet process distribution with base measure G_0 and concentration parameter λ. The sampling scheme for Dirichlet process mixtures we use is based on explicit cluster assignments which we will denote by C. Introducing a latent Gaussian vector leads to the following reformulation of X and Y:

$$\begin{aligned}X_j &= (F_{X_j|\theta})^{-1}(\Phi_{\Sigma_{jj}}(\tilde{X}_j)), j = 1, \ldots, p,\\ Y_j &= (F_{Y_j|\theta})^{-1}(\Phi_{\Sigma_{j+p,j+p}}(\tilde{Y}_j)), j = 1, \ldots, q,\end{aligned}$$

and the correlation matrix P can be obtained as

$$P = \mathcal{P}(\Sigma) = D\Sigma D, \quad \text{with} \quad D = diag(\Sigma_{11}^{-\frac{1}{2}}, \ldots, \Sigma_{p+q,p+q}^{-\frac{1}{2}}), \tag{4.8}$$

or equivalently

$$P_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}, i, j = 1, \ldots, p+q. \tag{4.9}$$

As explained in Section 3.5, the class of all Gaussian distributions with covariance Σ shares the same copula $C_{\mathcal{P}(\Sigma)}$ which makes the overparametrization possible. We can easily see from equation (4.9) that zero entries in Σ and P match, and the block matrix structure is preserved when using the new parametrization. Figure 4.3 gives a representation of the complete model. In the MCMC scheme we can easily sample Σ conditioned on (X, Y), $(\tilde{X}, \tilde{Y})$ and θ, since we can use the conjugacy property of prior and conditional likelihood. A sample of the correlation matrix can be obtained as $\mathcal{P}(\tilde{X}, \tilde{Y})$, the correlation matrix of the random vector $(\tilde{X}, \tilde{Y})$. The posterior updates of the parameters are detailed in Algorithm 4. We use a Metropolis-within-Gibbs scheme which introduces MH updates in a Gibbs sampling algorithm. The Gibbs scheme is composed of three steps and successively samples from the following conditional distributions:

$$\begin{aligned}\theta|\Sigma, (\tilde{X}, \tilde{Y}), (X, Y),\\ (\tilde{X}, \tilde{Y})|\theta, \Sigma, (X, Y),\\ \Sigma|(\tilde{X}, \tilde{Y}), \theta, (X, Y).\end{aligned}$$

Updates for the parameter θ are drawn for one dimension at a time, conditioned on the other dimensions, using a Metropolis-Hastings Algorithm (see Algorithm 1 in Chapter 2). The choice of the transition kernel Q used in the MH algorithm will be steered by the particular form of the marginal distribution considered. Updates of $(\tilde{X}, \tilde{Y})$ are obtained similarly using MH for one dimension at a time. A sample from the posterior of Σ is easily obtained since the Gaussian distribution of $(\tilde{X}, \tilde{Y})$ is conjugate to the inverse Wishart distribution of Σ. The notations $\theta_{\star j}, \mathcal{P}(\tilde{X})_{\star j}$ in Algorithm 4 are used to emphasize that the corresponding vector or matrix is considered as a function of $\theta_j, \tilde{X}_j$ and parameters for the other dimensions are treated as constants.

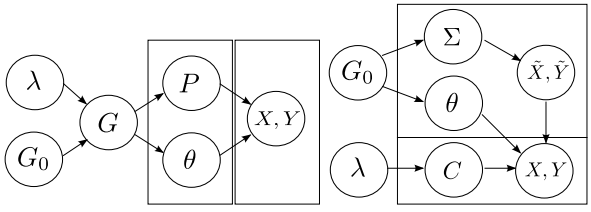

Figure 4.3: Graphical representation of the infinite copula mixture model with base measure G_0 and concentration λ. Left side: the original model, right side: the model augmented for sampling, where C denotes cluster assignment.

Algorithm 3 Markov Chain Sampling

$C_1, \ldots, C_n$ are the latent variables of the cluster assignments.
θ^{C_i} and P^{C_i} are the parameters specific to cluster C_i.
$n_{-i,c}$ is the number of datapoints in cluster c excluding observation i.
$C_{-i} = \{C_1, \ldots, C_{i-1}, C_{i+1}, \ldots, C_n\}$.
repeat
  for $i = 1, \ldots, n$ **do**
    if there exists k such that $C_k = C_i$ **then**
      Create a new cluster C_i^* with parameters θ^* and P^* drawn from G_0;
      Change C_i to C_i^* with probability $\min\left(1, \frac{\lambda}{n-1} \frac{f_{(X,Y)|\theta^*,P^*}(x,y)}{f_{(X,Y)|\theta^{C_i},P^{C_i}}(x,y)}\right)$;
    else
      Draw C_i^* from C_{-i} with $\mathrm{P}(C_i^* = c) = n_{-i,c}/(n-1)$. Change C_i to C_i^* with probability $\min\left(1, \frac{n-1}{\lambda} \frac{f_{(X,Y)|\theta^*,P^*}(x,y)}{f_{(X,Y)|\theta^{C_i},P^{C_i}}(x,y)}\right)$;
    end if
  end for
  for $i = 1, \ldots, n$ **do**
    if there exists k such that $C_k = C_i$ **then**
      Choose a new value for C_i with $\mathrm{P}(C_i^* = c) \propto \frac{n_{-i,c}}{(n-1)} f_{(X,Y)|\theta^c,P^c}(x,y)$;
    end if
  end for
  for $c \in \{C_1, \ldots, C_n\}$ **do**
    Update the parameters θ^c and P^c as described in Algorithm 4.
  end for
until stopping criterion

Algorithm 4 Posterior updates of $(\theta, P) \,|\, (X, Y)$

For clarity we omit the cluster index c.
1.*Sample* $\theta|\Sigma, (\tilde{X}, \tilde{Y}), (X, Y)$
for $j = 1, \ldots, p$ **do**
  Draw θ_j^x using Metropolis-Hastings;
  $\theta_j^x \sim f(\theta_j^x | \theta_{-j}^x, \tilde{X}, X) \propto \mathcal{M}(\theta_{\star j}^x, \mathcal{P}(\tilde{X}))\pi(\theta_j)$
end for
Apply the same procedure for Y;
2. *Sample* $(\tilde{X}, \tilde{Y})|\theta, \Sigma, (X, Y)$
for $j = 1, \ldots, p$ **do**
  Draw $\tilde{X}_j$ using Metropolis-Hastings;
  $\tilde{X}_j \sim f(\tilde{X}_j | \tilde{X}_{-j}, \theta, \Sigma, X) \propto \mathcal{M}(\theta, \mathcal{P}(\tilde{X})_{\star j})\mathcal{N}(0, \Sigma_x)$
end for
Apply the same procedure for $\tilde{Y}$;
3. *Sample* $\Sigma|(\tilde{X}, \tilde{Y})$, $\theta, (X, Y)$*:*
Draw $\Sigma_x \sim \mathcal{N}(0, \Sigma_x)\mathcal{IW}(p+1, I_p)$
$\sim \mathcal{IW}(I_p + \sum_{i=1}^n \tilde{X}_{(1p),i} \tilde{X}_{(1p),i}^T, p + 1 + n)$;
Apply the same procedure to obtain Σ_y.

4.4 Experiments

4.4.1 Simulated data

We simulate two different 2-dimensional multi-view data sets with Gaussian intra-view dependence structure. The marginal distributions are Gaussian in the first view, and beta or exponential in the second. Each data set is composed of two clusters which can be identified only by considering the inter-view dependencies. We first simulated data points with a single cluster structure in each view but a strong positive dependence between the first dimensions of the views, i.e. between X^1 and Y^1. In a second step we separated the data in two groups of unequal size and randomly permuted their order within groups to suppress any inter-view dependency within these groups. Figure 4.4 (bottom left panel) shows the resulting cluster structure in the joint space of the two views recovered by the copula mixture model. Parameters used for the simulations can be found in Table 4.1.

Table 4.1: Parameters used for the simulations.

Simulation 1	view 1: Normal	μ	$(0, 0)$
		σ^2	$(1, 1)$
		$(P_x)_{12}$	0.9
	view 2: Beta	α	$(3, 1)$
		β	$(1, 10)$
		$(P_y)_{12}$	-0.5
Simulation 2	view 1: Normal	μ	$(0, 0)$
		σ^2	$(1, 1)$
		$(P_x)_{12}$	0.9
	view 2: Exponential	λ	$(2.5, 2.5)$
		$(P_y)_{12}$	0.9

We compared the copula mixture (CM) with three other methods: a Dirichlet prior Gaussian mixture for dependency-seeking clustering (GM) as derived in Klami and Kaski (2007), a non-Bayesian mixture of canonical correlation models (CCM) Vrac (2010) Fern et al. (2005) and a variational Bayesian mixture of robust CCA models (RCCA) Viinikanoja et al. (2010). CCM and RCCA both assume that the number of clusters is known or can be determined as explained in Viinikanoja et al. (2010). In our comparison experiments we gave as input for both methods the correct number of clusters, giving them the advantage of this extra knowledge. Results presented in Figure 4.5 show that CM applied with the correct marginal distributions' form produces a better classification. GM does not perform well on those data sets because the number of clusters is overestimated; the model compensates for the inadequate Gaussian assumption by multiplying the number of components and additional clusters are created to approximate non-Gaussian distributions. Since the number of clusters in a Dirichlet prior Gaussian mixture can be reduced by imposing a too-strong prior on the variances, we modified the prior information to enforce artificially high variances in the second view until the mixture is forced to create no more than two clusters. We report both results obtained with less (GM1) and more (GM2) informative priors. As can be seen in Figure 4.4, when strong prior information is used to artificially reduce the number of clusters, the GM cannot recover the true cluster structure. CCM and RCCA used with the correct number of clusters as input perform comparatively, or better than the GM but clearly worse than CM for those data sets having non-linear inter-view dependencies.

4.4.2 Real data

We perform a combined analysis of two data sets providing information about the regulation of gene expression in yeast under heat shock; each data set being treated as one view. The first

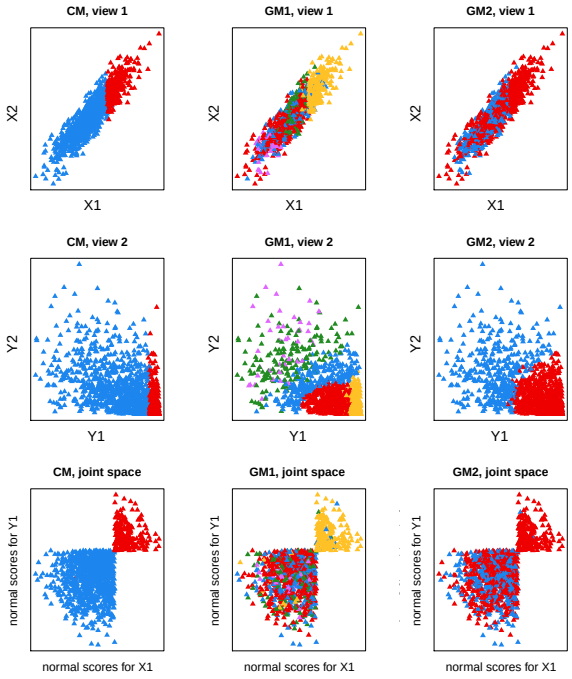

Figure 4.4: Scatterplot of the simulated data in the Gaussian view (first view, top panel), in the beta view (second view, middle panel) and in the joint space of the normal scores for the two views where the two clusters can be clearly identified (bottom panel). The clustering results are shown for the copula mixture (CM) and the Gaussian mixture with two different priors (GM1 and GM2). CM perfectly recovers the true cluster structure, whereas a model mismatch problem prevents GM to find the correct clustering.

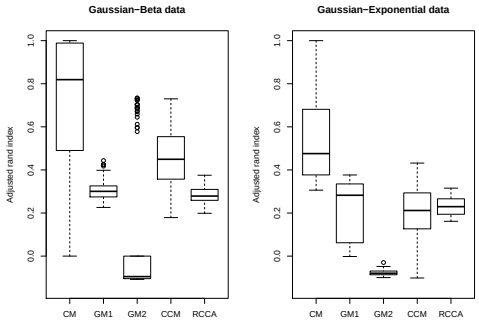

Figure 4.5: Boxplot of the adjusted rand index over 100 (Gaussian-beta data on the left panel) and 50 (Gaussian-exponential data on the right panel) simulations for the copula mixture (CM), the Gaussian mixture with two different priors (CM1 and CM2), the non-Bayesian mixture of CCA (CCM), and the robust CCA mixture (RCCA). Friedman's test with post-hoc analysis rejected, for both experiments, the null hypothesis of equal medians between CM and every other method (P-value < 0.005).

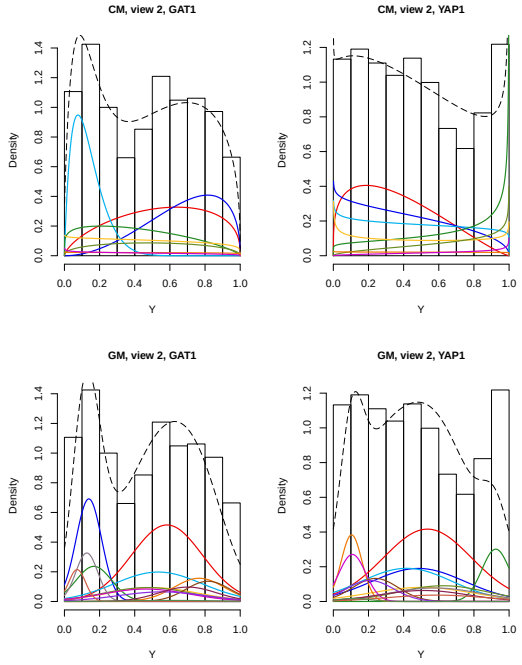

Figure 4.6: Histogram of the binding affinity scores for the binding factors GAT1 and YAP1. The estimated densities of the 8 clusters discovered by CM are represented as colored lines in the top panel. Estimated densities of the 14 clusters found by GM are shown in the bottom panel. The black dashed lines represent the total density resulting of the mixture.

data set (published in Gasch et al. (2000)) provides genes expression values measured at 4 time points. The second data set (given in Harbison et al. (2004)) contains binding affinity scores for interactions between these genes and 6 different binding factors. Similar data have already been analysed in Klami and Kaski (2007). 5360 genes present in both views are clustered using a Gaussian dependency-seeking clustering model (GM) and using the copula mixture (CM). CM uses Gaussian marginals in the first view and beta marginals in the second view. Here the choice of the beta distribution is motivated by the fact that observations in the second view are restricted to the $[0, 1]$ interval. For the univariate Gaussian margins we choose normal and inverse-gamma priors for mean and variance respectively, whereas for the beta margins both shape parameters have gamma priors. GM uses the standard conjugate prior [2].

For different values of the concentration parameter $\lambda \in \{0.01, 0.1, 1, 5, 10\}$, CM consistently estimates 8 clusters whereas GM estimated between 13 and 15 clusters. In this section we report the results obtained with $\lambda = 1$. As we observed with the simulated data more clusters need to be created by the Gaussian mixture to compensate for the model mismatch. This phenomenon is illustrated in Figure 4.6. The interpretation of the clustering then becomes very arduous since these additional clusters cannot be distinguished from those capturing the dependencies. Another interpretation problem clearly arises in the Gaussian model when we look at the estimated intra-view correlations. Two negative effects accumulate here; first correlation can be an inadequate dependence measure for non-normally distributed data, and second the additional split in many components can change the cluster-specific intra-view dependence as illustrated in Figure 4.7.

To understand what information one could gain by dependency-clustering, we perform three additional clustering of the same data: first we cluster the datapoints on each view separately, then

[2]The use of conjugate prior does not, in general, increase the number of clusters as shown in Rasmussen and Görür (2010).

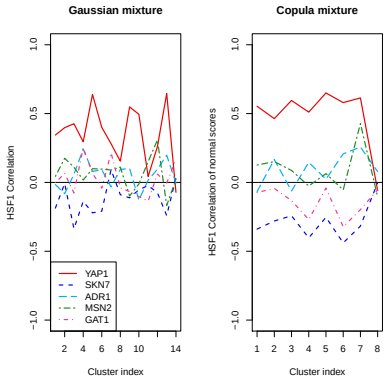

Figure 4.7: Correlations estimated with GM (left panel) and correlations of the normal scores estimated by CM (right panel) between HSF1 and the five other binding factors. In the Gaussian model the correlation between HSF1 and YAP1 seems to vary drastically with the clusters. In CM this correlation has stable positive values for all clusters with the exception of the last cluster. Since the binding factors HSF1 and YAP1 are both activated by the substance *menadione* as explained in Hohmann and Mager (2003), we can expect that their binding affinities are positively correlated and independent of the cluster.

we cluster them in the complete product space of the joint views, i.e. without imposing the constraint of a block structure on the correlation matrix. Priors and hyperparameters are kept constant over experiments. CM finds four clusters in the first view as well as in the second view. Clustering in the product space with full correlation matrix again leads to four groups. Figure 4.8 illustrates how the three main clusters found in the complete product space are further separated by dependency-seeking clustering, showing dependencies between the two views.

As mentioned in section 4.1, GM cannot exclusively focus on compact clusters because it needs to find a compromise between the cluster homogeneity and the approximation of a non-Gaussian mixture. As a result, non-homogenous clusters might emerge which are needed to fit the margins despite model mismatch. To test if this phenomenon is present here, we perform a gene ontology enrichment analysis (GOEA) using GOrilla Eden et al. (2009). GOEA is used to test if some of the biological processes associated with the genes are over-represented in the clusters, thereby providing a quality measure for the clustering. The analysis shows that 3 out of 14 clusters (these 3 clusters representing together 17,3% of the data points) found by GM do not express any significant enrichment. By contrast, all 8 clusters produced by CM express a highly significant enrichment and every cluster can be associated with a specific biological processes, e.g. the two largest clusters can be interpreted as groups of genes involved in organelle organization and meiosis respectively. The clear difference in the enrichment analysis results between GM and CM demonstrates that the quality of the clustering is indeed impaired when a model with inadequate margins is used.

4.5 Conclusion

A fundamental aspect in dependency-seeking clustering is that the partition possesses a semantic interpretation in terms of dependency: the dependencies are captured by the cluster structure. This interpretation is however only valid when the model is rich enough to properly fit each view, which can be particularly difficult to achieve for non-Gaussian data with existing models. This task becomes even more arduous when the dimensions of the views increase since the model then needs to adequately fit every margin while allowing for a sufficiently rich intra-view dependence

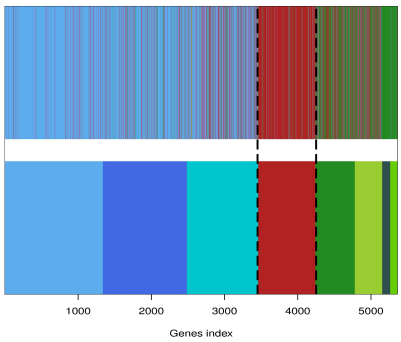

Figure 4.8: The bottom panel represents in different colors the cluster indices for all genes (re-ordered by cluster assignment) as obtained using dependency-seeking clustering with CM. The top panel shows the cluster indices obtained when clustering in the complete product space, i.e using CM with a full correlation matrix instead of a restricted block diagonal matrix. This illustrates how existing groups are further separated into smaller clusters expressing inter-view dependencies.

structure. The copula mixture model offers enough flexibility to cover both aspects: the margins can be specified separately for each dimension and the Gaussian copula allows for a wide range of intra-view dependencies. Using a Gaussian copula also facilitates the inference and we provide an efficient MCMC scheme. Experiments on simulated data show that the copula mixture model significantly improves the clustering results. In a large-scale real-world clustering problem of genes expression data and genes binding affinities, the dependency-seeking copula mixture model produces a clustering solution that significantly differs from those obtained on the single views or on the product space, and from that obtained by the standard Gaussian model which clearly suffered from model-mismatch problems. Detailed analysis of the functional annotation of the genes in the clusters discovered by dependency-seeking CM shows that the induced cluster structure allows a plausible biological interpretation in that the groups are clearly enriched by genes involved in distinct biological processes.

Chapter 5

The Information Bottleneck

5.1 Introduction

In the previous chapter we presented a copula-based model for dependency-seeking clustering which opened the scope of dependency-seeking models to a vast range of the new applications while retaining efficient inference due the properties of the Gaussian copula. Another interesting method which can benefit from the flexibility of copula models is the *Information Bottleneck* (IB). Before presenting in Chapter 6 a new copula-based model with attractive properties and showing the deep connections between copulas and the IB method, we dedicate this chapter to the presentation of the IB method.

The Information Bottleneck method was first introduced in Tishby et al. (1999) as a novel information compression technique. IB takes an original approach on compression by considering for the first time the *relevance* of information. In the early formulation of information theory (Shannon, 1948), the problem of information transmission is formulated independently from the notion of *meaningful* information. We, however, intuitively think of communication as the transmission of a certain meaning, which constitutes the heart of a message containing also less significant elements. The IB method formalises the intuitive idea that relevance of information could be crucial for effective compression. The tools best adapted to obtain a mathematical formulation of the concept of relevance actually already exist in information theory, and, as we will see below, take the form of mutual information. Compression, or more exactly lossy compression, was traditionally treated in the framework of *rate distortion theory* (see Cover and Thomas (1991) for more details). The main idea is that the loss incurred by compression can be expressed as an average distortion of the reconstructed signal, then a trade-off inevitably occurs between distortion and compression rate. Even if the choice of a distortion function is of decisive importance, it is not treated in rate distortion theory and standard functions such as the *Hamming distortion* or the *squared error distortion* are often used by default. As pointed out in Tishby et al. (1999), the distortion function implicitly performs feature selection on the transmitted information and an arbitrary choice might select irrelevant features. For many coding problems we however have an indirect knowledge of the relevant features, e.g. in a speech recognition task the features of interest are those which enable to identify the speaker. The central idea of IB is precisely to model this indirect knowledge using a *relevance variable*, in the speech recognition example this variable would encode the speaker's identity. The task then becomes to compress the signal while preserving information about the relevance variable, meaning preserving the relevant information.

5.2 The Information Bottleneck problem

Consider two random variables, possibly multivariate, X and Y with values in the measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. To be consistent with the IB literature, we use $p(x, y)$ to denote the joint probability mass function for discrete variables and the joint density for continuous variables. For simplicity, in both cases we refer to $p(x, y)$ as the joint "probability distribution" by a slight abuse of language. In this context X is interpreted as a signal for which we want to obtain a compressed version in the form of another random variable T. We assume that the joint distribution $p(x, y)$ is known and the task is to determine the compression T. Instead of considering a distortion function to guide the compression as in rate distortion theory, IB introduces the concept of relevant or meaningful information. The relevant information takes the form of the random variable Y, the relevance variable. The aim is then to construct a compressed representation T of X that is most informative about Y. To fully specify T we have to determine its joint distribution with X and Y denoted by $p(x, y, t)$. Since T is a compression of X it is independent of Y given X:

$$p(t|x, y) = p(t|x), \tag{5.1}$$

and the variables satisfy the following Markov relation

$$T \leftrightarrow X \leftrightarrow Y.$$

This relation also expresses the fact that the compression T cannot contain more information about Y than the original data X. The conditional independence property implies that the full joint distribution can be factorised as

$$p(x, y, t) = p(y, t|x)p(x) = p(y|x)p(t|x)p(x). \tag{5.2}$$

Since $p(x, y)$ is assumed known, determining $p(t|x)$ is enough to fully specify the joint distribution $p(x, y, t)$. The IB model is illustrated in Figure 5.1.

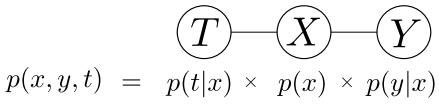

Figure 5.1: Graphical representation of the conditional independence structure of IB.

The IB problem is formulated using information-theoretical concepts only. Compression and relevance are both measured in terms of mutual information:

1. $I(X;T)$ measures how close T is to the original signal and a high value means low compression.
2. $I(Y;T)$ measures the information contained in T about Y and a high value means that the relevant information is well preserved.

The compression task can then be formulated as a variational problem which encompasses the need for compression and the information preservation target.

Definition 5.1 (IB variational problem)**.**

$$\min_{p(t|x)} \mathcal{L} \mid \mathcal{L} \equiv I(X;T) - \beta I(T;Y), \tag{5.3}$$

where the minimum is taken over all possible conditional distributions of T given X.

The Lagrange parameter $\beta > 0$ determines the trade-off between compression of X and preservation of information about Y, a large value will favour informativeness whereas a small value will put emphasis on compression.

The variational problem 5.1 is formulated in general terms, for discrete and continuous variables X, Y. However, in general, no analytical solution is available and Tishby et al. (1999) focuses on the discrete case for which they provide an iterative optimisation method. Their iterative algorithm is based on the characterisation of the stationary points of problem 5.1 given in the following theorem.

Theorem 5.1 (Formal solution to the IB problem)**.** *For a given β, a conditional distribution of T given X is a stationary point of* (5.3) *if and only if:*

$$p(t|x) = \frac{p(t)}{Z(\beta)} \exp\left(-\beta \mathrm{D}_{KL}(p(y|x) \parallel p(y|t))\right), \tag{5.4}$$

where $Z(\beta)$ denotes the normalisation constant.

A remarkable feature of Theorem 5.1 is that the Kullback-Leibler divergence between $p(y|x)$ and $p(y|t)$ appears on the right hand side, suggesting that it is the appropriate distortion measure for our compression problem. This theorem, however, cannot be used to find stationary points in practice since the term $p(t|x)$ also appears on the right hand side of equation 5.4:

$$p(t) = \sum_{x \in \mathcal{X}} p(t|x)p(x), \tag{5.5}$$

$$p(y|t) = \frac{1}{p(t)} \sum_{x \in \mathcal{X}} p(x, y)p(t|x). \tag{5.6}$$

The structure of equation (5.4) suggests using a fixed point method to obtain stationary points, and the *generalised Blahut-Arimoto* algorithm introduced in Tishby et al. (1999) alternates between self-consistent computations of $p(t|x)$, $p(t)$ and $p(y|t)$ as detailed in Algorithm 5. The resulting discrete T then defines "soft" clusters of X, the conditional probability mass function $p(t|x)$ giving the probability that the observation x is assigned to cluster t. Algorithm 5 converges to a stationary point of (5.3), see Slonim (2002) for a detailed proof, however, it is not guaranteed to attain the global optimum.

Algorithm 5 Generalised Blahut-Arimoto

Random initialisation of $p^{(m)}(t|x)$ with $m = 1$;
repeat
  Set $p^{(m)}(t) = \sum_{x \in \mathcal{X}} p^{(m)}(t|x)p(x)$;
  Set $p^{(m)}(y|t) = \frac{1}{p^{(m)}(t)} \sum_{x \in \mathcal{X}} p(x, y)p^{(m)}(t|x)$;
  Update $p^{(m+1)}p(t|x) = \frac{p^{(m)}(t)}{Z^{(m+1)}(\beta)} \exp\left(-\beta \mathrm{D}_{KL}(p(y|x) \parallel p^m(y|t))\right)$;
  Set m to $m + 1$;
until Convergence criterion satisfied.

In the case of continuous X and Y, a similar set of self-consistent equations for $p(t|x)$, $p(t)$ and $p(y|t)$ are obtained:

$$p(t) = \int_{\mathcal{X}} p(t|x)p(x)dx, \tag{5.7}$$

$$p(y|t) = \frac{1}{p(t)} \int_{\mathcal{X}} p(x, y)p(t|x)dx, \tag{5.8}$$

$$p(t|x) = \frac{p(t)}{Z(\beta)} \exp\left(-\beta \mathrm{D}_{KL}(p(y|x) \parallel p(y|t))\right), \tag{5.9}$$

which also translate into two coupled eigenvector problems for $\partial \log p(x|t)/\partial t$ and $\partial \log p(y|t)/\partial t$, but a direct solution of this problem is very difficult in practice and no iterative algorithm applicable to the general continuous case is known. This indicates that to make the IB method applicable in practice with continuous variables more assumptions are required. As we will see in the next section, the IB problem for continuous variables can be solved for one special case: when X and Y are jointly multivariate Gaussian distributed. The problem then becomes analytically tractable and obtaining T can be reduced to solving an eigenvalue problem.

5.3 Gaussian IB

Gaussian IB (GIB), introduced in Chechik et al. (2003) and developed in more details in Chechik et al. (2005), is the first IB method applicable to continuous variables. As explained in the previous section, a set of self-consistent equations also exist in the continuous case but no practical algorithm emerges. The particular properties of Gaussian distributions simplify the IB problem considerably and a solution can be derived analytically. Simplifications arise primarily from the closure properties of the Gaussian family which imply that an optimal compression T can always be found within the set of Gaussian variables. The solution T then provides a continuous compression of X instead of the soft clustering obtained in the discrete case. The analytical expression obtained for T can be readily used to solve problems of a real-world scale and, further, gives a precious insight into the IB mechanism. We present in this section the main theoretical results on GIB, proofs will often be briefly sketched as the complete exposition is rather technical and we instead focus on explaining the main ideas behind the results derived from Chechik et al. (2005) and Globerson and Tishby (2004).

Assume that the random vectors X and Y are jointly Gaussian distributed with zero mean and variance Σ:

$$(X, Y) \sim \mathcal{N}\left(0_{p+q}, \Sigma = \begin{pmatrix} \Sigma_x & \Sigma_{xy}^T \\ \Sigma_{xy} & \Sigma_y \end{pmatrix}\right), \tag{5.10}$$

where p is the dimension of X, q is the dimension of Y and 0_{p+q} is the zero vector of dimension $p+q$. A key property of the Gaussian case in that an optimal compression T can be found amongst the random variables which are jointly multivariate Gaussian distributed with the vector (X, Y). This result proven in Globerson and Tishby (2004) is expressed in the following lemma.

Lemma 5.1 (Gaussian optimality)**.** *When X and Y are jointly Gaussian, the global optimum of problem* (5.3) *is attained for a compression T which is also jointly Gaussian with (X, Y).*

The proof of Lemma 5.1 uses the entropy power inequality and is rather technical. However, the authors provides an interesting intuition about this result noting that, since the joint distribution of X, Y carries only second order correlations, potential higher order moments in the distribution of T would not bring additional information. Lemma 5.1 implies that T can be expressed as a noisy linear transformation of X

$$T = AX + \xi, \quad \xi \sim \mathcal{N}(0_p, \Sigma_\xi), \quad A \in \mathbb{R}^{p \times p}, \tag{5.11}$$

where the noise term ξ is independent of X. The optimal compression is then given by

$$T \sim \mathcal{N}(0_p, \Sigma_t) \quad \text{with} \quad \Sigma_t = A\Sigma_x A^T + \Sigma_\xi, \tag{5.12}$$

and the minimisation problem (5.3) is reduced to an optimisation task over A, Σ_ξ:

$$\min_{A, \Sigma_\xi} \mathcal{L} | \mathcal{L} \equiv I(X;T) - \beta I(T;Y). \tag{5.13}$$

Since the noise variance can be set to the identity matrix $\Sigma_\xi = I_p$, as shown in Lemma 5.2, the problem further simplifies to finally become an optimisation over A only.

Lemma 5.2. *Consider a pair (A, Σ_ξ) where $A \in \mathbb{R}^{p\times p}$ and Σ_ξ is a full-rank covariance matrix. Then there exists $\tilde{A} \in \mathbb{R}^{p\times p}$ such that $\mathcal{L}(\tilde{A}, I_p) = \mathcal{L}(A, \Sigma_\xi)$.*

Proof. The covariance matrix Σ_ξ can be decomposed as $\Sigma_\xi = LDL^T$ where D is a diagonal matrix and L is orthonormal. A straightforward calculation then verifies that $\tilde{A} := \sqrt{D^{-1}}LA$ satisfies $\mathcal{L}(\tilde{A}, I_p) = \mathcal{L}(A, \Sigma_\xi)$. □

The fact that optimisation problem 5.13 is ultimately independent of Σ_ξ can be intuitively understood: the noise term ξ do not carry any information about Y and therefore an optimal projection A should be determined independently. The noise variance can be fixed before optimisation is performed, it however remains a necessary component which fixes the scale of the problem. Having reduced the original IB problem to the optimisation task $\min_A \mathcal{L}|\mathcal{L} \equiv I(X;T) - \beta I(T;Y)$ we can compute a solution analytically as shown in the following theorem.

Theorem 5.2 (Gaussian IB solution)**.** *The transformation matrix A is given by:*

$$A = \begin{pmatrix} [0^T; \dots; 0^T] & 0 \leq \beta \leq \beta_1^c \\ [\alpha_1 v_1^T; 0^T; \dots, 0^T] & \beta_1^c \leq \beta \leq \beta_2^c \\ [\alpha_1 v_1^T; \alpha_2 v_2^T; 0^T; \dots; 0^T] & \beta_2^c \leq \beta \leq \beta_3^c \\ \vdots & \end{pmatrix}, \tag{5.14}$$

where $v_1^T, \dots, v_p^T$ are left eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$ sorted by their corresponding increasing eigenvalues $\lambda_1, \dots, \lambda_p$. The critical β values are $\beta_i^c = (1-\lambda_i)^{-1}$, and the α_i coefficients are defined by $\alpha_i = \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$ with $r_i = v_i^T \Sigma_x v_i$. In the above, 0^T is a p-dimensional row vector and semicolons separate rows of A.

We can see from equation (5.14) that the optimal transformation of X is a combination of weighted eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$. The number of selected eigenvectors, and thus the effective dimension of T, depends on the parameter β and changes at each critical point β_i^c. Since every coefficient α_i vanishes at the corresponding critical β_i^c, the change in A remains smooth as a function of β.

Proof. We give a proof sketch which underlines the main ideas of the proof. We first compute the conditional and unconditional covariance matrices involving T. Recalling that the matrices $\Sigma_x, \Sigma_y, \Sigma_{xy}$ are assumed known we obtain:

$$\Sigma_{xt} = \mathrm{cov}(X, T) = \mathrm{cov}(X, AX + \xi) = A\mathrm{cov}(X, X) + \mathrm{cov}(X, \xi) = A\Sigma_x, \tag{5.15}$$

$$\Sigma_{yt} = \mathrm{cov}(Y, T) = \mathrm{cov}(Y, AX + \xi) = A\mathrm{cov}(Y, X) + \mathrm{cov}(Y, \xi) = A\Sigma_{xy}, \tag{5.16}$$

$$\Sigma_{t|x} = \mathrm{cov}(AX + \xi|X) = \mathrm{cov}(AX|X) + \mathrm{cov}(\xi|X) = \Sigma_\xi = I, \tag{5.17}$$

$$\Sigma_{t|y} = \mathrm{cov}(AX + \xi|Y) = \mathrm{cov}(AX|Y) + \mathrm{cov}(\xi|Y) = A\Sigma_{x|y}A^t + \Sigma_\xi = A\Sigma_{x|y}A^t + I. \tag{5.18}$$

Using the following notation $\Sigma_{\bar{xt}} = \begin{pmatrix} \Sigma_x & \Sigma_{xt}^T \\ \Sigma_{xt} & \Sigma_t \end{pmatrix}$ we obtain expressions for the mutual informations involved:

$$\begin{aligned} \frac{1}{2}I(X;T) &= \log\left(\frac{|\Sigma_x||\Sigma_t|}{|\Sigma_{\bar{xt}}|}\right) = \log|\Sigma_t| + \log\left(\frac{|\Sigma_x|}{|\Sigma_x||\Sigma_t - \Sigma_{xt}\Sigma_x^{-1}\Sigma_{xt}^t|}\right) \\ &= \log|\Sigma_t| - \log|\Sigma_{t|x}| = \log(|A\Sigma_x A^t + I|), \\ \frac{1}{2}I(Y;T) &= \log|\Sigma_t| - \log|\Sigma_{t|y}| = \log(|A\Sigma_x A^t + I|) - \log(|A\Sigma_{x|y}A^t + I|). \end{aligned} \tag{5.19}$$

The objective function can then be rewritten as

$$\begin{aligned} \mathcal{L} &= \log|\Sigma_t| - \log|\Sigma_{t|x}| - \beta\log|\Sigma_t| + \beta\log|\Sigma_{t|y}| \\ &= (1-\beta)\log|A\Sigma_x A^t + I| + \beta\log|A\Sigma_{x|y}A^t + I|, \end{aligned} \tag{5.20}$$

and the derivative w.r.t. A is

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}A} = (1-\beta)(A\Sigma_x A^t + I)^{-1} 2A\Sigma_x + \beta(A\Sigma_{x|y}A^t + I)^{-1} 2A\Sigma_{x|y}. \tag{5.21}$$

Setting equation (5.21) to zero and rearranging the terms we obtain a necessary condition for the existence of a minimum:

$$\frac{\beta - 1}{\beta}\Big[(A\Sigma_{x|y}A^t + I)(A\Sigma_x A^t + I)^{-1}\Big] A = A\left(\Sigma_{x|y}\Sigma_x^{-1}\right). \tag{5.22}$$

We can recognise in equation (5.22) the form of a special eigenvalue problem for which the eigenvalues depend on A. The solution A must then be in the span of the eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$. This implies that A can expressed in the form $A = WV$ where the rows of V are the left normalised eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$ and W is a weight matrix. By substituting $A = WV$ in (5.22), and after more derivations detailed in Chechik et al. (2005), we find the form described in Theorem 5.2. □

Chapter 6

Meta-Gaussian Information Bottleneck

6.1 Introduction

The information bottleneck method (IB) introduced the concept of relevant information in the data compression problem, offering a new perspective on signal compression. Although the IB method beautifully formalises the compression problem under relevance constraints, the practical solution of this problem remains difficult, particularly in high dimensions. As mentioned in the previous chapter, the IB optimisation problem has no available analytical solution in the general case. When all variables are discrete, it can be solved iteratively using the generalized Blahut-Arimoto algorithm which, however, requires to estimate the joint distribution of the potentially high-dimensional variables X and Y. A formal analysis of the difficulties of this estimation problem was conducted in Shamir et al. (2010). In the continuous case, this iterative algorithm is not applicable in practice, moreover, estimation of multivariate densities becomes arduous and can be a major impediment to the practical application of IB. A notable exception is the case of joint Gaussian (X, Y) for which an analytical solution for the optimal representation T exists. The optimal T is jointly Gaussian with (X, Y) and takes the form of a noisy linear projection of eigenvectors of the normalised conditional covariance matrix. The existence of an analytical solution opens new application possibilities and IB becomes practically feasible in higher dimensions (Hecht et al., 2009). Finding closed form solutions for other continuous distribution families remains an open challenge. The practical usefulness of the Gaussian IB (GIB), on the other hand, suffers from its missing flexibility and the statistical problem of finding a robust estimate of the joint covariance matrix of (X, Y) in high-dimensional spaces.

With the aim of extending the GIB analytical solution to a larger class of models, we present in this chapter a reformulation of the IB problem for continuous variables in terms of copulas. Compression and relevance in IB are defined in terms of mutual information (MI) which, as we will see, bears an interesting relationship to copulas: mutual information equals negative copula entropy (Ma and Sun, 2008). Interestingly, although these two concepts were developped independently, they rejoin because both aim at capturing the "pure" dependency structure of random variables. In this work, we demonstrate that IB is completely independent of the marginal distributions of X, Y. The IB problem in the continuous case is in fact to find the optimal copula (or dependence structure) of T and X, knowing the copula of (X, Y). We focus on the case of Gaussian copulas and on the consequences of the IB reformulation for the Gaussian IB. We show that the analytical solution available for GIB can naturally be extended to multivariate distributions with Gaussian copula and arbitrary marginal densities, also called *meta-Gaussian* densities. Moreover, we show that the GIB solution depends only a correlation matrix, and not on the variance. This allows us to use robust rank correlation estimators instead of unstable

covariance estimators, and gives a robust version of GIB. It opens new possible applications of IB to continuous data and provides a solution more robust to outliers.

6.2 Copula and Information Bottleneck

6.2.1 Copula formulation of IB.

At the heart of the copula formulation of IB is the following identity: for a continuous random vector $Z = (Z_1, \dots, Z_d)$ with density $f(z)$ and copula density $c_Z(u)$, *multi-information* is the negative differential entropy of the copula density (Ma and Sun, 2008):

$$\begin{aligned} I(Z) \equiv D_{kl}(f(z) \parallel f_0(z)) &= \int_{\mathbb{R}^d} \log\left(\frac{f(z)}{f_0(z)}\right) f(z)\mathrm{d}z, \\ &= \int_{\mathbb{R}^d} \log\left(\frac{\prod_{j=1}^d f_j(z) c_Z(F_1(z), \dots, F_d(z))}{\prod_{j=1}^d f_j(z)}\right) \prod_{j=1}^d f_j(z) c_Z(F_1(z), \dots, F_d(z))\mathrm{d}z, \\ &= \int_{[0,1]^d} c_Z(u) \log c_Z(u)\mathrm{d}u = -H(c_Z), \end{aligned} \tag{6.1}$$

where $u = (u_1, \dots, u_d) \in [0,1]^d$, D_{kl} denotes the Kullback-Leibler divergence, and $f_0(z) = f_1(z_1)f_2(z_2)\dots f_d(z_d)$. For continuous multivariate X, Y and T, equation (6.1) implies that:

$$\begin{aligned} I(X;T) &= D_{kl}(f(x,t) \parallel f_0(x,t)) - D_{kl}(f(x)||f_0(x)) - D_{kl}(f(t)||f_0(t)), \\ &= -H(c_{XT}) + H(c_X) + H(c_T), \\ I(Y;T) &= -H(c_{YT}) + H(c_Y) + H(c_T), \end{aligned}$$

where c_{XT} is the copula density of the vector $(X_1, \dots, X_p, T_1, \dots, T_p)$ and the first equation follows from Proposition 2.3. The above derivation then leads to the following proposition.

Proposition 6.1. Copula formulation of IB
For continuous variables the Information Bottleneck minimisation problem 5.1 can be reformulated as

$$\min_{c_{XT}} \mathcal{L} \mid \mathcal{L} = -H(c_{XT}) + H(c_X) + H(c_T) - \beta\{-H(c_{YT}) + H(c_Y) + H(c_T)\}. \tag{6.2}$$

The minimisation problem defined in 5.1 is solved under the assumption that the joint distribution of (X, Y) is known, this now translates in the assumption that the copula copula density c_{XY} (and thus c_X) is assumed to be known. The density c_T is entirely determined by c_{XT}, and using the conditional independence structure, we show that c_{YT} is also determined by c_{XT} and c_{XY}. Since the joint density of (X, Y, T) decomposes as

$$f(x,y,t) = f(t,y|x)f(x) = f(t|x)f(y|x)f(x), \tag{6.3}$$

and using Theorem 3.9 (Elidan, 2010), we see that the corresponding copula density then also decomposes as

$$c_{XYT}(u_x, u_y, u_t) = R_{T|X}(u_x, u_t) R_{Y|X}(u_x, u_y) c_X(u_x), \tag{6.4}$$

where

$$R_{T|X}(u_x, u_t) = \frac{c_{XT}(u_x, u_t)}{c_X(u_x)}, \tag{6.5}$$

$$R_{Y|X}(u_x, u_y) = \frac{c_{XY}(u_x, u_y)}{c_X(u_x)}, \tag{6.6}$$

$$u_x \in [0,1]^p, u_y \in [0,1]^q, u_t \in [0,1]^p. \tag{6.7}$$

We can finally rewrite the copula density of (Y,T) as

$$c_{YT}(u_y,u_t)=\int c_{XYT}(u_x,u_y,u_t)\mathrm{d}u_x=\int \frac{c_{XT}(u_x,u_t)c_{XY}(u_x,u_y)}{c_X(u_x)}\mathrm{d}u_x, \tag{6.8}$$

which shows that c_{YT} is indeed fully determined by c_{XT} and c_{XY}. The IB optimisation problem then reduces to finding an optimal copula density c_{XT} and in order to construct the compression variable T, the only relevant aspect is the copula dependence structure between X,T and Y.

6.3 Meta-Gaussian IB

6.3.1 Meta-Gaussian IB formulation

The above reformulation of IB is of great practical interest when we focus on the special case of the Gaussian copula. The only known case for which a simple analytical solution to the IB problem exists is when (X,Y) are joint Gaussians. Equation (6.2) shows that actually an optimal solution does not depend of the margins but only on the copula density c_{XY}. From this observation the idea naturally follows that an analytical solution should also exist for any joint distribution of (X,Y) which has a Gaussian copula, and that regardless of its margins. We show below in Proposition 6.2 that this is indeed the case. The notation $\tilde{X}$ and $\tilde{Y}$ is used to represent the normal scores:

$$\tilde{X}=(\Phi^{-1}\circ F_{X_1}(X_1),\dots,\Phi^{-1}\circ F_{X_p}(X_p)). \tag{6.9}$$

Since copulas are invariant to strictly increasing transformations the normal scores have the same copulas as the original variables X and Y.

Proposition 6.2. Optimality of meta-Gaussian IB
Consider rv X,Y with a Gaussian dependence structure and arbitrary margins:

$$F_{X,Y}(x,y)\sim C_P(F_{X_1}(x_1),\dots,F_{X_p}(x_p),F_{Y_1}(y_1),\dots,F_{Y_q}(y_q)), \tag{6.10}$$

where F_{X_i},F_{Y_i} are the marginal distributions of X,Y and C_P is a Gaussian copula parametrized by a correlation matrix P. Then the optimum of the IB minimisation problem is obtained for $T\in\mathcal{T}$, where $\mathcal{T}$ is the set of all rv T such that (X,Y,T) has a Gaussian copula and T has Gaussian margins.

Before proving proposition 6.2 we give a short lemma.

Lemma 6.1. *$T\in\mathcal{T}\Leftrightarrow(\tilde{X},\tilde{Y},T)$ are jointly Gaussian.*

Proof.
1. If $T\in\mathcal{T}$ then (X,Y,T) has a Gaussian copula which implies that $(\tilde{X},\tilde{Y},T)$ also has a Gaussian copula. Since $\tilde{X},\tilde{Y},T$ all have normally distributed margins it follows that $(\tilde{X},\tilde{Y},T)$ has a joint Gaussian distribution.
2. If $(\tilde{X},\tilde{Y},T)$ are jointly Gaussian then $(\tilde{X},\tilde{Y},T)$ has a Gaussian copula which implies that (X,Y,T) has again a Gaussian copula. Since T has normally distributed margins, it follows that $T\in\mathcal{T}$.

□

Proposition 6.2 can now be proven by contradiction.

Proof of proposition 6.2. Assume there exists $T^* \notin \mathcal{T}$ such that:

$$\mathcal{L}(X,Y,T^*) := I(X;T^*) - \beta I(Y;T^*) < \min_{p(t|x),T\in\mathcal{T}} I(X;T) - \beta I(T;Y) \tag{6.11}$$

Since $(\tilde{X},\tilde{Y},T)$ has the same copula as (X,Y,T), we have that $I(\tilde{X};T) = I(X;T)$ and $I(\tilde{Y};T) = I(Y;T)$. Using Lemma 6.1 the right hand part of inequality (3.17) can be rewritten as :

$$\min_{p(t|x),T\in\mathcal{T}} \mathcal{L}(X,Y,T) = \min_{p(t|x),T\in\mathcal{T}} \mathcal{L}(\tilde{X},\tilde{Y},T) = \min_{p(t|\tilde{x}),(\tilde{X},\tilde{Y},T)\sim\mathcal{N}} \mathcal{L}(\tilde{X},\tilde{Y},T). \tag{6.12}$$

Combining equations (3.17) and (6.12) we obtain:

$$I(\tilde{X};T^*) - \beta I(\tilde{Y};T^*) < \min_{p(t|\tilde{x}),(\tilde{X},\tilde{Y},T)\sim\mathcal{N}} I(\tilde{X};T) - \beta I(T;\tilde{Y}).$$

This is in contradiction with the optimality of Gaussian information bottleneck, which states that the optimal T is jointly Gaussian with (X,Y). Thus the optimum for meta-Gaussian (X,Y) is attained for T with normal margins such that (X,Y,T) also is meta-Gaussian.

□

Corollary 6.1. *The optimal projection T^o obtained for $(\tilde{X},\tilde{Y})$ is also optimal for (X,Y).*

Proof. By the above we know that an optimal compression for (X,Y) can be obtained in the set of variables T such that $(\tilde{X},\tilde{Y},T)$ is jointly Gaussian, since $\tilde{\mathcal{L}} = \mathcal{L}$ it is clear that T^o is also optimal for (X,Y). □

As a consequence of Proposition 6.2, for any random vector (X,Y) having a Gaussian copula dependence structure, an optimal projection T can be obtained by first calculating the vector of the normal scores $(\tilde{X},\tilde{Y})$ and then computing $T = A\tilde{X} + \xi$. A is here entirely determined by the covariance matrix of the vector $(\tilde{X},\tilde{Y})$ which also equals its correlation matrix (the normal scores have unit variance by definition), and thus the correlation matrix P parametrizing the Gaussian copula C_P. In practice the problem is reduced to the estimation the Gaussian copula of (X,Y). In particular, for the traditional Gaussian case where $(X,Y) \sim \mathcal{N}(0,\Sigma)$, this means that we actually do not need to estimate the full covariance Σ but only the correlations. In summary, the main idea in MGIB is that the IB problem for meta-Gaussian data can be solved by applying GIB in the space of the normal scores $(\tilde{X},\tilde{Y})$, and we can apply GIB to meta-Gaussian data as long as we can make inference on two elements: P and the underlying Gaussian variables, which in the continuous case are simply the normal scores $(\tilde{X},\tilde{Y})$.

6.3.2 Meta-Gaussian mutual information

We derive in this section an expression for the mutual information of a meta-Gaussian pair X,Y. The multi-information for a meta-Gaussian random vector $Z = (Z_1,\ldots,Z_d)$ with copula C_{P_z}.

$$I(Z) = I(\tilde{Z}) = -\tfrac{1}{2}\log|\mathrm{cov}(\tilde{Z})| = -\tfrac{1}{2}\log|\Sigma_{\tilde{z}}| = -\tfrac{1}{2}\log|\mathrm{corr}(\tilde{Z})| = -\tfrac{1}{2}\log|P_z|, \tag{6.13}$$

where $|.|$ denotes the determinant. The mutual information between X and Y is then

$$I(X;Y) = -\tfrac{1}{2}\log|P| + \tfrac{1}{2}\log|P_x| + \tfrac{1}{2}\log|P_y|, \quad \text{with } P = \begin{pmatrix} P_x & P_{yx} \\ P_{xy} & P_y \end{pmatrix}. \tag{6.14}$$

It is obvious that the formula for the meta-Gaussian is similar to the formula for the Gaussian case

$$I_{\mathrm{Gauss}}(X;Y) = -\tfrac{1}{2}\log|\Sigma| + \tfrac{1}{2}\log|\Sigma_x| + \tfrac{1}{2}\log|\Sigma_y|, \tag{6.15}$$

but uses the correlation matrix parametrizing the copula instead of the data covariance matrix. The two formulas are equivalent when X, Y are jointly Gaussian.

In the following we additonally provide a direct derivation of the multi-information for a meta-Gaussian random vector $Z = (Z_1, \ldots, Z_d)$.

$$I(Z) = D_{kl}(f(z) \parallel f_0(z)) = \int_{[0,1]^d} c_Z(u) \log c_Z(u) \mathrm{d}u, \quad u = (u_1, \ldots, u_d),$$
$$= \int |P|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\Phi^{-1}(u))^T (P^{-1} - I)\Phi^{-1}(u)\right\} \log\left[|P|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\Phi^{-1}(u))^T (P^{-1} - I)\Phi^{-1}(u))\right\}\right] \mathrm{d}u,$$

where $\Phi^{-1}(u) = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d))$. We use the change of variable $g(\tilde{z}) = g(\tilde{z}_1, \ldots, \tilde{z}_d) := (\Phi(\tilde{z}_1), \ldots, \Phi(\tilde{z}_d)) = (u_1, \ldots, u_d)$. The Jacobian matrix of the transformation is diagonal with elements $\mathrm{D}g(\tilde{z})_{jj} = \Phi'(\tilde{z}_j)$ and its determinant is $\det(\mathrm{D}g) = 2\pi^{-d/2} \exp\{-\frac{1}{2}\tilde{z}^T I \tilde{z}\}$. We then obtain:

$$\begin{aligned}
I(Z) &= \int_{-\infty}^{+\infty} |P|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\tilde{z}^T (P^{-1} - I)\tilde{z}\right\} \log\left[|P|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\tilde{z}^T (P^{-1} - I)\tilde{z}\right\}\right] \det(\mathrm{D}g) \mathrm{d}\tilde{z}, \\
&= \int |P|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\tilde{z}^T P^{-1}\tilde{z}\right\} \exp\left\{\tfrac{1}{2}\tilde{z}^T I \tilde{z}\right\} \left[\log(|P|^{-\frac{1}{2}}) - \tfrac{1}{2}\tilde{z}^T (P^{-1} - I)\tilde{z}\right] 2\pi^{d/2} \exp\{-\tfrac{1}{2}\tilde{z}^T I \tilde{z}\} \mathrm{d}\tilde{z}, \\
&= \int 2\pi^{d/2} |P|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\tilde{z}^T P^{-1}\tilde{z}\right\} \left[\log(|P|^{-\frac{1}{2}}) - \tfrac{1}{2}\tilde{z}^T (P^{-1} - I)\tilde{z}\right] \mathrm{d}\tilde{z}, \\
&= \mathrm{E}_{\mathcal{N}(0,P)} \left[\log(|P|^{-\frac{1}{2}}) - \tfrac{1}{2}\tilde{z}^T (P^{-1} - I)\tilde{z}\right] \\
&= \log(|P|^{-\frac{1}{2}}) - \tfrac{1}{2}\mathrm{E}_{\mathcal{N}(0,P)}[\tilde{z}^T P^{-1}\tilde{z}] + \tfrac{1}{2}\mathrm{E}_{\mathcal{N}(0,P)}[\tilde{z}^T I \tilde{z}], \\
&= \log(|P|^{-\frac{1}{2}}) - \tfrac{1}{2}d + \tfrac{1}{2}d \\
&= -\tfrac{1}{2}\log|P|.
\end{aligned}$$

6.3.3 Semi-parametric copula estimation

As explained in Section 6.3.1, to apply MGIB we need to perform inference on the correlation matrix P and on the underlying standard Gaussian variables. In the continuous case, the hidden variables are given by the normal scores which suggests using the empirical cumulative distribution function and a semi-parametric estimation framework. Semi-parametric copula estimation has been studied in Genest et al. (1995), Tsukahara (2005) and Hoff (2007). The main idea is to combine non-parametric estimation of the margins with a parametric copula model, in our case the Gaussian copulas family. If the margins $F_1, \ldots, F_d$ of a random vector Z are known, P can be estimated by the matrix $\hat{P}$ with elements given by:

$$\hat{P}_{(k,l)} = \frac{\frac{1}{n}\sum_{i=1}^n \Phi^{-1}(F_k(z_{ik}))\Phi^{-1}(F_l(z_{il}))}{\left[\frac{1}{n}\sum_{i=1}^n \left[\Phi^{-1}(F_k(z_{ik}))\right]^2 \frac{1}{n}\sum_{i=1}^n \left[\Phi^{-1}(F_l(z_{il}))\right]^2\right]^{1/2}}, \tag{6.16}$$

where z_{ik} denotes the i-th observation of dimension k. $\hat{P}$ is assured to be positive semi-definite. If the margins are unknown we can instead use the rescaled empirical cumulative distributions:

$$\hat{F}_j(t) = \frac{n}{n+1}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{z_{ij} \leq t}\right). \tag{6.17}$$

The estimator resulting from using the rescaled empirical distributions (6.17) in equation (6.16) is given in the following definition.

Definition 6.1 (Normal scores rank correlation coefficient)**.** *The normal scores rank correlation coefficient is the matrix $\hat{P}^n$ with elements:*

$$\hat{P}^n_{(k,l)} = \frac{\sum_{i=1}^n \Phi^{-1}(\frac{R(z_{ik})}{n+1})\Phi^{-1}(\frac{R(z_{il})}{n+1})}{\sum_{i=1}^n \left(\Phi^{-1}(\frac{i}{n+1})\right)^2}, \tag{6.18}$$

where $R(z_{ik})$ denotes the rank of the i-th observation for dimension k. The estimator (6.18) have been studied in Boudt et al. (2012) showing good efficiency and robustness properties, also comparing favourably with Kendall and Spearman correlation measures. Using (6.18) we compute an estimate of the correlation matrix P parametrizing c_{XY} and obtain the transformation matrix A as detailed in Algorithm 6.

Algorithm 6 Construction of the transformation matrix A

1. Compute the normal scores rank correlation estimate $\hat{P}^n$ of the correlation matrix P parametrizing c_{XY}:
for $k, l = 1, \ldots, p+q$ **do**
Set the (k,l)-*th* element of $\hat{P}^n$ to $\frac{\sum_{i=1}^n \Phi^{-1}(\frac{R(z_{ik})}{n+1})\Phi^{-1}(\frac{R(z_{il})}{n+1})}{\sum_{i=1}^n \left(\Phi^{-1}(\frac{i}{n+1})\right)^2}$ as in equation (6.18) and where the i-th row of z is the concatenation of the i-th rows of x and y: $z_{i*} = (x_{i*}, y_{i*}) \in \mathbb{R}^{p+q}$.
end for
2. Compute the estimated conditional covariance matrix of the normal scores: $\hat{\Sigma}_{\tilde{x}|\tilde{y}} = \hat{P}^n_x - \hat{P}^n_{xy}(\hat{P}^n_y)^{-1}\hat{P}^n_{yx}$.
3. Find the eigenvectors and eigenvalues of $\hat{\Sigma}_{\tilde{x}|\tilde{y}}(\hat{P}^n_x)^{-1}$.
4. Construct the transformation matrix A as in equation (5.14).

6.4 Results

6.4.1 Simulations

We tested meta-Gaussian IB (MGIB) in two different setting, first when the data is Gaussian but contains outliers, second when the data has a Gaussian copula but non-Gaussian margins. We generated a training sample with $n = 1000$ observations of X and Y with dimensions fixed to $d_x = 15$ and $d_y = 15$. A covariance matrix was drawn from a Wishart distribution centered at a correlation matrix populated with a few high correlation values to ensure some dependency between X and Y. This matrix was then scaled to obtain the correlation matrix parametrizing the copula. In the first setting the data was sampled with $\mathcal{N}(0,1)$ margins. A fixed percentage of outliers, 8%, was then introduced to the sample by randomly drawing a row and a column in the data matrix and replacing the current value with a random draw from the set $[-6,-3]\cup[3,6]$. In the second setting data points were drawn from meta-Gaussian distributions with three different type of margins: Student with $df = 4$, exponential with $\lambda = 1$, and beta with $\alpha_1 = 0.5 = \alpha_2$. For each training sample two projection matrices A_G and A_C were computed, A_G was calculated based on the sample covariance $\hat{\Sigma}^n$ and A_C was obtained using the normal scores rank correlation $\hat{P}^n$. The compression quality of the projection was then tested on a test sample of $n = 10'000$ observations generated independently from the same distribution (without outliers). Each experiment was repeated 50 times. Figure 6.1 shows the information curves obtained by varying β from 0.1 to 200. The mutual informations $I(X;T)$ and $(Y;T)$ can be reliably estimated on the test sample using (6.13) and (6.18). The information curves start with a very steep slope, meaning that a small increase in $I(X;T)$ leads to a significant increase in $I(Y;T)$, and then slowly saturate to reach

their asymptotic limit in $I(Y;T)$. The best information curves are situated in the upper left corner of the figure, since for a fixed compression value $I(X;T)$ we want to achieve the highest relevant information content $(I;T)$. We clearly see in Figure 6.1 that MGIB consistently outperforms GIB in that it achieves higher compression rates.

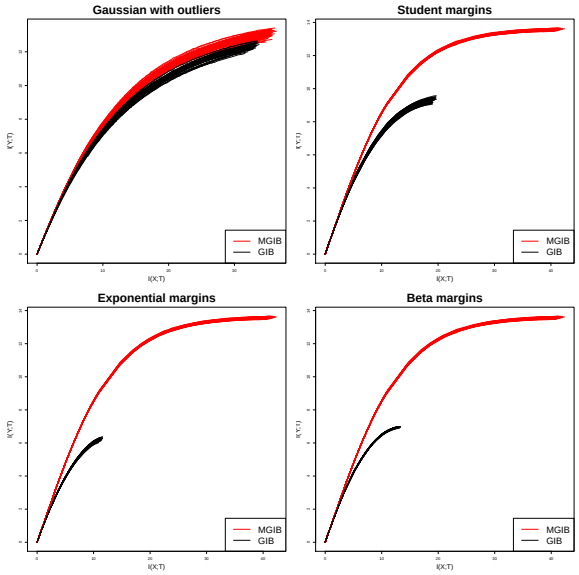

Figure 6.1: Information curves for Gaussian data with outliers, data with Student, Exponential and Beta margins. Each panel shows 50 curves obtained for repetitions of the MGIB (red) and the GIB (black). The curves stop when they come close to saturation. For higher values of β the information $I(X;T)$ would continue to grow while $I(Y;T)$ would reach its limit leading to horizontal lines, but such high beta values lead to numerical instability. Since GIB suffers from a model mismatch problem when the margins are not Gaussian, the curves saturate for smaller values of $I(Y;T)$.

6.4.2 Real data

We further applied MGIB to the *Communities and Crime* data set from the UCI repository [1]. The data set contains observations of predictive and target variables. After removing missing values we retained $n = 2195$ observations. In a pre-processing step we selected the $d_x = 10$ dimensions with the strongest absolute rank correlation to one of the relevance variables. Plotting empirical information curves as in the synthetic examples above was impossible, because even for this setting with drastically decreased dimensionality all mutual information estimates we tried (including the nearest-neighbor graph method in Pál et al. (2010)) were too unstable to draw empirical information curves. To still give a graphical representation of our results we show in Figure 6.2 non-parametric density estimates of the one dimensional compression T split in 5 groups according to corresponding values of the first relevance variable. We used GIB, MGIB and Principal Component analysis (PCA) to reduce X to a 1-dimensional variable. For PCA this is the first principal component, for GIB and MGIB we independently selected the highest value of β leading to a 1-dimensional compression. It is obvious from Figure 6.2 that the one-dimensional MGIB compression nicely separates the different target classes, whereas the GIB and PCA projections seem to contain much less information about the target variable. We conclude that similar to our synthetic examples above, the MGIB compression contains more information about the relevance variable than GIB at the same compression rate.

[1] http://archive.ics.uci.edu/ml/

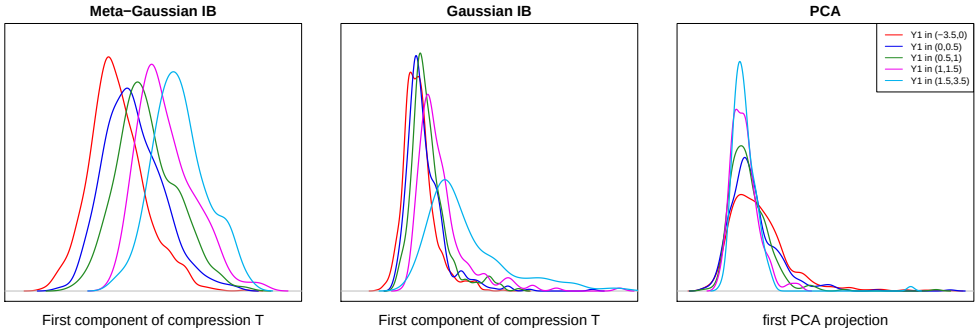

Figure 6.2: Parzen density estimates of the univariate projection of X split in 5 groups according to values of the first relevance variable. We see more separation between groups for MGIB than for GIB or PCA, which indicates that the projection is more informative about the relevance variable.

6.5 Conclusion

We present a reformulation of the IB problem in terms of copula which gives new insights into data compression with relevance constraints and opens new possible applications of IB for continuous multivariate data. Meta-Gaussian IB naturally extends the analytical solution of Gaussian IB to multivariate distributions with Gaussian copula and arbitrary marginal density. It can be applied to any type of continuous data, provided the assumption of a Gaussian dependence structure is reasonable, in which case the optimal compression can easily be obtained by semi-parametric copula estimation. Simulated experiments showed that MGIB clearly outperforms GIB when the marginal densities are not Gaussian, and even in the Gaussian case with a tiny amount of outliers MGIB has been shown to significantly benefit from the robustness properties of rank estimators. In future work, it would be interesting to see if the copula formulation of IB admits analytical solutions for other copula families.

Chapter 7

Sparse Meta-Gaussian information bottleneck

7.1 Introduction

Dimensionality reduction is an important domain of research for which a large variety of techniques have been developed. Given observations of a multivariate random variable X, the aim is to construct a new representation of the data with reduced dimensionality. Amongst the most prominent methods, Principal Component Analysis (PCA) and Canonical Component Analysis (CCA) have been extensively studied and extended. Other classical techniques include Factor Analysis and methods for manifold modeling. Every dimensionality reduction technique first needs to define what is important in the data and should therefore be preserved, e.g. PCA aims at preserving the variance. In this respect, dimensionality reduction is related to the data compression problem where the question of defining what is relevant is implicitly answered by the choice of a distortion function, i.e. a function evaluating the loss incurred by compression. As seen in Chapter 5, an interesting alternative to distortion functions is given by the Information bottleneck method (IB). Instead of evaluating the distortion between the compression T and original data X, IB introduces a relevance variable Y. An important advantage of IB is that compression and information are solely expressed in information-theoretic quantities. While some choices of distortion function or dependence measure (e.g. correlation) are not suitable for certain types of data, mutual information is a very general and theoretically well-founded dependence measure (Joe, 1989). In this chapter, we present a new sparse compression technique based on the information bottleneck principle, i.e. we perform feature selection with relevance information. This is achieved by introducing a sparse variant of IB in which T is built using only a few selected dimensions of the original data. Efficient IB algorithms were limited to discrete data before the introduction of Gaussian IB (GIB), which has later been generalised to the continuous meta-Gaussian distributions as explained in Chapter 6. However, the compression achieved by existing IB methods is usually not sparse and, therefore cannot be used for feature selection. Our model is an extension of MGIB, we impose sparsity on the projection obtained through MGIB and thereby select a sparse set of features. Our method shares some similarities with sparse regression techniques like Lasso (Tibshirani, 1996) but is based on different, less restrictive assumptions and achieves sparsity without imposing any norm penalty. To our knowledge, there exists no IB method to accommodate mixed distributions, i.e. distributions with continuous and discrete margins, without resorting to discretisation of the continuous dimensions. Besides introducing sparsity, we extend MGIB to mixed data by considering discrete margins as the result of a discretisation process of hidden continuous variables. This extension is motivated by the high prevalence of mixed data in many domains where feature selection is sought, especially in the medical and biological fields (de Leon and Chough, 2013).

7.2 Sparse IB

We first assume that the underlying Gaussian variables of our copula model and P are known. As explained in Chapter 6, we can assume that these hidden variables are centered and scaled, i.e. are $\mathcal{N}(0, P)$ distributed. Inference will be discussed in Section 7.3. To achieve sparse compression we consider the MGIB model, for which the optimal compression is given by $AX + \xi$, but restrict A to the class of diagonal matrices. If we denote by $a = (a_1, \ldots, a_p) \in \mathbb{R}^p$ the vector of the diagonal entries of A, the resulting projection is the vector $AX = (a_1 X_1, \ldots, a_p X_p)$. By varying the Lagrange parameter of the minimisation problem, the vector a becomes sparse and thereby selects only a few dimensions of X. As for the case of a full projection matrix we can assume that the noise components are uncorrelated with unit variance i.e. $\Sigma_\xi = I$. Our IB minimisation problem for standard normal variables is

$$\min_{A: A \in \mathbb{D}_p^+} \mathcal{L} | \mathcal{L} \equiv I(X;T) - \beta I(T;Y), \tag{7.1}$$

where $\mathbb{D}_p^+$ is the set of positive diagonal matrices as explained below and (see equation (5.19))

$$I(X;T) = 2\log(|AP_xA^t + I|), \tag{7.2}$$

$$I(Y;T) = 2\log(|AP_xA^t + I|) - 2\log(|AP_{x|y}A^t + I|). \tag{7.3}$$

Optimisation here is simplified by two convenient properties:

1. For any symmetric positive definite matrix B and diagonal matrix D, $\log|DBD + I| = \log|BD^2 + I|$ depends only on the squared entries D_{ii}^2. It is therefore sufficient to consider the space of positive diagonal matrices in $\mathbb{R}^p$, denoted by $\mathbb{D}_p^+$. In the following we use the notation $A := D^2$.
2. $\log|BA + I|$ is *concave* in A and strictly monotone increasing in every component A_{ii}. Concavity directly follows from the concavity of $\log|B|$ and the fact that $A \mapsto BA + I$ is an affine function.

The variational problem (7.1) can be rewritten as a constrained optimisation problem (see also (Chechik et al., 2005)):

$$\min_{A: A \in \mathbb{D}_p^+} I(X;T) \quad \text{s.t.} \quad I(T;Y) \geq \kappa', \tag{7.4}$$

for some $\kappa' \geq 0$. Using the Schur complement formula for conditional Gaussian covariance we obtain

$$\min_{A: A \in \mathbb{D}_p^+} \underbrace{\log|P_xA + I|}_{2I(X;T)} \quad \text{s.t.} \quad \underbrace{\log|P_xA + I| - \log|QA + I|}_{2I(T;Y)} \geq \kappa', \tag{7.5}$$

where $Q = P_x - P_{xy}P_y^{-1}P_{xy}^T$ is the conditional covariance matrix of X given Y. The corresponding Lagrangian then is

$$\mathcal{L}'(A, \beta) = \log|P_xA + I| - \sum_{j=1}^{p} \eta_j A_{jj} - \beta(\log|P_xA + I| - \log|QA + I| - \kappa'), \tag{7.6}$$

where η_j is the Lagrange parameter for the j-th non-negativity constraint. Nontrivial solutions for the original IB with full matrix A exist only for $\beta > 1$, see Chechik et al. (2005). We transform the IB problem into an equivalent form that exchanges objective function and constraint. Introducing a new Lagrange parameter $\lambda = (\beta - 1)/\beta$, $0 < \lambda < 1$, and dividing (7.6) by β, we arrive at the new Lagrangian

$$\mathcal{L}(A, \lambda) = \log|QA + I| - \sum_{j=1}^{p} \epsilon_j A_{jj} + \lambda(\kappa - \log|P_xA + I|), \tag{7.7}$$

with $\epsilon_j = \frac{\eta_j}{\beta}, \kappa \geq 0$. The corresponding minimisation problem can then be rewritten as

$$\min_{A: A \in \mathbb{D}_p^+} \underbrace{\log|QA + I|}_{:=f(a)} \quad \text{s.t.} \quad \underbrace{\log|P_x A + I|}_{:=g(a)} \geq \kappa, \tag{7.8}$$

which amounts to minimising a concave function $f(a)$ over a convex set $\{b \in \mathbb{R}^p | g(b) \geq \kappa\}$. Thus, the global minimum is attained at the boundary $g(a) = \kappa$. Note that for $\kappa = 0$ the constraint is always satisfied and there is a unique minimum at $a = 0$. While we cannot characterise all stationary points of the Lagrangian problem as formulated in (7.6), the minimisation problem (7.8) has a particular form (concave function over a convex set) for which algorithms with guaranteed convergence to a globally optimal solution exist (Benson and Horst, 1991). A Matlab code for this method is available online[1]. However, this algorithm is not efficient in higher dimensions and we therefore propose a log barrier interior point method detailed later in Algorithm 7.

The solution set S of optimisation problem (7.8) is defined as the set of points a^* in the non-negative orthant of $\mathbb{R}^p$ which are global minima for a certain value of κ:

$$S = \{a^* \in \mathbb{R}_+^p \text{ s.t. } \exists\, \kappa \geq 0 \text{ for which } a^* \text{ is a solution of (7.8)}\}.$$

In the following theorem, we show that for an interval $[0, \kappa_2^c]$, S is a curve parametrized by κ, meaning that to every κ in this interval corresponds a unique point in S. In the following we assume that P_x and P_y are random matrices of maximal rank and write $\Phi := Q^{-1}, \Psi := P_x^{-1}$. We denote points in S by $a^* = (a_1^*, \ldots, a_p^*)$.

Theorem 7.1. *With probability one, there exist $\kappa_2^c > \kappa_1^c > 0$ such that:*

1. *If $\kappa \in [0, \kappa_1^c]$ then*
$$a_i^* = \begin{cases} e^\kappa - 1 & \text{if } i = \operatorname{argmin}_j(Q_{jj}) =: i_f, \\ 0 & \text{else.} \end{cases}$$
2. *If $\kappa \in [\kappa_1^c, \kappa_2^c]$ then*
$$a_i^* = \begin{cases} G(\kappa; \Psi, \Phi) & \text{if } i = i_s, \\ c_1 a_{i_s}^* + c_0 & \text{if } i = i_f, \\ 0 & \text{else} \end{cases}$$

where $c_1 = \frac{|\Psi| - \Phi_{22}}{|\Psi| - \Phi_{11}}$, $c_0 = \frac{|\Psi|(\Phi_{11} - \Phi_{22})}{|\Psi| - \Phi_{11}}$ and

$$G(\kappa; \Psi, \Phi) = -\frac{|\Psi|(1 + c_1) + c_0}{2c_1} + \left(\frac{(|\Psi|(1 + c_1) + c_0)^2}{4c_1^2} - \frac{|\Psi|(c_0 + 1 - e^\kappa)}{c_1} \right)^{0.5},$$

The value of i_s can be determined by searching over the $p - 1$ possible combinations (i_f, i) and choosing the dimension i_s which gives the minimal value of $f(a)$.

We use the term critical values to designate κ_1^c, κ_2^c. We call i_f the most informative dimension of X and i_s the second most informative dimension. Theorem 7.1 tells us that, for small enough κ values, the solution set is a curve parametrized by κ which starts at the point zero, runs along the i_f-axis until c_0 is reached, and then takes the form of a straight line with slope c_1, see Figure 7.1 for an illustration. We therefore call S the solution path. We prove this result in three steps given by Lemma 7.1, Lemma 7.2 and Lemma 7.3.

Lemma 7.1 (Most informative dimension)**.** *The most informative dimension is the dimension i_f with the smallest corresponding entry in Q, i.e. $i_f = \operatorname{argmin}_i(Q_{ii})$. Moreover, when only one component i_f of a is non-zero, $a_{i_f}^* = e^\kappa - 1$.*

[1] http://www.mathworks.com/matlabcentral/fileexchange/36247-function-for-global-minimisation-of-a-concave-function

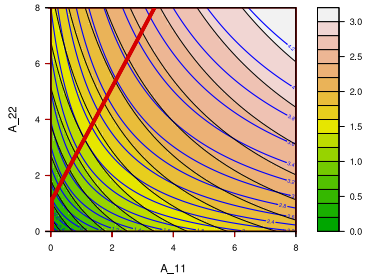

Figure 7.1: Objective function f (filled contours, colour coded) and constraint g (thick blue contour lines). The solution path is depicted by the red line. Above the critical constraint value (in this example the $\kappa_1^c \approx 0.8$ contour line), all solutions lie on the line $a_2^* = c_1 a_1^* + c_0$ with positive slope c_1, below the critical value, the solution path is a vertical line at the origin, i.e. solutions are sparse ($a_1^* = 0$).

Proof. For every possible choice of $i = 1, \ldots, p$, the value of a_i is uniquely determined by the constraint:

$$\kappa = g(a_i) = \log((P_x)_{ii} a_i + 1) = \log(a_i + 1), \tag{7.9}$$

which implies that $a_i = e^\kappa - 1$. The optimal i_f is then determined by the minimum value of $f(a_i) = \log(Q_{ii} a_i + 1) = \log(Q_{ii}(e^\kappa - 1) + 1)$. □

Lemma 7.2 bellow provides an analytical solution in the case $p = 2$. In the following we denote the partial derivatives of a real function f of a by $\frac{\partial f}{\partial a_i}(a) = f_{a_i}$.

Lemma 7.2 (2-dimensional case)**.** *Assume $p = 2$ and, w.l.o.g. that the most informative dimension is $i_f = 2$. Let $c_1, c_0, G(\kappa; \Psi, \Phi)$ be defined as in Theorem 7.1. The first critical value is $\kappa_1^c = \log(c_0 + 1)$, and with probability one, $\kappa_1^c > 0$. Moreover, the optimal a^* is given*

1. *for every $\kappa \in [\kappa_1^c, \infty)$ by*

$$a_2^* = c_1 a_1^* + c_0, \tag{7.10}$$
$$a_1^* = G(\kappa; \Psi, \Phi), \tag{7.11}$$

2. *for every $\kappa \in [0, \kappa_1^c]$ by $a_2^* = e^\kappa - 1, a_1^* = 0$.*

Proof. We first determine the set of stationary points with strictly positive components. When $a_i > 0, i = 1, 2$, the non-negativity constraints are inactive and $\epsilon_i = 0, i = 1, 2$. Stationary points are characterised by a vanishing Lagrangian gradient $\nabla \mathcal{L} = 0$, which here means that $\nabla f(a) = \lambda \nabla g(a)$. When $p = 2$, this proportionality condition is equivalent to the orthogonality condition $-g_{a_2} f_{a_1} + g_{a_1} f_{a_2} = 0$. Adding the constraint $g(a) = \kappa$ leads to a system of 2 equations in 3 variables (a_1, a_2 and κ) which implicitly defines the set of all stationary points with strictly positive components:

$$\begin{cases} \tilde{H}_{12} & : \quad f_{a_2} g_{a_1} - f_{a_1} g_{a_2} = 0 \\ \tilde{H}_0 & : \quad g(a) - \kappa = 0 \end{cases}, \tag{7.12}$$

To solve the above system we first need to to compute the partial derivatives f_{a_i}, g_{a_i}. Rewriting $f(a) = \log(|Q||Q^{-1} + A|) = \log|Q| + \log|\Phi + A|$ and $g(a) = \log|\mathrm{corr}_x| + \log|\Psi + A|$, we directly obtain

$$f_{a_j} = \frac{|\Phi + A|^{[-j]}}{|\Phi + A|}, \quad g_{a_j} = \frac{|\Psi + A|^{[-j]}}{|\Psi + A|}. \tag{7.13}$$

From expressions 7.13 we can see that $\tilde{H}_{12}$ can advantageously be replaced by

$$H_{12} : |\Phi + A||\Psi + A|(f_{a_2} g_{a_1} - f_{a_1} g_{a_2}) = 0. \tag{7.14}$$

Further, $\tilde{H}_0$ can be replaced by $H_0 : |P_x||\Psi + A| - e^{\kappa} = 0$. We solve system (7.12) in two steps. First, using H_{12} we obtain an expression for a_2 as a function of a_1, leading to equation (7.10). Second, we replace a_2 in H_0 by expression (7.10), leading to equation (7.11). Finally, we can compute the critical value κ_1^c by solving $a_1^* = G(\kappa; \Psi, \Phi) = 0$ for κ.

A straightforward derivation leads to

$$H_{12} = (\Phi_{11} + a_1)(\Psi_{22} + a_2) - (\Phi_{22} + a_2)(\Psi_{11} + a_1).$$

Setting H_{12} to zero gives:

$$a_2 = \left(\frac{\Psi_{22} - \Phi_{22}}{\Psi_{11} - \Phi_{11}}\right) a_1 + \frac{\Phi_{11}\Psi_{22} - \Phi_{22}\Psi_{11}}{\Psi_{11} - \Phi_{11}},$$

from which we can identify

$$c_1 = \frac{\Psi_{22} - \Phi_{22}}{\Psi_{11} - \Phi_{11}}, \quad c_0 = \frac{\Phi_{11}\Psi_{22} - \Phi_{22}\Psi_{11}}{\Psi_{11} - \Phi_{11}}. \tag{7.15}$$

Noting that $\Psi_{11} = \Psi_{22} = |\Psi|$, we find the final expressions for c_0 and c_1.

We further set H_0 to zero and replace a_2 by $c_1 a_1 + c_0$ to obtain:

$$\begin{aligned}
&|P_x||\Psi + A| = |\Psi|^{-1}|\Psi + A| = e^{\kappa}, \\
&(\Psi_{11} + a_1)(\Psi_{22} + c_1 a_1 + c_0) - \Psi_{12}^2 = |\Psi| e^{\kappa}, \\
&a_1^2 + a_1 \frac{|\Psi|(1 + c_1) + c_0}{c_1} + \frac{|\Psi|(c_0 + 1 - e^{\kappa})}{c_1} = 0.
\end{aligned} \tag{7.16}$$

We can recognise in (7.16) a quadratic equation of the form $a_1^2 + r a_1 + s = 0$. Since we assumed that dimension 2 is the most informative we have that $c_0 > 0$ and the solution for a_1 is then given by $a_1 = -\frac{r}{2} + \left(\frac{r^2}{4} - s\right)^{0.5}$ leading directly to equation (7.11). Finally, we can compute the critical value κ_1^c by noting that $a_1 = 0$ is equivalent to $|\Phi|(c_0 + 1 - e^{\kappa})/c_1 = 0$ which implies that $c_0 + 1 - e^{\kappa_1^c} = 0$.

Solutions for $\kappa \leq \kappa_1^c$ are given by Lemma 7.1. The probability of c_0 being equal to zero is null, i.e. $\mathbb{P}\{c_0 = 0\} = 0$, since $\mathbb{P}\{Q_{11} = Q_{22}\} = 0$ for random correlation matrices P_x and P_y. This implies that $\kappa_1^c > 0$ with probability one and there always exist values of κ for which the 1-dimensional solution is optimal. □

We come back to the general p-dimensional case with the following result.

Lemma 7.3 (General p-dimensional case)**.** *In the p-dimensional problem, with probability one, the following statements hold:*

1. *There is an non-empty interval $[\kappa_1^c, \kappa_3^c]$ for which the global optimum is attained with two active dimensions.*
2. *There is an non-empty interval $[\kappa_1^c, \kappa_2^c]$ for which the global optimum is attained in a fixed 2-dimensional subspace.*

Proof. Assume w.l.o.g. that the most informative dimension is $i_f = 1$. As seen in the proof of Lemma 7.2, stationary points with strictly positive components are characterised by $\nabla f(a) = \lambda \nabla g(a)$. A new dimension $j \neq 1$ becomes active when $f_{a_j} = \lambda g_{a_j}$.

We prove the first statement by contradiction. Assume that there exists no global optimum having exactly two non-zero components, this means that when varying κ, solutions "jump" from one to

(at least) three non-zero components[2]. In particular, at the jumping point $a^* = (a_1^*, 0, \ldots, 0)$ two equations of the form

$$\begin{cases} H_{1m} & : \quad f_{a_m} g_{a_1} - f_{a_1} g_{a_m} = 0 \\ H_{1s} & : \quad f_{a_s} g_{a_1} - f_{a_1} g_{a_s} = 0 \end{cases} \tag{7.17}$$

for $m, s \neq 1$ with $m \neq s$ must be fulfilled. When only the component a_1^* is non-zero, both equations in (7.17) are linear in a_1^*. We can therefore eliminate a_1^* from the above system and are left with one equation in P_x and Q only.

$$\begin{aligned} &(P_x)_{1m}^2 Q_{11}(Q_{11} - Q_{ss}) + (P_x)_{1s}^2 Q_{11}(-Q_{11} + Q_{mm}) + [Q_{11}(Q_{1s}^2 - Q_{11}Q_{ss}) + \\ &Q_{11}(-Q_{1m}^2 - Q_{mm} + Q_{11}Q_{mm}) + (Q_{1m}^2 Q_{ss} - Q_{1s}^2 Q_{mm})] = 0 \end{aligned} \tag{7.18}$$

However, since P_x, P_y are random matrices, the probability that equation (7.18) holds is null. We can therefore conclude that with probability one there is only one other dimension $m \neq 1$ for which $f_{a_m} = \lambda g_{a_m}$. This implies that there exist $\kappa_3^c > \kappa_1^c \geq 0$ such that exactly two dimensions are active.

We now prove the second statement. We restrict ourself w.l.o.g to the interval $[0, \kappa_3^c]$ where the global optimum is attained only with one or two dimensions. The most informative dimension i_f being fixed, there are $p-1$ different possible 2-dimensional subspaces. We can apply Lemma 7.2 to each subspace separately to obtain $p-1$ different values of the first critical κ: $0 < \kappa_{1,1}^c < \cdots < \kappa_{1,p}^c$. Since each $\kappa_{1,i}^c$ is a function of different entries in Ψ and Φ (see Lemma 7.2), and recalling that P_x, P_y are random matrices, we can see that all values $\kappa_{1,1}^c, \ldots, \kappa_{1,p}^c$ are indeed distinct. The solution path S leaves the i_f axis when the first solution with two non-zero components becomes optimal, i.e. when $\kappa = \kappa_{1,1}^c$ and we can finally set $\kappa_1^c := \kappa_{1,1}^c$. In the interval $[\kappa_{1,1}^c, \kappa_{1,2}^c[$ only one 2-dimensional subspace can contain global optima, namely the subspace having first critical value $\kappa_{1,1}^c$. We can therefore set $\kappa_2^c := \kappa_{1,2}^c$.

□

We solve optimisation problem (7.8) using a log barrier interior point method detailed in Algorithm 7. Algorithm 7 starts by minimising f for a large value of the constraint κ such that all dimensions are selected, and then successively decreases κ until finally a unique dimension remains active. Even if we cannot prove that the solution path obtained with Algorithm 7 connects only global minima, we can verify that it reaches the globally optimal 2-dimensional subspace. This was indeed in the case in all simulations conducted.

Additional check for Algorithm 7. As an additional check of the path of stationary points obtained with Algorithm 7, we verify that this path does not have any bifurcations. We thereby insure that no other path connecting stationary points rejoins or diverges from the obtained path. A classical way to study bifurcations in 1-dimensional manifolds is provided by the *Implicit function theorem*. We first need to derive a set of equations which characterise the set of stationary points. For a stationary point a^* with strictly positive components, the non-negativity constraints are inactive and $\epsilon_j = 0, \ \forall j$. Stationary points are characterised by a vanishing Lagrangian gradient $\nabla \mathcal{L} = 0$, meaning that $\nabla f(a^*) = \lambda \nabla g(a^*)$. This proportionality condition can be translated into an orthogonality condition which eliminates λ: $\nabla f(a^*)$ must be orthogonal to the $(p-1)$-dimensional hyperplan orthogonal to $\nabla g(a^*)$. Constructing a basis $(g_\perp^1, \ldots, g_\perp^{p-1})$ of this hyperplan we obtain $p-1$ orthogonality conditions: $\nabla f \cdot g_\perp^i = 0, i = 1, \ldots, p-1$. Adding the constraint $g(a) = \kappa$ leads to a set of p equations in $p+1$ variables (a and κ). In the following we denote the partial derivatives of a real function f of a by $\frac{\partial f}{\partial a_i}(a) = f_{a_i}$, and the matrix of partial derivatives for a vector-valued function $\mathcal{F}$ by $J_a \mathcal{F}$. We further assume that $P_x, P_{x|y}$ have full rank and write $\Phi := P_{x|y}^{-1}, \Psi := P_x^{-1}$. In the p-dimensional case, the hyperplan orthogonal to ∇g is

[2]The case of a jump from $a = (0, \ldots, 0)$ to a least three active dimensions can similarly be excluded.

Algorithm 7 Optimisation of sparse MGIB

1. Denote the entries of A by $a \in \mathbb{R}^p$ and fix the set of κ values: $\kappa \in \{\kappa_0 > \kappa_1 > \cdots > \kappa_m\}$;
2. Initialisation step with $\kappa = \kappa_0$:
   compute $a^{\max} \in \mathbb{R}^p$: $a_j^{\max} = (e^{\kappa} - 1)/(P_x)_{jj}$;
   set $a^{\mathrm{m}} := 1/\sum_j (a_j^{\max})^{-1}$;
   compute $\lambda_1, \ldots, \lambda_p$ the eigenvalues of P_x;
   compute $c = \mathrm{argmin}_{[0,a^{\mathrm{m}}]} f_1(v), v \in \mathbb{R}$ with $f_1(v) = \left[\sum_j \log(\lambda_j) + \sum_j \log(\lambda_j^{-1} + v) - \kappa\right]^2$;
   set $a = (c + \delta, \ldots, c + \delta)$ for a small $\delta > 0$;
3. Optimisation for $\kappa \geq \kappa_0$:

for $\kappa \in \{\kappa_0, \ldots, \kappa_m\}$ **do**
  for $\epsilon \to 0$ **do**
    Set $a^* = \mathrm{argmin} f_2(w)$ where $w \in \mathbb{R}^p$, W is the diagonal matrix with elements w and $f_2(w) = \log|P_{x|y}W + I| - \epsilon \log\left[\log|P_x W + I| - \kappa\right] - \epsilon \sum_j \log(w_j)$;
  end for
  Exclude the dimensions corresponding to zero elements in a^* from the minimisation;
end for

$(p-1)$-dimensional and a basis for it is given by $g_\perp^1, \ldots, g_\perp^{p-1}$, where the vectors $g_\perp^i$ have $-g_{a_{i+1}}$ at position i, g_{a_i} at position $i+1$ and 0 otherwise. The set of stationary points is then implicitly defined by the equation $H = 0$, where $H : \mathbb{R}^{p+1} \to \mathbb{R}^p$ is defined by

$$H(a,\kappa) = \begin{pmatrix} H_1(a,\kappa) \\ \vdots \\ H_{p-1}(a,\kappa) \\ H_p(a,\kappa) \end{pmatrix} = \begin{pmatrix} \nabla f(a) \cdot g_\perp^1(a) \\ \vdots \\ \nabla f(a) \cdot g_\perp^{p-1}(a) \\ g(a) - \kappa \end{pmatrix}. \tag{7.19}$$

By the Implicit function theorem we know that if $|J_a H(a^*)| \neq 0$ for some point $a^* \in S$, then in a neighbourhood of a^* the solution path S has no bifurcation. While running Algorithm 7 we therefore regularly check that this determinant remains non-zero: the algorithm proceeds by successive optimisation steps with decreasing κ values $\{\kappa_0 > \cdots > \kappa_m\}$ and for each value obtains an optimum $a^*(\kappa)$, for every such optimum we can then verify that $|J_a H(a^*)| \neq 0$. This operation can be efficiently conducted since the computation of all partial derivatives $\partial H_i / \partial a_j$ requires only two matrix inversions. Indeed, for $i = 1, \ldots, p-1$ we have

$$\begin{aligned} \frac{\partial H_i}{\partial a_j}(a) &= f_{a_{i+1},a_j} g_{a_i} + f_{a_{i+1}} g_{a_i,a_j} \\ &\quad - f_{a_i,a_j} g_{a_{i+1}} - f_{a_i} g_{a_{i+1},a_j}, \\ f_{a_i} &= (\Phi + A)_{ii}^{-1}, \quad g_{a_i} = (\Psi + A)_{ii}^{-1}, \\ f_{a_i,a_j} &= (-1)^{i+j}(\Phi + A)_{ij}^{-1} - (\Phi + A)_{ii}^{-1}(\Phi + A)_{jj}^{-1}, \\ g_{a_i,a_j} &= (-1)^{i+j}(\Psi + A)_{ij}^{-1} - (\Psi + A)_{ii}^{-1}(\Psi + A)_{jj}^{-1}, \end{aligned} \tag{7.20}$$

where $f_{a_i,a_j} = \frac{\partial^2 f_i}{\partial a_i \partial a_j}(a)$ and $g_{a_i,a_j} = \frac{\partial^2 g_i}{\partial a_i \partial a_j}(a)$. The remaining elements of the Jacobian are given by $\partial H_p / \partial a_j(a) = g_{a_j} = -((\Psi + A)_{jj}^{-1})^2$ for $j = 1, \ldots, p$.

7.3 Inference for mixed continuous-discrete data.

To make our model applicable to mixed data we use Gaussian hidden variables and, as in MGIB, apply GIB to these underlying variables. This approach is based on the fact that every ordinal categorical variable can be assumed to be a quantised version of an underlying continous one

(Joe, 1989). Our model can be applied to data with any combination of continuous and ordinal discrete margins. Copula models have mainly been used with continuous margins since for any continous mutlivariate cdf $F = (F_1, \ldots, F_n)$ Sklar's theorem ensures the existence and the unicity of the copula C such that $F(v_1, \ldots, v_n) = C(F_1(v_1), \ldots, F_n(v_n))$. However, the copula construction $C(F_1(.), \ldots, F_n(.))$ stills leads to a valid cdf when all or some of the margins are discrete. As explained in Chapter 3 (Section 3.6), the main difficulty in copula modeling with discrete margins arises from the fact that uniqueness of the copula is guaranteed only on the range of the margins and traditional estimation techniques face an unidentifiability problem (Genest and Nešlehová, 2007). Despite this unidentifiability issue, efficient methods for copula estimation have recently been developed (Pitt et al., 2006), (Smith and Khaled, 2012). We follow the Bayesian semiparametric approach for Gaussian copula introduced in Hoff (2007) and consider the following model:

$$(\bar{X}_1, \ldots, \bar{X}_p, \bar{Y}_1, \ldots, \bar{Y}_q) \sim \mathcal{N}(0, P), \tag{7.21}$$

$$X_j = F_{X_j}^{-1}(\Phi(\bar{X}_j)), \;\; Y_l = F_{Y_l}^{-1}(\Phi(\bar{Y}_l)), \quad j = 1, \ldots, p, \;\; l = 1, \ldots, q, \tag{7.22}$$

where $F_{X_j}^{-1}$, $F_{Y_l}^{-1}$ are the generalised inverse of arbitrary, continuous or discrete, cdfs. We assume a parametric form for the copula, namely Gaussian, but not for the margins which are treated nonparametrically. Equation (7.22) imply that $X_j \sim F_{X_j}$, $Y_l \sim F_{Y_l}$. Unlike in the continuous case, the underlying Gaussian variables $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_p)$, $\bar{Y} = (\bar{Y}_1, \ldots, \bar{Y}_q)$ are not of the form $(\Phi^{-1} \circ F_{X_1}(X_1), \ldots, \Phi^{-1} \circ F_{X_p}(X_p))$. When some margins are discrete, estimates based on the empirical marginal distributions (Liu et al., 2009) or on the rank correlation cannot be used. To make inference on P and $(\bar{X}, \bar{Y})$ we use the marginal likelihood method introduced in Hoff (2007) and detailed below. This method has the advantage of being able to handle missing values and rank-deficient correlation matrices when the number of observations is smaller than the total number of dimensions: $n < p + q$.

Inference method detailed. To simplify the notation we write $Z := (X, Y)$ for the $(p+q)$-dimensional random variable formed by concatenation of X and Y. Similarly, we use $\bar{Z} := (\bar{X}, \bar{Y})$. The observed data matrix is denoted by (z_{ij}), where the i^{th} row $z_{i*} = (x_{i1}, \ldots, x_{ip}, y_{i1}, \ldots, y_{iq})$ is the i^{th} observation of (X, Y), and the corresponding unobserved realisations of $\bar{Z}$ are denoted by $(\bar{z}_{ij})$. Even without assuming any knowledge about the margins, the observed data (z_{ij}) provides some information about $(\bar{z}_{ij})$. Since the marginal cdfs are non-decreasing, a certain ordering of the observations must be preserved and the following holds:

$$z_{mj} < z_{nj} \Rightarrow \bar{z}_{mj} < \bar{z}_{nj}, \;\; \forall m, n, j. \tag{7.23}$$

Note that the converse implication does not hold since in the case of discrete observations ties can be involved. From (7.23) we can conclude that observations of $\bar{Z}$ must lie in the set $\mathcal{D}$:

$$\mathcal{D} := \left\{ (\bar{z}_{ij}) \in \mathbb{R}^{n \times (p+q)} | a_{ij} < \bar{z}_{ij} < b_{ij} \right\}, \tag{7.24}$$

where the bounds for each $\bar{z}_{ij}$ are defined as

$$a_{ij} = \max_{l \neq i} \{ \bar{z}_{lj} : z_{lj} < z_{ij} \} \tag{7.25}$$

$$b_{ij} = \min_{l \neq i} \{ \bar{z}_{lj} : z_{lj} > z_{ij} \}. \tag{7.26}$$

The set $\mathcal{D}$ is the set of all latent observations compatible with the (per column) ordering of the observed data. For continuous data, $\mathcal{D}$ reduces to the set of latent observations having the same ranks (for every dimension) as the observed data. Since the event $\bar{Z} \in \mathcal{D}$ occurs with probability one whenever Z is observed, the distribution of the observed data can then be written as:

$$p(Z|P, F_X, F_Y) = p(Z, \bar{Z} \in \mathcal{D}|P, F_X, F_Y) = \Pr(\bar{Z} \in \mathcal{D}|P) \; p(Z|\bar{Z} \in \mathcal{D}, P, F_X, F_Y), \tag{7.27}$$

where $F_X = \{F_{X_1}, \ldots, F_{X_p}\}, F_Y = \{F_{Y_1}, \ldots, F_{Y_q}\}$. The last equality in (7.27) follows from the fact that the event $\bar{Z} \in \mathcal{D}$ depends only on the copula parameters P and is independent of the

margins. Equation (7.27) provides a decomposition of the data distribution as a product of two terms. Importantly, the first term depends only on the parameter of interest P, the dependence of the data on the margins being confined to the second term. The marginal likelihood approach then estimates P using $\Pr(\bar{Z} \in \mathcal{D}|P)$ only, treating F_X and F_Y as nuisance parameters. Conducting Bayesian inference, we are interested in the posterior distribution,

$$p(P|\bar{Z} \in \mathcal{D}) \propto p(P)\ p(\bar{Z} \in \mathcal{D}|P), \tag{7.28}$$

where $p(P)$ denotes the prior distribution of the correlation matrix P. Since $p(P|\bar{Z} \in \mathcal{D})$ cannot be analytically calculated, we use a Monte Carlo Markov Chain algorithm to obtain samples from the posterior. The Gibbs sampling scheme alternates between sampling the hidden variables and sampling the correlation matrix. As explained in Hoff (2007), it is considerably simpler to obtain samples for P in an indirect fashion by first sampling from the posterior distribution of a covariance matrix Σ which is then projected onto the corresponding correlation matrix. Consider a model based on a covariance matrix Σ and a multivariate Gaussian variable $K \sim \mathcal{N}(0, \Sigma)$. We impose that Σ admits P as underlying correlation matrix ($\mathcal{P}(\Sigma) = P$). Then the variable $\bar{Z}$ is simply a scaled version of K and it is clear that $\bar{Z} \in \mathcal{D}$ and $K \in \mathcal{D}$ are equivalent. Posterior inference for $\mathcal{P}(\Sigma)|K \in \mathcal{D}$ is equivalent to inference for $P|\bar{Z} \in \mathcal{D}$ as long as the chosen prior for Σ induces the same prior for P. By introducing the covariance matrix Σ we overparametrize our model, we introduce a variance component which is unnecessary for the model specification but simplifies considerably the sampling procedure since we can use the conjugacy properties of Gaussian likelihood and Wishart prior. When required, samples of $\bar{Z}$ can simply be obtained by scaling samples of K. For computational efficiency, we reparametrize the algorithm from Hoff (2007) in terms of precision matrix, thereby avoiding repetitive matrix inversions in Algorithm 8. Algorithm 8 implements the Bayesian inference procedure described in further details below. We denote the precision matrix by $B = \Sigma^{-1}$. The prior distribution of B is Wishart: $p(B) \sim \text{Wishart}(\nu, B_0)$. The following summaries our basic modelling assumptions:

$$B \sim \text{Wishart}(\nu, B_0), \tag{7.29}$$

$$P_{ij} = \frac{B^{-1}_{ij}}{\sqrt{B^{-1}_{ii} B^{-1}_{jj}}},\ \forall i, j, \tag{7.30}$$

$$K_{1*}, \ldots, K_{n*}|B \overset{\text{iid}}{\sim} \mathcal{N}(0, B^{-1}), \tag{7.31}$$

where $K_{i*} = (K_{i1}, \ldots, K_{i(p+q)})$. Since B and K have conjugate distributions the posterior distribution of $B|K$ is again Wishart:

$$B|K \sim \text{Wishart}\left(\nu + n, \left(B_0 + K^T K\right)^{-1}\right), \tag{7.32}$$

where n is the number of observations. Gibbs sampling consists of three steps:

1. Sample $K|B, K \in \mathcal{D}$.
2. Sample $B|K, K \in \mathcal{D} = B|K$ following equation (7.32).
3. Compute P with components $P_{ij} = \frac{B^{-1}_{ij}}{\sqrt{B^{-1}_{ii} B^{-1}_{jj}}}$.

Whereas steps 2 and 3 above are straightforward, the sampling of K is more involved. In step 1, we sample K iteratively over observations $i = 1, \ldots, n$ and dimensions $j = 1, \ldots, p+q$. Although $K_{1*}, \ldots, K_{n*}|B$ are iid, by conditioning on the event $K \in \mathcal{D}$ both the independence property and the equal distribution property are lost. Indeed, the condition $K \in \mathcal{D}$ expresses a certain ordering over $K_{1*}, \ldots, K_{n*}$ imposed by the observed data Z. The imputed realisations of the hidden variables no longer are iid Gaussian but follow each a different truncated Gaussian distribution:

$$K_{i*}|B, K_{-i*}, K \in \mathcal{D} \sim \mathcal{TN}(0, B^{-1}, a_{i*}, b_{i*}), \tag{7.33}$$

where $a_{i*} = (a_{i1}, \ldots, a_{i(p+q)})$ and similarly $b_{i*} = (b_{i1}, \ldots, b_{i(p+q)})$. We sample one dimension of K_{i*} at a time, conditioned on the others. Useful closure properties of the multivariate Gaussian remain valid in the case of a truncated distribution and, in particular, the one-dimensional conditional distributions are still Gaussian. We precisely sample from the following conditionals:

$$K_{ij}|B, K \in \mathcal{D}, K_{-i,-j} \sim \mathcal{TN}\left(\mu_{ij}, \sigma_j^2, a_{ij}, b_{ij}\right),$$

where $\mu_{ij} = k_{i,-j}B_{-j,j}/(-B_{jj})$ and $\sigma_j^2 = 1/B_{jj}$. Here $k_{i,-j}$ denotes the i^{th} observation from the previous sweep from which dimension j has been removed, similarly $B_{-j,j}$ denotes the j^{th} column of B from which the row j has been removed.

Algorithm 8 Imputation of K and P.

0. The prior distribution of B is Wishart(ν, B_0);
1. Update K:
for $j = 1, \ldots, p+q$ **do**
  Set $\sigma_j := 1/B_{jj}$;
  for $r \in \text{unique}\{z_{1j}, \ldots, z_{nj}\}$ **do**
    set lower bound to $a := \max\{k_{ij}|z_{ij} < r\}$;
    set upper bound to $b := \min\{k_{ij}|z_{ij} > r\}$;
    for $i \in \{1, \ldots, n\}|z_{ij} = r$ **do**
      compute $\mu_{ij} := k_{i,-j}B_{-j,j}/(-B_{jj})$;
      sample $k_{ij} \sim \mathcal{TN}(\mu_{ij}, \sigma_j^2, a, b)$ from a truncated Gaussian;
    end for
  end for
end for
2. Sample $B \sim \text{Wishart}\left(\nu + n, \left(B_0 + K^T K\right)^{-1}\right)$.
3. Compute P: $P_{ij} = (B^{-1})_{ij}/\sqrt{(B^{-1})_{ii}(B^{-1})_{jj}}$.

7.4 Experiments

7.4.1 Simulated data

Simulation: Comparison between different IB methods. We generate training samples with $n = 1000$ observations $(x_i, y_i), i = 1, \ldots, n$ and dimensions fixed to $p = 15$, $q = 15$. The samples are drawn from a meta-Gaussian distribution (i.e. in the form of (6.10)) in two steps. First, we generate the Gaussian hidden variables $(\bar{X}, \bar{Y})$, then using the margin transformations (7.22) we obtain samples of (X, Y). P is obtained by scaling a covariance matrix drawn from a Wishart distribution. We use a Wishart distribution centered at a correlation matrix P_0 populated with some high correlation values to ensure some dependency between X and Y. In our first experiment we compare the following methods:

1. *MGIB bound*: MGIB is applied to observations of the continuous hidden variables $\bar{X}, \bar{Y}$. These hidden variables are not observable in practice and MGIB bound provides an upper bound on achievable information curves.
2. *MGIB*: We apply MGIB to observations $\{(x_i, y_i), i = 1, \ldots, n\}$ of the mixed variables without adjustment for mixed data i.e. using the model introduced in Chapter 6.
3. *Sparse MGIB*: We apply sparse MGIB (no adjustment for mixed data) to $\{(x_i, y_i)\}$.
4. *MMGIB*: Mixed Meta-Gaussian IB is MGIB for mixed data applied to $\{(x_i, y_i)\}$.
5. *Sparse MMGIB*: Sparse version of MMGIB applied to $\{(x_i, y_i)\}$.

We assess the efficiency of the different compression matrices A obtained by the above methods on a test set with 5000 observations. The compression T is calculated using the projection matrix

A obtained on the training data and the mutual information $I(\bar{X};T)$, $I(\bar{Y};T)$ are calculated using the formula for meta-Gaussian variables given in Chapter 6. Simulations are conducted with mixed margins for X and Y. For each dimension we use one of the following distributions: Student t_4, Binomial$(2, 0.5)$ or Binomial$(10, 0.5)$. By varying the parameter κ between 0.1 and 80 we can represent $I(Y;T)$ as a function of $I(X;T)$ and obtain the information curves. Each experiment is repeated to obtain the 50 curves for each method (shown in the left panel of Figure 7.2). We can see that some information was lost during the discretisation process $I(X;Y) < I(\bar{X},\bar{Y})$, and therefore the most effective compression (green curves) is obtained when applying MGIB to $\bar{X},\bar{Y}$ (not observable in practice). MMGIB (blue curves) performs clearly better than MGIB (red curves) on mixed data and achieves a compression rate closer to the MGIB bound. The left panel of Figure 7.2 also illustrates the difference between sparse and traditional IB. The information curves for sparse IB lie slightly below the traditional IB curves since less information can be captured when A is restricted to a diagonal matrix. However, sparse and traditional IB curves tend to the same value $I(Y;T)$ as $I(X;T)$ increases.

Simulation: Feature selection. To test the efficiency of sparse MMGIB in selecting relevant dimensions of X, we generate data such that only some dimensions of X were informative about Y. This is achieved by using a correlation matrix P_0 with only a few non-zero entries and a Wishart distribution with high degrees of freedom. The margins of X and Y were again mixed, following either a Beta$(0.5, 0.5)$ or a Binomial$(10, 0.5)$ distribution. The center and right panels of Figure 7.2 show simulations conducted with two different choices for P_0. On the left panel three dimensions of X are strongly informative about Y with corresponding values of 0.8 in P_0 and the remaining dimensions represent noise with correlations in P_0 close to 0. The right panel represents a more difficult situation: three dimensions of X have corresponding values of 0.8 in P_0, three other dimensions 0.6, and finally three more dimensions 0.4. Both panels show the solution paths for the 15 entries of A, each line corresponding to one dimension of X. The information curve obtained is shown in red with the corresponding κ values. On the center panel, the most informative dimensions (green curves) are selected first and can be clearly distinguished from the noise (gray dashed curves). On the right panel, the most informative dimensions (green curves) appear first again. Then, as κ increases, the remaining informative dimensions (blue and lilac curves) are selected as well. The noise dimensions are always selected last, when the value of κ becomes higher, to allow $I(X;T)$ to reach the required level of 0.5κ.

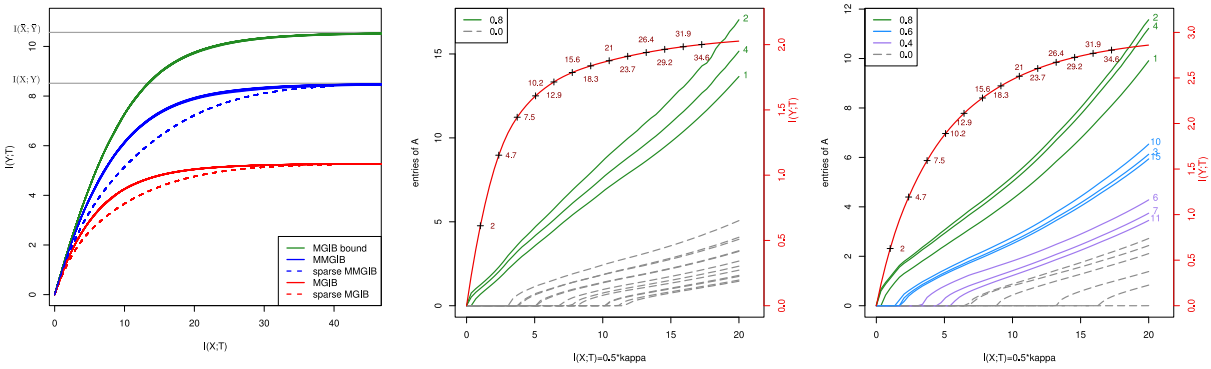

Figure 7.2: **Left:** the figure shows 50 (overlapping) information curves for each method. MGIB bound (green curves) provides a benchmark. MMGIB (blue curves) achieves a better compression than MGIB (red curves). Sparsity in T is achieved at the price of a small decrease in efficiency (dashed curves). **Center and right:** Solution paths for the 15 entries of A and corresponding information curve with κ values. The two panels represent results for different P_0, the more complex configuration can be seen on the right. Green curves correspond to very informative dimensions of X (correlation 0.8 in P_0), blue and lilac curves to informative dimensions (corr. of 0.6 and 0.4 in P_0, respectively), and gray dashed curves represent noise (corr. 0 in P_0).

7.4.2 Real data

We consider a real world problem from computational biology. A recurring and important task in medical prognosis is to identity biomarkers relevant to the disease evolution (Fuchs and Buhmann, 2011). Identification of a small number of key variables provides additional insight into the disease's mechanisms, and is crucial for cost-effective prognosis and therapy optimisation. We consider this selection problem in the context of cutaneous malignant melanoma (MM), the most common cause for fatalities in skin cancer. A first promising approach to identify biomarkers important for survival prediction was reported in (Meyer et al., 2012). Data was available in the form of immunohistochemical (IHC) expressions of 70 candidate biomarkers measured for 364 patients. Additionally, 9 different clinical observations were available which reflected experts' opinion about the stage of the tumor or directly characterised the severeness of the disease in terms of survival times. Focusing exclusively on survival information (and ignoring all other clinical attributes), a *7 marker signature* which is used to separate the patients into a low-risk and high-risk group has been proposed in (Meyer et al., 2012). In particular, the signature is defined via a risk-score of the form

$$\text{score}(x) = \frac{\sum_{i=1}^{7}(\beta_i x_i)\alpha_i}{\sum_{i=1}^{7}\alpha_i}, \quad \alpha_i = \begin{cases} 1 & \text{if } x_i \text{ is measured} \\ 0 & \text{if } x_i \text{ is missing} \end{cases}$$

where β_i are the regression coefficients of a univariate Cox model and x_i are are the IHC expression measurements of the 7 markers. A convincing statistical interpretation of the selected markers, however, remains unclear: the selection proceeded in several stages where only univariate tests have been used. Moreover, the relation of the biomarkers to established prognosis-related clinical observations like the pathological Tumor-Node-Metastasis (pTNM) staging or the Clark level was ignored in the model.

Our sparse MGIB model is best adapted to this problem, since the data falls into two groups which nicely fit into our framework: the 70 markers constitute the candidate features X, whereas the 9 clinical observations can be used for the target variable Y. Defining a *signature* of molecular markers might be seen as finding the best sparse compression of the biomarkers' expression on a molecular level which is still informative with respect to the clinical data in the second group. Further, the technical specifications of this dataset also perfectly fit into our mixed-data Bayesian framework: most of the expression levels are represented as ordered factors in a semi-quantitative scoring system with 5 levels, but other variables like survival times are continuous in nature, and roughly 10% of all values are missing. Repeated experiments conducted to a final selection of 6 markers as explained in the left panel of Figure 7.3.

Interestingly, our information-theoretic analysis method which is not particularly tailored to survival prediction could nicely reproduce the survival regression results in (Meyer et al., 2012): using our set of 6 markers (containing 4 markers which were already part of the original 7 markers signature) in the same "signature" formalism also led to clear separation of low-risk and high-risk patients on an independent test cohort. The right panel of Figure 7.3 shows Kaplan-Meier estimates of two different patient groups separated by using the risk-score defined above applied to *our* signature. The β_i coefficients were calculated on the training set described above but the 6 markers expressions were measured on an external test cohort of 221 patients. We see that the 6 markers selected using our method indeed provide highly effective prognosis predictors. Given the relatively small number of patients, the differences in p-values ($2 \cdot 10^{-4}$ for our signature vs. $2 \cdot 10^{-5}$ reported in panel A of Figure 6 in (Meyer et al., 2012)) are probably not too relevant. This observation is corroborated by the fact that the effect of regular updates on the the patients' censoring status from new follow-up reports have led to even bigger differences for the individual signatures in the past. We conclude that our sparse IB model can indeed by used as a high-quality prognosis predictor, even though it was not specifically designed for survival regression. The real advantage of the IB model, however, is its capability to extract many more details about the interaction between markers and a rich set of clinical measurements. Much could be said about the interpretation of the joint (X, Y)-correlation matrix obtained within the semi-parametric copula

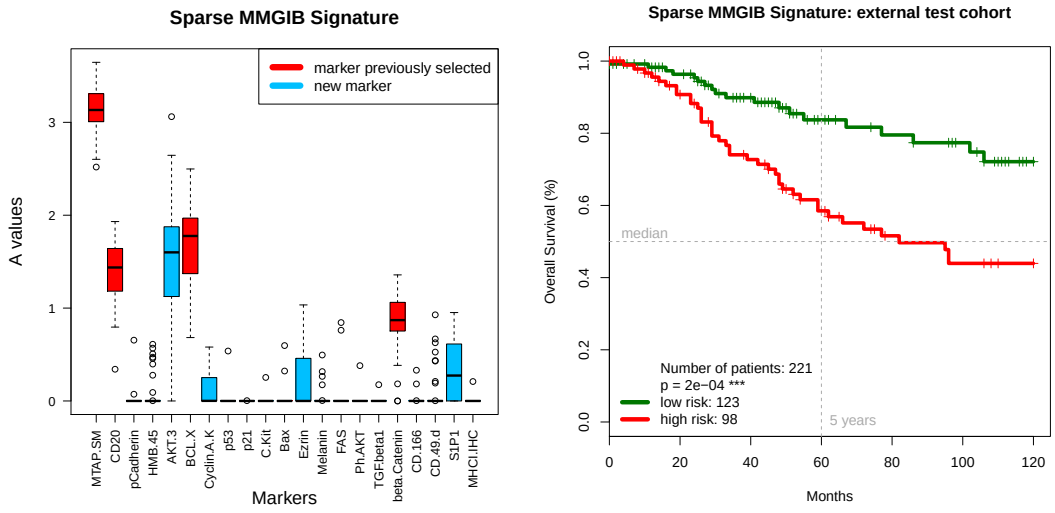

Figure 7.3: **Top:** boxplots of the elements A_{ii} in the diagonal projection matrix computed on the basis of 50 consecutive samples from the posterior distribution of the correlation matrix in the Gibbs sampler (features that are identically zero are not shown). The solution path was always cut at the same compression level (which typically lead to 6-8 selected features). Red boxes represent markers already identified in (Meyer et al., 2012). For our signature we select the 6 markers with a median above zero. **Bottom:** Kaplan-Meier plots of the two patient groups from the test cohort resulting from thresholding the *risk score* computed from the markers {MTAP, CD20, AKT.3, BCL.X, beta.Catenin, S1P1} at the median.

framework. Due to space constraints, however, we focus here only on one particularly interesting aspect, namely the *reversal* of the roles of X and Y, resulting in a sparse compression of the clinical variables Y subject to a constraint of preserving information about the molecular markers X. Figure 7.4 shows the corresponding solutions paths. Interestingly, it turns out that the *disease specific event status for overall survival* contains by far the most information about the molecular markers. This dominance of the disease specific event status over classical prognosis-related quantities like the *T score* or the *clark level* is interesting for the following reason: it basically shows that the classical clinical indicators fail to capture all relevant disease-specific information contained in the molecular data, showing that quantitative analysis of the biomarkers' expression patterns indeed adds valuable information about prognosis and survival in addition to the expert's macroscopic scoring/staging estimates.

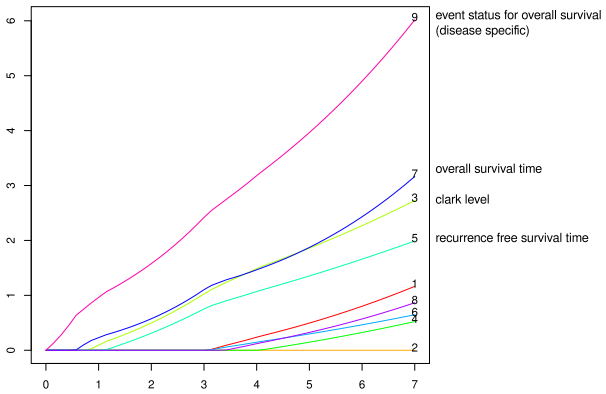

Figure 7.4: Solution paths for 9 diagonal entries of A when the roles of X and Y are reversed.

7.5 Conclusion

Sparse Meta-Gaussian IB provides a very flexible method for sparse compression with side information. By assuming a Gaussian copula, it encompasses a wide class of distributions with arbitrary non-Gaussian margins (continuous or discrete). Using relevance information, we do not need to impose any norm penalty to obtain sparsity. Our Bayesian framework for copula estimation can handle large-scale high dimensional datasets with potentially missing values. Our log barrier interior point algorithm is efficient even in high dimensions. Moreover, we prove that the two globally best input features can be found in arbitrary dimensions in an efficient way, thereby also providing an additional validation of the results obtained with our algorithm. Finally, we demonstrate in a clinical application that our model can compete with state-of-the-art survival prediction methods, while additionally allowing for an in-depth analysis of relevance- and interaction patterns between molecular markers and clinical measurements. We conclude that the proposed model is a highly flexible analysis tool which has the potential to significantly advance the field of exploratory data analysis within a well-defined information-theoretic framework.

Chapter 8

Conclusion

Although copulas are beginning be seen as a useful tool in machine learning, their use remains marginal. The goal of this thesis is to show how copulas can be incorporated into several established machine learning, resulting in wider applicability and improved results on non-Gaussian data. We summarize the contributions of this thesis to the solutions of three different problems below.

Detection of dependencies. We considered the problem of detecting dependencies between two sets of co-occurring samples, concentrating on meta-Gaussian data in which we infer a cluster structure that has a semantic interpretation in terms of dependencies. We build on the dependency-seeking clustering method for Gaussian variables of Klami and Kaski (2007) which is based on the idea of discovering clusters to capture the dependencies. However, when applied to non-Gaussian data, the clusters found by this method are used to approximate the non-Gaussian structure of the true clusters and thus lose their interpretation in terms of dependence. We proposed a Bayesian non-parametric generalization of this approach which overcomes the model mismatch issues occurring when the Gaussian assumption is too restrictive by enlarging the class of distributions covered, offering a wider range of application for dependency-seeking clustering.

Compression with relevance information. We then turned our attention to the Information Bottleneck approach to the problem of compression in presence of relevance information. We established a strong connection between IB method and copulas by showing that the method can be reformulated entirely in terms of a copula. This view allows us to avoid the difficult multivariate density estimation problem by focusing on the underlying copula, resulting in improved robustness and efficiency. Although an analytical solution is available for Gaussian distributions, solving the general problem can only be achieved by an iterative procedure. As a result, applications of the IB method were previously limited to Gaussian and low-dimensional discrete and Gaussian data. By taking advantage of the flexibility of a Gaussian copula model, we generalised the efficient analytical solution to meta-Gaussian variables and thereby provided an efficient IB method for a much larger class of continuous data.

Feature selection with relevance information. Building on our IB method for meta-Gaussian data, we proposed a new approach to feature selection in presence of relevance information, which identifies the most informative features w.r.t. the relevance variable. We adapted the classic IB problem to perform hard feature selection by constraining the diagonal of the projection matrix to be sparse and forcing the off-diagonal elements to be zero. We solve the resulting contrained IB problem using a tailored interior point method which remains efficient in high dimensions and

derive interesting convergence properties for it. The resulting Bayesian estimation technique can handle target copulas for discrete, continuous or mixed distributions and easily deal with missing values.

Bibliography

E. C. Anderson. Monte carlo methods and importance sampling. *Lecture Notes for Statistics*, Statistical Genetics(578C), 1999.

F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. *Technical report 688, Department of Statistics, University of California, Berkeley*, 2005.

J. Barnard, R. McCulloch, and X. Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10:1281–1311, 2000.

M. J. Bayarri and J. O. Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

T. Bedford and R. M. Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, (32): 245–268, 2001.

C. B. Bell. Mutual information and maximal correlation as measures of dependence. *The Annals of Mathematical Statistics*, 33(2):587–595, 1962.

H. P. Benson and R. Horst. A branch and bound-outer approximation algorithm for concave minimization over a convex set. *Journal of Computers and Mathematics with Applications*, 21 (6/7):67–76, 1991.

J. O. Berger, J. M. Bernardo, and D. Sun. The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938, 2009.

J. M. Bernardo. Modern bayesian inference: Foundations and objective methods. *Philosophy of Statistics*, pages 263–306, 2011.

D. Blackwell. Discreteness of ferguson selections. *Annals of Statistics*, 1(2):356–358, 1973.

D. Blackwell and J. B. MacQueen. Ferguson distributions via pólya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

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

L. Breiman. Statistical modeling: the two cultures. *Statistical Science*, 16(3):199–231, 2001.

S. Cambanis, S. Huang, and G. Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11:365–385, 1981.

H. Carley. Maximum and minimum extensions of finite subcopulas. *Communications in Statistics - Theory and Methods*, 31:2151–2166, 2002.

E. Çinlar. *Probability and Stochastics*. Graduate Texts in Mathematics. Springer, 2011.

G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information bottleneck for Gaussian variables. *Proceedings of the Neural Information Processing Systems (NIPS)*, 2003.

G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005.

U. Cherubini, E. Luciano, and W. Vecchiato. *Copula methods in finance*. Chichester. John Wiley & Sons Ltd, 2004.

S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

G. Consonni and M. Scarsini. Lo studio della concordanza nel contesto della teoria della variabilità superficiale. *Statistica*, 42:69–77, 1982.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. John Wiley & Sons, 1991.

A. R. de Leon and K. Carriére Chough. *Analysis of mixed data: methods and applications*. Chapman and Hall, 2013.

A. Dolati and M. Úbeda Flores. Measures of multivariate concordance. *Journal of Probability and Statistical Science*, 4(2):147–163, 2006.

E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, 2009.

G. Elidan. Copula bayesian networks. *Proceedings of the Neural Information Processing Systems (NIPS)*, 2010.

G. Elidan. *Copulas and Machine Learning*. Springer Berlin Heidelberg, 2013.

P. Embrechts. Copulas: a personal view. *J. Risk Ins.*, 76(3):639–650, 2009.

P. Embrechts and M. Hofert. A note on generalized inverses. *Preprint*, 2013.

T. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2): 209–230, 1973.

X. Z. Fern, C. E. Brodley, and M. A. Friedl. Correlation clustering for learning mixture of canonical correlation models. *Accepted for SIAM International Conference on Data Mining*, 2005.

R. Féron. Sur les tableaux de corrélation dont les marges sont données: cas de l'espace a trois dimensions. *Publ. Inst. Statist. Univ. Paris*, 5:3–12, 1956.

M. Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon*, 14 (3):53–77, 1951.

T.J. Fuchs and J.M. Buhmann. Computational pathology: challenges and promises for tissue analysis. *Journal of Computerized Medical Imaging and Graphics*, 35(7):515–530, April 2011. ISSN 0895-6111. doi: DOI: 10.1016/j.compmedimag.2011.02.006. URL `http://www.sciencedirect.com/science/article/pii/S0895611111000383`.

A. P. Gasch, P. T. Spellman, and et al. C. M. Kao. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.

H. Gebelein. Das statistische problem der correlation also variations und eigenwert problem. *Zeitschrift für Angewandte Mathematik und Mechanik*, 21:364–379, 1941.

A. E. Gelfand. Gibbs sampling. *Journal of the American Statistical Association*, 95(452):1300–1304, 2000.

A. E. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal densities. *Journal of American Statistics Association*, 85(410):398–409, 1990.

C. Genest and J. Nešlehová. A primer on copulas for count data. *Astin Bulletin*, 37(2):475–515, 2007.

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

C. Genest, M. Gendron, and M. Bourdeau-Brien. The advent of copulas in finance. *The European Journal of Finance*, 15(7-8):609–618, 2009.

A. Globerson and N. Tishby. On the optimality of the Gaussian information bottleneck curve. *Hebrew University Technical Report*, 2004.

R. M. Gray. *Entropy and Information Theory*. Springer, 1990.

C. T. Harbison, D. B. Gordon, and et al. T. I. Lee. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

R. V. L. Hartley. Transmission of information. *Bell System Technical Journal*, page 535, 1928.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

R. M. Hecht, E. Noor, and N. Tishby. Speaker recognition by Gaussian information bottleneck. *INTERSPEECH*, pages 1567–1570, 2009.

W. Hoeffding. Masstabinvariante korrelations-theorie. *Schriften Math. Inst. Univ. Berlin*, 5:181–233, 1940.

P. D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, 1(1):273, 2007.

S. Hohmann and W. H. Mager. *Yeast Stress Responses*. Topics in Current Genetics, Vol. 1. Springer, 2003.

S. Ihara. *Information theory for continuous systems*. Word scientific, 1993.

H. Joe. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405):157–164, 1989.

H. Joe. Multivariate concordance. *Journal of Multivariate Analysis*, 35:12–30, 1990.

H. Joe. *Multivariate models and dependence concepts*. Monographs on Stastistics and applied probability. Chapman and Hall, 1997.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

S. Kirshner. Learning with tree-averaged densities and distributions. *Proceedings of the Neural Information Processing Systems (NIPS)*, 2007.

A. Klami and S. Kaski. Local dependent components. *Proceedings of the 24th International Conference on Machine Learning*, 2007.

A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, (72):39–46, 2008.

A. Klami, S. Virtanen, and S. Kaski. Bayesian exponential family projections for coupled data sources. *Uncertainty in Artificial Intelligence*, 2010.

A. N. Kolmogorov. Foundations of the theory of probability. *Chelsea Publ. Co.*, 1950.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

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

J. Ma and Z. Sun. Mutual information is copula entropy. *arXiv:0808.0845v1*, 2008.

A. J. McNeil and J. Nešlehová. Multivariate archimedean copulas, d-monotone functions and l1-norm symmetric distributions. *The Annals of Statistics*, (37):3059–3097, 2009.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management*. Princeton Series in Finance. Princeton University Press, 2005.

N. Metropolis, A. W. Rosenbluth, M. N. Teller, and E. Teller. Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1091, 1953.

S. Meyer, T. J. Fuchs, and P. J. Wild. A seven-marker signature and clinical outcome in malignant melanoma: a large-scale tissue-microarray study with two independen patient cohorts. *PLoS ONE*, 7(6), 2012.

A. C. Micheas and K. Zografos. Measuring stochastic dependence using ϕ-divergence. *Journal of Multivariate Analysis*, 97:765–784, 2006.

T. Mikosch. Copulas: tales and facts. *Extremes*, 9:3–20, 2006.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Technical report 9815, Department of Statistics, University of Toronto*, 2011.

R. B. Nelsen. *An introduction to copulas*. Lecture notes in Statistics. Springer, 1999.

J. Nešlehová. On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis*, 98:544–567, 2007.

H. Nyquist. Certain topics in telegraph transmission theory. *A.I.E.E. Trans.*, 47:617, 1928.

P. Orbanz. Infinte-dimensional exponential families in cluster analysis of structured data. *Phd Thesis, ETH Zürich*, 2008.

D. Pál, B. Póczos, and C. Szepesvári. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. *Proceedings of the Neural Information Processing Systems (NIPS)*, 2010.

E. Parzen. On estimation of a probability density function on mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

M. Pitt, D. Chan, and R. Kohn. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006.

C. E. Rasmussen and D. Görür. Dirichlet process Gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):615–626, 2010.

A. Rényi. On measures of dependence. *Acta. Math. Acad. Sci. Hung.*, 10:441–451, 1959.

M. Scarsini. On measures of concordance. *Stochastica*, 8:201–218, 1984.

B. Schweizer and A. Sklar. Probabilistic metric spaces. *New-York: North-Holland/Elsevier*, 1983.

B. Schweizer and E. F. Wolff. On nonparametric measures measures of dependence for random variables. *The Annals of Stastistics*, 9(4):879–885, 1981.

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

O. Shamir, S. Sabato, and N. Tishby. Learning and generalization with the information bottleneck. *Theor. Comput. Sci.*, 411(29-30):2696–2711, 2010.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

A. Sklar. Random variables, distribution functions, and copulas: a personal look backward and forward. 76(3):639–650, 2009.

N. Slonim. The information bottleneck: Theory and applications. *Ph.D. Thesis, the Hebrew University*, 2002.

M. S. Smith and M. A. Khaled. Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association*, 107(497):290–303, 2012.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

M. D. Taylor. Multivariate measures of concordance. *Annals of the Institute for Statistical Mathematics*, 59(4):789–806, 2007.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *The 37th annual Allerton Conference on Communication, Control, and Computing*, (29-30):368–377, 1999.

H. Tsukahara. Semiparametric estimation in copula models. *The Canadian Journal of Statistics*, 33(3):357–375, 2005.

J. Viinikanoja, A. Klami, and S. Kaski. Variational Bayesian mixture of robust CCA models. *Principles of Data Mining and Knowledge Discovery*, pages 370–385, 2010.

M. Vrac. *CCMtools: Clustering through "Correlation Clustering Model" (CCM) and cluster analysis tools.*, 2010. URL `http://CRAN.R-project.org/package=CCMtools`. R package version 1.0.