High Dimensional Dependence

Copulae, Sensitivity, Sampling

Daniel Lewandowski

High Dimensional Dependence Copulae, Sensitivity, Sampling

Daniel Lewandowski

Invitation

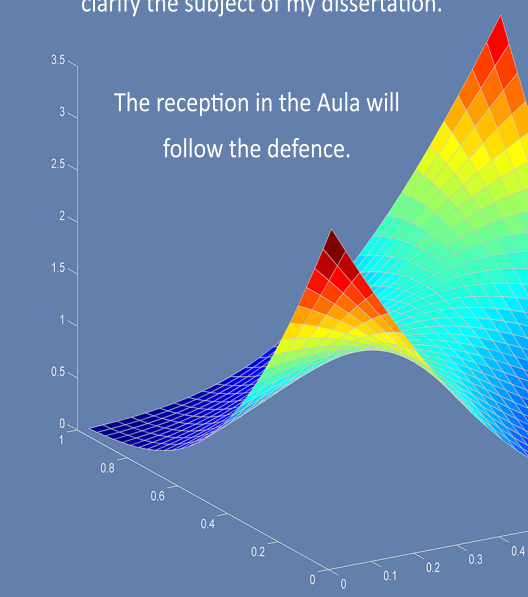
to attend the public defence of my
dissertation and theses

High Dimensional Modelling

Copulae, Sensitivity, Sampling

on Monday December 15, 2008
at 12.30 in the Aula of
Delft University of Technology,
Mekelweg 5, Delft.
Prior to the defence, at 12:00
I shall give a short presentation to
clarify the subject of my dissertation.

The reception in the Aula will
follow the defence.



Daniel Lewandowski

Prof. Henketstraat 25
2628 KK Delft
+31 624 191 292
daniel.lewandowski@gmail.com

Propositions

accompanying the thesis

High Dimensional Dependence: Copulae, Sensitivity, Sampling

Daniel Lewandowski

1. Let $c(x, y)$ be the density of a Ferguson's (generalized diagonal band) copula generated with density $g(z)$, $z \in [0, 1]$. If g is bounded on $[0, 1]$, has finite number of discontinuities and

$$g(0) - g^-(1) \geq 0,$$

where

$$\frac{d}{dx} g^-(x) = \max\left\{-\frac{d}{dx} g(x), 0\right\}, \quad g^-(0) = 0,$$

then $c(x, y)$ is a density of a mixture of diagonal band copulae.

see Chapter 3 of this dissertation

2. The vine method of generating random correlation matrices allows us to generate correlation matrices conditional on correlation values in an arbitrary tree.

see Chapter 5 of this dissertation

3. Let $G = G(\mathbf{X})$, where $\mathbf{X}$ is a random vector of length n, and let $\eta^2(G|X_i)$ denote the correlation ratio of G and X_i, $i = 1, \ldots, n$. Then

$$\eta^2(G|X_i) \geq \eta^2(G|h(X_i)),$$

where $h(X_i)$ is a function of X_i such that $\sigma^2_{h(X_i)} < \infty$.

see Chapter 6 of this dissertation

4. Let $G = G(\mathbf{X})$, where $\mathbf{X}$ is a random vector of length n and $\sigma^2_G < \infty$. Then

$$\arg\min_f \mathbf{E}\left[(G - f(X_i))^2\right] = \mathbf{E}[G|X].$$

see Chapter 6 of this dissertation

5. Based on numerical research the bivariate Gaussian copula has lower relative information with respect to the independent copula than the bivariate Frank's copula for rank correlations less in absolute values than 0.445. Otherwise, the Frank's copula has lower relative information.

6. Let $D(K)$ denote the determinant of a Hermitian correlation matrix of random variables indexed by the set K. Then the following holds

$$1 - \rho_{ij;L}\bar{\rho}_{ij;L} = \frac{D(\{i,j,L\})D(\{L\})}{D(\{i,L\})D(\{j,L\})},$$

where $\bar{\rho}_{ij;L}$ denotes the conjugate of partial correlation $\rho_{ij;L}$ and L is a set of indices such that $i, j \notin L$.

7. The crucial fact for studying connections between (Hermitian) correlation matrices and corresponding partial correlations matrices is that the inverse of the correlation matrix $\mathbf{R}$ is equal to the (conjugate) transpose of its cofactor matrix divided by the determinant of $\mathbf{R}$.

8. Let partial correlations in each tree of a regular vine be equal. That is

$$\rho_{ij;K} = \rho_{lm;L}$$

if $|L| = |K|$. Then

$$\rho_{ij} = \rho_{lm}.$$

Having such a specification on a D-vine leads to Toeplitz product moment correlation matrices.

9. One of the most critical and deepest texts about religion in general and its significance for all human beings is "Baudolino" by Umberto Eco.

10. Mathematicians and politicians both make generalizations; the difference is that the former's are often deep while the latter's are often stupid. Poland exhibits both extremes (Prof. Banach vs President Kaczyński).

11. Best ideas come to mind after the second beer, but they are gone after the third.

These propositions are considered opposable and defendable and as such have been approved by the supervisor, Prof. dr. R.M. Cooke.

Stellingen

behorende bij het proefschrift

Hoog-Dimensionale Afhankelijkheden: Copula's, Gevoeligheden, Trekkingen

Daniel Lewandowski

1. Zij $c(x, y)$ de verdeling zijn van een Ferguson's (algemene diagonale band) copula gegenereerd met de verdeling $g(z)$, $z \in [0, 1]$. Als g begrensd is op het interval $[0, 1]$, een eindig aantal discontinuiteiten heeft, en

$$g(0) - g^-(1) \geq 0,$$

met

$$\frac{d}{dx}\, g^-(x) = \max\left\{ -\frac{d}{dx}\, g(x), 0 \right\}, \quad g^-(0) = 0,$$

dan is $c(x, y)$ een verdeling van een mengsel van diagonale band copulae.

zie Hoofstuk 3 van deze dissertatie

2. De vine methode voor het genereren van stochastische correlatiematrices biedt de mogelijkheid tot het genereren van correlatie matrices geconditioneerd op correlatiewaarden in een willekeurige boom.

zie Hoofstuk 5 van deze dissertatie

3. Zij $G = G(\mathbf{X})$, met $\mathbf{X}$ een stochastische vector van lengte n, waarbij de correlatie ratio van G en X_i, $i = 1, \ldots, n$ wordt gegeven door $\eta^2(G|X_i)$. Dan geldt

$$\eta^2(G|X_i) \geq \eta^2(G|h(X_i)),$$

met $h(X_i)$ een functie van X_i zodanig dat $\sigma^2_{h(X_i)} < \infty$.

zie Hoofstuk 6 van deze dissertatie

4. Zij $G = G(\mathbf{X})$, met $\mathbf{X}$ een stochastische vector van lengte n en $\sigma^2_G < \infty$. Dan geldt

$$\arg\min_f \mathbf{E}\left[(G - f(X_i))^2\right] = \mathbf{E}[G|X].$$

zie Hoofstuk 6 van deze dissertatie

5. Gebaseerd op numerieke onderzoek, voor rangcorrelaties lager dan 0.445 de Gaussische copula heeft een lagere relatieve informatie ten opzichte van de onafhankelijke copula dan Frank's copula. Voor hogere rangcorrelaties, heeft Frank's copula een lagere relatieve informatie.

6. Zij $D(K)$ de determinant van een Hermitische correlatiematrix van stochasten met een indexverzameling K. Dan geldt:

$$1 - \rho_{ij;L}\bar{\rho}_{ij;L} = \frac{D(\{i,j,L\})D(\{L\})}{D(\{i,L\})D(\{j,L\})},$$

waarbij $\bar{\rho}_{ij;L}$ de geconjugeerde is van de partiële correlaties $\rho_{ij;L}$, en L een verzameling is van indices zdd $i, j \notin L$.

7. Het cruciale gegeven bij de studie van relaties tussen (Hermitische) correlatiematrices en de corresponderende partiële correlatiematrices is, dat de inverse correlatiematrix $\mathbf{R}$ gelijk is aan de (geconjugeerd) getransponeerde van zijn cofactormatrix, gedeeld door de determinant van $\mathbf{R}$.

8. Veronderstel dat de partiële correlaties in elke boom van een reguliere vine gelijk zijn, dwz

$$\rho_{ij;K} = \rho_{lm;L}$$

als $|L| = |K|$. Dan geldt

$$\rho_{ij} = \rho_{lm}.$$

Een dergelijke specificatie voor een D-vine heeft tot gevolg dat de correlatiematrix een Toeplitz matrix is.

9. Een van de diepste en meest kritische teksten over religie in het algemeen, en over haar betekenis voor de mensen is "Baudolino" van Umberto Eco.

10. Zowel wiskundigen als politici maken generalisaties, maar met dit verschil; generalisaties van wiskundigen zijn vaak diep terwijl die van politici vaak stompzinnig zijn. Polen geeft voorbeelden van beide (Prof. Banach en President Kaczynski).

11. De beste ideeën komen na het tweede biertje, maar vertrekken na het derde.

Deze stellingen worden opponeerbaar en verdedigbaar geacht en zijn als zodanig goedgekeurd door de promotor, Prof. dr. R.M. Cooke.

High Dimensional Dependence

Copulae, Sensitivity, Sampling

High Dimensional Dependence

Copulae, Sensitivity, Sampling

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. dr. ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen
op maandag 15 december 2008 om 12.30 uur

door

Daniel LEWANDOWSKI

wiskundig ingenieur

geboren te Kożuchów, Polen.

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. R.M. Cooke

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. R.M. Cooke	Technische Universiteit Delft, promotor
Dr. D. Kurowicka	Technische Universiteit Delft, copromotor
Prof. dr. T. Bedford	Strathclyde University, Glasgow, UK
Prof. dr. F.M. Dekking	Technische Universiteit Delft
Prof. dr. H. Joe	University of British Columbia, Canada
Prof. dr. T.A. Mazzuchi	George Washington University, Washington D.C., USA
Prof. dr. J. Misiewicz	University of Zielona Góra, Poland

Dit onderzoek kwam tot stand met steun van NWO.

NWO
Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Het Stieltjes Instituut heeft bijgedragen in de drukkosten van het proefschrift.

Thomas Stieltjes Institute
for Mathematics

ISBN 978-90-8570-317-4

On the cover: Minimum information copula

Typeset by the author with the LaTeX Documentation System
Printed in The Netherlands by: Wöhrmann Print Service

To my parents, my sister, and my beloved wife Agnieszka

Acknowledgements

It seems like it was only yesterday, when I entered the office of Prof. Roger M. Cooke at TU Delft for the first time. Since that very first moment many people have contributed to the research summarized in this thesis as well as to my personal development. There are not enough pages to list them all here, but some deserve a special mention. First of all, I would like to thank Prof. Cooke for being such a great mentor encouraging me to pursue my academic carrier and to make it happen. There is probably no other person I have learned so much from. We have also shared some good laughs and I can only hope we can call ourselves good friends. From the first day in the Netherlands I could also count on Dr. Dorota Kurowicka, both in personal and scientific matters. Of course, all these eight years abroad would never have become reality without Prof. Jolanta Misiewicz from the University of Zielona Gora, Poland, who sent me as an exchange student to the Delft University of Technology. I would like to thank her for this tremendous opportunity.

I am also very grateful to the members of the doctoral committee. Prof. Tim Bedford gave me twice the chance to work with him at the University of Strathclyde in Glasgow, UK, and this was such a great experience to me. Also I will never forget the time spent in Washington D.C., USA, where I was invited on several occasions by Prof. Tom A. Mazzuchi. His hospitality as well as his fantastic family made all stays in Washington unforgettable. I would like to thank Prof. Harry Joe for his patient explanations of mathematical issues and crucial contribution to our jointly written publication. This thesis has contributed substantially from comments and remarks of Prof. M. Dekking and for this I thank him.

Throughout the years I spent at the Department of Mathematics of TU Delft I have met many wonderful people whom I call friends now. The summary of this thesis has been translated to Dutch by Maarten-Jan Kallen. I am very grateful for his hard work. My officemates, Dan and Sebastian, did everything possible to provide positive vibration in our work environment. Anca is the person who

was the driving force making me to finalize this thesis. I would also like to thank fellow PhD students — Oswaldo, Patrycja and Rabin (thanks for translating the propositions). There are simply too many of you to list you all here, but you know whom I am talking about. Thank you for all the work we have done together and the parties we attended to have a good time. Hopefully, we shall have many opportunities in the future to get together and share some laughs.

My wife and my family in Poland have always been very supportive and encouraging in those not so rare moments of doubt. This work I dedicate to you.

Delft, *Daniel Lewandowski*
December 2008

Contents

Main Content

1 Introduction **1**
- 1.1 Uncertainty analysis . . . 2
- 1.2 Dependence modelling . . . 2
  - 1.2.1 Product moment correlation matrices . . . 2
  - 1.2.2 Copulae . . . 3
  - 1.2.3 Dependence trees . . . 4
  - 1.2.4 Vines . . . 5
- 1.3 Sensitivity analysis . . . 6
- 1.4 Expert judgement . . . 7

2 Review of multivariate copulae **9**
- 2.1 Introduction . . . 9
- 2.2 Prerequisites . . . 10
  - 2.2.1 Spearman's ρ and Kendall's τ for copulae . . . 10
  - 2.2.2 Partial and conditional correlations . . . 10
  - 2.2.3 Tail dependence . . . 12
- 2.3 Elliptical and Archimedean copulae . . . 13
  - 2.3.1 Gaussian (normal) copula . . . 14
  - 2.3.2 Normal vines . . . 15
  - 2.3.3 t-Copula . . . 16
  - 2.3.4 Archimedean copulae . . . 17
- 2.4 Dirichlet-type copula as an example of a multivariate copula . . . 18
  - 2.4.1 Generalized Dirichlet distribution . . . 20
  - 2.4.2 Generalized Gamma distributions . . . 21
  - 2.4.3 Multivariate copulae . . . 23
- 2.5 Conclusions . . . 25

3 Generalized diagonal band copulae **27**

3.1 Introduction . . . 27
3.2 Construction and properties of the generalized diagonal band copula 28
3.3 Mixtures of diagonal band copulae . . . 32
3.4 Minimally informative GDB copula . . . 35
3.4.1 Minimum information copula . . . 35
3.4.2 Approximation to the minimally informative GDB copula given the correlation constraint . . . 36
3.5 Examples of GDB copulae . . . 38
3.5.1 Triangular generating function . . . 39
3.5.2 Truncated exponential distribution as the generating function 39
3.5.3 The ogive distribution as the generating function . . . 41
3.5.4 Other copulae . . . 44
3.6 Relative information of various copulae . . . 45
3.7 Conclusions . . . 47

4 Building discretized minimally-informative copulae with given constraints 49
4.1 Introduction . . . 49
4.2 The D_1AD_2 algorithm . . . 50
4.3 The DAD algorithm for the 3-dimensional case . . . 52
4.4 Constructing minimally informative copula with the D_1AD_2 algorithm . . . 53
4.5 Software program for interactive expert assignment of minimally-informative copulae . . . 57
4.5.1 Algorithm searching for feasible values . . . 58
4.5.2 Example: Several observables . . . 59
4.6 Implementation issues . . . 60
4.7 Conclusions . . . 62

5 Generating random correlation matrices with vines and Onion method 63
5.1 Introduction . . . 63
5.2 Generating random correlation matrices with partial correlations regular vines . . . 64
5.2.1 Partial and multiple correlations . . . 66
5.2.2 Jacobian of the transformation from unconditional correlations to the set of partial correlations . . . 67
5.2.3 Partial derivatives . . . 69
5.2.4 Algorithm for generating correlation matrices with vines . . 73
5.3 Onion method . . . 75
5.3.1 Background results . . . 75
5.3.2 Algorithm for generating random correlation matrices . . . 76
5.3.3 Derivation of the normalizing constant . . . 77
5.4 Computational time analysis . . . 79
5.5 Conclusions . . . 80

6 Sample-based estimation of correlation ratio with polynomial approximation 83
6.1 Introduction . . . 83
6.2 Global sensitivity measures . . . 85
6.3 Definition of correlation ratio . . . 86
6.4 Properties of correlation ratios . . . 87
6.5 Standard methods of estimating correlation ratio . . . 89
6.5.1 Bayesian approach . . . 90
6.5.2 State Dependent Parameter models . . . 90
6.5.3 Sobol' method . . . 90
6.5.4 Kendall-Stuart method . . . 91
6.6 Polynomial approximation methods . . . 92
6.6.1 Polynomial fit . . . 92
6.6.2 Prevention of overfitting . . . 94
6.7 Simulations and Results . . . 96
6.7.1 Influence of sample size . . . 97
6.7.2 Overfitting . . . 98
6.7.3 Robustness . . . 100
6.7.4 The analytic function of Oakley and O'Hagan . . . 102
6.8 Conclusions and discussion . . . 103

7 Conclusions 105

References 109

Appendicies

A More examples of GDB copulae 115
A.1 Beta distribution as the generating function . . . 115
A.2 Distribution based on cosine function as the generating function . 116
A.3 Relative information of the GDB copula in terms of its generating function . . . 118

B Mixtures of diagonal band copulae with discontinuous mixing measures 121
B.1 Introduction . . . 121
B.2 Determining the continuous part of the mixing measure . . . 122
B.3 Determining the discontinuous part of the mixing measure . . . 123
B.4 Formulation of the theorem . . . 123

C Computer source code for generating random correlation matrices 127

Summary 131

Samenvatting 133

Curriculum Vitae **137**

CHAPTER 1

Introduction

> Certainty is the mother of quiet and repose, and uncertainty the cause of variance and contentions.
>
> *Edward Coke*

There are many sources of uncertainty — lack of knowledge, noise in data, chaotic nature of systems, etc. Whatever the source is, the uncertainty cannot remain untackled. This is sometimes a regulatory requirement, and sometimes it just pays off, as in optimization of industrial processes. A variety of ways of dealing with uncertainties include increasing the predictability of the system by taking control over some of its parameters, or at least *measuring* them, whatever "measuring" may mean, as long as it gives us a new and useful information about the state of these parameters. This, however, is not always feasible. Nobody can claim to control the wave height of the ocean, or exactly measure temperature on the surface of Sun. There is not much more to do than just to "tame" the uncertainty — deal with it within a well established framework of reference. This is where *statistics* comes to play the leading role and this thesis is the result of four years of studying its concepts and methods focused on high dimensional modelling. Before we proceed to the presentation of new results in the following chapters, a few general words will be said in this introductory chapter about various fields of statistics. This will help putting the results in a broader context.

In general we talk about a model instead of a system, as this suits better the mathematical nature of this dissertation. We denote the model as G, where G is a function of inputs $X_1, X_2, \ldots, X_n$. The variable G is the *explanandum* (the variable to be explained) and the variable X_i, $i = 1, \ldots, n$, is the *explanans* (the variable doing the explaining). The input variables need not be independent.

1.1 Uncertainty analysis

First of all, analysis of any model must include identification of its input factors (X_i's). Sometime this is quite easy, for instance when G is a clearly defined physical phenomena with a mathematical formulation given in an analytical form. On the other hand the identification process may require approaching experts on the given subject of study and using their knowledge to come up with a *reasonable* set of input parameters. The word *reasonable* itself allows for some subjectivity in the selection of X_i's. This is not necessarily a problem as long as it appropriate methods (*structured expert judgement*) are used to elicit this expertise.

Having selected the input parameters, their probability distributions must be determined. Data–rich areas like banking, insurance, or finance are privileged in this regard. One can simply sample from the set of reported realizations or fit a parametric distribution and use this one in further analysis. The latter solution allows to account for *unexpected* realizations, that is realizations not reported in the data, but still possible to occur.

The uncertainties in inputs are being propagated through the model to obtain the distribution of the output. However, it would be very unreasonable to assume independence between the input factors and treat all of them as not influencing each other. The dependence may significantly affect the output distribution and in the end make the whole analysis unrealistic if not accounted for.

1.2 Dependence modelling

When talking about dependence modelling we distinguish three subareas worth deeper analysis. First of all: what is *dependence*? How do we define it and measure it? There is a great deal of literature on various measures of dependence. Among them [Joe, 1997, Mari and Kotz, 2001] provide a good and extensive overview of dependence concepts. Many scientists have different views on the concept of dependence as reflected in the measures they employ. Pearson's product moment correlation, Spearman's rank correlation, Kendall's tau, tail dependence are among the key dependence measures concepts. This variety shows that the world we try to model is too complex to subjugate it to one measure only, although they are not mutually independent. Choosing one specific measure follows very simple reasoning: take the one you know how to cope with (analytically and numerically). Product moment correlations are well established in this regard.

1.2.1 Product moment correlation matrices

Product moment correlation is often chosen as a measure of dependence between two random variables. Usually denoted by Greek letter ρ it indicates the strength ($0 \leq |\rho| \leq 1$) and the direction ($\text{sgn}(\rho)$) of a linear relationship between two random variables. Simple to compute from data it has some obvious flaws however. Take $X \sim U[-1, 1]$ and $Y = |X|$. Once a realization of X is known, we know exactly the value of Y, hence clearly X and Y are dependent. However $\rho(X, Y) = 0$ indicating two uncorrelated random variables. In fact taking any

even function of X as Y results in their product moment correlation being zero. This small example shows that in general lack of correlation does not translates into independence (a notable exception is the joint normal distributions).

Product moment correlation measures the strength of linear association of two random variables. The counterpart of product moment correlation of two random variables for multivariate models is a product moment correlation matrix. This matrix contains product moment correlations computed for every possible pair of random variables involved. Given a model consisting of n variables, there are $\binom{n}{2}$ such pairs, which can be arranged in a form of a square symmetric matrix of dimension $n \times n$. The symmetrical characteristic of the matrix reflects the symmetricity property of the product moment correlation itself $\rho(X, Y) = \rho(Y, X)$. Entries of such a matrix are not algebraically independent as this would mean that every square matrix with one's on the main diagonal and values between -1 and 1 on off-diagonal is a correlation matrix. This is not true since not every such matrix is positive semi-definite. In fact, treating correlations in the correlation matrix as random variables, allows noticing some very interesting and complex dependencies between them. Chapter 5 explores this idea further. Also it describes two ways of generating product moment correlation matrices and extends them. The generating can be done such that the joint density of correlations (treated as random variables) is uniform for instance. Hence we can sample uniformly from the set of positive semi-definite square matrices with 1's on the diagonal and $\rho_{ij} \in [-1, 1]$ off-diagonal. This can be helpful in model testing to see how they behave in various scenarios.

The form of a matrix is very convenient for the following reason. Keeping the limitations in mind one can develop a methodology for sampling dependent random variables with correlations specified in a given correlation matrix. The simplest is the following. Let $\mathbf{U}$ be a random vector of length n distributed uniformly on the surface of the unit sphere in $\mathbb{R}^n$. Then $\mathbf{X} = A\mathbf{U}$ has covariance matrix $\Sigma = AA^T$. We call $\mathbf{X}$ a rotationally invariant random vector. Further, if $\mathbf{Y} = R\mathbf{X}$, where R is a non-negative random variable, then $\mathbf{Y}$ has still the same covariance matrix Σ and is called elliptically contoured. If $R^2 \sim \chi^2_2$ (Chi-squared distribution with 2 degrees of freedom), then $\mathbf{Y}$ follows the multivariate normal distribution. This is easy to implement, but the family of distributions obtainable in this way may not suit our needs. The vine-copula method presented later in this chapter provides much more robust techniques of generating samples from multivariate joint distributions with correlated marginals.

1.2.2 Copulae

Introducing correlations does not necessarily have to rely on the multivariate normal model mentioned above. The concept of copulae can be successfully applied instead. They represent a natural tool for modeling high-dimensional distributions with Markov dependence trees and a recent generalization thereof called vines [Bedford and Cooke, 2002](see section 1.2.4 for a brief description of vines) in which a multivariate distribution is built from bivariate pieces with given rank and conditional rank correlations.

Early accounts of bivariate distributions with uniform marginals can be traced back to the early 40's of the last century. Hoeffding [1940] studies such distributions in the square $[-1/2, 1/2]^2$. Formally the notion of copula was first formulated by Sklar [1959]. A copula is a joint distribution C on the unit hypercube with uniform marginals. Any m-variate continuous distribution F has an associated copula, which is the distribution C on the unit hypercube $[0,1]^m$ of the vector of uniform random variables $(F_1(x_1), \ldots, F_m(x_m))$ where F_i, $i = 1, 2, \ldots, m$, is the i-th univariate marginal distribution of F. The functional form of $C : [0,1]^m \to [0,1]$ is

$$C(\mathbf{u}) = F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)),\ \mathbf{u} \in [0,1]^m.$$

Conversely, if random quantities have known continuous marginals and a specified continuous copula then the joint distribution is specified by the formula

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_2(x_2)).$$

When $\mathbf{U}$ is a vector of m independent random variables then it is easy to see that their copula is the uniform distribution on the unit hypercube, $C(\mathbf{u}) = \prod_{i=1}^{m} u_i$. Markov trees and vines are coupled with bivariate copulae. The dependence structure can be introduced by choosing a copula for each edge of the tree or vine.

In practical applications the choice of a copula is mostly determined by the copula's efficiency in coping with certain problems. For example, software applications like UNICORN, a tool for carrying out uncertainty analysis developed at the Delft University of Technology, make use of a set of fixed parametric families of copulae, including the diagonal band copula [Cooke and Waij, 1986], the elliptical copula [Kurowicka et al., 2001] and the minimum information copula [Bedford and Meeuwissen, 1997], leaving the final choice to the user. This methodology has been successfully applied in uncertainty analysis combined with expert judgment [Cooke, 1991]. [Nelsen, 2007] offers an extensive overview of copulae families.

Due to the nature of copulae we have control over the the rank correlation between X and Y rather than the product moment correlation. This follows from the fact the parameter of the copula usually corresponds to the product moment correlation between F_X and F_Y. Applying the inverse cumulative distribution functions F_X^{-1} and F_Y^{-1} respectively gives the random variables of interest. This may pose some problems as in many cases a closed form expression for the relationship between the rank and the product moment correlation for a given copula is simply not known and hard to establish.

Chapter 3 describes the concept of Generalized Diagonal Band (GDB) copulae. We also find the minimally informative GDB copula with respect to the uniform background measure given the correlation constraint and introduce a new class of multivariate copulae, namely the Dirichlet-type copulae in chapter 2.

1.2.3 Dependence trees

Markov dependence trees have long been known as a simple and intuitive graphical model for dependence representation. They consist of nodes (random variables) and edges joining selected nodes expressing correlations between bivariate

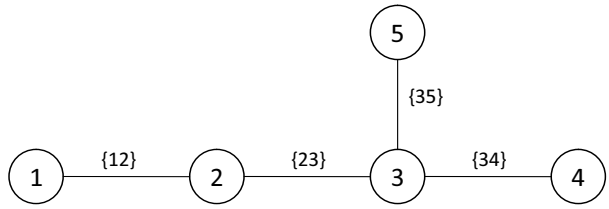


Figure 1.1: *Example of a Markov dependence tree on 5 variables.*

margins of a multivariate joint distribution. It implicitly assumes conditional independence between each pair of random variables not directly connected, but rather with a path leading through other intermediate nodes. The independence is conditional on the variables on the path between these two variables. For instance, in Figure 1.1 variables 1 and 5 are conditionally independent given variables 2 and 3.

Their main usage is coupled with Monte Carlo sampling. When analytical methods of determining joint distributions fail, the only solution may be generating a set of *scenarios* for input variables and studying the distribution of the output variable based only on the obtained data. If dependence between input variables is assumed and specified in a form of rank correlations, the scenarios can be sampled with the use of dependence trees. The procedure is quite simple. Pick a root (any node in Figure 1.1) and follow the path determined by edges connecting the nodes until all nodes are visited. Each edge of the tree is assigned a constraint set (a doubleton of indices of variables reachable from a given edge), a rank correlation and a bivariate copula. The rank correlation is a parameter for the copula joining the ranks (normalized to interval $[0, 1]$) of original random variables. This allows to sample two variables uniform on $[0, 1]$ representing the ranks with the copula as their joint distribution. Knowing the marginal distributions of the variables on the tree one can apply their respective inverse cumulative distribution functions to eventually obtain the proper quantiles.

Rank correlations on all edges of the tree are algebraically independent (thanks to the fact that the tree is an acyclic graph) hence they can be freely changed to any value between -1 and 1.

The conditional independence statements are quite strong assumptions. Therefore a generalization of trees has been introduced called *vines* [Cooke, 1997]. We describe this concept in the next section.

1.2.4 Vines

Dependence vines are generalizations of dependence trees. Here the conditional independence statements implied by the structure of the tree have been replaced with the conditional dependence statements.

A vine is a set of nested trees, that is, edges of tree T_i are nodes of tree T_{i+1}. Consider the set of nodes $\{1, 2, 3\}$ representing random variables in Figure 1.2. Edges $\{1, 2\}$ and $\{2, 3\}$ of tree T_1 are nodes of tree T_2. They can be joined

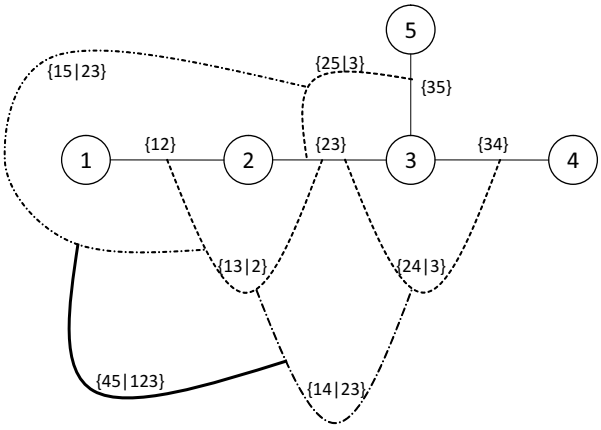


Figure 1.2: *Example of a regular vine on 5 variables.*

by an edge in T_2, namely the edge denoted as $\{13|2\}$. This edge is assigned a conditional rank correlation. Here $\{1,3\}$ is the conditioned set and $\{2\}$ is the conditioning set. Arcs between two nodes in a tree T_i can be drawn only if these nodes share a common node in tree T_{i-1}. This is called the *regularity* condition and it ensures that the conditioned sets are doubletons. Regular vines offer a very convenient tool for modelling dependence in the sense that each edge of the vine can be assigned a rank or conditional rank correlations which are algebraically independent. Quite understandably sampling a dependence vine is much more complicated then sampling a dependence tree. The dependence structure is more complicated now and involves conditional rank correlations. Algorithms for sampling a dependence vine can be found in Kurowicka and Cooke [2006a].

Dependence vines are described in greater detail in section 5.

1.3 Sensitivity analysis

If uncertainty analysis is the first step in studying complex statistical models then the sensitivity analysis is a natural follow-up. It answers the question of importance of input factors for the output result. As in the case of measuring dependence, one first has to determine what the term *important variable* means and how this importance can be quantified. We elaborate on this in the following paragraphs. The decision to carry out sensitivity analysis if very simple to justify from practical point view. Sensitivity analysis can be easily translated directly to saving money and/or reducing risks. After all, observing only selected, most important factors means fewer resources used.

However these factors must first be determined. Sensitivity analysis does this in a methodological manner with well established theorems. As in the case of dependence measures, there is no one general sensitivity measure. A particular

choice will depend on what aspect interests us most, whether it a global picture of influences or maybe harder to study nuisances better investigated with local sensitivity measures.

Section 6 introduces a new estimation method of the global sensitivity measure, called *correlation ratio*. It is a variance based measure, which is designed to explain to what degree the variance of the output follows from the variance of inputs. Computing this quantity analytically depends on the complexity of a given model and in most cases an estimation must be computed.

1.4 Expert judgement

The vine-copula method of studying complex multivariate models requires specifying rank and conditional rank correlations for the edges of the vine and selecting (conditional) copulae. Assessing all these correlations can be a tremendous task given the fact that the number of correlations to be specified for a model increases dramatically as the dimension of the model increases. Not all of them can be computed from data and the ultimate solution becomes the use of expert judgement (apart from setting the missing correlations to 0, effectively making more assumptions than one intended).

A number of methods for eliciting experts' knowledge exist and Cooke [1991] provides an excellent overview of these methods, as well as it introduces new ones used in many applications. Section 4 introduces a new methodology based on the minimum information principle that provides a much more flexible approach to defining copulae. During the elicitation a copula is built rather than a correlation elicited. Any quantity (or a group of quantities) depending on two variables in question can be considered as quantiles specified by an expert. The system can give guidance on the range of values for each quantile that are compatible with the specifications already made by the expert. The method uses a D_1AD_2 algorithm to build the copula that minimizes the information function given the constraints. This method has been implemented in a MATLAB code and is illustrated by an example.

CHAPTER 2

Review of multivariate copulae

> We used to think that if we knew one, we knew two, because one and one are two. We are finding that we must learn a great deal more about 'and'.
>
> *Sir Arthur Eddington*

2.1 Introduction

Building multivariate distributions can be very effectively done by using the vine-copula method, which "couples" bivariate pieces of this distribution in order to get the full multivariate joint distribution. The dependence structure is provided by specifying rank and conditional rank correlations on the edges of the corresponding vine. This specification is simplified by the fact, that correlations on a vine are algebraically independent. Alternatively one can use a method of sampling a multivariate distribution which immediately follows from Sklar's theorem. Every continuous multivariate distribution has its unique copula representative. Hence knowing how to sample from a multivariate copula allows us to obtain samples from the corresponding multivariate distribution with given dependence structure. This section introduces concepts necessary to fully understand the notion of multivariate copulae and their significance for generating samples from multivariate distributions. Also, a new multivariate copula derived from the generalized Dirichlet distribution has been introduced.

2.2 Prerequisites

There are many dependence concepts that can be discussed in conjunction with copulae. We list some of the most widely used here.

2.2.1 Spearman's ρ and Kendall's τ for copulae

Copulae correlate *percentiles* of univariate margins of joint distributions. Therefore we mostly talk about rank correlations in case of copulae, instead of product moment correlations. Two popular measures of such association are the Spearman rho, denoted as ρ_r, and Kendall's τ. They can be calculated for a given copula $C(u, v)$ as follows [Hoeffding, 1940]

$$\begin{aligned} \rho_r &= 12 \iint_{[0,1]^2} C(u,v)\, du\, dv - 3, \\ \tau &= 4 \iint_{[0,1]^2} C(u,v) dC(u,v)\, du\, dv - 1. \end{aligned}$$

They both take values in the interval $[-1, 1]$, with the sign indicating a negative or positive dependence.

Kendall's τ is easier to calculate for well known copulae, like the ones mentioned below in section 2.3. First of all, very simple analytic expressions exist for converting parameters of Archimedean copulae, like Clayton or Gumbel copulae, to this association coefficient. Spearman's ρ_r has to be numerically estimated in these cases, as closed form expressions do not exist. Secondly, in case of elliptical distributions (like Gaussian and t-Student distributions) Kendall's τ depends only on the product moment correlation between pairs of its univariate margins, where Spearman's ρ_r depends also on the specific type of this distribution.

2.2.2 Partial and conditional correlations

The partial correlation $\rho_{12;3,\ldots,n}$ can be interpreted as the correlation between the orthogonal projections of random variables X_1 and X_2 on the plane orthogonal to the space spanned by $X_3, \ldots, X_n$.

Definition 2.2.1 (Partial correlation)**.** *The partial correlation of random variables X_1 and X_2 with $X_3, \ldots, X_n$ held constant is*

$$\rho_{12;3,\ldots,n} = -\frac{C_{21}}{\sqrt{C_{11}C_{22}}},$$

where $C_{i,j}$ denotes the (i, j)th cofactor of the n-dimensional product moment correlation matrix; that is, the determinant of the submatrix gotten by removing row i and column j.

Partial correlations can be calculated recursively with the following formula [Yule and Kendall, 1965]

$$\rho_{ij;kL} = \frac{\rho_{ij;L} - \rho_{ik;L}\rho_{jk;L}}{\sqrt{(1-\rho^2_{ik;L})(1-\rho^2_{jk;L})}}, \tag{2.1}$$

where L is a set of indices, possibly empty, distinct from $\{i, j, k\}$. They can be assigned to the edges of a regular vine, such that conditioned and conditioning sets of the edges and those of partial correlations coincide. Every such assignment uniquely parameterizes a product moment correlation matrix.

The conditional rank correlation of X and Y given random variables indexed by set L is the rank correlation computed with the conditional distributions of X, Y given L. They can be assigned to the edges of a regular vine in the same way we do it for partial correlations. This is a more natural association as parameters of copulae can be expressed in terms of rank correlations.

The vine-copula method gains significantly from knowing the relationship between partial correlations and conditional rank correlations for a given copula. This is of great value for models with constant conditional rank correlations only, or in cases where one is willing to violate this assumption. The following heuristics works for vines with Gaussian copulae used for coupling bivariate piece of the joint distribution realized by this vine. Suppose one wants to change one vine to another (in other words change the conditional rank correlation specification). This will not be possible to achieve unless conditional rank correlations on the original vine can be somehow *translated* to conditional rank correlations on the new vine. One of the solutions is to employ partial correlations as intermediate step of the transformation. Consider this algorithm

Algorithm 2.2.1 (Changing conditional rank specification on a normal vine).

1. Convert conditional rank correlations on the original vine to corresponding conditional product moment correlations (the transformation of course depends on the chosen copula).

2. Convert conditional product moment correlations on the original vine to corresponding partial correlations (the transformation also depends on the chosen copula).

3. Use recursive formula (2.1) on partial correlations to obtain the unconditional product moment correlation matrix.

4. Use recursive formula (2.1) on the unconditional product moment correlation matrix to obtain partial correlations for the edges of the new vine.

5. Convert these new partial correlations to corresponding product moment correlations with the inverse of the transformation used in step 2.

6. Finally, convert conditional product moment correlations to conditional rank correlations with the inverse of the transformation used in step 1.

The main obstacle in using this algorithm is the lack of knowledge of the transformations used in steps 1 (and inverse thereof in step 6) and 2 (5). There are cases when they can be easily derived however and the following results of Baba et al. [2004] are useful for determining multivariate distributions for which the partial and the conditional product moment correlations coincide.

Theorem 2.2.1. *For any random vectors* $\mathbf{X} = (X_1, X_2)$ *and* $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$ *the following two conditions are equivalent*

- $\mathbf{E}(\mathbf{X}|\mathbf{Y}) = \alpha + B\mathbf{Y}$ *for a vector* α *and a matrix* B*,*
- $\Sigma_{\mathbf{XX};\mathbf{Y}} = \mathbf{E}\left(\Sigma_{\mathbf{XX}|\mathbf{Y}}\right)$*,*

where $\Sigma_{\mathbf{XX};\mathbf{Y}}$ *is the partial covariance matrix of* $\mathbf{X}$ *with* $\mathbf{Y}$ *fixed, and* $\Sigma_{\mathbf{XX}|\mathbf{Y}}$ *is the corresponding conditional covariance matrix.*

and

Corollary 2.2.2. *For any random vectors* $\mathbf{X} = (X_1, X_2)$ *and* $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$*, if there exists a vector* α *and a matrix* B*, such that*

$$\mathbf{E}(\mathbf{X}|\mathbf{Y}) = \alpha + B\mathbf{Y} \text{ and } \rho_{12|\mathbf{Y}} \text{ does not depend on } \mathbf{Y},$$

then $\rho_{12;\mathbf{Y}} = \rho_{12|Y}$ *almost surely.*

Elliptically contoured distributions immediately come to mind when considering distributions complying with the assumptions of Corollary 2.2.2.

2.2.3 Tail dependence

So far only correlation coefficients have been discussed as concepts of dependence. They measure *average* dependence over the domain of variables of interest. There is however a measure that tries to capture the dependence more locally rather than globally, in the tails (lower and/or upper) of distributions. We introduce the notion of tail dependence:

Definition 2.2.2 (Upper tail dependence)**.** *Let* $\mathbf{X} = (X_1, X_2)$ *be a random vector. We say that* $\mathbf{X}$ *is upper tail dependent if*

$$\lambda_U = \lim_{v \to 1} \mathbf{P}\{X_1 > F_1^{-1}(v) | X_2 > F_2^{-1}(v)\} > 0,$$

if the limit λ_U *exists.*

Conversely, we define the lower tail dependence

Definition 2.2.3 (Lower tail dependence)**.** *Let* $\mathbf{X} = (X_1, X_2)$ *be a random vector. We say that* $\mathbf{X}$ *is lower tail dependent if*

$$\lambda_L = \lim_{v \to 0} \mathbf{P}\{X_1 \leq F_1^{-1}(v) | X_2 \leq F_2^{-1}(v)\} > 0,$$

if the limit λ_L *exists.*

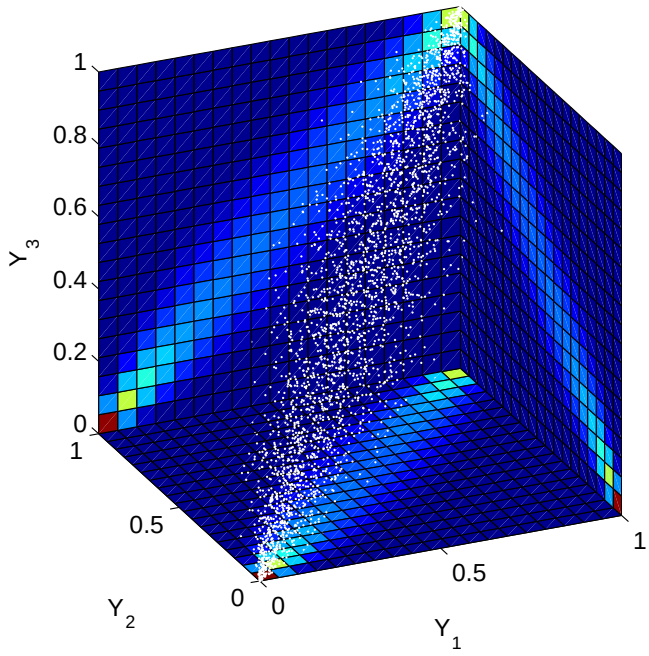


Figure 2.1: *Scatter plot of samples from a 3-dimensional normal copula with projections of 2-dimensional margins (all correlations are equal to 0.95).*

These limits have simpler representation for copulae

$$\begin{aligned} \lambda_U &= \lim_{v \to 1} \frac{1 - 2v + C(v, v)}{1 - v}, \\ \lambda_L &= \lim_{v \to 0} \frac{C(v, v)}{v}, \end{aligned}$$

where C is the cdf of a copula.

The next section gives examples of multivariate copulae and how the introduced concepts of dependence can be used to describe the differences between them.

2.3 Elliptical and Archimedean copulae

This section studies the most widely used multivariate copulae, serving as a point of reference for the copula we develop in section 2.4. As an example we show a scatter plot of samples from a 3-dimensional normal copula in Figure 2.1. White dots represent the samples. Intensity of the colours of the plots in the background represents relative height of the corresponding two dimensional marginals. Rank

correlations between every pair of random variables is $\rho_r(X,Y) = \rho_r(Y,Z) = \rho_r(X,Z) = 0.95$.

2.3.1 Gaussian (normal) copula

Finance and insurance industries are already very familiar with the multivariate normal copula, constructed from the multivariate normal distribution via Sklar's theorem. While the normality condition is satisfied in many applications, the main reason for the normal copula being used so widely is its tractability. Quite often solutions for problems involving this copula exist analytically (closed form relationships between various correlation coefficients for that matter). The multivariate normal copula is in fact the joint normal transform. Since most computations are done in the Gaussian space we can take advantage of linear conditional expectations (regressions). This is important for the equivalence of the conditional and partial correlations for the reasons explained in section 2.2.2.

Using this copula also means that the whole correlation matrix must be specified and this involves the previously mentioned problem of compatibility of correlations (see section 1.2.1). Also, if a vector $\mathbf{X}$ of n random variables yields rank correlation matrix R_r, then one has to transform this matrix to a product moment correlation matrix using the following transformation

$$\rho(X_i, X_j) = 2 \sin\left(\frac{\pi}{6}\rho_r(X_i, X_j)\right), \tag{2.2}$$

where $i, j = 1, \ldots, n$, $i \neq j$. This step is necessary since the input for the normal copula is a product moment correlation matrix. Unfortunately, as it has been pointed out in [Kurowicka and Cooke, 2006a, chapter 4.2], this causes a large percentage of semi-positive definite matrices R_r to be transformed to non-positive definite matrices, with the percentage decreasing to 0 as the dimension increases. We can cope with this problem by applying the notion of the partial correlation, as we show in the next section.

For the sake of completeness we give the formula for the product moment correlation as as function of Kendall's τ for the normal copula [Fang et al., 2002]

$$\rho(X_i, X_j) = \sin\left(\frac{\pi}{2}\tau(X_i, X_j)\right). \tag{2.3}$$

This formula however is even more prone to transforming a Kendall's τ matrix into a non positive definite matrix, as it has been shown in Figure 2.2.

The multivariate normal distribution is an example of a distribution complying with the assumptions of Corollary 2.2.2. In fact, the corollary holds for a much broader class of distributions, namely elliptically contoured distributions (of which the normal distribution is a member). The equivalency of the partial and conditional product moment correlation allows to develop another method of generating correlated variables with the help of multivariate normal copula, called *normal vines* introduced in [Kurowicka and Cooke, 2006a].

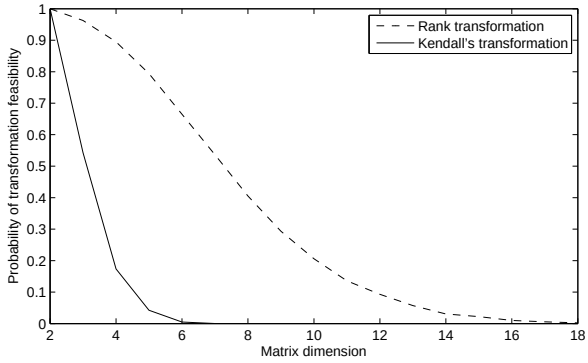


Figure 2.2: *Probability of feasibility of rank correlation and Kendall's τ to product moment correlation transformation.*

2.3.2 Normal vines

Suppose that conditional rank correlations are specified on the edges of a regular vine. Under the normality assumptions these conditional rank correlations can be transformed to conditional product moment correlations with the relationship 2.2, which in turn, are equal to the partial correlations. This way we obtain a partial correlation vine for a joint normal distribution. Any partial correlation specification on a vine characterizes a unique product moment correlation matrix and this is the input for the multivariate normal copula. This method has been implemented in UNICORN, software for uncertainty analysis with correlations developed at the Department of Mathematics of the Delft University of Technology.

Example of generating correlated samples with normal vine Consider a random vector $\mathbf{X} = (X_1, X_2, X_3)$, where X_i are iid $\mathcal{N}(0,1)$, $i = 1, 2, 3$. Suppose the following vine representation is given:

$$A_r = \begin{bmatrix} \rho_{r12} & \rho_{r13} \\ & \rho_{r23|1} \end{bmatrix} = \begin{bmatrix} 0.5 & 0.6 \\ & -0.8 \end{bmatrix}.$$

The conditional rank correlation matrix A_r can be transformed with eq.(2.2) to the conditional product moment correlation matrix

$$A = \begin{bmatrix} \rho_{12} & \rho_{13} \\ & \rho_{23|1} \end{bmatrix} = \begin{bmatrix} 0.5176 & 0.618 \\ & -0.8135 \end{bmatrix}.$$

Knowing $\rho_{ij|k} = \rho_{ij;k}$ allows us to determine the product moment correlation matrix R from the conditional product moment correlation matrix A with the help of the recursive formula (2.1).

$$R = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.5176 & 0.618 \\ 0.5176 & 1 & -0.2272 \\ 0.618 & -0.2272 & 1 \end{bmatrix},$$

and this one is always positive definite as any partial correlation matrix uniquely parameterizes a product moment correlation matrix. The resulting lower triangular matrix L, such that $LL^T = R$ is

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0.5176 & 0.8556 & 0 \\ 0.618 & -0.6395 & 0.4572 \end{bmatrix}.$$

Now we draw 600 000 samples of $\mathbf{X}$ and calculate $\mathbf{Y} = L\mathbf{X}$. The result of this operation, the random vector $\mathbf{Y}$, is a correlated normal random vector with the following sample product moment correlation matrix

$$\widehat{R} = \begin{bmatrix} 1 & 0.5184 & 0.618 \\ 0.5184 & 1 & -0.2269 \\ 0.618 & -0.2269 & 1 \end{bmatrix}.$$

This is a very accurate approximation of the original correlation matrix R. Having $\widehat{R}$ allows to compute the sample partial correlation matrix

$$\widehat{A} = \begin{bmatrix} 0.5184 & 0.618 \\ & -0.8141 \end{bmatrix}.$$

2.3.3 t-Copula

t-Copula as the name suggests corresponds to the multivariate Student's t-distribution. It is more powerful than the Gaussian copula in the sense that it has an extra parameter, ν, called the degree of freedom. In fact, the Gaussian copula can be seen as a limiting case of the t-copula as $\nu \to \infty$. Therefore it is not surprising, that the upper tail dependence for the $t$copula decreases to 0 as $\nu \to \infty$, since the Gaussian copula is tail independent ($\lambda_U = \lambda_L = 0$)[Embrechts et al., 2002].

Recently the t-copula has become a more popular choice for financial and actuarial stochastic modelling; namely because of its close relation to the familiar Gaussian copula. [Demarta and McNeil, 2005] extensively cover many variations of the t-copula.

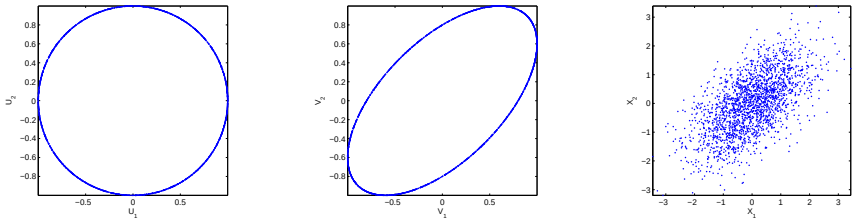
Both the Gaussian and the t-copulae are examples of elliptical copulae, that is, copulae of elliptically contoured distributions. We characterize the elliptical distributions as follows. Let vector $\mathbf{U} \in \mathbb{R}^n$ have a sign-symmetric Dirichlet distribution. In other words $\mathbf{U}$ is uniformly distributed on a unit sphere. The joint density of the first k components of $\mathbf{U}$, $k < n$, is of the form

$$f_k(\mathbf{u}) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-k}{2}\right)\Gamma\left(\frac{1}{2}\right)^k}\left(1 - \sum_{i=1}^{k} u_i^2\right)_+^{\frac{n-k}{2}-1},$$

where $(a)_+ = a$ whenever $a \geq 0$ and $(a)_+ = 0$ otherwise.

Definition 2.3.1. *A vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is said to be elliptically contoured with parameters μ and Σ if it has the stochastic representation*

$$\mathbf{X} = \mu + RA\mathbf{U}, \tag{2.4}$$



(a) $\mathbf{U} = (U_1, U_2)$ **(b)** $\mathbf{V} = A\mathbf{U}$ **(c)** $\mathbf{X} = R\mathbf{V}$

Figure 2.3: *Scatter plots of vectors uniform on a sphere* $\mathbf{U} = (U_1, U_2)$*, rotationally invariant* $\mathbf{V}$*, and elliptically contoured* $\mathbf{X}$*.*

where $\mu \in \mathbb{R}^n$ *(*$n \times 1$ *vector of means),* $R \geq 0$ *is a random variable independent of* $\mathbf{U}$*, and* A *is an* $n \times n$ *constant matrix, such that* $AA^T = \Sigma$*.*

Vector $\mathbf{X}$ has mean vector μ and covariance matrix Σ. Figure 2.3 shows three stages of generating samples from an elliptically contoured distribution starting from a distribution on a sphere in $\mathbb{R}^2$ and resulting in correlation $\rho(X_1, X_2) = 0.8$. The *radius* random variable R is Chi-square distributed with two degrees of freedom, resulting in $\mathbf{X}$ being joint normal distribution. It is quite easy to see why elliptical distributions have linear regressions.

Figure 2.4 shows samples generated from various copulae presented in this section. Tails of the t-copula exhibit higher concentration of samples than in case of the Gaussian copula indicating that it is both lower and upper tail dependent ($\lambda_L > 0$ and $\lambda_U > 0$).

2.3.4 Archimedean copulae

The normal copula is very popular choice for generating correlation variables adopted by such credit risk models as KMV or CreditMetrics. However historical data usually suggests that other types of copulae may be more suitable as they fit data better. A wide family of copulae are the so-called Archimedean copulae.

Definition 2.3.2 (Archimedean copula)**.** *A copula* $C(u_1, u_2, \ldots, u_n)$ *is called an Archimedean copula if its joint cumulative distribution function can be written as*

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

for all $0 \leq u_1, u_2, \ldots, u_n \leq 1$ *and where* ϕ *is a generator function satisfying*

- $\phi(1) = 0$;
- ϕ *is decreasing and convex.*
- *the inverse function* $f = \phi^{-1}$ *is completely monotonic, ie.* $(-1)^{-n} f^{(n)} \geq 0$ *for* $n = 0, 1, 2, \ldots$.

Some well known families of Archimedean copula include

- Clayton copula: $\phi(t) = t^{-\theta} - 1$, $\theta > 0$;
- Gumbel copula: $\phi(t) = (-\ln(t))^{\theta}$, $\theta \geq 1$;
- Frank copula: $\phi(t) = -\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$, $\theta \in \mathbb{R} \setminus \{0\}$.

The particular choice of a copula depends on how fits data. Every Archimedean copula behaves differently with respect to the tail dependence, a very important factor in dependence modelling.

As it has been already mentioned there is no analytical expression for the Spearman's ρ for Archimedean copulae. On the other hand, Kendall's τ can be computed with this formula [Genest and MacKay, 1986]

$$\tau = 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)}dt.$$

Different coefficients for the left and right tail dependence are desirable in many models, where complex dependencies occur. The Gumbel copula exhibits higher correlation in the right tail, whereas the Clayton copula shows tighter concentration of mass in the left tail (see Figure 2.4c and 2.4d). Quite often copulas can be *flipped*, i.e., instead of the original copula variable U we take $1 - U$, which means that a left tail dependent copula becomes a right tail dependent copula. This works only if a given copula is asymmetric obviously. The Frank copula is the only bivariate Archimedean copula symmetric about the main diagonal and the anti–diagonal of its domain, hence it has these coefficients equal (it is one of the copulae implemented in UNICORN). It should be also noted that the Clayton and the Gumbel copula in their standard forms realize only positive correlations. [Joe, 1997] and [Venter, 2002] provide a good overview of tail dependence for Archimedean copulae.

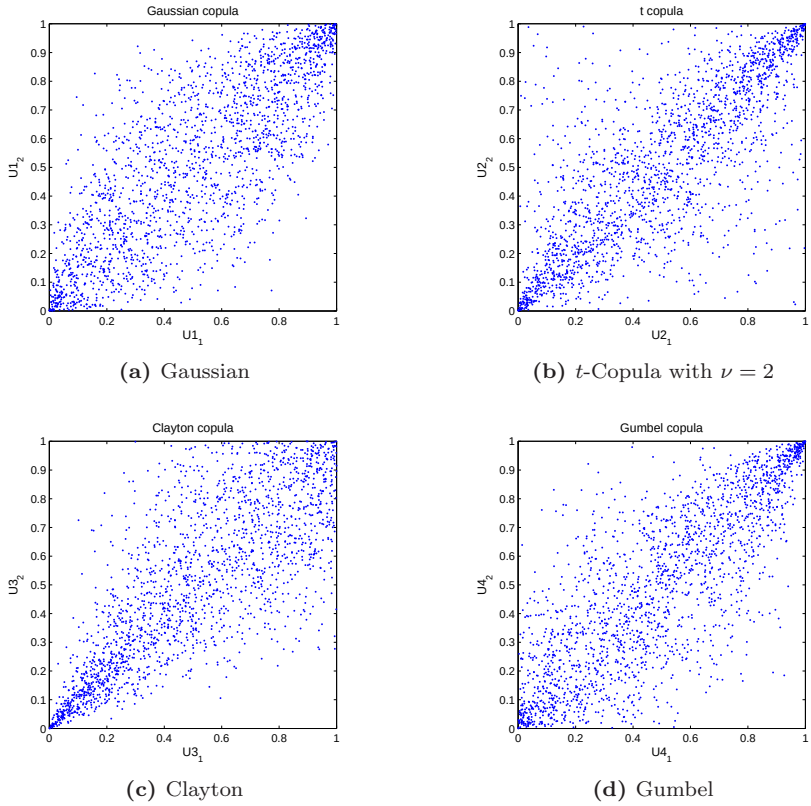
Often bivariate Archimedean copulae can be extended to higher dimensions by making use of their associativity property

$$C(C(u_1, u_2), u_3) = C(u_1, C(u_2, u_3)).$$

2.4 Dirichlet-type copula as an example of a multivariate copula

This section introduces a new copula of the Dirichlet type. The Dirichlet distribution has been studied mainly as a conjugate prior for the multinomial distribution in Bayesian analysis [see Gustafson and Walker, 2003, for instance]. Indeed, the standard Dirichlet distribution is defined on the n-simplex and as such it can represent vectors of probabilities since the sum of the components of this vector is unity.

[1]This section is based on a manuscript written jointly with Prof. Jolanta Misiewicz from University of Zielona Góra, Zielona Góra, Poland.



(a) Gaussian **(b)** t-Copula with $\nu = 2$

(c) Clayton **(d)** Gumbel

Figure 2.4: *Scatter plots of various copulae realizing rank correlation* $\rho_r = 0.8$.

Definition 2.4.1. *A random vector* $(D_1, D_2, \ldots, D_n)$ *has a Dirichlet distribution with positive parameters* $\beta_1, \beta_2, \ldots, \beta_n$ *(notation* $D(\beta_1, \ldots, \beta_n)$*) if*

1. $D_i \in [0,1]$, $i = 1, 2, \ldots, n$;
2. $\sum_{i=1}^{n} D_i = 1$ *almost surely;*
3. *the joint density function of* $D(\beta_1, \ldots, \beta_{n-1})$ *is*

$$\frac{\Gamma(\beta_1 + \ldots + \beta_n)}{\Gamma(\beta_1) \cdot \ldots \cdot \Gamma(\beta_n)} \left[\prod_{i=1}^{n-1} d_i^{\beta_i - 1} \right] \left(1 - \sum_{i=1}^{n-1} d_i \right)_+^{\beta_n - 1},$$

where $(c)_+ = \max(c, 0)$.

The j-th one-dimensional marginal density of the Dirichlet distribution is of the form

$$f_1(d_j) = \frac{\Gamma(\sum_{i=1}^{n} \beta_i)}{\Gamma(q_i)\Gamma(\beta_i)} d_i^{\beta_i - 1} (1 - d_i)_+^{q_i - 1},$$

where $q_i = \sum_{k \neq i} \beta_k$ and hence is Beta distributed with parameters β_i and q_i for all $j = 1, \ldots, n$. The marginals are uniform only for $\beta_i = q_i = 1$, and this means that the Dirichlet distribution has all marginals uniform only for $n = 2$. The class of Dirichlet distributions will be extended in the next section. The generalization allows to obtain uniform marginals for arbitrary dimension n.

This section is organized as follows. We generalize the class of Dirichlet distributions in section 2.4.1. Section 2.4.2 links the Gamma distribution with the Dirichlet distribution. We introduce a new multivariate copula based on Dirichlet distribution in section 2.4.3 and proceed with conclusions in section 2.5.

2.4.1 Generalized Dirichlet distribution

We generalize the Dirichlet distribution following Gupta et al. [1996]:

Definition 2.4.2 (Generalized Dirichlet distribution)**.** *We say that a random vector $(D_1, \ldots, D_n)$ follows the generalized Dirichlet distribution with positive parameters $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ (notation $D(\alpha_1, \ldots, \alpha_n; \beta_1, \ldots, \beta_n)$) if*

1. *$D_i \in [0, 1]$, $i = 1, \ldots, n$;*
2. *$\sum_{i=1}^{n} D_i^{\alpha_i} = 1$ almost surely;*
3. *the joint density function of $(D_1, \ldots, D_{n-1})$ is*

$$\frac{\Gamma(\delta_1 + \ldots + \delta_n)}{\Gamma(\delta_1) \cdot \ldots \cdot \Gamma(\delta_n)} \left[\prod_{i=1}^{n-1} \alpha_i \, d_i^{\beta_i - 1} \right] \left(1 - \sum_{i=1}^{n-1} d_i^{\alpha_i} \right)_+^{\delta_n - 1},$$

where $\delta_i = \frac{\beta_i}{\alpha_i}$ for $i = 1, 2, \ldots, n$.

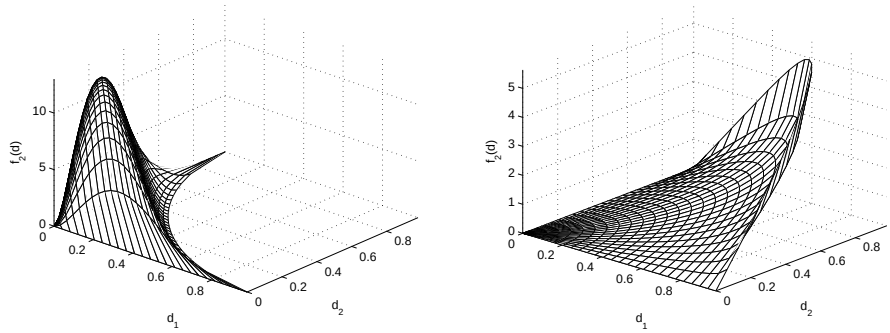
The Dirichlet distribution can also be symmetrized:

Definition 2.4.3 (Symmetrized Generalized Dirichlet distribution)**.** *A random vector $(D_1, \ldots, D_n)$ has a sign-symmetric Dirichlet-type distribution with parameters $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ (we use notation $D_s(\alpha_1, \ldots, \alpha_n; \beta_1, \ldots, \beta_n)$) if*

1. *D_i, $i = 1, \ldots, n$, is a symmetric random variable;*
2. *$\sum_{i=1}^{n} |D_i|^{\alpha_i} = 1$ almost surely;*
3. *the joint density function of $(D_1, \ldots, D_{n-1})$ is*

$$\frac{\Gamma(\delta_1 + \ldots + \delta_n)}{\Gamma(\delta_1) \cdot \ldots \cdot \Gamma(\delta_n)} \left(\prod_{i=1}^{n-1} \frac{\alpha_i}{2} |d_i|^{\beta_i - 1} \right) \left(1 - \sum_{i=1}^{n-1} |d_i|^{\alpha_i} \right)_+^{\delta_n - 1}.$$

If $\alpha_i \equiv 2$ and $\beta_i \equiv 1$ then the random vector $(D_1, \ldots, D_n)$ has uniform distribution on the unit sphere in $\mathbb{R}^n$. This special random vector will be denoted by $U^{(n)} = (U_{1,n}, \ldots, U_{n,n})$ and we denote the distribution of $U^{(n)}$ by ω_n.



(a) $\alpha = (0.6, 0.6, 0.6)$, $\beta = (2, 2, 2)$ **(b)** $\alpha = (3, 3, 3)$, $\beta = (2, 2, 2)$

Figure 2.5: *Examples of densities of Dirichlet distributed random vectors with given parameters.*

The marginal density function of $(D_1, \ldots, D_k)$, $k < n$, of the generalized Dirichlet distribution $D(\alpha_1, \ldots, \alpha_n; \beta_1, \ldots, \beta_n)$ at the point $\mathbf{d} = (d_1, d_2, \ldots, d_k)$ is given by the following

$$f_k(\mathbf{d}) = \frac{\Gamma\left(\sum_{i=1}^{n} \delta_i\right)}{\Gamma\left(\sum_{i=k+1}^{n} \delta_i\right)\left[\prod_{i=1}^{k} \Gamma(\delta_i)\right]} \left(\prod_{i=1}^{k} \alpha_i d_i^{\beta_i - 1}\right) \left(1 - \sum_{i=1}^{k} d_i^{\alpha_i}\right)_+^{\sum_{i=k+1}^{n} \delta_i - 1}.$$

The expectation of a product $\prod_{i=1}^{k} D_i^{h_i}$, $h_i \in \mathbb{R}$, can be expressed as [see Gupta et al., 1996]

$$\mathbf{E}\left(\prod_{i=1}^{k} D_i^{h_i}\right) = \frac{\Gamma\left(\sum_{i=1}^{n} \frac{\beta_i}{\alpha_i}\right)}{\Gamma\left(\sum_{i=1}^{k} \frac{\beta_i + h_i}{\alpha_i} + \sum_{i=k+1}^{n} \frac{\beta_i}{\alpha_i}\right)} \prod_{i=1}^{k} \frac{\Gamma\left(\frac{\beta_i + h_i}{\alpha_i}\right)}{\Gamma\left(\frac{\beta_i}{\alpha_i}\right)}. \tag{2.5}$$

The formula for the covariance, variance and product moment correlation of D_i and D_j can be easily determined from eq.(2.5). For instance

$$\mathrm{Cov}(D_i, D_j) = \frac{\Gamma(\delta_0)\left[\Gamma\left(\delta_0 + \frac{1}{\alpha_i}\right)\Gamma\left(\delta_0 + \frac{1}{\alpha_j}\right) - \Gamma(\delta_0)\Gamma\left(\delta_0 + \frac{1}{\alpha_i} + \frac{1}{\alpha_j}\right)\right]}{\Gamma\left(\delta_0 + \frac{1}{\alpha_i} + \frac{1}{\alpha_j}\right)\Gamma(\delta_i)\Gamma(\delta_j)},$$

where $\delta_0 = \sum_{i=1}^{n} \delta_i$.

2.4.2 Generalized Gamma distributions

The Dirichlet distribution has many connections with the Gamma and Beta distributions. The relation between the standard Dirichlet and the generalized Dirichlet distribution is analogous to the relation between the standard Gamma

and the generalized Gamma distribution. The well known Gamma distribution with positive parameters β, λ has the probability density function

$$f(x;\beta,\lambda) = \frac{\lambda^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-\lambda x}\, \mathbf{1}_{[0,\infty)}(x),$$

where $\Gamma(\beta)$ is the gamma function by $\Gamma(\beta,\lambda)$. The generalized Gamma distribution has an extra parameter $\alpha > 0$.

Definition 2.4.4 (Generalized Gamma and Symmetrized Generalized Gamma distribution)**.** *A random variable X follows the generalized Gamma distribution $\Gamma(\alpha,\beta,\lambda)$, $\alpha,\beta,\lambda > 0$, if it has density*

$$\frac{\alpha\lambda^\beta}{\Gamma(\beta/\alpha)} x^{\beta-1} e^{-(\lambda x)^\alpha}.$$

A random variable X has a generalized and symmetrized Gamma distribution $\Gamma_s(\alpha,\beta,\lambda)$, $\alpha,\beta,\lambda > 0$, if it has density

$$\frac{\alpha\lambda^\beta}{2\Gamma(\beta/\alpha)} |x|^{\beta-1} e^{-|\lambda x|^\alpha}.$$

Notice that $\Gamma(\beta,\lambda) = \Gamma(1,\beta,\lambda)$ and $\Gamma_s(\beta,\lambda) = \Gamma_s(1,\beta,\lambda)$. Moreover we have that if $X \sim \Gamma(\beta/\alpha,1) = \Gamma(1,\beta/\alpha,1)$, then $X^{1/\alpha}$ is distributed as $\Gamma(\alpha,\beta,1)$. If θ_0 is a random variable with probability distribution $\mathbf{P}\{\theta_0 = 1\} = \mathbf{P}\{\theta_0 = -1\} = \frac{1}{2}$, and is independent of X then the product $X^{1/\alpha}\theta_0$ has the generalized Gamma distribution $\Gamma_s(\alpha,\beta,1)$. It follows that a random variable Y with the generalized Gamma distribution $\Gamma_s(\alpha,\beta,\lambda)$ has the representation $Y \stackrel{d}{=} (X/\lambda)^{1/\alpha}\theta_0$.

Dirichlet distributed random variables can be generated by transforming Gamma random variables.

Proposition 2.4.1. *Let $X_1,\ldots,X_n$ be independent random variables with distributions $\Gamma(\alpha_i,\beta_i,1)$ respectively. Then the random vector*

$$\left(\frac{X_1}{(\sum_{i=1}^n X_i^{\alpha_i})^{1/\alpha_1}},\ldots,\frac{X_n}{(\sum_{i=1}^n X_i^{\alpha_i})^{1/\alpha_n}}\right)$$

has the generalized Dirichlet distribution $D(\alpha_1,\ldots,\alpha_n;\beta_1,\ldots,\beta_n)$.
Let $X_1,\ldots,X_n$ be independent random variables with distributions $\Gamma_s(\alpha_i,\beta_i,1)$ respectively. Then the random vector

$$\left(\frac{X_1}{(\sum_{i=1}^n |X_i|^{\alpha_i})^{1/\alpha_1}},\ldots,\frac{X_n}{(\sum_{i=1}^n |X_i|^{\alpha_i})^{1/\alpha_n}}\right)$$

has the sign-symmetric Dirichlet-type distribution $D_s(\alpha_1,\ldots,\alpha_n;\beta_1,\ldots,\beta_n)$.

This gives a very convenient way to generate Dirichlet distributed random vectors based on independent Gamma random variables.

2.4.3 Multivariate copulae

One of the features of the generalized Dirichlet distribution is that it can be a basis for a multidimensional copula constructed using two different techniques. The first one relies on applying the Sklar's theorem, that is transforming arbitrary marginal distributions to uniform. An interesting alternative exists however — the uniform distribution is in the class of all marginals of the generalized Dirichlet distribution. For certain values of its parameters such that all the marginals are uniform we obtain the Dirichlet-type copula. In other words, a high dimensional copula is a special case of the generalized Dirichlet distribution. It can be shown that in order to get uniform distributions for all one-dimensional marginals one has to take

$$\beta_1 = \cdots = \beta_n = 1, \quad \text{and} \quad \alpha_1 = \cdots = \alpha_n = n - 1.$$

Proposition 2.4.2. *A multidimensional generalized Dirichlet distribution, denoted as* $D(\alpha_1, \ldots, \alpha_n; \beta_1, \ldots, \beta_n)$*, is a copula if and only if* $\alpha_i = n - 1$ *and* $\beta_i = 1$ *for every* $i = 1, \ldots, n$*.*

From now on we will use the notation $D^n = (D_1^n, \ldots, D_n^n)$ for the random vector with distribution $\mathcal{D}(n-1, \ldots, n-1; 1, \ldots, 1)$ and the notation $D^{n,s} = (D_1^{n,s}, \ldots, D_n^{n,s})$ for the random vector with the distribution $\mathcal{D}_s(n-1, \ldots, n-1; 1, \ldots, 1)$.

Since the random vector $D^{n,s}$ has the joint density with contours symmetric about the origin, it follows that

$$\operatorname{Cov}\left(D_i^{n,s}, D_j^{n,s}\right) = 0, \quad \text{for every } i, j = 1, \ldots, n;\ i \neq j,$$

thus the copula $\mathcal{D}_s(n-1, \ldots, n-1; 1, \ldots, 1)$ will not be further investigated in this thesis.

On the other hand, given a random vector D^n and using eq.(2.5) allows to find

$$\tau_n \stackrel{def}{=} \mathbf{E}\left(D_i^n D_j^n\right) = \frac{\Gamma\left(\frac{n}{n-1}\right)\Gamma\left(\frac{2}{n-1}\right)^2}{\Gamma\left(\frac{n+2}{n-1}\right)\Gamma\left(\frac{1}{n-1}\right)^2}$$

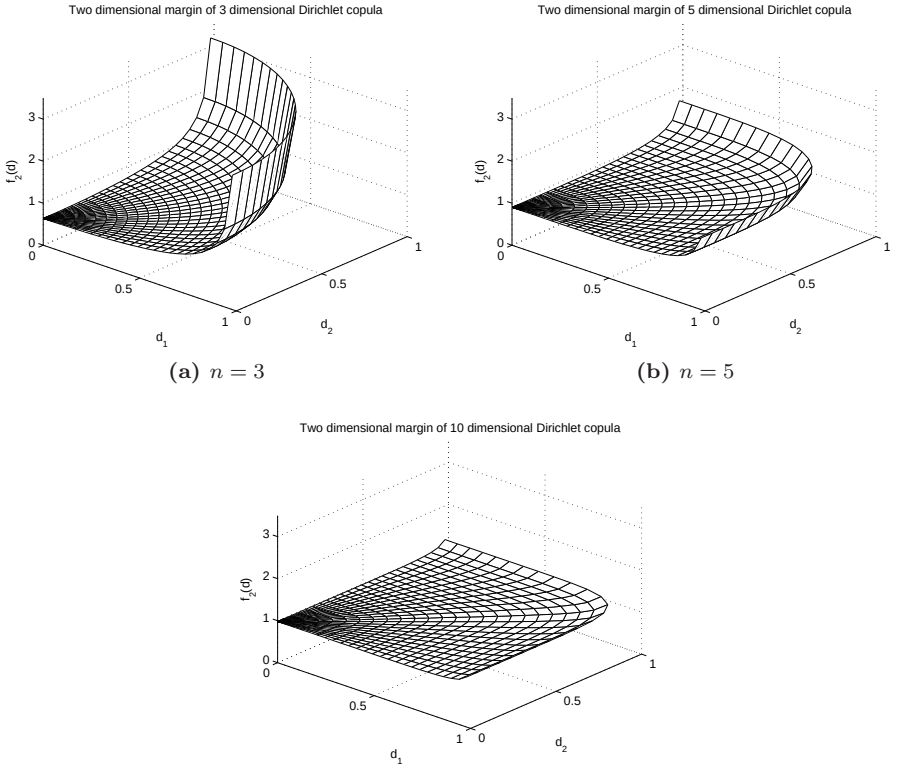
for $i \neq j$. Utilizing the relation $\Gamma(1+r) = r\Gamma(r)$ and substituting $p = 1/(n-1)$ we obtain

$$\tau_n = \frac{\Gamma(2p)^2}{3\Gamma(3p)\Gamma(p)}.$$

For $n = 2$ we find $\tau_2 = -1/6$, thus $\rho(D_1^2, D_2^2) = -1$. For $n = 3$ we have $\tau_3 = 2/(3\pi)$, thus $\rho(D_i^3, D_j^3) = (8 - 3\pi)/\pi$. We write for further reference

$$\rho_n \stackrel{def}{=} 12(\tau_n - 3),$$

where ρ_n is the product moment correlation of any pair of random variables D_i^n, D_j^n, $i \neq j$.



(a) $n = 3$ **(b)** $n = 5$

(c) $n = 10$

Figure 2.6: *Examples of Dirichlet type copulae.*

Figure 2.6 shows two dimensional marginals of n-dimensional Dirichlet-type copulae. The densities clearly converge to the uniform distribution on the unit square indicating decreasing correlation between the random variables as n increases.

The problem of having a fixed correlation given the dimension of the Dirichlet distributed random vector D can be overcome (to some extent) by applying partial symmetrization with respect to the diagonals and convex combinations. The construction is based on the fact, that if the random variable X is uniformly distributed on $[0, 1]$ then also $(1 - X)$ has a uniform distribution on $[0, 1]$. It is also easy to see that the random variable

$$X(\theta) = \theta X + (1 - \theta)(1 - X),$$

where θ independent of X has distribution $\mathbf{P}\{\theta = 1\} = p = 1 - \mathbf{P}\{\theta = 0\}$, has uniform distribution on $[0, 1]$, and the same we can say about the random variable $Z(\theta) = \theta X + (1 - \theta)(1 - X')$, where X' independent of θ is an independent copy

of X. Generalizing this simple remark we see that if

$$\mathbf{X}(\overline{\theta}) = (D_1^n(\theta_1), \ldots, D_n^n(\theta_n)),$$

where $\overline{\theta} = (\theta_1, \ldots, \theta_n)$ is a random vector independent of D^n taking values in $\{0,1\}^n$, then $\mathbf{X}(\overline{\theta})$ has a copula distribution. Moreover, for each fixed value of $\overline{\theta} = (\theta_1, \ldots, \theta_n)$, a choice of $i, j \in \{1, \ldots, n\}$ ensures

$$\begin{aligned}\left(D_i^n(\theta_i), D_j^n(\theta_j)\middle|(\theta_1, \ldots, \theta_n)\right) \stackrel{d}{=} \\ \left(D_i^n(\theta_i), D_j^n(\theta_j)\middle|(1, \ldots, 1, \theta_i, 1, \ldots, 1, \theta_j, 1, \ldots, 1)\right),\end{aligned}$$

where $\stackrel{d}{=}$ denotes equality of distributions. Assume that the distribution of the random vector $\overline{\theta}$ is given. Let

$$p_{i,j}(\varepsilon_1, \varepsilon_2) = \mathbf{P}\left\{\theta_i = \varepsilon_1, \theta_j = \varepsilon_2\right\};\ \varepsilon_1, \varepsilon_2 \in \{0,1\}$$

and

$$p_{ij} = p_{i,j}(1,1) + p_{i,j}(0,0).$$

Now we can calculate

$$\begin{aligned}\mathbf{E}\left(D_i^n(\theta_i)D_j^n(\theta_j)\right) &= \mathbf{E}\left(D_i^n D_j^n p_{i,j}(1,1)\right) + \mathbf{E}\left(D_i^n(1-D_j^n)p_{i,j}(1,0)\right) + \\ &\quad + \mathbf{E}\left((1-D_i^n)D_j^n p_{i,j}(0,1)\right) + \mathbf{E}\left((1-D_i^n)(1-D_j^n)p_{i,j}(0,0)\right) \\ &= \tau_n p_{i,j}(1,1) + \left(\frac{1}{2} - \tau_n\right) p_{i,j}(1,0) + \left(\frac{1}{2} - \tau_n\right) p_{i,j}(0,1) + \tau_n p_{i,j}(0,0) \\ &= \tau_n\left(2p_{i,j}(1,1) + 2p_{i,j}(0,0) - 1\right) + \frac{1}{2}\left(1 - p_{i,j}(1,1) - p_{i,j}(0,0)\right) \\ &= \tau_n\left(2p_{ij} - 1\right) + \frac{1}{2}\left(1 - p_{ij}\right).\end{aligned}$$

Finally we can calculate the product moment correlation

$$\rho\left(D_i^n(\theta_i), D_j^n(\theta_j)\right) = 12\left(\mathbf{E}\left(D_i^n(\theta_i)D_j^n(\theta_j)\right) - \tfrac{1}{4}\right) = (12\tau_n - 3)\left(2p_{i,j} - 1\right).$$

Attaining bounds of ρ_{ij} corresponds to setting $p_{ij} = \pm 1$. This however does not mean that the problem of finding the range of correlations obtainable with the Dirichlet-type copula can be directly translated into the problem of finding the range of obtainable correlations for Bernoulli distributed random vector $\overline{\theta}$. It would have been equivalent if the vector $(p_1, \ldots, p_n)$ of the means of $(\theta_1, \ldots, \theta_n)$ was fixed and $p_{i,j}(1,1) = \mathbf{E}(\theta_i\theta_j) = p_{ij}$, since then p_{ij} would fully control the correlation between θ_i and θ_j. In our case we can freely choose expectations of $\overline{\theta}$ and distribute the probability mass p_{ij} between $p_{i,j}(0,0)$ and $p_{i,j}(1,1)$.

2.5 Conclusions

Dirichlet-type copulae are examples of another type of distributions in the sparse class of multivariate copulae. Their construction is quite simple as they are special

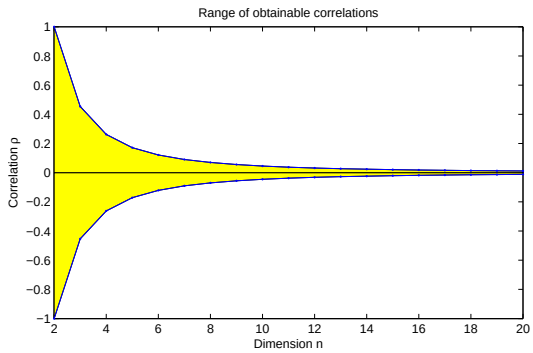


Figure 2.7: *Bounds on the range of obtainable correlations with the Dirichlet-type copula in a given dimension n.*

cases of generalized Dirichlet distributions. This simplicity is reflected also in the number of parameters controlling the copula, namely one — its dimension. The correlation between every pair of margins of this copula is the same and converges quickly to 0 as the dimension n increases. Because contours of two-dimensional margins of Dirichlet-type copula of dimension n are parts of unit spheres in L_n-space the correlation ρ_n is negative. In order to overcome the problem of having fixed correlation we apply the convex convolution of measures. This allows to obtain any correlation in $[-\rho_n, \rho_n]$.

We also briefly presented some other multivariate copulae in this chapter, namely Gaussian, t-copula and some examples of Archimedean copulae. Although various multivariate copulae share such desirable properties like tractability and are easy to cope with (see Gaussian copula as a prime example), they also suffer from a limited set of dependency structures they can satisfy. The t-copula is a welcome extension of the Gaussian copula in this sense. Archimedean copulae in their standard form also exhibit the limitations of Gaussian and t-copula — limited dependence structure. Various generalization of Archimedean copula are aiming on reducing this problem. Among the most interesting is the concept of *fully nested* Archimedean copulae, mentioned already in [Joe, 1997]. Bivariate margins of this copula are still positive quadrant dependent, but at least different degree of this dependence is allowed for different bivariate margins.

All these known results on multivariate extensions of standard copulae suggests, that going this road does not allow to introduce flexible, satisfying a wide range of dependence structures multivariate copulae. A better idea maybe is to *couple* bivariate margins having the properties we would like to introduce. With the vine-copula method we are not restricted to the use of one copula type for all bivariate margins and the process of constructing a copula is well defined.

CHAPTER 3

Generalized diagonal band copulae

> If I have ever made any valuable discoveries, it has been owing more to patient attention, than to any other talent.
>
> *Isaac Newton*

3.1 Introduction

Suppose we want to build a multivariate distribution based only on very limited information, such as univariate marginals and rank correlations elicited from experts. Assuming that these constraints are feasible, it would then be natural to choose that distribution which adds as little information as possible beyond them. The notion of mutual information between two continuous random variables can help make this statement more precise.

Let $f(x, y)$ denote the joint density of a random pair (X, Y), and denote by f_X and f_Y the corresponding marginals. The amount $I(X|Y)$ of information that each variable contains about the other may then be defined as

$$I(X|Y) = \iint f(x,y) \log \left\{ \frac{f(x,y)}{f_X(x) f_Y(y)} \right\} dx\, dy \geq 0. \tag{3.1}$$

In the subsequent references to this formula we write $I(f_X|f_Y)$ to also denote the relative information of X with respect to Y.

In the special case where f_X and f_Y are uniform on the interval $(0, 1)$, eq.(3.1) represents the relative information of the copula density $f(x, y)$ with respect to

[2]This chapter is based on the publication *Generalized diagonal band copulas* by Daniel Lewandowski, published in *Insurance: Mathematics and Economics*, Volume 37, pages 49–67, 2005.

the independent copula $f_X(x)f_Y(y) = 1$. More generally, $I(X|Y)$ vanishes if, and only if, X and Y are independent.

Bedford and Meeuwissen [1997] showed that the density $f(x, y)$ of the minimally informative copula with given rank correlation $\rho(\theta)$ is of the form

$$f_\theta(x, y) = \kappa\left(x - \frac{1}{2}\right)\kappa\left(y - \frac{1}{2}\right)e^{\theta(x-1/2)(y-1/2)}, \tag{3.2}$$

where $\kappa\,(x - 1/2)$ is even around $x = 1/2$. The correlation induced by this copula is controlled by the parameter θ. Although a Taylor series expansion for (3.2) is available, this minimally informative copula is not tractable and must be numerically approximated for each value of θ through a discretized optimization problem. An additional difficulty associated with the use of (3.2) is that no analytical form is generally available for its conditional cumulative distribution functions and their inverses. Accordingly, simulating from the least informative copula is inconvenient.

In this section, the search for a minimally informative copula satisfying correlation constraints is not considered in full generality, but rather within the broad class of generalized diagonal band (GDB) copulae introduced by Ferguson [1995]. This family of copulae, described in Section 3.2, extends the class of diagonal band (DB) copulae first considered by Cooke and Waij [1986]. Those GDB copulae that can be recovered by mixing only DB copulae are characterized in Section 3.3. In Section 3.4.2, we then deal specifically with the problem of approximating minimally informative GDB copulae with given correlation. Section 3.5 contains three examples of GDB copulae generated using Ferguson's method for constructing distributions in the convex closure of DB copulae; two other copulae already implemented in a software for uncertainty modeling called UNICORN are also described there. These five classes of copulae are then compared in Section 3.6 in terms of their relative information with respect to the uniform background measure under given correlation constraint. Finally, Section 3.7 contains conclusions.

3.2 Construction and properties of the generalized diagonal band copula

Introduced by Ferguson [1995], the generalized diagonal band (GDB) copula of a pair (X, Y) of uniform random variables on the unit interval is defined as follows.

Definition 3.2.1 (Generalized Diagonal Band copula)**.** *Let Z be a continuous random variable on the interval $[0, 1]$ with density g. An absolutely continuous copula C is a generalized diagonal band (GDB) copula if its associated density is of the form*

$$c(x, y) = \frac{1}{2}\left\{g(|x - y|) + g(1 - |1 - x - y|)\right\}. \tag{3.3}$$

In the sequel, g is called the generating density of the GDB copula.

In his paper, Ferguson [1995] emphasized that each GDB copula may be seen as a mixture of bivariate uniform densities on the boundaries of rectangles with

corners $(z,0)$, $(0,z)$, $(1-z,1)$ and $(1,1-z)$. The weight of each of the densities is given by $g(z)$, $z \in [0,1]$. Most of the basic properties of GDB copulae stem from this fact. In particular, note the symmetries

$$g(x) = c(x,0) = c(1-x,1) \quad \text{and} \quad g(y) = c(0,y) = c(1,1-y)$$

and the fact that for any $(x,y) \in A = \{(x,y)|0 \leq y \leq 1/2, y \leq x \leq -y+1\}$, Equation (3.3) simplifies to

$$c(x,y) = \frac{g(x-y) + g(x+y)}{2}\,.$$

For our purposes, one major advantage of the class of GDB copulae is that the mutual information associated with any copula of the form (3.1) can be expressed in terms of its generating density g. To see this, first observe that in the light of the symmetry of c and the above identity, we have

$$I(c|u) = 4 \iint_A c(x,y) \log\{c(x,y)\}\, dx\, dy,$$

where u denotes the uniform density on the unit square. Now if we substitute $x+y=v$ and $x-y=t$ (with Jacobian 1/2), we get

$$\begin{aligned} I(c|u) &= \int_0^1 \int_0^v \{g(v)+g(t)\} \log\{g(v)+g(t)\}\, dt\, dv \\ &\quad - \int_0^1 \int_0^v \{g(v)+g(t)\} \log(2)\, dt\, dv. \end{aligned}$$

Since

$$\int_0^1 \int_0^v \{g(v)+g(t)\}\, dt\, dv = 4 \iint_A \frac{g(x-y)+g(x+y)}{2}\, dx\, dy = 1$$

we get

$$I(c|u) = \int_0^1 \int_0^v \{g(v)+g(t)\} \log\{g(v)+g(t)\}\, dt\, dv - \log(2). \tag{3.4}$$

A second major advantage of the GDB class of copulae for our purposes stems from the simple relationship between the generating density g of a GDB copula C and the value of Spearman's rho for the associated pair (X,Y). To be specific, Ferguson [1995] showed that if Z is distributed as g, then

$$\rho_r(X,Y) = 1 - 6\mathbf{E}\left(Z^2\right) + 4\mathbf{E}\left(Z^3\right). \tag{3.5}$$

Using this fact and the above mentioned symmetries, one can thus check that if the GDB copula generated by $g(z)$ has correlation ρ, then $g(1-z)$ generates a GDB copula with correlation $-\rho$. Furthermore, the following relationships between the two GDB copulae hold:

$$\begin{aligned} c(x,y;\rho) &= c(1-x,y;-\rho), \\ C_{Y|X=x}(y;\rho) &= C_{Y|X=1-x}(y;-\rho), \\ C^{-1}_{Y|X=x}(y;\rho) &= C^{-1}_{Y|X=1-x}(y;-\rho). \end{aligned} \tag{3.6}$$

The third major asset of GDB copulae is the ease with which they can be generated, using explicit forms for their conditional and inverse conditional cumulative distribution functions. Particularly useful in this regard is the work of Bojarski [2001], who developed GDB copulae independently of Ferguson. Whereas Ferguson used a single generating function $g(z)$, Bojarski's generating density g_θ is a symmetric function whose support $[-1+\theta, 1-\theta]$ depends on a parameter $\theta \in [0,1]$. Nevertheless, the two approaches yield the same class of copulae, and the two generating functions are very closely related. Indeed, one can see that

$$g(z) = \begin{cases} g_\theta(-z) + g_\theta(z), & \text{if } z \in [0, 1-\theta]; \\ 0, & \text{otherwise.} \end{cases}$$

As implied by the work of Bojarski, the conditional density $c_\theta(y|x)$ is given by

$$c_\theta(y|x) = g_\theta(y-x) + \mathbf{1}_{\{y+x<1-\theta\}} g_\theta(-y-x) + \mathbf{1}_{\{y+x\geq 1+\theta\}} g_\theta(2-y-x),$$

which often leads to closed-form formulas for conditional and inverse conditional cumulative distribution functions. It also suggests a simple GDB copula sampling algorithm, namely:

Algorithm 3.2.1 (Sampling from a GDB copula).

1. Simulate independently a single x and y' according to the uniform distribution on $[0,1]$.
2. Calculate $y^* = G_\theta^{-1}(y'-x)$, where G_θ^{-1} is the inverse cumulative distribution function of the random variable with probability density g_θ.
3. If $y^* < 0$, then let $y = -y^*$; else if $y^* > 1$, then take $y = 1 - y^*$.
4. The pair (x, y) is then an observation from GDB copula density c associated with g.

The regression of Y given $X = x$, where X and Y are random variables joined by the GDB copula, can be calculated with the formula

$$\begin{aligned} \mathbf{E}(Y|X=x) &= \frac{1}{2}\left(\mathbf{E}[|U-x|] + 1 - \mathbf{E}[|1-U-x|]\right) \\ &= \frac{1}{2}\left[\int_0^x (x-u)\, G(u) du + \int_x^1 (u-x)\, G(u) du + 1 - \right. \\ &\quad \left. - \int_0^{1-x} (1-u-x)\, G(u)\, du + \int_{1-x}^1 (u+x-1)\, G(u)\, du \right] \end{aligned} \tag{3.7}$$

We will refer to this expression when showing examples of GDB copulae in the further sections.

The general form of the joint cumulative distribution function of a GDB copula can be expressed as a set of integrals. If $x \leq y$ and $x + y \leq 1$ then

$$F(x,y) = \int_0^{y-x} x\, G(u)\, du + \int_{y-x}^{x+y} \frac{x+y-u}{2}\, G(u)\, du. \tag{3.8}$$

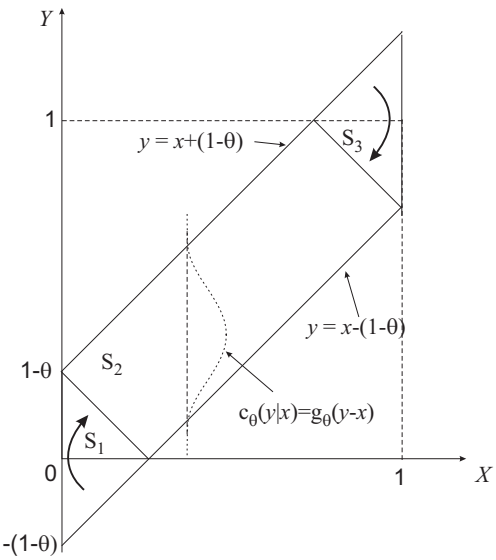


Figure 3.1: *Construction of the GDB copula using Bojarski's method.*

Use the fact that $F(x,y) = F(y,x)$ to obtain the distribution function for $x \geq y$ and $x + y \leq 1$. If $x \leq y$ and $x + y \geq 1$

$$F(x,y) = \tag{3.9}$$
$$= \int_0^{y-x} x\, G(u)\, du + \int_{y-x}^{2-x-y} \tfrac{x+y-u}{2}\, G(u)\, du + \int_{2-x-y}^{1} (x+y-1)\, G(u)\, du$$

and use the same fact $F(1-x, 1-y) = F(1-y, 1-x)$ to obtain the distribution function for $x \geq y$ and $x + y \geq 1$.

Before closing this section, let us mention diagonal band (DB) copulae as one of the simplest, yet very flexible one-parameter subclass of GDB copulae. A DB copula with parameter $\theta \in [0,1]$ (and hence $\rho \in [0,1]$) is given by

$$d_\theta(x,y) = \begin{cases} \dfrac{1}{1-\theta}, & \text{if } (x,y) \in S_1 \cup S_3; \\ \dfrac{1}{2(1-\theta)}, & \text{if } (x,y) \in S_2; \\ 0, & \text{elsewhere.} \end{cases}$$

Here,

$$\begin{aligned} S_1 &= \{(x,y) \in [0,1]^2 \,|\, x + y \leq 1 - \theta\}, \\ S_2 &= \{(x,y) \in [0,1]^2 \,|\, x - (1-\theta) \leq y \leq x + (1-\theta), x + y > 1 - \theta, x + y < 1 + \theta\}, \\ S_3 &= \{(x,y) \in [0,1]^2 \,|\, x + y \geq 1 + \theta\}, \end{aligned}$$

as displayed in Figure 3.1.

3.3 Mixtures of diagonal band copulae

In some circumstances, step 1 of Algorithm 3.2.1 is difficult to apply because the inverse cumulative distribution function G_θ^{-1} cannot be expressed in analytical form. We present here another approach to generating a wide subclass of GDB copulae that may sometimes solve this problem.

The content of this section is largely based on the work of Meeuwissen [1993], but with some corrections and extensions. We characterize a class of mixtures of DB copulae C_M, which cover a wide subset of the class of GDB copulae. To avoid unnecessary complications, we only deal here with absolutely continuous mixtures $c_M(x, y)$. Details concerning the treatment of the more general case including discontinuities are available in Appendix B.

Let $M(\theta)$ be a probability distribution on $[-1, 1]$ with discrete mass $1-p$ at the origin and the rest of the probability spread on $[-1, 1]$ according to a continuous function $m(\theta) \geq 0$, such that

$$\int_{-1}^{1} m(\theta)\, d\theta = p \in [0, 1].$$

We call this distribution a mixing function. Note, that the discrete atom of $M(\theta)$ with probability $1 - p$ corresponds to the independent copula being used in the mixture, hence the minimum value of the density of C_M is $1 - p$. A mixture of DB copulae may be defined as follows.

Definition 3.3.1 (Mixture of DB densities)**.** *A mixture $c_M(x, y)$ of DB densities $d_\theta(x, y)$ is given by*

$$c_M(x, y) = \int_{-1}^{1} d_\theta(x, y)\, dM(\theta).$$

For such mixtures the following hold:

a) $c_M(x, y) = c_M(y, x)$;

b) $c_M(x, y) = c_M(1 - y, 1 - x)$;

c) $c_M(x, y) = \frac{1}{2}\{c_M(|x - y|, 0) + c_M(1 - |1 - x - y|, 0)\}$.

Indeed, DB copulae have these properties for any $\theta \in [-1, 1]$, and the latter are obviously preserved under mixing.

In view of the above, mixtures of DB densities are in the class of GDB copulae. Our intention in this section is to show that reciprocally, a wide subclass of GDB copulae can be recovered from mixtures of DB copulae.

We begin by observing that since the density $c_M(x, y)$ of a mixture of DB copulae is uniquely determined by its conditional density $c_M(x, 0)$, the problem of mixing densities of DB copulae can be simplified to mixing conditional densities of DB copulae, which are step functions of the form

$$d_\theta(x, 0) = \begin{cases} 0, & \text{if } x \in [0, -\theta]; \\ \dfrac{1}{1 + \theta}, & \text{if } x \in (-\theta, 1]; \end{cases} \tag{3.10}$$

for $\theta \leq 0$ and

$$d_\theta(x,0) = \begin{cases} \dfrac{1}{1-\theta}, & \text{if } x \in [0, 1-\theta]; \\ 0, & \text{if } x \in (1-\theta, 1]; \end{cases} \tag{3.11}$$

for $\theta \geq 0$. These functions are finite for any $\theta \in (-1,1)$. For $\theta = 1$ or $\theta = -1$, we obtain the so-called Fréchet–Hoeffding bounds, and $d_{-1}(x,0)$ and $d_1(x,0)$ are Dirac delta functions.

We show in Theorem 3.3.1 below that for a generating density g satisfying certain conditions, there exists a mixing function M such that

$$g(x) = \int_{-1}^{1} d_\theta(x,0)\, dM(\theta) = \int_{-1}^{1} d_\theta(x,0) m(\theta)\, d\theta + (1-p)\, d_0(x,0). \tag{3.12}$$

Before stating the result, let us introduce two differentiable functions g^+ and g^-, whose derivatives with respect to x are as follows:

$$\frac{d}{dx} g^+(x) = \max\left\{ \frac{d}{dx} g(x), 0 \right\}, \quad g^+(0) = 0, \tag{3.13}$$

$$\frac{d}{dx} g^-(x) = \max\left\{ -\frac{d}{dx} g(x), 0 \right\}, \quad g^-(0) = 0. \tag{3.14}$$

Then $\frac{d}{dx} g(x) = \frac{d}{dx} g^+(x) - \frac{d}{dx} g^-(x)$ and $g(x) = g(0) + g^+(x) - g^-(x)$. See Figure 3.2 for an example of $g(x)$ with corresponding $g^+(x)$ and $g^-(x)$.

It can be shown that if $g(x) = c_M(x,0)$ is a conditional density of a mixture $c_M(x,y)$ of diagonal band copulae, then the continuous part of the mixing function M is

$$m(\theta) = \begin{cases} -(1+\theta) \dfrac{d}{d\theta} g^+(-\theta), & \theta < 0; \\ -(1-\theta) \dfrac{d}{d\theta} g^-(1-\theta), & \theta > 0. \end{cases} \tag{3.15}$$

We are now in a position to formulate the main theorem.

Theorem 3.3.1. *Let $c(x,y)$ be the density of a GDB copula generated with generating density $g(z)$, $z \in [0,1]$. If g is absolutely continuous and*

$$g(0) - g^-(1) \geq 0, \tag{3.16}$$

then $c(x,y)$ may be expressed as the density of a mixture of DB copulae.

Proof. We prove the result by showing that there exists a mixing function M given the stated assumptions. By the construction of the DB copula, as per equations (3.10) and (3.11), we have $\theta = -x$ if $\theta < 0$ and $\theta = 1 - x$ when $\theta > 0$. Hence in

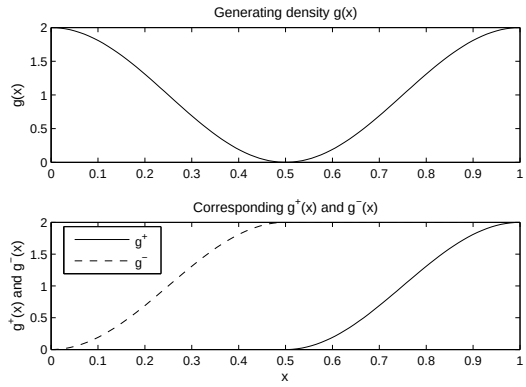


Figure 3.2: *Construction of $g^+(x)$ and $g^-(x)$.*

the light of (3.15),

$$\begin{aligned}
\int_{-1}^{1} m(\theta)\,d\theta &= \int_0^1 m(-x)\,dx + \int_0^1 m(1-x)\,dx \\
&= \int_0^1 (1-x)\frac{d}{dx}\,g^+(x)\,dx + \int_0^1 x\frac{d}{dx}\,g^-(x)\,dx \\
&= \int_0^1 \frac{d}{dx}\,g^+(x)\,dx - \int_0^1 x\left\{\frac{d}{dx}\,g^+(x) - \frac{d}{dx}\,g^-(x)\right\}dx \\
&= g^+(1) - \int_0^1 x\frac{d}{dx}\,g(x)\,dx = g^+(1) - g(1) + 1.
\end{aligned}$$

Also, as observed earlier, it follows from (3.13) and (3.14) that $g(x) = g^+(x) + g(0) - g^-(x)$, and hence in particular when $x = 1$. Thus

$$\int_{-1}^{1} m(\theta)\,d\theta = g^+(1) + 1 - g^+(1) - g(0) + g^-(1) = 1 - g(0) + g^-(1).$$

Now by assumption, we have $g(0) - g^-(1) \geq 0$, while $g(0) - g^-(1) \leq 1$ follows from the fact that g is a probability density on the interval $[0, 1]$. Consequently, we have

$$0 \leq \int_{-1}^{1} m(\theta)\,d\theta = p\,.$$

Define

$$g^\star(x) = \int_1^{-1} d_\theta(x, 0) m(\theta)\,d\theta.$$

It can be shown that

$$g^\star(x) = g^+(x) - g^-(x) + g^-(1),$$

which in light of (3.12) leads to $g(x) - g^\star(x) = g(0) - g^-(1) = 1 - p$. Therefore

$$g(x) = \int_1^{-1} d_\theta(x,0) m(\theta)\, d\theta + 1 - p = \int_1^{-1} d_\theta(x,0) m(\theta)\, d\theta + (1-p)\, d_0(x,0).$$

■

We call $g^-(1) \geq 0$ the total decrement of function g. All monotonic density functions g satisfy condition (3.16), but also many non-monotonic functions can be expressed as a mixture of the form (3.12). Note that parameter p can be easily determined as

$$1 - \int_{-1}^{1} m(\theta)\, d\theta = 1 - p = \min_{u \in [0,1]} g(u).$$

Mixtures of DB copulae are very easy to sample. The procedure is as follows:

Algorithm 3.3.1 (Sampling from mixtures of DB copulae).

1. Simulate a single θ according to $M(\theta)$.
2. Simulate a single observation according to d_θ.
3. To generate a pseudo-random sample of size n, repeat the above steps n times.

Hence if it is problematic to derive $G_\theta^{-1}(z)$, but easy to obtain $M^{-1}(\theta)$ in analytical form, then one can use algorithm 3.3.1, instead of algorithm 3.2.1. In order for this to work, of course, $g_\theta(z)$ must generate a GDB copula which is also a mixture of DB copulae.

3.4 Minimally informative GDB copula

In many situations we are interested in adding as little extra information to the studied problem as possible. The copula ensuring this is the constrained minimum information copula — it has the lowest value of the relative information with respect to the independent copula (uniform) given constraints. We will constrain on correlations. We present the constrained minimally informative copula below and show how to obtain minimally informative mixtures of diagonal band copulae.

3.4.1 Minimum information copula

Suppose we want to reconstruct a multivariate distribution with given marginals and correlation structure elicited from experts. If this is the only information we have, then it is desired to choose the least *informative* distribution among all the other multivariate distributions with the same marginals and correlation structure, provided it exists. If one wants to obtain the minimally informative joint distribution given the correlation specification on a dependence tree or vine, then the constrained minimally informative copula assigned to the edges of the tree/vine will ensure that (see Theorems 7 and 12 in [Cooke, 1997]).

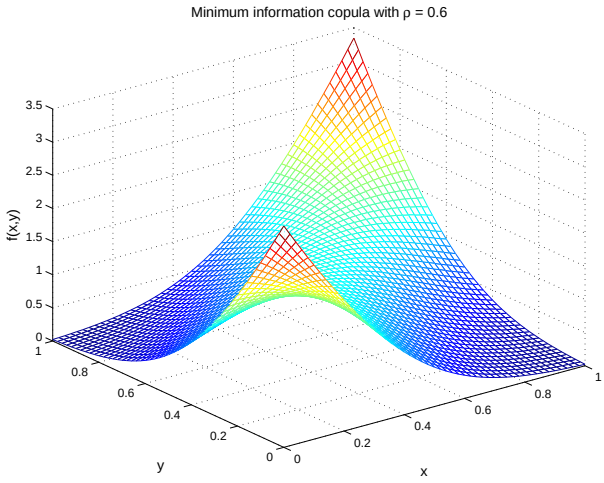


Figure 3.3: *Minimum information copula with correlation* $\rho = 0.6$.

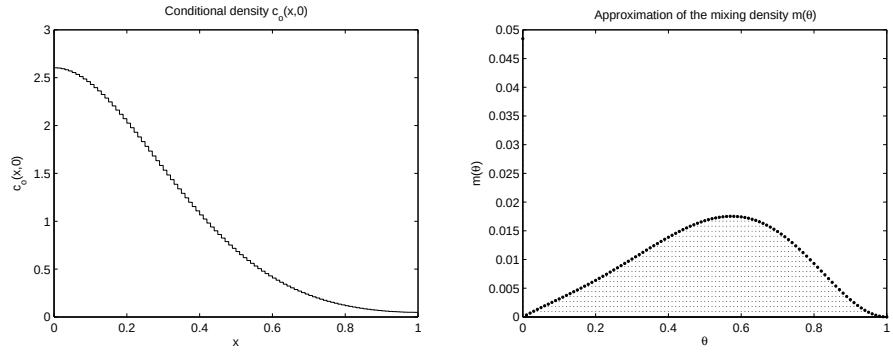
There has been a great deal of effort concentrated on determining the expression for the least informative copulae with given correlation (see Figure 3.3 for an example of a minimum information copula) and Bedford and Meeuwissen [1997] solved this problem. They showed that the density $f(x, y)$ of the minimally informative copula has functional form

$$f(x,y) = \kappa\left(x - \frac{1}{2}\right)\kappa\left(y - \frac{1}{2}\right)e^{\theta(x-\frac{1}{2})(y-\frac{1}{2})}$$

where $\kappa\left(x - \frac{1}{2}\right)$ is a function even around $x = \frac{1}{2}$, for which a Taylor series expansion has been determined. This solution is not easily tractable and for each value of the rank correlation a discretized optimization problem must be solved in order to obtain a numerical approximation. This is not efficient from the computational point of view and can affect the accuracy. Therefore an approximation to the minimum information copula is needed, which would provide an analytical form for the conditional cumulative distribution function and inverse conditional distribution function for this copula.

3.4.2 Approximation to the minimally informative GDB copula given the correlation constraint

Part of Meuwissen's research on mixtures of DB copulae concerned solving a discretized optimization problem. The solution of this problem was a discretized conditional density (step function) that met the conditions of Theorem 3.3.1 generalized to the non-continuous case. Therefore, it could be considered a finite mixture of DB densities. This density generated a GDB copula with minimal mutual information with respect to the uniform distribution under a correlation constraint ρ. Although GDB copulae had not been formally introduced at that



(a) The conditional density $c_o(x, 0)$. **(b)** Mixing weights for parameter θ.

Figure 3.4: *The conditional density $c_o(x, 0)$ of the minimally informative GDB copula with $\rho = 0.6$ and the corresponding mixing weights as the solution of the system of nonlinear equations (3.18)-(3.20) where $n = 100$.*

time, Meeuwissen made use of the unique property of DB copulae (see eq. (3.3)), which was later to be discovered by Ferguson [1995], allowing him to introduce the entire class of copulae.

Let us briefly describe the approach Meeuwissen took in order to solve this optimization problem. Assume that we are looking for an optimal GDB copula $c_o(x, y)$ with minimal relative information with respect to the independent copula under the correlation constraint generated by a step function $c_o(x, 0)$. For $i = 1, \ldots, n$, let $c_i(x)$ denote the value of the step function at point $x \in (x_{i-1}, x_i]$, where $x_0, \ldots, x_n$ is a partition of $x \in [0, 1]$ into n intervals of equal length $1/n$ ($x_0 = 0$ and $x_n = 1$). Hence the solution is a vector of length n of non-negative real numbers.

Meeuwissen [1993] showed that the relative information of a GDB copula $c_o(x, y)$ generated by $c_o(x, 0)$ with respect to the uniform background measure is

$$I(c_o|u) = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \frac{c_i + c_j}{n^2} \log\left(\frac{c_i + c_j}{2}\right) + \sum_{i=1}^{n} \frac{c_i}{n^2} \log(c_i), \tag{3.17}$$

and the correlation ρ realized by $c_o(x, y)$ is

$$\rho(X, Y) = 1 + \sum_{i=1}^{n} c_i \left\{ \left(x_i^4 - x_{i-1}^4\right) - 2\left(x_i^3 - x_{i-1}^3\right) \right\}.$$

Meeuwissen differentiated eq. (3.17) with respect to each of $c_1, \ldots, c_n$ in order to obtain n equations necessary to solve the problem. Then the solution is the

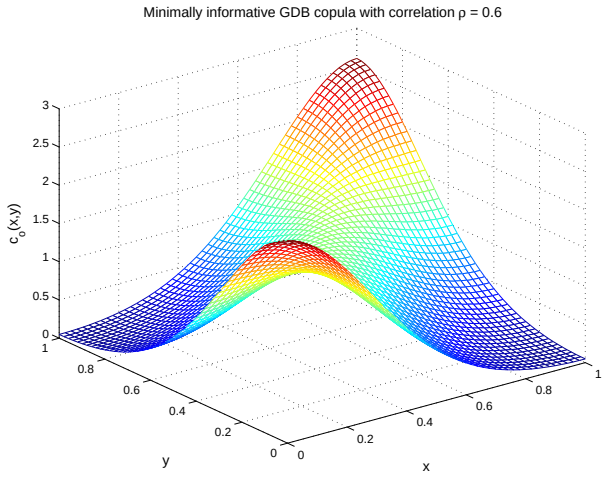


Figure 3.5: *Minimally informative GDB copula with correlation* $\rho = 0.6$.

result of solving the following system of $n + 2$ non-linear equations

$$\frac{1}{n}\sum_{j=1}^{n}\log(c_i + c_j) + \lambda\left\{\left(x_i^4 - x_{i-1}^4\right) - 2\left(x_i^3 - x_{i-1}^3\right)\right\} + \mu = \log(2) - 1, \quad (3.18)$$

$$\sum_{j=1}^{n} c_j\left\{\left(x_j^4 - x_{j-1}^4\right) - 2\left(x_j^3 - x_{j-1}^3\right)\right\} = \rho - 1, \quad (3.19)$$

$$\frac{1}{n}\sum_{j=1}^{n} c_i = 1. \quad (3.20)$$

The mixing weights for the mixture of DB copulae are extracted from the conditional distribution $c_o(x, 0)$. These weights for minimally informative mixture of DB copulae with $\rho = 0.6$ are presented in Figure 3.4b. Notice that the weight for $\theta = 0$ is equal to the minimum of $c_o(x, 0)$ and directly corresponds to the contribution of the uniform bivariate distribution.

As one can see in Figure 3.5, the numerically derived conditional distribution $c(x, 0)$ converges to a solution whose derivative with respect to x equals 0 for $x = 0$ and $x = 1$. This provides *smoothness* of the copula, i.e., differentiability everywhere on the unit square, even along both the diagonals.

3.5 Examples of GDB copulae

As we have already mentioned, our main goal is to find a copula within the class of GDB copulae which (i) approximates the minimum information copula; (ii) is capable of realizing any correlation $\rho \in (-1, 1)$; and (iii) provides an analytical form for the conditional cumulative distribution function and its inverse. The DB copula meets most of these criteria, except for the *small informativeness*. In fact,

it is not hard to find a family of GDB copulae having less information than the DB copula. However, the additional conditions, in particular the availability of an analytical form for the inverse cumulative distribution function, make this problem more complicated. There are few probability distributions that are flexible enough to generate a GDB copula with an arbitrary correlation, and at the same time that have a simple enough form to allow for various kinds of analytical transformations.

In this section, we introduce three families of GDB copulae which have an analytical form and which, to some extent, comply with the above mentioned desiderata. These copulae can be generated by applying Ferguson's approach and achieve non-negative correlations. If negative correlations are desired, one need only make use of property (3.6).

3.5.1 Triangular generating function

Assume the following generating function with non-negative parameter a

a) if $0 \leq a \leq 2$,

$$g_a(z) = -az + 1 + a/2\,, \quad z \in [0,1],$$

b) if $a \geq 2$

$$g_a(z) = \begin{cases} -az + \sqrt{2a}, & \text{if } z \in \left[0, \sqrt{2/a}\right]; \\ 0, & \text{if } z \in \left[\sqrt{2/a}\,, 1\right]. \end{cases}$$

Based on these equations for the generating density of the GDB copula, the conditional and inverse conditional cumulative distribution functions can be determined and formulated in closed form expressions. However, we shall not mention them here, in view of their complexity. Figure 3.6 shows the density of this copula and the corresponding mixing function $m(\theta)$.

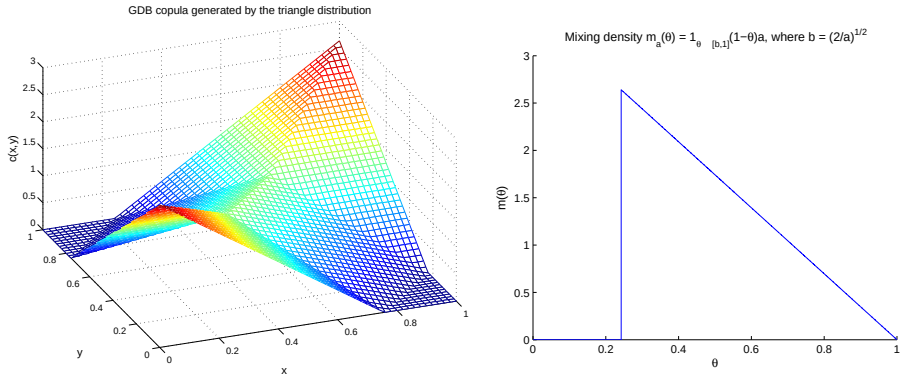
3.5.2 Truncated exponential distribution as the generating function

A GDB copula will now be presented that is generated by the truncated exponential distribution with truncation parameter equal to 1. The probability density of a truncated exponential random variable Z with truncation parameter 1 is given by

$$g_\lambda(z) = \frac{\lambda\, e^{-\lambda z}}{1 - e^{-\lambda}}\,, \quad \text{for } z \in [0,1]. \tag{3.21}$$

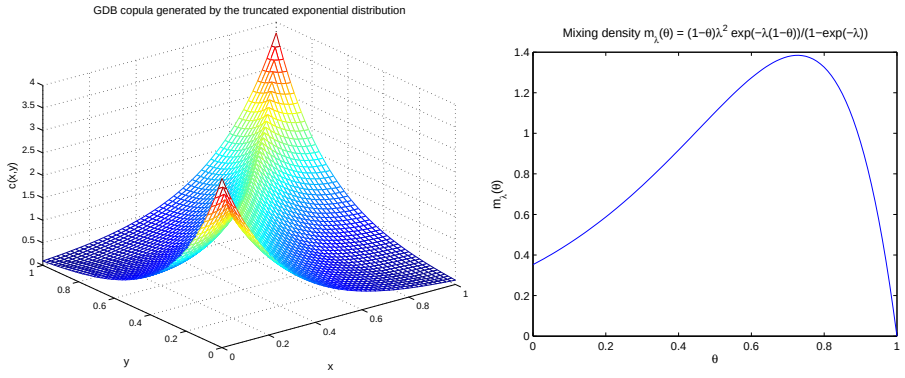
The derivative of $g_\lambda(z)$ with respect to z is not 0 at $z = 0$ and $z = 1$, hence the generated copula density will not be differentiable on the diagonals.

This family of copulae does not include a copula for which $\rho = 0$, since the exponential distribution family does not include the uniform distribution or, at least, a distribution symmetric about point $z = 1/2$. However, the independence case $\rho = 0$ is a limiting case corresponding to $\lambda = 0$ since $g_\lambda(z) \to 1$ as $\lambda \to 0$ for all $z \in [0,1]$. The conditional and inverse conditional cumulative distribution function have closed form expressions for GDB copulae generated by (3.21) (see Figure 3.7 for an example of the density function of this copula).



(a) Density **(b)** Mixing weights for parameter θ.

Figure 3.6: *GDB copula with correlation* $\rho = 0.6$ *generated by the triangle distribution with parameter* $a = 3.4849$.



(a) Density **(b)** Mixing weights for parameter θ.

Figure 3.7: *GDB copula with correlation* $\rho = 0.6$ *generated by the truncated exponential distribution with parameter* $\lambda = 3.6673$.

A relationship between correlation ρ and parameter λ is needed now. The equation we have to solve for λ is given below

$$1 + 2\,\frac{-6\,e^{\lambda}\lambda + \lambda^3 - 6\,\lambda + 12\,e^{\lambda} - 12}{\lambda^3\,(e^{\lambda} - 1)} = \rho.$$

This equation can be solved numerically. It can be shown (we use eq.(3.7)) that for the truncated exponential distribution case the regression curve is given as follows

$$\mathbf{E}(Y|X = x) = \frac{e^{\lambda}\lambda\,x - \lambda + \lambda\,x + e^{-\lambda\,(x-1)} - e^{\lambda\,x}}{\lambda\,(e^{\lambda} - 1)}.$$

The joint cdf and pdf expressions for this copula can be calculated using

eq.(3.8) and (3.9)

$$F(x,y)=\begin{cases}\frac{2xe^{\lambda}\lambda-e^{\lambda(1+x-y)}+e^{-\lambda(x+y-1)}}{2\lambda(e^{\lambda}-1)}, & \text{if } x\le y \text{ and } x+y\le 1;\\ \frac{2ye^{\lambda}\lambda-e^{\lambda(1-x+y)}+e^{-\lambda(x+y-1)}}{2\lambda(e^{\lambda}-1)}, & \text{if } x\ge y \text{ and } x+y\le 1;\\ \frac{2\lambda\left(xe^{\lambda}-x-y+1\right)+e^{\lambda(x+y-1)}-e^{\lambda(1+x-y)}}{2\lambda(e^{\lambda}-1)}, & \text{if } x\le y \text{ and } x+y\ge 1;\\ \frac{2\lambda\left(ye^{\lambda}-x-y+1\right)+e^{\lambda(x+y-1)}-e^{\lambda(1-x+y)}}{2\lambda(e^{\lambda}-1)}, & \text{if } x\ge y \text{ and } x+y\ge 1.\end{cases}$$

$$f(x,y)=\begin{cases}\frac{\lambda\left(e^{\lambda(1+x-y)}+e^{\lambda(1-x-y)}\right)}{2(e^{\lambda}-1)}, & \text{if } x\le y \text{ and } x+y\le 1;\\ \frac{\lambda\left(e^{\lambda(1-x+y)}+e^{\lambda(1-x-y)}\right)}{2(e^{\lambda}-1)}, & \text{if } x\ge y \text{ and } x+y\le 1;\\ \frac{\lambda\left(e^{\lambda(1+x-y)}+e^{\lambda(-1+x+y)}\right)}{2(e^{\lambda}-1)}, & \text{if } x\le y \text{ and } x+y\ge 1;\\ \frac{\lambda\left(e^{\lambda(1-x+y)}+e^{\lambda(-1+x+y)}\right)}{2(e^{\lambda}-1)}, & \text{if } x\ge y \text{ and } x+y\ge 1.\end{cases}$$

The conditional cumulative distribution functions can be expressed as follows. First assume positive correlations. If $x\in[0,1/2]$ then

$$F_{Y|X}(y)=\begin{cases}\frac{e^{\lambda(1-x)}\sinh\lambda y}{e^{\lambda}-1}, & y\in[0,x];\\ \frac{e^{\lambda}\left[1-e^{-\lambda y}\cosh\lambda x\right]}{e^{\lambda}-1}, & y\in(x,1-x];\\ 1-\frac{e^{\lambda x}\sinh\lambda(1-y)}{e^{\lambda}-1}, & y\in(1-x,1],\end{cases}$$

where sinh and cosh are hyperbolic sine and cosine, respectively. If $x\in(1/2,1]$ then

$$F_{Y|X}(y)=\begin{cases}\frac{e^{\lambda(1-x)}\sinh\lambda y}{e^{\lambda}-1}, & y\in[0,1-x];\\ \frac{e^{-\lambda x}\left[-2e^{\lambda x}+e^{\lambda y}\left(e^{\lambda}+e^{\lambda(2x-1)}\right)\right]}{2(e^{\lambda}-1)}, & y\in(1-x,x];\\ 1-\frac{e^{\lambda x}\sinh\lambda(1-y)}{e^{\lambda}-1}, & y\in(x,1].\end{cases}$$

In order to derive the inverse cumulative distribution functions for negative correlations use the following property of the GDB copula

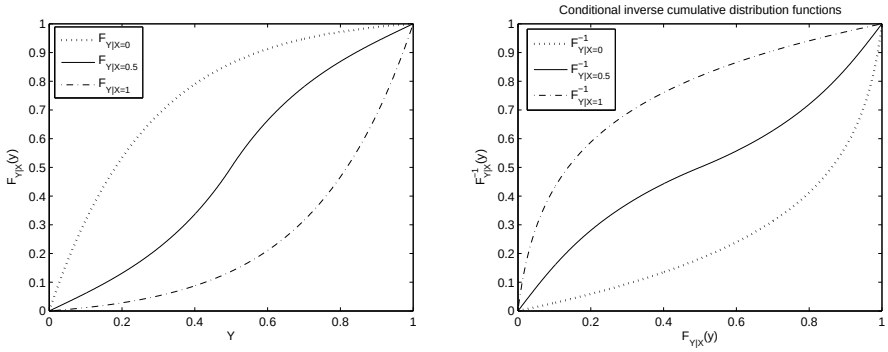
$$F_{Y|X=x}(y;\rho)=F_{Y|X=1-x}(y;-\rho).$$

Based on the above one can derive the inverse cumulative distribution functions necessary for use with the vine-copula method.

3.5.3 The ogive distribution as the generating function

The optimization problem solved in Section 3.4.2 gives a characterization of an optimal generating density providing a GDB copula with minimal relative information under the correlation constraint. Unfortunately, it is quite problematic to find a probability density g such that

$$g'(0)=g'(1)=0, \tag{3.22}$$



(a) Conditional cdf **(b)** Inverse conditional cdf

Figure 3.8: *Conditional and inverse conditional cdf's for the GDB copula generate by the truncated exponential distribution with parameter* $\lambda = 2.1624$ *(*$\rho = 0.4$*)*

and which can be easily integrated and differentiated. Kotz and van Dorp [2004] proposed the so-called *ogive* distribution, whose density function is given by

$$f(z;q) = \frac{2q(1-z)^{q-1}\left\{2q-1-(q-1)(1-z)^q\right\}}{3q-1}, \quad q>1. \tag{3.23}$$

For $q > 2$, density (3.23) has the property (3.22). The GDB copula generated by the ogive distribution with $q = 2$ has correlation $\rho \approx 0.34$. Hence in order to generate *smooth* GDB copulae achieving lower correlations, we must mix the ogive distribution with the uniform density as follows:

$$g(z;p,q) = p + (1-p)f(z;q). \tag{3.24}$$

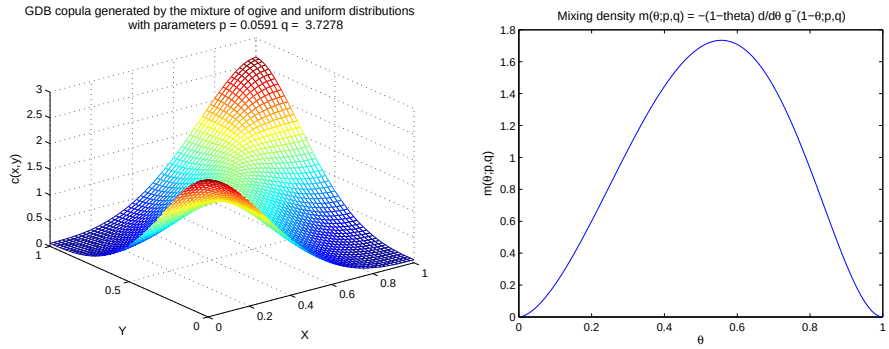
Here $p \in [0,1]$ and $q > 1$ are parameters. This generating function ensures smoothness of the generated copula along the diagonals, and hence lower relative information compare to copulae generated by the truncated exponential or triangular density functions. In Figure 3.9a, we present a GDB copula with correlation $\rho = 0.6$ generated by the mixture of ogive and uniform density (3.24) with parameters $p = 0.0591$ and $q = 3.7278$.

Substituting the second and third central moment of random variable Z with probability density (3.24) into (3.5) yields

$$\rho = (1-p)\left\{1 - \frac{90\,q^3+168\,q^2+18\,q-36}{(3\,q-1)\,(1+q)\,(q+2)\,(q+3)\,(2\,q+3)}\right\},$$

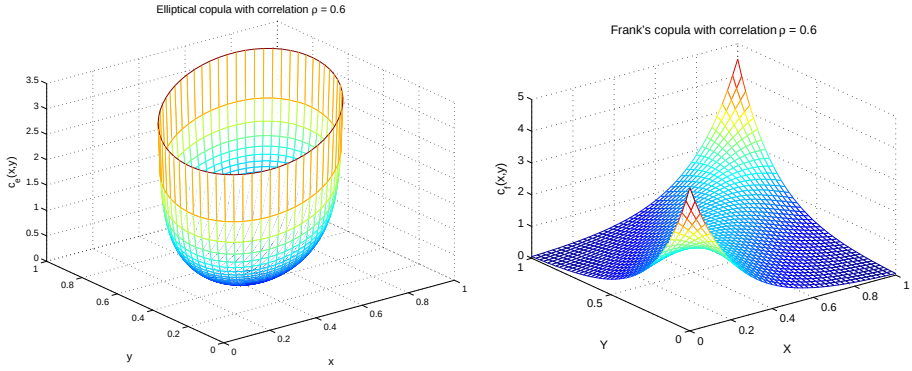
and solving this for q allows us to find an analytically given relationship between correlation ρ and parameters q and p, viz.

$$\begin{aligned} p &= \frac{(6\,\rho-6)\,q^5+(43\,\rho-43)\,q^4+(105\,\rho-15)\,q^3}{(1-q)\,(2\,q-1)\,(3\,q^3+26\,q^2+45\,q+18)} \\ &\quad + \frac{(95\,\rho+73)\,q^2+(9\,\rho+9)\,q-18\,\rho-18}{(1-q)\,(2\,q-1)\,(3\,q^3+26\,q^2+45\,q+18)}. \end{aligned}$$



(a) Density **(b)** Mixing weights for parameter θ.

Figure 3.9: *GDB copula with correlation* $\rho = 0.6$ *generated by the the mixture of ogive and uniform distributions with parameters* $p = 0.0591$ *and* $q = 3.7278$.



(a) "Elliptical copula" with correlation $\rho = 0.6$. **(b)** Frank's copula with correlation $\rho = 0.6$ ($\alpha = 4.4658$).

Figure 3.10: *Other copulae implemented in* UNICORN.

We choose parameters p and q such that the copula generated by (3.24) with these two parameters has minimal information among all other copulae in its class realizing the same rank correlation ρ. Complexity of the expression representing the mutual information (3.4) for this copula does not allow to solve this problem analytically. As an alternative, we implemented a numerical routine searching for optimal values of the parameters given the correlation constraint.

Unfortunately, the conditional cumulative distribution functions for this copula are not analytically invertible. However, with one extra step in the algorithm of Section 2, we can sample easily from this distribution. Simply, first *sample* the generating density for GDB copula, the uniform distribution with probability p or the ogive TSP distribution with probability $1 - p$. Then follow the general approach for sampling from a GDB copula with given generating density, since

both distributions have invertible cumulative distribution functions.

As we show in the next section, this copula contains less relative information for a given correlation than any other GDB copula introduced in this paper given in analytical form.

3.5.4 Other copulae

We briefly present also two other families of copulae implemented already in UNICORN (UNcertainty analysis wIth CORelations), a software for dependence modeling with correlations developed at the Department of Mathematics of Delft University of Technology. We compare these with the just generated GDB copulae in terms of their relative information with respect to the uniform distribution under fixed correlation constraint. The first copula implemented in UNICORN was the diagonal band copula. Later the minimum information copula was implemented in the form of precomputed tables stored in memory. This solution was neither memory efficient nor very accurate.

Elliptical copula

The "elliptical copula" [see Kurowicka et al., 2001] is an absolutely continuous, centrally symmetric copula with linear regression that can realize any correlation in $(-1, 1)$ (Fig.3.10a). Let

$$e(x, y; \rho) = \left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2 - 2\rho\left(x - \frac{1}{2}\right)\left(y - \frac{1}{2}\right).$$

Then the copula's density $c_e(x, y)$ with correlation $\rho \in (-1, 1)$ is

$$c_e(x, y; \rho) = \begin{cases} \dfrac{1}{\pi\sqrt{\frac{1}{4}(1-\rho^2) - e(x, y; \rho)}}, & (x, y) \in B; \\ 0, & (x, y) \notin B; \end{cases}$$

where

$$B = \left\{ (x, y) \in [0, 1]^2 \,\middle|\, e(x, y; \rho) < \frac{1}{4}(1 - \rho)^2 \right\}$$

This "elliptical copula" is a particular case of the multivariate Pearson type II distribution and should not be confused with the notion of meta-elliptical copulae widely used in the literature [see Fang et al., 2002, Abdous et al., 2005].

The main disadvantage of this copula is its high mutual information coefficient for given correlation and the fact that it does not include the independent copula.

Frank's copula

Frank's copulae is the only class of centrally symmetric Archimedean copulae (Fig.3.10b). It is a one-parameter distribution with density

$$c_f(x, y; \alpha) = \frac{\alpha\left(1 - e^{-\alpha}\right)\left(e^{-\alpha(x+y)}\right)}{\left\{1 - e^{-\alpha} - (1 - e^{-\alpha x})(1 - e^{-\alpha y})\right\}^2} \tag{3.25}$$

for non–zero parameter α. Parameter $\alpha > 0$ corresponds to positive correlations and vice versa. The independent copula $u(x, y) = 1$ can be seen as the limit

$$u(x, y) = \lim_{\alpha \to 0} c_f(x, y; \alpha).$$

For additional details about this family of copulae, see [Frank, 1979, Nelsen, 1986, Genest, 1987].

This copula can be problematic to implement in software. The inverse conditional cumulative distribution function of this copula includes the term $1 - \exp(-\alpha)$, of which a natural logarithm is computed. Unfortunately, most computer floating-point number representations would consider this term to be simply equal 0 for $\alpha > 37.429$. This means that in practice, the highest correlation that this copula can realize is of the order of 0.987. A similar term occurs in the denominator of eq.(3.25), which also causes numerical problems for large values of α, and x and y close to 1.

For the two copulae presented in this section, conditional and inverse conditional cumulative distribution functions are given in closed form expressions. However, they differ substantially from each other in the mutual information they contain with respect to the independent copula under the correlation constraint.

3.6 Relative information of various copulae

In this section, we compare the mutual information values for the presented copulae as functions of the rank correlation. For any GDB copula with density $c(x, y)$ generated by the generating function g, the relative information of c with respect to the uniform density u can be calculated with the following formula

$$I(c|u) = 4 \int_0^{\frac{1}{2}} \int_y^{1-y} \left(\frac{g(x+y) + g(x-y)}{2} \right) \log \left(\frac{g(x+y) + g(x-y)}{2} \right) dx\, dy.$$

Equivalently if we substitute $x + y = v$ and $x - y = t$ (the Jacobian is 1/2) one can use this expression instead

$$I(c|u) = \int_0^1 \int_0^v (g(v) + g(t)) \log \left(g(v) + g(t) \right) dt\, dv - \log(2). \tag{3.26}$$

or the following proposition

Proposition 3.6.1. *For any GDB copula $C(x, y)$ with density $c(x, y)$ generated by the generating function g, the relative information of c with respect to the uniform density u is*

$$I(c|u) \quad = \quad \int_0^1 g(u) \log(g(u) + g(1)) du - \int_0^1 \int_0^1 \frac{\frac{dg(v)}{dv}\, v\, g(u)}{g(u) + g(v)}\, dv\, du - \log(2)$$

Proof. Solve eq.(3.26) by parts. ■

Table 3.1: *The relative information of the minimum information copula for given rank correlation (col. A) and percent increment for other copulae.*

Rank correlation	Copula								
r	A	B	C	D	E	F	G	H	I
0.1	0.00498	0.80	3.70	3.82	5.02	5.02	0.04	538.76	11013.25
0.2	0.02016	0.94	3.48	3.77	5.16	5.26	0.20	247.37	2721.58
0.3	0.04630	1.10	3.15	3.97	5.62	5.81	0.64	150.60	1186.33
0.4	0.08489	1.35	2.68	4.57	7.03	6.34	1.52	99.41	648.72
0.5	0.13853	1.69	2.15	6.83	7.54	7.09	2.65	69.40	399.72
0.6	0.21212	2.02	1.52	5.28	6.19	7.74	4.02	51.03	263.74
0.7	0.31526	2.35	1.01	3.85	4.78	8.23	5.58	37.39	180.75
0.8	0.47140	2.45	0.86	2.80	3.64	8.12	7.10	27.45	124.70
0.9	0.75820	2.18	1.35	2.07	2.76	7.09	8.25	18.58	81.85
0.95	1.06686	1.45	1.32	1.34	1.91	5.54	8.01	14.24	60.51
0.99	1.82640	1.16	0.98	0.47	0.85	3.27	6.54	8.84	37.26

A - minimum information copula
B - Frank's copula
C - minimally informative GDB copula
D - GDB copula generated with the mixture of uniform and TSP distribution
E - GDB copula generated with the triangle distribution
F - GDB copula generated with the truncated exponential distribution
G - Gaussian copula
H - DB copula
I - Elliptical copula

The relative information values presented in Table 3.1 have been calculated numerically by first generating a given copula density on a grid of 500 by 500 cells, and then approximating the relative information based on this density. We used this method, because there is no closed form expression for the density of the minimum information copula. Therefore, we decided to apply the same numerical method to all copulae considered. The order of the copulae in the table reflects their performance in terms of the relative information coefficient, in comparison with the minimum information copula given the rank correlation. The percentages in columns B–I express the increase in the relative information coefficient relatively to this value for the minimum information copula.

We have one remark concerning the results in Table 3.1. We have shown, that the constrained minimally informative mixture of DB copulae, with the correlation being the constraint, is also a constrained minimally informative copula in the class of all GDB copulae. Therefore numbers in columns D–F should never be lower than in column C of the table. There is however such case for $\rho = 0.99$. This is clearly a numerical error resulting from approximating the relative information coefficient from discretized densities. With ρ approaching 1 some discretized densities may exhibit numerical instabilities as their values increase to infinity in the corners on the domain, causing errors in the estimates. All of the densities

converge to the upper Fréchet-Hoeffding bound, but do it in a slightly different way.

3.7 Conclusions

GDB copulae form a very large family of copulae, with the class of mixtures of DB copulae as an important subset. Theorem 3.3.1 gives a characterization of this subclass. In this paper, we systematized and extended current knowledge on the class of GDB copulae. The main appeal of the GDB copula is its simple and intuitive construction. Ferguson's method provides a simple expression for the rank correlation coefficient and straightforward sampling routine, whereas the Bojarski's method allows to simplify determining the conditional and inverse conditional cumulative distribution functions for a given GDB copula.

The copulae presented in this paper can be used to approximate the minimum information copulae. The copulae generated by the triangle density and the mixture of uniform and ogive TSP distribution can achieve any correlation $\rho \in (-1, 1)$; the same holds for the copula generated by the truncated exponential density. Further research on sampling with vines should concentrate on overcoming the requirement of having analytical forms for the conditional cumulative distribution functions and their inverses giving more freedom in choosing a copula.

Algorithms for generating samples from GDB copulae (satisfying certain conditions) have been proposed. If the inverse cumulative G_θ^{-1} of the generating density G_θ is given in analytical form, then algorithm 3.2.1 can be applied. If this is not the case, but one can sample from $M(\theta)$, then algorithm 3.3.1 is available for use.

CHAPTER 4

Building discretized minimally-informative copulae with given constraints

> The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them.
>
> *Sir William Bragg*

4.1 Introduction

Decision support for problems involving uncertainty necessarily involves the modelling of those uncertainties using joint probability distributions. Expert assessment is usually used — sometimes in combination with statistical data — to assess those distributions. Bayesian Belief Networks (see [Jensen, 2001]) provide one possible method of modelling joint probability distributions and have become very popular with the advent of easy to use software such as HUGIN [Andreassen et al., 1989], GENIE and recently introduced yet already very powerful UNINET. The latter has been developed at the Department of Mathematics of Delft University of Technology. One of the limitations of Bayesian Belief Networks (BBN) is however that the elicitation burden for experts is rather *heavy* when the natural quantification route of marginals and conditionals is taken.

Much research has been carried out about eliciting marginal distributions [Cooke, 1991, Cooke and Goossens, 2000]. Much less has been written about eliciting joint information. Common strategies involve making assumptions about the joint distribution in order to reduce the required information to the elicitation of a

[3]This chapter is based on a manuscript written jointly with Prof. Tim Bedford from University of Strathclyde, Glasgow, UK.

correlation coefficient [Clemen et al., 2000]. For example, the method of Iman and Conover [1982] assumes that after coordinate transformation of the marginals to make them normal marginals, the joint distribution has become joint normal. Copulae provide another route to quantifying joint distributions. They have become very popular, with the Archimedean family in particular being used frequently [Smith, 2003, Genest and Rivet, 1993]. Bedford and Meeuwissen [1997] take the copula which has minimal information with respect to the uniform (independent) copula amongst all those with a given (expert specified) Spearman rank correlation.

The use of rank correlation as a measure of the degree of association between two variables is a good first step but has some clear limitations, the most obvious being that the interpretation is very difficult for domain experts (who are not necessarily expert in the subtleties of the many varieties of correlation). Alternatively one can elicit information about observable quantities and infer the dependency structure to be consistent with the observables [Kraan and Bedford, 2003, Kraan, 2002]. Another reason to consider observables in preference to rank correlations is that correlations between more than two variables have to satisfy algebraic relations which may not be obvious to the expert. Cooke [1997] and Bedford and Cooke [2002] use an alternative parametrization via a vine structure to define a correlation matrix without the problem of algebraic relations. This is at the cost of having to elicit more complex conditional rank correlations. However this issue has been addressed in [Morales-Napoles et al., 2007].

This chapter sets out to show that we can use the minimum information techniques from [Bedford and Meeuwissen, 1997] in conjunction with expert elicitation of observables, to define a copula that represents the decision makers uncertainty about the joint distribution of two random variables. This method differs from previous methods also in that it allows interactive elicitation of expert opinions by giving guidance as to what values of uncertain quantities are compatible with the assessments already made. The method is based on using a D_1AD_2 algorithm to determine the copula based on potentially asymmetric information about the two variables. This contrasts with the simpler DAD algorithm used to determine copulae with given rank correlation in [Bedford and Meeuwissen, 1997], as rank correlation information is intrinsically symmetric information about the unknown quantities.

Section 4.2 and 4.3 show the principles of using the D_1AD_2 algorithms for 2 and 3 dimensional cases, respectively. The usage of the 2-dimensional version of the algorithm has been illustrated in section 4.4. Finally, in section 4.5 we give an example of a software for interactive expert elicitation implementing the D_1AD_2 algorithm for constructing a minimally informative copula given constraints. A simple expert elicitation is carried out for an artificial data set.

4.2 The D_1AD_2 algorithm

Bedford and Meeuwissen [1997] applied a so-called DAD algorithm to produce discretized minimally informative copula with given rank correlation. This algo-

rithm works because we know the general form taken by the copula, but relies on the fact that the correlation is determined by the mean of the symmetric function UV. The same approach can be used whenever we wish to specify the expectation of any symmetric function of U and V. In order to have asymmetric specifications we need to use the more general approach provided in [Borwein et al., 1994]. It states that if $\mathbf{A}$ is a positive square matrix (called a kernel), then we can find row vectors $\mathbf{D}_1$ and $\mathbf{D}_2$ such that the cell-wise product of $\mathbf{D}_1^T\mathbf{D}_2$ and $\mathbf{A}$ is doubly stochastic. The theory exists also for continuous functions and, indeed in much more generality.

Suppose there are two random variables X and Y, with cumulative distribution functions F_X and F_Y respectively. These are the variables of interest that we would like to correlate by introducing constraints based on some knowledge about functions of these variables. Suppose there are k of these functions, namely $h'_1(X,Y), h'_2(X,Y), \ldots, h'_k(X,Y)$, and that the expert wishes to specify mean values $e_1, \ldots, e_k$ for all these functions respectively. We can find corresponding functions of the copula variables U and V, defined by $h_1(U,V) = h'_1(F_1^{-1}(U), F_2^{-1}(V))$, etc., and clearly these should also have the specified expectations $e_1, \ldots, e_k$. Let u denote the realization of U and v the realization of V. We form the kernel

$$A(u,v) = \exp(\lambda_1 h_1(u,v) + \ldots + \lambda_k h_k(u,v)). \tag{4.1}$$

For practical implementations we have to discretize the set of (u,v) values such that the whole domain of the copula is covered. This means that the kernel A described above becomes a 2-dimensional matrix $\mathbf{A}$ and that we seek row vectors $\mathbf{D}_1$ and $\mathbf{D}_2$. Together they allow computing a doubly stochastic matrix $\mathbf{B}$ over $[0,1]^2$, that is a discretized copula density

$$\mathbf{B} = \mathbf{D}_1^T\mathbf{D}_2.\mathbf{A}, \tag{4.2}$$

where . (dot) denotes the cell-wise product operator applied to same size matrices.

For each vector $(\lambda_1, \ldots, \lambda_k)$ we can use the D_1AD_2 algorithm to generate a unique joint density with uniform marginals. This copula gives the vector of functions $(h_1, \ldots, h_k)$ an expected value vector which we call $\phi(\lambda_1, \ldots, \lambda_k)$. Now, general theory [see Borwein et al., 1994] says that this copula is always the unique minimum information copula (with respect to the uniform distribution) giving the expected value vector $\phi(\lambda_1, \ldots, \lambda_k)$. Furthermore, the mapping ϕ maps $\mathbb{R}^k$ onto the set of achievable expected value vectors. That set of possible expected value vectors is a convex set, but little else can be said about it in general.

Suppose that both U and V are discretized into n points, respectively u_i, and v_j, $i,j = 1, \ldots, n$. Then we write $\mathbf{A} = (a_{ij})$, $\mathbf{D}_1 = \left(d_1^{(1)}, \ldots, d_n^{(1)}\right)$, $\mathbf{D}_2 = \left(d_1^{(2)}, \ldots, d_n^{(2)}\right)$, where $a_{ij} = A(u_i, v_j)$, $d_i^{(1)} = D_1(u_i)$, $d_j^{(2)} = D_2(v_j)$. The double stochasticity of D_1AD_2 with the extra assumption of uniform marginals means, that

$$\forall_{i=1,\ldots,n} \textstyle\sum_j d_i^{(1)} d_j^{(2)} a_{ij} = n, \text{ and}$$
$$\forall_{j=1,\ldots,n} \textstyle\sum_i d_i^{(1)} d_j^{(2)} a_{ij} = n,$$

since for any given i and j the selected cell size in the unit square is $1/n^2$. Hence

$$d_i^{(1)} = \frac{n}{\sum_j d_j^{(2)} a_{ij}} \text{ and } d_j^{(2)} = \frac{n}{\sum_i d_i^{(1)} a_{ij}}.$$

The D_1AD_2 algorithm works by fixed point iteration and is closely related to iterative proportional fitting algorithms [Csiszar, 1975]. The idea is very simple - start with arbitrary positive initial vectors for $\mathbf{D}_1$ and $\mathbf{D}_2$. Then successively define new vectors by iterating the maps

$$d_i^{(1)} \mapsto \frac{n}{\sum_j d_j^{(2)} a_{ij}} \quad (i = 1, \ldots, n), \quad d_j^{(2)} \mapsto \frac{n}{\sum_i d_i^{(1)} a_{ij}} \quad (j = 1, \ldots, n).$$

This iteration converges geometrically to give us the vectors required.

4.3 The DAD algorithm for the 3-dimensional case

The principles described in section 4.2 can be successfully adopted for constructing 3-dimensional and higher dimensional counterparts of minimally informative copulae derived in the previous section. For the 3-dimensional case, if $\mathbf{A}$ is a positive cube matrix (called a kernel), then we have to find three row vectors $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$ such that the cell-wise product $\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{D}_3$, $\mathbf{A}$ is triply stochastic. The continuous version of the kernel has the following form:

$$A(u, v, t) = \exp(\lambda_1 h_1(u, v, t) + \ldots + \lambda_k h_k(u, v, t)). \tag{4.3}$$

The triple stochasticity requirement means that

$$\sum_{i,j} d_i^{(1)} d_j^{(2)} d_s^{(3)} a_{ijs} = n^2, \quad \sum_{i,s} d_i^{(1)} d_j^{(2)} d_s^{(3)} a_{ijs} = n^2$$
$$\text{and } \sum_{j,s} d_i^{(1)} d_j^{(2)} d_s^{(3)} a_{ijs} = n^2.$$

Hence

$$d_i^{(1)} = \frac{n^2}{\sum_{j,s} d_j^{(2)} d_s^{(3)} a_{ijs}}, \; d_j^{(2)} = \frac{n^2}{\sum_{i,s} d_i^{(1)} d_s^{(3)} a_{ijs}} \text{ and } d_s^{(3)} = \frac{n^2}{\sum_{i,j} d_j^{(1)} d_j^{(2)} a_{ijs}}.$$

Similarly to the 2-dimensional case, the $D_1D_2D_3A$ algorithm works by fixed point iteration as follows. Start with arbitrary starting vectors for D_1, D_2 and D_3. Then successively define new vectors by iterating the maps

$$d_i^{(1)} \mapsto \frac{n^2}{\sum_{j,s} d_j^{(2)} d_s^{(3)} a_{ijs}} \quad (i = 1, \ldots, n), \quad d_j^{(2)} \mapsto \frac{n^2}{\sum_{i,s} d_i^{(1)} d_s^{(3)} a_{ijs}} \quad (j = 1, \ldots, n)$$
$$\text{and} \quad d_s^{(3)} \mapsto \frac{n^2}{\sum_{i,j} d_j^{(1)} d_j^{(2)} a_{ijs}} \quad (s = 1, \ldots, n).$$

Using the analogy a similar algorithm for higher dimensional minimally informative copulae can be derived.

4.4 Constructing minimally informative copula with the D_1AD_2 algorithm

The mapping from the set of vectors of λ's onto the set of vectors of resulting expectations of functions $(h_1, \ldots, h_k)$ has to be found numerically. We employ optimization techniques for achieving the result. Experts specify expectations e_i of k functions of variables X and Y

$$E[h_i'(X,Y)] = E[h_i(U,V)] = e_i, \;\; i = 1, 2, \ldots, k.$$

The discretized copula density $\mathbf{B}$ is given by eq.(4.2). Hence, if one wants to determine λ's satisfying expert's assessments, then the following set of equations has to be solved

$$l_i(\lambda_1, \ldots, \lambda_k) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{b} \mathbf{B}(u_j, v_k) h_i(u_j, v_k) - e_i, \;\; i = 1, 2, \ldots, k. \tag{4.4}$$

The left hand sides of the above equations are just functions of λ's and with optimization algorithms their roots can be found. One of the possible solvers for this task would be FSOLVE - MATLAB's optimization routine. It implements various root finding techniques allowing for choosing the one suiting our problem best. However we also obtained good results by using another of MATLAB's optimization procedures in the example below, namely FMINSEARCH, which implements the Nelder-Mead simplex method [Lagarias et al., 1998]. The minimized function is

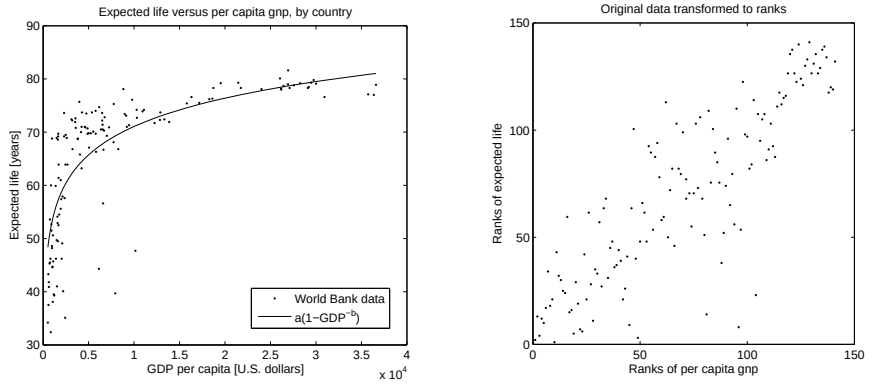
$$l_{sum}(\lambda_1, \ldots, \lambda_k) = \sum_{i=1}^{k} l_i^2(\lambda_1, \ldots, \lambda_k).$$

Example - World Bank data Consider the World Bank data on life expectancy at birth collected in years 2000-2005 (variable X) and GDP per capita collected in year 2002 (variable Y). The relation between between these two variables is of considerable interest for social planning. Rather than seeking a functional relation between these variables, we construct a copula that represents their joint distribution.

The data is available for 141 countries from around the world and is shown in Figure 4.1. This plot includes a regression fit of the form

$$\bar{Y} = a(1 - X^{-b})^2, \tag{4.5}$$

where $a = 11900$ and $b = 0.0006503$. This model type is known as a constant relative risk aversion model (CRRA) and has been widely used in climate change studies [Nordhaus, 2008]. The empirical cumulative distribution function for random variables X and Y with realizations x and y are denoted by F_X and F_Y respectively. It should be noted that all presented results were obtained based on discretized copula densities computed on a grid of 500 by 500 equally spaced points.



(a) Original data **(b)** Ranks of the original data

Figure 4.1: *World Bank data on life expectancy at birth versus GDP per capita for 141 countries.*

A minimally informative copula (denoted as C_1) under constraints will be constructed to model the World Bank data. Suppose the objective function for the D_1AD_2 algorithm is chosen to be $h_1'(X,Y) = XY$. We fix the value of this expectation

$$e_1 = \frac{1}{141}\sum_{i=1}^{141} x_i \cdot y_i = 634\,100.$$

This means that in fact the covariance and the product moment correlation are being fixed, since $E[X]$, $E[Y]$, $Var[X]$ and $Var[Y]$ are given. The resulting minimally informative copula density $c_1(u,v)$ for X and Y given $E[XY]$ fixed is presented in Figure 4.2. In order to make the plot clearer each copula variable has been discretized into only 30 equally spaced points.

One could introduce more constraints like higher order cross moments $E[X^pY^q]$. Our simulation results have shown however that the simulated samples given only the $E[XY]$ constraint exhibit high level of concordance of higher order cross moments with their counterparts for the original World Bank data already. More constraints can be added in order to get a better fit, but there is no need for adding more constraints of this type.

A second constraint is added to model better the regression eq.(4.5) in the original data. We also fix the expectation of $h_2'(X,Y) = (Y - a(1 - X^{-b}))^2$. The value of $E[h_2'(X,Y)]$ estimated from the World Bank data is 62.8242 and we take this value as the second constraint in our optimization problem. Both objective functions are shown in Figure 4.3 together with their counterparts in the copula space.

The minimum information copula C_2 with respect to the uniform distribution given two constraints $e_1 = E[h_1(U,V)] = 634\,100$ and $e_2 = E[h_2(U,V)] = 62.8242$ has been constructed on the same grid of 500 by 500 points. The Lagrange

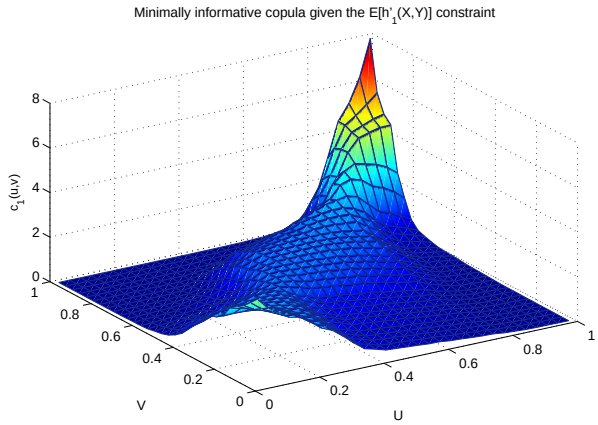


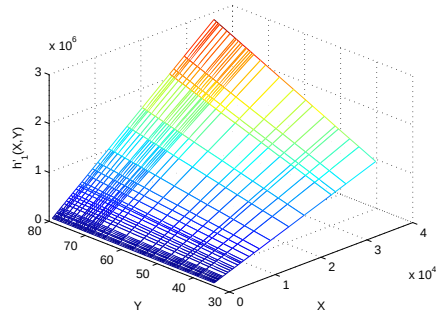
Figure 4.2: *The minimally informative copula for the World Bank data given the* $E[h_1'(X,Y)]$ *constraint.*

Table 4.1: *Comparison of various statistics for constructed copulae.*

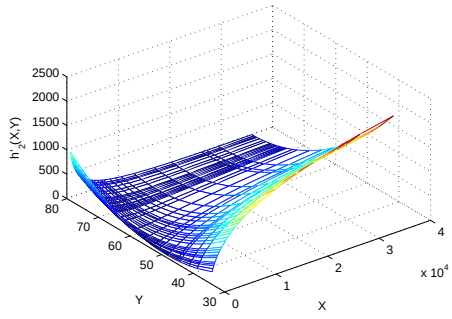
	C_1	C_2	$Frank_{maxlike}$	$Frank_{con}$
$E[h_1'(X,Y)]$	634 100	634 100	632 640	634 100
$E[h_2'(X,Y)]$	67.0508	62.8242	60.5371	56.0238
LogLikelihood	85.3314	86.1270	88.5635	87.8631
Relative information	0.6268	0.6322	0.6501	0.7221

multipliers in this case are $\lambda_1 = 3.7856 \cdot 10^{-5}$ and $\lambda_2 = -0.0026811$. Figure 4.4 shows the resulting minimum information copula density $c_2(u,v)$ on a grid of 30 by 30 points. It differs slightly from the copula in Figure 4.2. Although both perfectly realize the first constraint, only the second one realizes the second constraint (see Table 4.1).

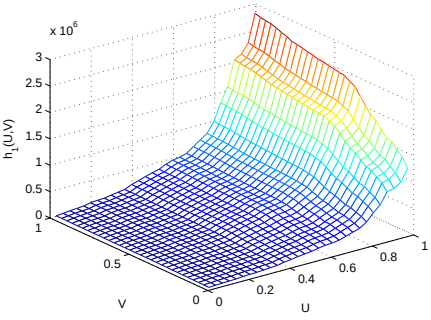
The D_1AD_2 approach is very well suited for constructing minimally informative copulae with imposed constraints with respect to the uniform background measure. However it does not guarantee to fit the data better than other copulae as measured by the likelihood score. In fact, we found a parametric copulae that yields higher likelihood score than C_1 or C_2 and it is the Frank copula $Frank_{maxlike}$, whose parameter τ is chosen to maximize the likelihood of the World Bank data (see Table 4.1). None of the constraints $E[h_1(U,V)]$ and $E[h_2(U,V)]$ apply to this copula. For comparison we also find Frank's copula $Frank_{con}$, such that the first constraint is satisfied. This lowers the likelihood score albeit it is still higher than what C_1 or C_2 achieved. Also, none of the other maximum likelihood parametric copulae tested (Gaussian, t-copula, Gumbel, Clayton) scores higher on the likelihood than C_1 or C_2 for this data set. This shows that while keeping the relative information low, C_1 and C_2 provide a very good fit to the data. We expect the D_1AD_2 approach to perform even better with respect to the likelihood score for less symmetric data, where standard centrally-



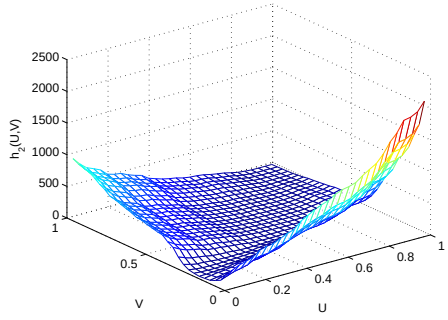
(a) $h_1'(x,y) = xy$



(b) $h_2'(x,y) = (y - a(1 - x^{-b}))^2$



(c) $h_1(u,v) = F_X^{-1}(u)F_Y^{-1}(v)$



(d) $h_2(u,v) = (F_Y^{-1}(v) - a(1 - (F_X^{-1}(u))^{-b}))^2$

Figure 4.3: *Plots of objective functions over their domains.*

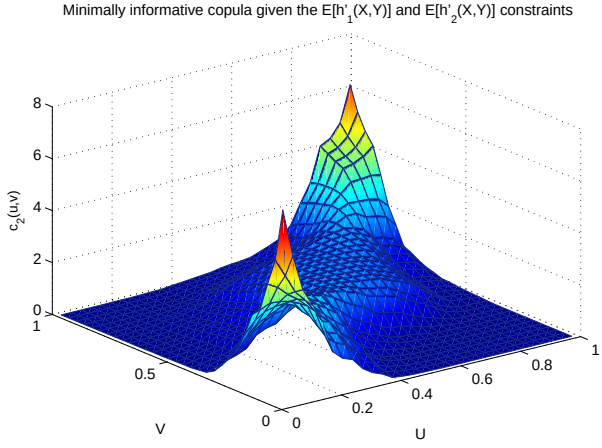


Figure 4.4: *The minimally informative copula for the World Bank data given* $E[h_1'(X,Y)]$ *and* $E[h_2'(X,Y)]$ *constraints.*

symmetric copulae are simply bound to fail to provide a decent fit. The likelihood for copulae was computed using the method. Namely the data points were first converted to uniform random variables using their respective empirical cumulative distribution functions. After that the copulae densities were interpolated with the bicubic method to obtain values of the densities at data points.

Computationally this method of constructing minimally informative copulae is not very demanding for currently available computers. The search for λ's carried out with the FMINSEARCH procedure was finished after 226 iterations (28 seconds) for the copula C_2 with the starting vector for λ's equal (0.00001, 0.00001). The termination tolerances on the function l_{sum} value and the λ's were set to be very restrictive at value of 10^{-12}. This ensured very accurate estimation of the Lagrange multipliers for the minimally informative copula given the two constraints.

4.5 Software program for interactive expert assignment of minimally-informative copulae

The minimally informative copulae given constraints can be the end result of expert elicitation procedure with the use of the theory described in sections 4.2–4.4. For this purpose a MATLAB script has been written. The script is a tool with graphical user interface and is able of presenting results in a form of plots. Figure 4.5 presents the interface of the program.

The entire process of elicitation can be described by the following algorithm available for use with the DAD algorithm for 3-dimensional case:

Algorithm 4.5.1 (Interactive expert elicitation).

1. Define the target function of the variables of interest (for example, $X+Y-Z$, $X-Y$, YZ etc.).
2. If experts are to assess a percentile, then specify which percentile.
3. Search for possible values of the quantity to be assessed (with taking into account the information supplied by the expert in previous iterations) and present the results to experts.
4. Ask experts for their assessments (within the specified bounds).
5. If you want to include more information on the variables of interest, then go to step 1.

Bounds obtained in step 3 will depend on the assessments given by experts in previous iterations of the algorithm. It may happen, that the set of achievable values of the quantity of interest will be empty. In such a situation the expert has to reassess the value she/he proposed in the previously. Notice that in i-th iteration the problem has to be solved for i Lagrange multipliers. This is the crucial step, because the computer program has to determine the range of achievable values for the assessed quantities. The next section describes our approach to

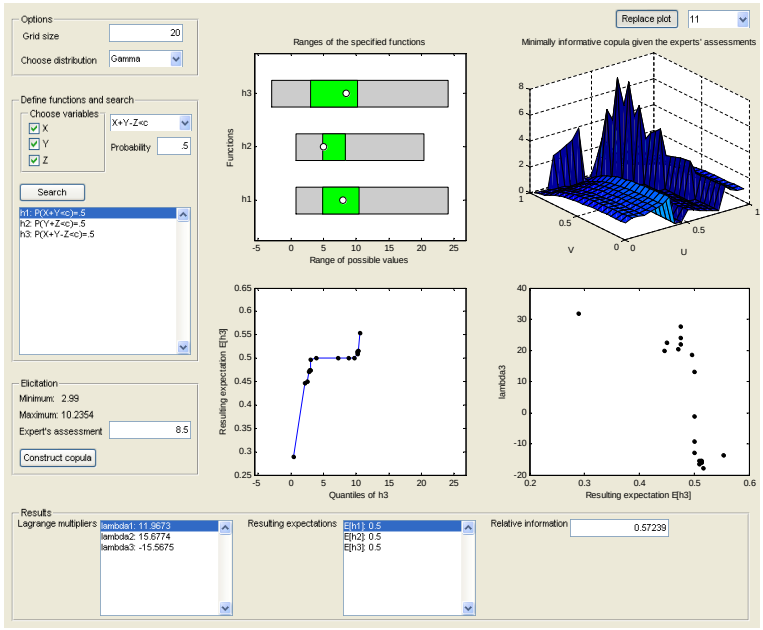


Figure 4.5: *Graphical user interface of* MATLAB *program for interactive expert elicitation.*

dealing with this problem, although there is plenty of other possibilities that can be successfully applied.

4.5.1 Algorithm searching for feasible values

A solver of non-linear equations is necessary for use in this elicitation method. Many results will depend on the quality of this solver and its ability to find the optimal solution, ie. for any given set of expectations find corresponding Lagrange multipliers λ's. Suppose an expectation elicited by expert is given and we know that there is a corresponding λ, but the solver does not converge. This particular value of expectation will be considered as *not* feasible in such situations. It may not be a problem if we deal with only one expectation to elicit. One can always derive the relationship between λ and the resulting expectation and numerically invert it. But the situation becomes complicated when dealing with a k-dimensional problem (assessment of k expectations). Now each of the k Lagrange multipliers depends on all of the specified expectations. Hence in order to determine the multipliers (solve the system of equations (4.4)), one has to apply a solver of a system of non-linear equations, such as, for instance, FSOLVE implemented in MATLAB. Therefore it useful to incorporate the solver in the process of searching for the bounds on the achievable values of expectations, because then only the value of the i-th observable is sampled and combined with all the previously assessed $i-1$ observables. Next, the system of non-linear equations

(4.4) of k variables is solved for the corresponding vector of λ's. If the solver converges, then we have found one of the achievable values of the i-th observable quantity. In the simplest case, one can just sample a number of values of this i-th observable and check, for which values the algorithm converges. The presented MATLAB script implements a more efficient procedure, based on the bisection method, which starts from some initial value, solves (4.4) for Lagrange multipliers, checks the resulting expectations, and then based on some specified criteria alters the initial value by Δ and repeats these steps. The magnitude of the step Δ decreases by half with each iteration, hence the maximum error of the estimation of the lower and upper bound is $2^{-n}(b-a)$, where $[a,b]$ is the domain of the i-th observable and n is the total number of iterations.

4.5.2 Example: Several observables

Consider expectations of various types of observable quantities (functions h_i's). The $D_1D_2D_3A$ approach allows to explore the set of simultaneously achievable values that may be taken by the expectations of these observables, under different minimally informative distributions.

We illustrate the above approach with an example. Suppose we want to model relationships between three random variables. Let the variables be gamma distributed independent random variables X, Y and Z with different parameters, ie. $X \sim \Gamma(1,2)$, $Y \sim \Gamma(2,3)$ and $Z \sim \Gamma(1,1)$. We ask experts to give us their judgement on expectations of three functions of X, Y and Z. These bits of information are used to build a copula for the joint of the gamma distributions. The copula is given in a discretized form over a grid of size $n = 20$ per variable.

We start with asking the expert to asses the median percentile of the distribution of $X+Y$. The bounds on achievable values of this percentile have been found to be $[4.85, 10.45]$, which is a considerably smaller interval than the whole domain of $X+Y$, namely $[0.7773, 24.0927]$. Suppose, that the expert assessed the value of the median to be rather conservative in the middle of the feasible values interval as $c_1 = 8$. The corresponding value of λ_1 is $\lambda_1 = -1.336$. The relative information of the resulting joint distribution with respect to the independent copula is $R = 0.0514$, which means that the expert's assessment indeed added some information to the joint distribution of X, Y and Z. It can be shown, that the relative information would not increase if the expert had assessed the median of $X+Y$ to be 7.0645. By the construction of the minimally informative copula, the bivariate margin $f_{UV|T}(u,v)$ of the constructed joint copula $f(u,v,t)$ does not depend on T, hence the conditional density $f_{UV|T}(u,v)$ stays the same for all values of T and is presented in Figure 4.6.

Next we ask the expert to assess the median of the distribution of $Y+Z$. The interval of achievable values for the median is given the bounds $[4.8397, 8.3805]$. This time the expert is less conservative and assesses the median of $X+Y$ to be $c_2 = 5$, closer to the lower bound of achievable values. We should expect a significant increase in the relative information coefficient. Given the expert's estimate, the problem is solved for a pair of λ's, which are $\lambda_1 = -1.366, \lambda_2 = 10.4849$. Notice, that λ_1 did not change by introducing the additional information on the

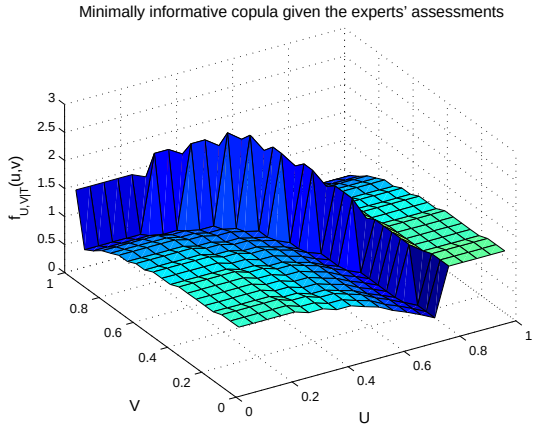


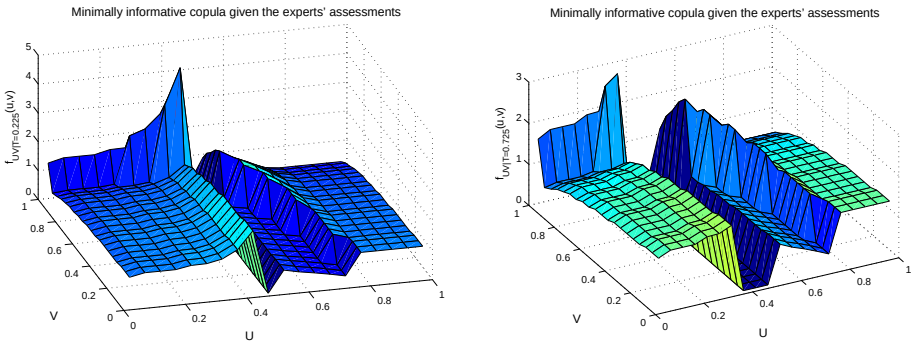
Figure 4.6: *Minimally informative copula given the expert's assessment on the median of* $X + Y$.

distribution of $Y + Z$. This is because both assessments concern functions which share only one variable, namely Y, and as such they can be assessed independently. The relative information increases to $R = 0.2464$. Since the copula density depends now on all of the variables, it is impossible to show the whole 3-dimensional distribution in one plot. We present only two conditional densities of $f(u, v, t)$ in Figure 4.7, for $t = 0.225$ (22.5-th percentile of Z) and $t = 0.725$ (72.5-th percentile of Z).

Finally, the expert has to assess the median of $X + Y - Z$. The domain of the distribution of $X + Y - Z$ is interval $[-2.9116, 24.0674]$. The bounds on achievable values of the median of this distribution are $[2.99, 10.2354]$. Suppose, that the expert assess the median to be $c_3 = 8.5$. Then the corresponding λ's are: $\lambda_1 = 11.9673$, $\lambda_2 = 15.6774$, $\lambda_3 = -15.5675$. Now λ_1 and λ_2 changed their values, because the assessment of the median of $X+Y-Z$ affects the previously assessed quantities of interest. The relative information with respect to the independent 3-dimensional copula increases again and now its value achieves $R = 0.57239$. Again we show only two conditional densities of $f(u, v, t)$ in Figure 4.8.

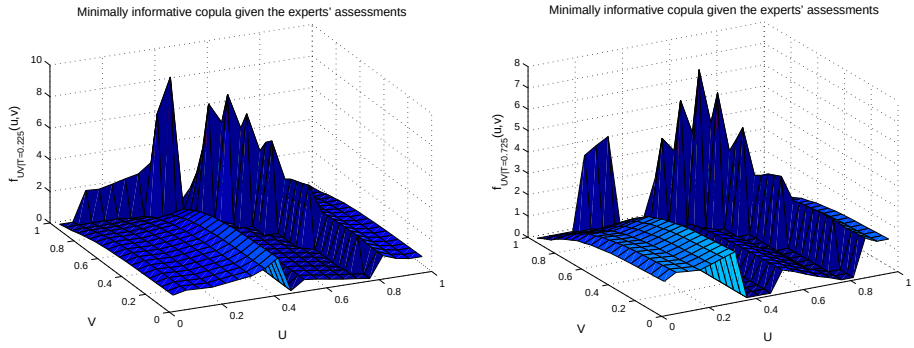
4.6 Implementation issues

Most problems with practical implementations of various algorithms in computer software are related to the limitations of representing floating-point numbers. The most common representation of floating-point numbers in computers is so-called double-precision format, which means that the minimum and maximum positive values that one can represent are $2.22507 \cdot 10^{-308}$ and $1.79769 \cdot 10^{308}$, respectively. If we take the logarithm of these numbers that we obtain the minimum



(a) The conditional density $f_{UV|T=0.225}(u,v)$. **(b)** The conditional density $f_{UV|T=0.725}(u,v)$.

Figure 4.7: *Conditional densities* $f_{UV|T}$.



(a) The conditional density $f_{UV|T=0.225}(u,v)$. **(b)** The conditional density $f_{UV|T=0.725}(u,v)$.

Figure 4.8: *Conditional densities* $f_{UV|T}$.

(≈ -708.3964) and maximum (≈ 709.7827) values for equation

$$\lambda_1 h_1(u,v) + \cdots + \lambda_k h_k(u,v), \tag{4.6}$$

which is part of the kernel function (4.3). It turns out, that while searching for achievable values of our quantity of interest (see section 4.5.1), we encounter this situation quite often, especially when dealing with multidimensional optimization problems. A simple solution is to constantly monitor the value of eq.(4.6) and change it to, say 700, if it is greater than 700, before the value of kernel function (4.3) is computed. Otherwise, this number will be treated as either 0 or infinity and the *DAD* algorithm will not converge and will not give any sensible results.

Furthermore, a very important requirement for the elicitation method introduced in this paper to work is ensuring proper performance of a solver of a system of nonlinear equations. As most of modern mathematical software packages (MATLAB, MAPLE, MS EXCEL, etc.) include such optimization routines, the choice

of the implementation platform becomes rather an issue of personal preferences. The solver FSOLVE implemented in MATLAB did not cause any problems in our implementation and was giving good results without any need to interfere with the optimization process. Much of attention should also be concentrated on the convergence of the DAD algorithm to ensure, that the density (4.3) has indeed uniform marginal distributions.

At last we would like to point out an issue that may not be a big problem in general, nevertheless can cause serious numerical errors. Namely empirical results show, that choosing values close to boundaries of the range of allowable values results in copula densities that have rather irregular shapes (high peaks, many areas with the density being equal zero). Similar situation takes place when several observables (3 or more) are being assessed by experts. Algorithms generating samples can be susceptible to numerical errors during sampling from such densities, and in result produce samples that don't reflect the information given by experts.

4.7 Conclusions

One of the most frequently employed method of experts elicitation is to ask them to assess median values of some quantities of interest. Then based on those assessments, the rank correlation between pairs of the variables is estimated. This rank correlation can be treated as a parameter of some predetermined copula (mostly centrally-symmetric) and samples are generated from this copula. For experts, who are not trained in statistics, the notion of correlation may be problematic to understand. A more natural approach to the problem of the elicitation is to ask experts questions, that occur in their professional work on a daily basis. We propose the approach which complies with this recommendation.

We have introduced the $D_1D_2D_3A$ algorithm to show how non-symmetric functions can be used for the subjective specification of copulae. A key difference with earlier work using the rank correlation is that the set of allowable values for observable expectations depends on the full set of observables under discussion. Hence an interactive system is needed for the expert in ensuring that such values are chosen coherently. A software programme has been written for this purpose and presented with simulation results. We show step by step how additional information can be nested and used for constructing minimally informative copula with respect to the uniform background measure. The copula method can be easily employed for generating dependent samples of the variables of interest. We show that the achieved results proved the method to be useful, tractable and intuitive. Future research may include implementations of other measures of dependence as, for example, Kendall's tau.

CHAPTER 5

Generating random correlation matrices with vines and Onion method

> Whenever you are asked if you can do a job, tell 'em, 'Certainly I can!' Then get busy and find out how to do it.
>
> *Theodore Roosevelt*

5.1 Introduction

In his recent work Joe [2006] introduced a new method for generating random correlation matrices uniformly from the space of positive definite correlation matrices. The method is based on an appropriate transformation of partial correlations to ordinary product moment correlations. The partial correlations can be assigned to edges of a regular vine — an extension of the concept of Markov dependence trees. Joe based his method on the so-called D-vine. We show that his methodology can be applied to any regular vine and argue that another type of regular vine, namely the C-vine, is more suitable for generating random correlation matrices. They require less computational time since the transformation of a set of partial correlations on a C-vine to a corresponding set of unconditional correlations operates only on partial correlations that are already specified on that vine. Please see [Bedford and Cooke, 2002] for more details on dependence vines.

An alternative method of sampling correlation matrices called *onion* method has been proposed by Ghosh and Henderson [Ghosh and Henderson, 2003]. This method can be explained in terms of elliptical distributions, and it does not involve

[4]This chapter is based on the manuscript *Generating random correlation matrices based on vines and extended Onion method* by Daniel Lewandowski, Dorota Kurowicka, and Harry Joe accepted for publication in *Journal of Multivariate Analysis*.

partial correlations. We extend it to allow generating random correlation matrices with the joint density of the correlations being proportional to a power of the determinant of the correlation matrix.

The chapter is organized as follows. Section 5.2 generalizes the method of generating correlation matrices proposed by Joe. In section 5.3 we extend the onion method. We carry out a computational time analysis of both methods in section 5.4. This is followed by the conclusions in section 5.5.

5.2 Generating random correlation matrices with partial correlations regular vines

The main idea of Joe's method [see Joe, 2006] to generate a correlation matrix of size $d \times d$ is to sample values of $\binom{d}{2}$ appropriately chosen partial correlations. The distribution of a given partial correlation is a Beta$\left(\frac{d-k}{2}, \frac{d-k}{2}\right)$ distribution on $(-1, 1)$, where the value k is the cardinality of the set of conditioning variables for the partial correlation. For a 4-dimensional correlation matrix Joe's choice of partial correlations become the following

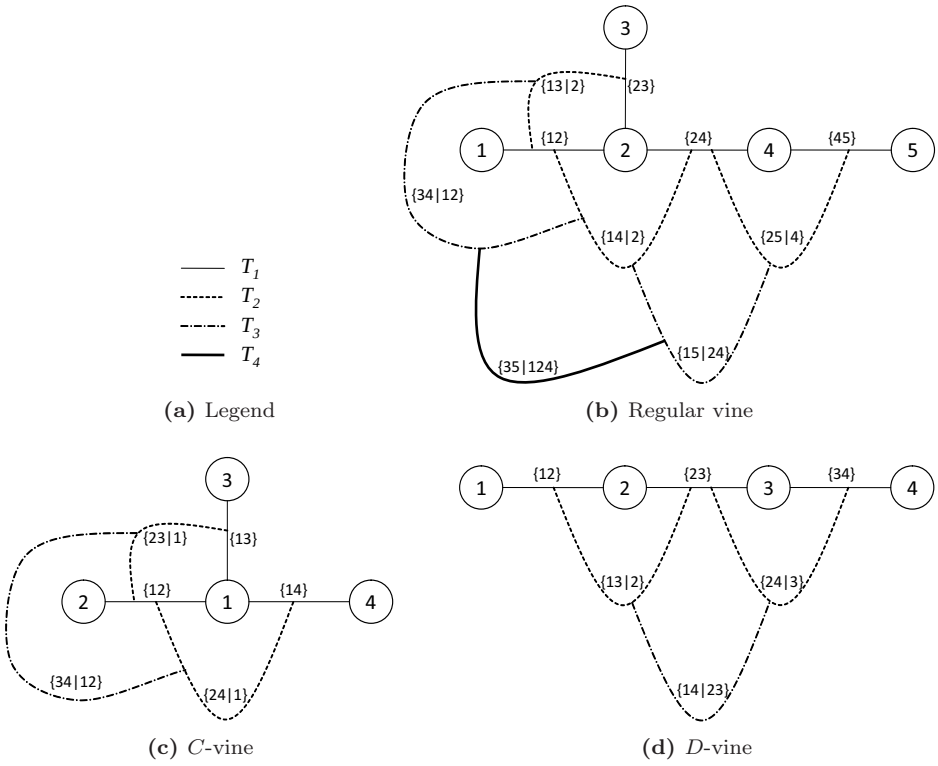
$$\rho_{12}, \rho_{23}, \rho_{34}, \rho_{13;2}, \rho_{24;3}, \rho_{14;23}. \tag{5.1}$$

However we extend the method to allow different choices for $\binom{d}{2}$ partial correlations. All choices of sets of partial correlations required for the method to work can be described using the notion of the partial correlation regular vine [Bedford and Cooke, 2002].

A vine $\mathcal{V}$ on d variables is a nested set of connected trees $\mathcal{V} = \{T_1, \ldots, T_{d-1}\}$ where the edges of tree T_i are the nodes of tree T_{i+1}, $i = 1, \ldots, d-2$. We denote the set of all edges in tree T_i by E_i. A *regular* vine is a vine in which two edges in tree T_i are joined by an edge in tree T_{i+1} only if these edges share a common node, $i = 1, \ldots, d-2$. Figure 5.1b shows an example of a regular vine on five variables. According to the regularity condition edges $\{1, 2\}$ and $\{4, 5\}$ of this vine cannot be joined by an edge in tree T_2, however this is possible for edges $\{2, 3\}$ and $\{2, 4\}$. For each edge e of a vine we define the *constraint* set U_e, the *conditioned* set $\{C_{1e}, C_{2e}\}$ and the *conditioning* set D_e of this edge as follows: the variables reachable from e are called the constraint set of this edge. When two edges are joined by an edge of the next tree, the intersection of the respective constraint sets form the conditioning set, and the symmetric difference of the constraint sets is the conditioned set. The regularity condition ensures the conditioned set to be a doubleton. In Figure 5.1 a symbol of the general form $\{L|K\}$ denotes a constraint set with conditioned set L and conditioning set K. The degree of node e is $\#D_e$.

Two distinct subtypes of regular vines are so-called C-vines (each tree T_i has a unique node of degree $d - i$; see Figure 5.1c) and D-vines (each node in T_1 has degree at most 2, see Figure 5.1d). This chapter aims on employing the C-vine in further analysis to generate random correlation matrices. Theorems presented here will be illustrated on an example of a regular vine $\mathcal{V}_5$ shown in Fig. 5.1b.

We define two concepts allowing expressing some properties of regular vines.



(a) Legend **(b)** Regular vine

(c) *C*-vine **(d)** *D*-vine

Figure 5.1: *Examples of various vines types.*

Definition 5.2.1 (m-child, m-descendent)**.** *If node e of a regular vine is an element of node f, we say that e is an* **m-*child*** *of f; similarly, if e is reachable from f via the membership relation: $e \in e_1 \in \ldots \in f$, we say that e is an* ***m-descendent*** *of f.*

A few of the properties of regular vines are [see Kurowicka and Cooke, 2006a]:

Property 1 There are $\binom{d}{2}$ edges in a regular vine on d variables.

Property 2 If $\mathcal{V}$ is a regular vine on d variables, then for all $i = 1, \ldots, d-1$ and all $e \in E_i$, the conditioned set associated with e is a doubleton and $\#D_e = i - 1$.

Property 3. If the conditioned sets of nodes e and f in a regular vine are equal, then $e = f$.

Property 4. For any node e in one of the trees $T_2, \ldots, T_{d-1}$ in a regular vine, if variable i is a member of the conditioned set of e, then i is a member of the conditioned set of exactly one of the m-children of e, and the conditioning set of an m-child of e is a subset of the conditioning set of e.

We add to this list one more property.

Lemma 5.2.1. *Let $e \in E_i$, $i > 1$, be the node with constraint set $\{1, \ldots, i+1\}$ and $\{s,t\} \subset D_e$. There exists $f \in E_j$, $j < i$, such that $\{C_{1f}, C_{2f}\} = \{s,t\}$.*

Proof. Node f is an m-descendent of e. The cardinality of the constraint set U_e of e is $i+1$, thus there are $\binom{i+1}{2}$ distinct doubletons in this set. Note also that there are $\binom{i+1}{2}$ edges in the subvine on nodes $\{1, \ldots, i+1\}$ by Property 1. By Property 4 the conditioned sets of all m-descendants of e are subsets of the constraint set of e and by Property 3 these conditioned sets are all different. Therefore one of the m-descendants of e must have the conditioned set $\{s,t\}$. ■

As an example, Property 4 means that for node $\{35;124\}$ of vine $\mathcal{V}_5$, variable 3 or 5 can occur only in the conditioned set of one of the m-children of this node, that is in either $\{34;12\}$ or $\{15;24\}$, never in both at the same time. According to Lemma 5.2.1 there should be three m-descendants of node $\{35;124\}$ with conditioned sets being doubleton subsets of its conditioning set $\{124\}$. These are nodes $\{12\}$, $\{24\}$ and $\{14;2\}$.

5.2.1 Partial and multiple correlations

One can notice that Joe's choice of partial correlations in eq.(5.1) corresponds to a partial correlation specification on the D-vine (compare with Fig. 5.1d). However the best choice for computing ordinary product moment correlations from partial correlations is a C-vine. For example, determining ρ_{34} from $\rho_{34;12}$ in the C-vine in Fig. 5.1c can be done recursively in two steps with eq.(2.1) solved for $\rho_{ij;L}$ as follows:

$$\begin{aligned} \text{step 1:} \quad \rho_{34;1} &= \rho_{34;12}\sqrt{(1-\rho^2_{23;1})(1-\rho^2_{24;1})} + \rho_{23;1}\rho_{24;1}, \\ \text{step 2:} \quad \rho_{34} &= \rho_{34;1}\sqrt{(1-\rho^2_{13})(1-\rho^2_{14})} + \rho_{13}\rho_{14}. \end{aligned}$$

Notice that only partial correlations specified in the vine appear in the formulae. This is not the case with the partial correlations specified on a D-vine.

We adopt the notation $D(\{L\})$ for the determinant of the correlation matrix with random variables indexed by the set L.

Definition 5.2.2 (Multiple correlation)**.** *The multiple correlation $R_{d\{d-1,\ldots,1\}}$ of variable X_d with respect to $X_{d-1}, \ldots, X_1$ is given by:*

$$1 - R^2_{d\{d-1,\ldots,1\}} = \frac{D(\{1,\ldots,d\})}{C_{dd}},$$

where $D(\{1,\ldots,d\})$ is the determinant of the correlation matrix $\mathbf{R}$ and C_{dd} is the (d,d) cofactor of $\mathbf{R}$. By permuting indices, other multiple correlations in d variables are defined.

The multiple correlation satisfies [see Kendall and Stuart, 1961]:

$$\begin{aligned} 1 - R^2_{d\{d-1,\ldots,1\}} &= (1 - R^2_{d\{d-2,\ldots,1\}})(1 - \rho^2_{d,d-1;d-2,\ldots,1}) \\ &= (1-\rho^2_{d,1})(1-\rho^2_{d,2;1})(1-\rho^2_{d,3;2,1})\ldots(1-\rho^2_{d,d-1;d-2,\ldots,1}). \end{aligned} \tag{5.2}$$

The determinant of a correlation matrix for d random variables can be expressed as a product of terms involving multiple correlations [Kendall and Stuart, 1961]:

$$\begin{aligned} D(\{1,\ldots,d\}) &= (1 - R^2_{d\{d-1,\ldots,1\}})(1 - R^2_{d-1\{d-2,\ldots,1\}})\ldots(1 - R^2_{2\{1\}}) \\ &= (1 - R^2_{d\{d-1,\ldots,1\}})D(\{1,\ldots,d-1\}). \end{aligned} \tag{5.3}$$

Lemma 5.2.2. *Let $i, j \notin L$.*

$$1 - \rho^2_{ij;L} = \frac{D(\{i,j,L\})D(\{L\})}{D(\{i,L\})D(\{j,L\})}.$$

Proof. From eq.(5.2) with permuted indices we have

$$1 - \rho^2_{ij;L} = \frac{1 - R^2_{i\{j,L\}}}{1 - R^2_{i\{L\}}}.$$

Use eq.(5.3) to simplify the terms on the right hand side to obtain the result. This simplifies the proof of Lemma 2 in [Joe, 2006]. ■

5.2.2 Jacobian of the transformation from unconditional correlations to the set of partial correlations

We investigate the Jacobian matrix for the transform T of a vector of ordinary product moment correlations $\mathbf{Q}$ (all cells of the upper triangle part of a correlation matrix $\mathbf{R}$ arranged in a row vector form) to a vector $\mathbf{P}$ of partial correlations on a regular vine. Both of these vectors have the same length by the construction of a regular vine. The elements of $\mathbf{P}$ are

$$P_i = \rho_{C_{1i},C_{2i};D_i}, \;\; i = 1,\ldots,\binom{d}{2}.$$

Let the partial correlations in $\mathbf{P}$ be ordered lexicographically as follows: first order partial correlations in the top tree T_1 lexicographically, then order partial correlations in the tree T_2 lexicographically, and so on. Reorder the product moment correlations in $\mathbf{Q}$ correspondingly simply by removing the conditioning sets from the partial correlations. Hence for the partial correlation specification on the regular vine in Figure 5.1b we have defined subsets $\mathbf{P}^{(i)}$ and $\mathbf{Q}^{(i)}$, $i = 1, 2, 3, 4$, of $\mathbf{P}$ and $\mathbf{Q}$ respectively as

$$\begin{aligned} \mathbf{P}^{(1)} &= \{\rho_{12}, \rho_{23}, \rho_{24}, \rho_{45}\}, & \mathbf{Q}^{(1)} &= \{\rho_{12}, \rho_{23}, \rho_{24}, \rho_{45}\}, \\ \mathbf{P}^{(2)} &= \{\rho_{13;2}, \rho_{14;2}, \rho_{25;4}\}, & \mathbf{Q}^{(2)} &= \{\rho_{13}, \rho_{14}, \rho_{25}\}, \\ \mathbf{P}^{(3)} &= \{\rho_{15;24}, \rho_{34;12}\}, & \mathbf{Q}^{(3)} &= \{\rho_{15}, \rho_{34}\}, \\ \mathbf{P}^{(4)} &= \{\rho_{35;124}\}, & \mathbf{Q}^{(4)} &= \{\rho_{35}\}. \end{aligned}$$

This order will be advantageous for deriving the Jacobian of the transformation T in a simple form. In the following pages we derive the appropriate conditions for this transformation to ensure the joint density of product moment correlations to be proportional to a power of $\det(\mathbf{R})$ with the uniform distribution as a special case.

We show the relationship between the form of the determinant of the correlation matrix and the determinant of the Jacobian [Kurowicka and Cooke, 2006b].

Theorem 5.2.3. *Let* $\mathbf{R}$ *be a* d*-dimensional correlation matrix and* $\mathbf{P}$ *the corresponding vector of partial correlations on a regular vine. One has then*

$$det(\mathbf{R}) = \prod_{i=1}^{\binom{d}{2}}(1 - P_i^2) = \prod_{i=1}^{\binom{d}{2}}(1 - \rho^2_{C_{1i},C_{2i};D_i}). \tag{5.4}$$

This is an important theorem as it allows us to express the determinant of a product moment correlation matrix as a product of 1 minus squared partial correlations on any regular vine. Joe [2006] provides the special case of this formula for D-vines. We show that the Jacobian of the transformation T also includes the same partial correlations as in eq.(5.4).

Lemma 5.2.4. *Let* $\rho_{ij;L}$ *be a partial correlation of order* $|L|$*. There is no other partial correlation* $\rho_{st;D_{st}}$ *of order* $|L|$ *in the regular vine, such that*

$$\frac{\partial \rho_{st;D_{st}}}{\partial \rho_{ij}} \neq 0.$$

Proof. The partial derivative $\partial \rho_{st;D_{st}}/\partial \rho_{ij} \neq 0$ if and only if set $\{i, j\}$ is in the constraint set $\{s, t, D_{st}\}$. By Property 3, $\{s, t\} \neq \{i, j\}$, thus either one of the elements, i or j, must be in $\{s, t\}$ and the other in D_{st}, or both $\{i, j\} \subset D_{st}$. In case of the first situation assume without loss of generality that $s = i$ and $j \in D_{st}$. That means that one of the m-children of $\rho_{st;D_{st}}$ has constraint set $\{i, j, D_{st} \backslash \{j\}\}$. This cannot happen because of Lemma 5.2.1. The second situation when $\{i, j\} \subset D_{st}$ also cannot happen because of Property 3 and Lemma 5.2.1. ■

Theorem 5.2.5. *The Jacobian matrix* $\mathbf{J}$ *of the transform from* $\mathbf{Q}$ *to* $\mathbf{P}$ *has the form*

$$\mathbf{J} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A} & \mathbf{B} \end{bmatrix},$$

where $\mathbf{I}$ *is the identity matrix of size* $(d-1)\times(d-1)$*,* $\mathbf{0}$ *is the matrix of* 0*'s of size* $(d-1)\times(d-1)(d-2)/2$*,* $\mathbf{A}$ *is a rectangular matrix of size* $(d-1)(d-2)/2\times(d-1)$ *and* $\mathbf{B}$ *is a square lower triangular matrix of size* $(d-1)(d-2)/2\times(d-1)(d-2)/2$*.*

Proof. Let J_{ij} denote the partial derivative of P_i with respect to Q_j. The elements P_i and Q_i are equal, $i = 1, \ldots, d-1$, and are not functions of any correlations other than themselves, and hence for $i = 1, \ldots, d-1$ and $j = 1, \ldots, d(d-1)/1$

$$J_{ij} = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{otherwise.} \end{cases}$$

This gives the identity matrix $\mathbf{I}$ and the matrix of zeros $\mathbf{0}$ as the upper parts of the Jacobian matrix. By Definition 2.2.1 an element of $\mathbf{P}^{(i)}$ is a function of product moment correlations in $\cup_{k\leq i}\mathbf{Q}^{(k)}$ only. Combining this result with Lemma 5.2.4 gives matrices $\mathbf{A}$ and $\mathbf{B}$, and $\mathbf{B}$ is lower triangular. ■

Corollary 5.2.6. *The determinant* $det(\mathbf{J})$ *of the Jacobian matrix* $\mathbf{J}$ *is*

$$det(\mathbf{J}) = \prod_{i=1}^{\binom{d}{2}} \frac{\partial P_i}{\partial Q_i}. \tag{5.5}$$

The proof follows from $\mathbf{B}$ being lower triangular. For $i = 1, \ldots, d-1$ the partial derivative $\partial P_i/\partial Q_i = 1$, hence the product in eq.(5.5) can start from $i = d$.

5.2.3 Partial derivatives

We derive the expression for the partial derivative of partial correlation $\rho_{ij;L}$ with respect to its corresponding unconditional correlation ρ_{ij}.

Lemma 5.2.7. *Let* L *be a nonempty set with indices distinct from* $\{i, j\}$. *Then*

$$\frac{\partial \rho_{ij;L}}{\partial \rho_{ij}} = \frac{1}{\sqrt{1 - R^2_{i\{L\}}}\sqrt{1 - R^2_{j\{L\}}}}. \tag{5.6}$$

Proof. The lemma will be proved by induction. If $L = \{l\}$ then from (2.1) we have

$$\begin{aligned} \frac{\partial \rho_{ij;l}}{\partial \rho_{ij}} &= \frac{\partial}{\partial \rho_{ij}} \left(\frac{\rho_{ij} - \rho_{il}\rho_{jl}}{\sqrt{(1-\rho^2_{il})(1-\rho^2_{jl})}} \right) \\ &= \frac{1}{\sqrt{(1-\rho^2_{il})(1-\rho^2_{jl})}} = \frac{1}{\sqrt{(1-R^2_{i\{l\}})(1-R^2_{j\{l\}})}} \end{aligned}$$

and the lemma holds. Assume that eq.(5.6) holds for the conditioning set L containing d nodes. Extend now the conditioning set to include $d+1$ nodes, ie. $\{k, L\}$. The corresponding partial derivative thanks to the Chain Rule and the recursive formula (2.1) can be expressed as

$$\frac{\partial \rho_{ij;kL}}{\partial \rho_{ij}} = \frac{\partial \rho_{ij;kL}}{\partial \rho_{ij;L}} \frac{\partial \rho_{ij;L}}{\partial \rho_{ij}} = \frac{1}{\sqrt{1-\rho^2_{ik;L}}\sqrt{1-\rho^2_{jk;L}}} \frac{1}{\sqrt{1-R^2_{i\{L\}}}\sqrt{1-R^2_{j\{L\}}}}.$$

This can be expanded further by using Lemma 5.2.2

$$\frac{\partial \rho_{ij;kL}}{\partial \rho_{ij}} = \sqrt{\frac{1-R^2_{i\{L\}}}{1-R^2_{i\{kL\}}}} \sqrt{\frac{1-R^2_{j\{L\}}}{1-R^2_{j\{kL\}}}} \frac{1}{\sqrt{1-R^2_{i\{L\}}}\sqrt{1-R^2_{j\{L\}}}}.$$

Simplifying this equation yields

$$\frac{\partial \rho_{ij;kL}}{\partial \rho_{ij}} = \frac{1}{\sqrt{1 - R^2_{i\{kL\}}}\sqrt{1 - R^2_{j\{kL\}}}}.$$

■

Joe [2006] published a similar result:

$$\frac{\partial \rho_{1d;2\ldots d-1}}{\partial \rho_{1d}} = \frac{D(\{2,\ldots,d-1\})}{\sqrt{D(\{1,\ldots,d-1\})D(\{2,\ldots,d\})}} = \frac{1}{\sqrt{1 - R^2_{1\{2,\ldots,d-1\}}}\sqrt{1 - R^2_{d\{2,\ldots,d-1\}}}}$$

for one specific ordering of nodes using the properties of partial correlations on a D-vine. We gave a more general proof with no reference to any specific type of vine. This lemma shows that the partial derivative $\partial \rho_{35;124}/\partial \rho_{35}$ in case of $\mathcal{V}_5$ can be expressed as

$$\begin{aligned}\frac{\partial \rho_{35;124}}{\partial \rho_{35}} &= \left((1 - R^2_{3\{124\}})(1 - R^2_{5\{124\}})\right)^{-\frac{1}{2}} \\ &= \left((1 - \rho^2_{34;12})(1 - \rho^2_{13;2})(1 - \rho^2_{23}) \cdot (1 - \rho^2_{15;24})(1 - \rho^2_{25;4})(1 - \rho^2_{45})\right)^{-\frac{1}{2}}\end{aligned}$$

Only partial correlations specified in $\mathcal{V}_5$ appear in this product.

Lemma 5.2.8. *Suppose variable d is in the conditioned set of the top node of a regular vine. Then there is a permutation $(j_1, \ldots, j_{d-1})$ of $(1, \ldots, d-1)$ such that the product of all partial derivatives involving variable d is equal to*

$$\left[D(\{d-1,\ldots,1\}) \prod_{i=2}^{d-1} \left(1 - R^2_{d\{j_{i-1},\ldots,j_1\}}\right)\right]^{-\frac{1}{2}}.$$

Proof. Let $\{d, j_{d-1}; j_{d-2}, \ldots, j_1\}$ be the constraint set of the single node e of the top most tree T_{d-1}. Collect all m-descendants of e containing variable d. By Property 4, d occurs only in the conditioned set of m-descendent nodes of e and the conditioning set of a m-child is a subset of the conditioning set of its m-parent. By Property 3, variable d occurs exactly once with every other variable $\{d-1,\ldots,1\}$ in the conditioned set of some node. Hence there is some permutation $(j_1, \ldots, j_{d-1})$ of $(1, \ldots, d-1)$, such that in tree T_i $(i = 1, \ldots, d-1)$ there is a partial correlation associated with one of the edges of the tree with the constraint set $\{d, j_i; j_{i-1}, \ldots, j_1\}$. By Lemma 5.2.7 the product of all partial derivatives of partial correlations involving node d can be expressed as

$$\begin{aligned}\prod_{i=2}^{d-1} \frac{\partial \rho_{dj_i;j_{i-1},\ldots,j_1}}{\partial \rho_{dj_i}} &= \prod_{i=2}^{d-1} \left[1 - R^2_{d\{j_{i-1},\ldots,j_1\}}\right]^{-\frac{1}{2}} \cdot \prod_{i=2}^{d-1} \left[1 - R^2_{j_i\{j_{i-1},\ldots,j_1\}}\right]^{-\frac{1}{2}} \\ &= \left[\prod_{i=2}^{d-1} \left(1 - R^2_{d\{j_{i-1},\ldots,j_1\}}\right) \cdot D(\{d-1,\ldots,1\})\right]^{-\frac{1}{2}},\end{aligned}$$

where the last equality follows from the definition of the multiple correlation coefficient via $1-R^2_{j_i\{j_{i-1},\ldots,j_1\}} = D(\{j_i, j_{i-1},\ldots,j_1\})/D(\{j_{i-1},\ldots,j_1\})$. If $i=1$, then $\partial\rho_{dj_i;j_{i-1},\ldots,j_1}/\partial\rho_{dj_i} = 1$ and therefore there is no need to include this term in the above product. ■

The determinant $D(\{d-1,\ldots,1\})$ does not depend in any particular way on the indexing of the nodes $\{d-1,\ldots,1\}$. Let $|\mathbf{J}_d|$ denote the determinant of the Jacobian of the transform of $\mathbf{Q}$ to $\mathbf{P}$ for a regular vine on d nodes.

Lemma 5.2.9. *Suppose variable d is in the conditioned set of the top node of a regular vine. Then there is a permutation $(j_1,\ldots,j_{d-1})$ of $(1,\ldots,d-1)$ such that the recursive formula for the determinant $|\mathbf{J}_d|$ of the Jacobian for the transform of $\mathbf{Q}$ to $\mathbf{P}$ is:*

$$|\mathbf{J}_d| = |\mathbf{J}_{d-1}|\left[D(\{d-1,\ldots,1\})\prod_{i=2}^{d-1}\left(1-R^2_{d\{j_{i-1},\ldots,j_1\}}\right)\right]^{-\frac{1}{2}}.$$

Proof. By Corollary 5.2.6

$$|\mathbf{J}_d| = \prod_{i=1}^{\binom{d}{2}}\frac{\partial P_i}{\partial Q_i} = \prod_{i\in\mathcal{A}}\frac{\partial P_i}{\partial Q_i}\cdot\prod_{i\in\mathcal{B}}\frac{\partial P_i}{\partial Q_i},$$

where $\mathcal{A}$ is the set of all partial correlations on a regular vine without node d in the constraint set, and $\mathcal{B}$ is the set of all partial correlations with d in the conditioned set. By Corollary 5.2.6, the first product is $|\mathbf{J}_{d-1}|$. By Lemma 5.2.8, the second product simplifies and the claimed result is obtained. ■

Next is the main theorem of this chapter.

Theorem 5.2.10. *The determinant $|\mathbf{J}_d|$ of the Jacobian for the transform of $\mathbf{Q}$ to $\mathbf{P}$ is*

$$|\mathbf{J}_d| = \left[\prod_{i=1}^{\binom{d}{2}-1}(1-\rho^2_{C_{1i},C_{2i};D_i})^{d-\#D_i-2}\right]^{-\frac{1}{2}}. \tag{5.7}$$

Proof. Without loss of generality, assume variable d is in the conditioned set of the top node. Let $(j_1,\ldots,j_{d-1})$ be the permutation of $(1,\ldots,d-1)$ from Lemma 5.2.8.

The proof goes by induction. For $d=3$, the P_i for $i=1,2,3$ are $\rho_{j_1j_2}$, ρ_{3j_1} and $\rho_{3j_2;j_1}$, respectively. We have by Lemma 5.2.9

$$|\mathbf{J}_3| = \frac{|\mathbf{J}_2|}{\sqrt{1-\rho^2_{j_1j_2}}\sqrt{1-\rho^2_{j_13}}} = \frac{1}{\sqrt{1-\rho^2_{j_1j_2}}\sqrt{1-\rho^2_{j_13}}}$$

and the theorem is satisfied. Assume that eq.(5.7) holds for $d-1$. Then again by Lemma 5.2.9 for d we have

$$|\mathbf{J}_d| = |\mathbf{J}_{d-1}|\left[D(\{d-1,\ldots,1\})\prod_{i=2}^{d-1}\left(1-R^2_{d\{j_{i-1},\ldots,j_1\}}\right)\right]^{-\frac{1}{2}}.$$

However with Theorem 5.2.3 and induction,

$$
\begin{aligned}
&|\mathbf{J}_{d-1}|\, D(\{d-1,\ldots,1\})^{-\frac{1}{2}} = \\
&= \left[\prod_{i=1}^{\binom{d-1}{2}-1} (1-\rho^2_{C_{1i},C_{2i};D_i})^{d-\#D_i-3} \cdot \prod_{i=1}^{\binom{d-1}{2}} (1-\rho^2_{C_{1i},C_{2i};D_i})\right]^{-\frac{1}{2}} \\
&= \left[\prod_{i=1}^{\binom{d-1}{2}} (1-\rho^2_{C_{1i},C_{2i};D_i})^{d-\#D_i-2}\right]^{-\frac{1}{2}} . \qquad (5.8)
\end{aligned}
$$

The above product contains all terms with partial correlation from the vine on nodes $\{d-1,\ldots,1\}$ raised to the appropriate power. There are $d-2$ terms missing in order to obtain the claimed result. These are the terms involving all partial correlations with d in the conditioned set. They are obtained from $\prod_{i=2}^{d-1}\left(1-R^2_{d\{j_{i-1},\ldots,j_1\}}\right)$. By eq.(5.2)

$$
\prod_{i=2}^{d-1}\left(1-R^2_{d\{j_{i-1},\ldots,j_1\}}\right) = \prod_{i=2}^{d-1}(1-\rho^2_{d,j_{i-1};j_{i-2},\ldots,j_1})^{d-(i-2)-2}. \qquad (5.9)
$$

Notice that $i-2$ in the exponent is the cardinality of the conditioning set. Hence by combining eq.(5.8) with (5.9) we prove the theorem. ■

The product in eq.(5.7) contains terms with all the partial correlations assigned to the edges of a regular vine taken to the appropriate power depending on the cardinality of the conditioning set. It does not explicitly include the term with the top most partial correlation with the highest cardinality of the conditioning set, i.e., for $i=\binom{d}{2}$, but its exponent according to the formula would be 0 anyway, hence index i can go safely from 1 to $\binom{d}{2}$ in eq.(5.7).

The above calculations can also be carried out in a simplified form for C-vines. Let $\mathcal{V}$ be a C-vine on d nodes with node 1 as the root of the vine. Then one can introduce a partial correlation specification on the nodes of this vine and present them in the form of a matrix:

$$
\mathbf{R} = \begin{bmatrix}
1 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,d-1} & \rho_{1,d} \\
 & 1 & \rho_{2,3;1} & \cdots & \rho_{2,d-1;1} & \rho_{2,d;1} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
 & & & & 1 & \rho_{d-1,d;1\ldots d-2} \\
 & & & & & 1
\end{bmatrix}.
$$

The partial derivative of $\rho_{d-1,d;1\ldots d-2}$ with respect to $\rho_{d-1,d}$ is

$$
\frac{\partial \rho_{d-1,d;1\ldots d-2}}{\partial \rho_{d-1,d}} = \prod_{i=1}^{d-2} \frac{\partial \rho_{d-1,d;1\ldots i}}{\partial \rho_{d-1,d;1\ldots i-1}} = \prod_{i=1}^{d-2} \left[(1-\rho^2_{d,i;1\ldots i-1})(1-\rho^2_{d-1,i;1\ldots i-1})\right]^{-\frac{1}{2}},
$$

where we assume the conditioning set $\{1, \ldots, i-1\}$ for $i = 1$ to be the empty set. For the lower order partial correlations one has

$$\frac{\partial \rho_{j,j+n;1\ldots j-1}}{\partial \rho_{j,j+n}} = \prod_{i=1}^{j-1} \left[(1-\rho^2_{j,i;1\ldots i-1})(1-\rho^2_{j+n,i;1\ldots i-1})\right]^{-\frac{1}{2}}$$

for $1 \leq j \leq d-1$ and $2 \leq j+n \leq d$.

The determinant $|\mathbf{J}_d|$ of the Jacobian for the transform of $\mathbf{Q}$ to $\mathbf{P}$ for the partial correlations on a C-vine is

$$|J_d| = \left[\prod_{k=1}^{d-2} \prod_{i=k+1}^{d} \left(1-\rho^2_{k,i;1,\ldots,k-1}\right)^{d-k-1}\right]^{-\frac{1}{2}}. \tag{5.10}$$

All partial correlations from the correlation matrix $\mathbf{R}$ except $\rho_{d-1,d;1,\ldots,d-2}$ appear in the expression (5.10). However this term can also be added safely because its exponent would be 0 ($d-k-1$, where $k = d-1$). Therefore k in the first product in (5.10) can increase up to $d-1$ instead of $d-2$. We make this adjustment in the subsequent calculations.

5.2.4 Algorithm for generating correlation matrices with vines

We show how to use the theorems to generate random correlation matrices such that the density of the random correlation matrix is invariant under the choice of partial correlation vine. Following the calculations of Joe [2006] we employ the linearly transformed Beta(α, α) distribution on the interval $(-1, 1)$ to simulate partial correlations. The density g of this random variable is

$$g(x;\alpha) = \frac{2^{-2\alpha+1}}{B(\alpha,\alpha)}(1-x^2)^{\alpha-1} = \frac{2^{-2\alpha+1}\Gamma(2\alpha)}{\Gamma^2(\alpha)}(1-x^2)^{\alpha-1}, \tag{5.11}$$

where B is the beta function.

Suppose $\rho_{C_{1i},C_{2i};D_i}$ has a Beta(β_i, β_i) density on $(-1, 1)$ and its realization is denoted by $p_{C_{1i},C_{2i};D_i}$. Similarly, let the realization of an ordinary product moment correlation $\rho_{C_{1i},C_{2i}}$ be denoted by $q_{C_{1i},C_{2i}}$. Then the joint density f of ordinary product moment correlations in $\mathbf{R}$ is proportional to

$$\begin{aligned} &f(q_{C_{1i},C_{2i}}; 1 \leq i \leq d(d-1)/2) \propto \\ &\prod_{j=1}^{\binom{d}{2}} g(p_{C_{1j},C_{2j};D_j};\beta_j) \cdot |\mathbf{J}_d| = \prod_{j=1}^{\binom{d}{2}} (1-p_{C_{1j},C_{2j};D_j})^{\beta_j - \frac{d-\#D_j}{2}}. \end{aligned} \tag{5.12}$$

The exponent $\beta_j - \frac{d-\#D_j}{2}$ is a function of $\#D_j = n$ for a given d. In order to make this exponent equal to a constant $\eta - 1$, β_j will be replaced by α_n so that $\alpha_n - (d-n)/2 = \eta - 1$; thus $\alpha_n = \eta + \frac{d-n-2}{2}$. We replace β_j with α_n in eq.(5.12) and use Theorem 5.2.3 to obtain

$$f(q_{C_{1i},C_{2i}}; 1 \leq i \leq d(d-1)/2) = c_d \prod_{j=1}^{\binom{d}{2}} (1-p_{C_{1j},C_{2j};D_j})^{\eta-1} = c_d\, det(\mathbf{R})^{\eta-1}, \tag{5.13}$$

where c_d is the normalizing constant depending on the dimension d. The uniform density is obtained for $\eta = 1$, which means that the marginal densities for partial correlations $p_{C_{1i},C_{2i};D_i}$ are Beta$\left(\frac{d-\#D_i}{2}, \frac{d-\#D_i}{2}\right)$ on $(-1,1)$, for $i = 1, \ldots, d(d-1)/2$,

For the C-vine the above reasoning has the following implications. By eq.(5.10) the joint density f of the ordinary product moment correlations is

$$f(q_{ij}, 1 \le i < j \le d) = c_d \prod_{k=1}^{d-1} \prod_{l=k+1}^{d} (1 - p_{kl;1,\ldots,k-1}^2)^{\alpha_{k-1} - 1 - \frac{d-k-1}{2}}. \qquad (5.14)$$

The exponent $\alpha_{k-1} - 1 - \frac{d-k-1}{2}$ is of the form $\beta_j - \frac{d-\#D_j}{2}$ as in eq.(5.12) with $\#D_j = k-1$. Thus the density (5.14) is uniform if $\alpha_{k-1} = \frac{d-k+1}{2}$ and the marginal densities for partial correlations $\rho_{kl;1,\ldots,k-1}$ $(1 \le k \le d-1$ and $k+1 \le l \le d)$ in the matrix $\mathbf{R}$ are Beta$\left(\frac{d-k+1}{2}, \frac{d-k+1}{2}\right)$ on $(-1,1)$.

The normalizing constant c_d for eq.(5.13) and (5.14) has the same formula as the one derived in [Joe, 2006] since it does not depend on the specific vine used in the calculations

$$\begin{aligned} c_d &= 2^{\sum_{k=1}^{d-1}(2\eta-2+d-k)(d-k)} \times \\ &\quad \times \prod_{k=1}^{d-1} \left[B(\eta + \tfrac{1}{2}(d-k-1), \eta + \tfrac{1}{2}(d-k-1))\right]^{d-k}. \end{aligned} \qquad (5.15)$$

If $\eta = 1$ this equation simplifies to

$$2^{\sum_{k=1}^{d-1} k^2} \cdot \prod_{k=1}^{d-1} \left[B\left(\frac{k+1}{2}, \frac{k+1}{2}\right)\right]^k.$$

We denote the realization of random matrix $\mathbf{R}$ by $\mathbf{r}$. Elements of $\mathbf{r}$ are r_{ij}, $1 \le i, j \le d$. The algorithm for generating correlation matrices with density proportional to $[\det(\mathbf{r})]^{\eta-1}$, $\eta > 1$ is quite simple using the vine method based on a C-vine.

Algorithm 5.2.1 (Generating random correlation matrices with C-vines).

1. Initialization $\beta = \eta + (d-1)/2$.
2. Loop for $k = 1, \ldots, d-1$.
   a) $\beta \leftarrow \beta - \frac{1}{2}$;
   b) Loop for $i = k+1, \ldots, d$;
      i) generate $p_{k,i;1,\ldots,k-1} \sim \text{Beta}(\beta, \beta)$ on $(-1,1)$;
      ii) use recursive formula (2.1) on $p_{k,i;1,\ldots,k-1}$ to get $q_{k,i} = r_{k,i} = r_{i,k}$.
3. Return $\mathbf{r}$, a $d \times d$ correlation matrix.

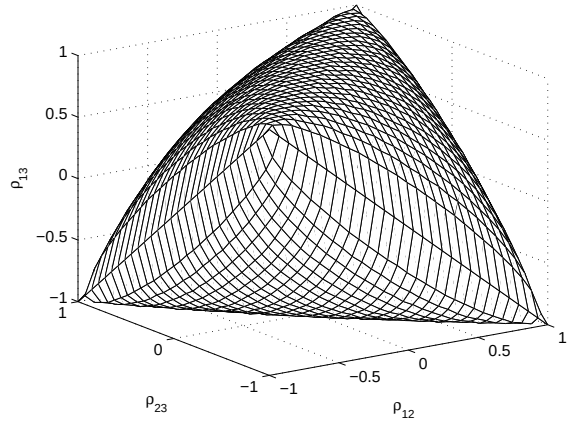


Figure 5.2: *Boundary of the set of all triples* $(\rho_{12}, \rho_{13}, \rho_{23})$ *leading to semi-positive definite 3-dimensional correlation matrices.*

Because the partial correlations in a regular vine can independently take values in the interval $(-1, 1)$, one could more generally assign an arbitrary density g_i, supported on $(-1, 1)$, to $\rho_{C_{1i},C_{2i};D_i}$, and get a joint density for the correlation matrix by multiplying $\prod_{i=1}^{\binom{d}{2}} g_i(p_{C_{1i},C_{2i};D_i})$ by the Jacobian. This density in general is not invariant under the choice of partial correlation vine, but by choosing the vine and the g_i appropriately, one could get random correlation matrices that have larger correlations at a few particular pairs.

5.3 Onion method

Another interesting method of sampling uniformly from the set of correlation matrices was the method proposed in [Ghosh and Henderson, 2003]. We give a simpler explanation of their method, together with an extension to random correlation matrices with density proportional to $[\det(\mathbf{r})]^{\eta-1}$ for $\eta > 0$. With the derivation, we check that the normalization constant is the same as that given in [Joe, 2006].

5.3.1 Background results

We start with some background results on the elliptically contoured distributions [see Joe, 1997]. Consider the spherical density $c(1 - \mathbf{w}^T\mathbf{w})^{\beta-1}$ for $\mathbf{w} \in \mathbb{R}^k$, $\mathbf{w}^T\mathbf{w} \leq 1$, where c is the normalizing constant. If $\mathbf{W}$ has this density, then it has the stochastic representation $\mathbf{W} = V\mathbf{U}$ where $V^2 \sim \text{Beta}(k/2, \beta)$ and $\mathbf{U}$ is uniform on the surface of the k-dimensional hypersphere. If $\mathbf{Z} = \mathbf{AW}$, where $\mathbf{A}$ is a $k \times k$ non-singular matrix, then the density of $\mathbf{Z}$ is

$$c[\det(\mathbf{AA}^T)]^{-1/2}(1 - \mathbf{z}^T[\mathbf{AA}^T]^{-1}\mathbf{z})^{\beta-1}$$

over $\mathbf{z}$ such that $\mathbf{z}^T[\mathbf{A}^T\mathbf{A}]^{-1}\mathbf{z} \leq 1$.

Lemma 5.3.1. *The normalization constant c of the spherically contoured density $c(1-\mathbf{w}^T\mathbf{w})^{\beta-1}$ is*

$$c = \Gamma(\beta + k/2)\pi^{-k/2}/\Gamma(\beta).$$

Proof. From known results on elliptical densities, the density of the radial direction V is

$$cS_k(1-v^2)^{\beta-1}v^{k-1}, \quad 0 < v < 1,$$

where $S_k = 2\pi^{k/2}/\Gamma(k/2)$. The density of $Y = V^2$ is

$$cS_k(1-y)^{\beta-1}y^{(k-1)/2} \cdot \tfrac{1}{2}y^{-1/2} = \tfrac{1}{2}cS_k y^{k/2-1}(1-y)^{\beta-1}, \quad 0 < y < 1,$$

This is a Beta$(k/2, \beta)$ density, so that

$$\tfrac{1}{2}cS_k = \frac{\Gamma(\beta + k/2)}{\Gamma(k/2)\,\Gamma(\beta)} \qquad \text{or} \qquad c = \frac{\Gamma(\beta + k/2)}{\pi^{k/2}\Gamma(\beta)}.$$

■

The onion method is based on the fact that any correlation matrix of size $(k+1) \times (k+1)$ can be partitioned as

$$\mathbf{r}_{k+1} = \begin{bmatrix} \mathbf{r}_k & \mathbf{z} \\ \mathbf{z}^T & 1 \end{bmatrix},$$

where $\mathbf{r}_k$ is an $k \times k$ correlation matrix and $\mathbf{z}$ is a k-vector of correlations. From standard results we have $\det(\mathbf{r}_{k+1}) = \det(\mathbf{r}_k) \cdot \left(1 - \mathbf{z}^T\mathbf{r}_k^{-1}\mathbf{z}\right)$. Let the upper case letter of $\mathbf{r}_k$, $\mathbf{z}$, $\mathbf{r}_{k+1}$ denote random vectors and matrices and let $\beta, \beta_k > 0$ be two known parameters. If $\mathbf{R}_k$ has density proportional to $[\det(\mathbf{r}_k)]^{\beta_k - 1}$, and $\mathbf{Z}$ given $\mathbf{R}_k = \mathbf{r}_k$ has density proportional to $[\det(\mathbf{r}_k)]^{-1/2}(1 - \mathbf{z}^T\mathbf{r}_k^{-1}\mathbf{z})^{\beta-1}$ (hence it is elliptically contoured), then the density of $\mathbf{R}_{k+1}$ is proportional to $[\det(\mathbf{r}_k)]^{\beta_k - 3/2}(1 - \mathbf{z}^T\mathbf{r}_k^{-1}\mathbf{z})^{\beta-1}$. If one sets $\beta_k = \beta + \frac{1}{2}$, then the density of $\mathbf{R}_{k+1}$ is proportional to $[\det(\mathbf{r}_{k+1})]^{\beta-1}$.

Because the density in eq.(5.11) is proportional to $(1-u^2)^{\alpha-1}$, which is a power of $\det \begin{pmatrix} 1 & u \\ u & 1 \end{pmatrix} = 1 - u^2$, it can be used to generate $\mathbf{r}_2$.

5.3.2 Algorithm for generating random correlation matrices

Combining the above results yields the following algorithm for the extended onion method to get random correlation matrices in dimension d with density proportional to $[\det(\mathbf{r})]^{\eta-1}$, $\eta > 1$

Algorithm 5.3.1 (Generating random correlation matrices with the Onion method).

1. Initialization. $\beta = \eta + (d-2)/2$, $r_{12} \leftarrow 2u - 1$, where $u \sim$ Beta(β, β), $\mathbf{r} \leftarrow \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$.

2. Loop for $k = 2, \ldots, d-1$.

   a) $\beta \leftarrow \beta - \frac{1}{2}$;

   b) generate $y \sim \text{Beta}(k/2, \beta)$

   c) generate $\mathbf{u} = (u_1, \ldots, u_k)^T$ uniform on the surface of k-dimensional hypersphere;

   d) $\mathbf{w} \leftarrow y^{1/2}\mathbf{u}$, obtain $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^T = \mathbf{r}$, set $\mathbf{z} \leftarrow \mathbf{A}\mathbf{w}$;

   e) $\mathbf{r} \leftarrow \begin{bmatrix} \mathbf{r} & \mathbf{z} \\ \mathbf{z}^T & 1 \end{bmatrix}$.

3. Return $\mathbf{r}$, a $d \times d$ correlation matrix.

In step c), it should be numerically faster to use $\mathbf{A}$ from the Cholesky decomposition of $\mathbf{r}$ rather than $\mathbf{r}^{1/2}$ based on the singular value decomposition. The latter is indicated in [Ghosh and Henderson, 2003].

5.3.3 Derivation of the normalizing constant

As in case of the vine method, every off-diagonal element of the random correlation matrix $\mathbf{R}$ has a marginal density $\text{Beta}(\eta + [d-2]/2,\ \eta + [d-2]/2)$ on $(-1, 1)$. For the special case of $\eta = 1$ leading to uniform over the space of correlation matrices, the marginal density of every correlation is $\text{Beta}(d/2, d/2)$ on $(-1, 1)$.

In the kth step of the algorithm, $\beta = \eta + [d-1-k]/2$. Using Lemma 5.3.1 and eq.(5.11), the reciprocal normalizing constant is

$$\begin{aligned} c_d{}' &= 2^{2\eta+d-3}\frac{\Gamma^2(\eta+\frac{d}{2}-1)}{\Gamma(2\eta+d-2)}\prod_{k=2}^{d-1}\frac{\pi^{\frac{k}{2}}\Gamma(\eta+\frac{d-1-k}{2})}{\Gamma(\eta+\frac{d-1-k}{2}+\frac{k}{2})} \\ &= 2^{2\eta+d-3}\frac{\Gamma^2(\eta+\frac{d}{2}-1)}{\Gamma(2\eta+d-2)}\prod_{k=2}^{d-1}\frac{\pi^{\frac{k}{2}}\Gamma(\eta+\frac{d-1-k}{2})}{\Gamma(\eta+\frac{d-1}{2})}. \end{aligned} \tag{5.16}$$

We show that the expressions for the normalizing constants (5.15) and (5.16) are equivalent. The proof makes use of the duplication formula relation [see Abramowitz and Stegun, 1964, Duplication formula, pp. 256]

$$\frac{\Gamma(2t)}{\Gamma(t)} = 2^{(2t-1)}\frac{\Gamma(t+\frac{1}{2})}{\Gamma(\frac{1}{2})} \Longrightarrow \frac{\Gamma^2(t)}{\Gamma(2t)}2^{2t-1} = \frac{\pi^{\frac{1}{2}}\Gamma(t)}{\Gamma(t+\frac{1}{2})} \tag{5.17}$$

Proof. We start with eq.(5.15). By the duplication formula (5.17) with $t = \eta +$

$(d-1-k)/2$ we have

$$\begin{aligned} c_d &= \prod_{k=1}^{d-1}\left[2^{2(\eta+(d-1-k)/2)-1}\frac{\Gamma^2(\eta+\frac{d-1-k}{2})}{\Gamma(2\eta+d-1-k)}\right]^{d-k} \\ &= \prod_{k=1}^{d-1}\left[\frac{\pi^{\frac{1}{2}}\Gamma(\eta+\frac{d-1-k}{2})}{\Gamma(\eta+\frac{d-1-k}{2}+\frac{1}{2})}\right]^{d-k} \\ &= \frac{\prod_{k-1}^{d-1}\pi^{\frac{k}{2}}}{\Gamma^{d-1}\left(\eta+\frac{d-1}{2}\right)}\prod_{k=1}^{d-1}\Gamma^{d-k}\left(\eta+\frac{d-1-k}{2}\right)\cdot \\ &\quad \cdot\prod_{k=2}^{d-1}\Gamma^{-(d-k)}\left(\eta+\frac{d-1-k}{2}+\frac{1}{2}\right). \end{aligned}$$

Start the indexing in the second product from 1 instead of 2 and increase k by 1. The upper limit can stay $d-1$ because $-(d-k)+1=0$ for $k=d-1$.

$$\begin{aligned} c_d &= \frac{\prod_{k-1}^{d-1}\pi^{\frac{k}{2}}}{\Gamma^{d-1}\left(\eta+\frac{d-1}{2}\right)}\prod_{k=1}^{d-1}\Gamma^{d-k}\left(\eta+\frac{d-1-k}{2}\right)\cdot \\ &\quad \cdot\prod_{k=1}^{d-1}\Gamma^{-(d-k)+1}\left(\eta+\frac{d-1-k}{2}\right) \\ &= \frac{\prod_{k-1}^{d-1}\pi^{\frac{k}{2}}}{\Gamma^{d-1}\left(\eta+\frac{d-1}{2}\right)}\prod_{k=1}^{d-1}\Gamma\left(\eta+\frac{d-1-k}{2}\right) \\ &= \prod_{k=1}^{d-1}\frac{\pi^{\frac{k}{2}}\Gamma(\eta+\frac{d-1-k}{2})}{\Gamma(\eta+\frac{d-1}{2})}. \end{aligned}$$

This is the expression for $c_d{}'$ with

$$2^{2\eta+d-3}\frac{\Gamma^2(\eta+\frac{d}{2}-1)}{\Gamma(2\eta+d-2)}=\frac{\pi^{\frac{1}{2}}\Gamma(\eta+\frac{d}{2}-1)}{\Gamma(\eta+\frac{d}{2}-\frac{1}{2})}=\frac{\pi^{\frac{k}{2}}\Gamma(\eta+\frac{d-1-k}{2})}{\Gamma(\eta+\frac{d-1}{2})}.$$

where $k=1$. ■

The expression for the normalizing constant can be further simplified for $\eta=1$.

Theorem 5.3.2. *If $\eta=1$ then the normalizing constant c_d can be expressed as*

$$c_d=\begin{cases}\pi^{(d^2-1)/4}\ \frac{\prod_{k=1}^{(d-1)/2}\Gamma(2k)}{2^{(d-1)^2/4}\,\Gamma^{d-1}\left(\frac{d+1}{2}\right)}, & \text{if } d \text{ is odd;}\\[2ex] \pi^{d(d-2)/4}\ \frac{2^{(3d^2-4d)/4}\,\Gamma^d\left(\frac{d}{2}\right)\prod_{k=1}^{(d-2)/2}\Gamma(2k)}{\Gamma^{d-1}(d)}, & \text{if } d \text{ is even.}\end{cases}$$

Proof. We rearrange terms in eq.(5.16) with $\eta=1$:

$$c_d'=\frac{\pi^{d(d-1)/4}}{\Gamma^{d-1}\left(\frac{d+1}{2}\right)}\prod_{k=1}^{d-1}\Gamma\left(\frac{d-k+1}{2}\right)=\frac{\pi^{d(d-1)/4}}{\Gamma^{d-1}\left(\frac{d+1}{2}\right)}\prod_{k=1}^{d-1}\Gamma\left(\frac{k}{2}+\frac{1}{2}\right). \qquad (5.18)$$

If d is odd then by using the duplication formula (5.17) we obtain

$$\prod_{k=1}^{d-1} \Gamma\left(\frac{k}{2}+\frac{1}{2}\right) = \prod_{k=1}^{(d-1)/2} \Gamma(k)\,\Gamma\left(k+\tfrac{1}{2}\right)$$
$$= \prod_{k=1}^{(d-1)/2} \Gamma(k)\,\frac{\Gamma(2k)\,\pi^{\frac{1}{2}}}{\Gamma(k)\,2^{2k-1}} = \frac{\pi^{(d-1)/4}}{2^{\sum_{k=1}^{(d-1)/2} 2k-1}} \prod_{k=1}^{(d-1)/2} \Gamma(2k). \qquad (5.19)$$

Substituting eq.(5.19) in to eq.(5.18) yields the claimed result. If d is even then

$$\prod_{k=1}^{d-1} \Gamma\left(\frac{k}{2}+\frac{1}{2}\right) = \Gamma\left(\frac{d}{2}\right) \prod_{k=1}^{(d-2)/2} \Gamma(k)\,\Gamma\left(k+\tfrac{1}{2}\right) = \frac{\Gamma\left(\frac{d}{2}\right)\pi^{(d-2)/4}}{2^{\sum_{k=1}^{(d-2)/2} 2k-1}} \prod_{k=1}^{(d-2)/2} \Gamma(2k). \qquad (5.20)$$

Substitute eq.(5.20) in to eq.(5.18) gives

$$c'_d = \frac{\pi^{(d^2-2)/4}\,\Gamma\left(\frac{d}{2}\right)}{2^{(d-2)^2/4}\,\Gamma^{d-1}\left(\frac{d+1}{2}\right)} \prod_{k=1}^{(d-2)/2} \Gamma(2k).$$

Apply the duplication formula to $\Gamma^{d-1}(\frac{d+1}{2})$ and cancel common terms to obtain the final result. ■

All arguments of the gamma functions in the formulae presented in Theorem 5.3.2 are integers and hence can be replaced with factorials. Note that the exponent of π in Theorem 5.3.2 for an odd number d is the same as that for the next largest even number; for $d = 3, 4, \ldots$, the exponents are respectively $2, 2, 6, 6, 12, 12, 20, 20, \ldots$.

5.4 Computational time analysis

Both the vine method and the onion method have been implemented in computer software and compared in terms of time required to generate a given number of random correlation matrices. Two different software platforms were used for this task, namely the scripting language of MATLAB and a low level programming language C. We used the built-in functions of MATLAB to generate Beta and Gaussian distributed random variables and to compute the Cholesky decomposition of correlation matrices required by the onion method. These functions of MATLAB are compiled and cannot be edited. The onion method implemented in MATLAB computes the full Cholesky decomposition at each iteration of the generating procedure. However the amount of calculations can be limited by implementing a Cholesky decomposition computed incrementally — that is a new row is added at each stage when a new $\mathbf{z}$ is generated. We took this approach when implementing the onion method in C; without the incremental Cholesky decomposition, the onion method was much slower than the vine method in the C programming language. It does not save any computational time in MATLAB compared to the built-in Cholesky decomposition function because the advantage of having fewer

Table 5.1: *Time in seconds required to generate 5000 correlation matrices of given dimension.*

Dimension	compiled C code with full optimization enabled (/0x)			m-script in Matlab 2007b		
	onion	C-vine	D-vine	onion	C-vine	D-vine
5	0.015	0.016	0.031	1.422	0.775	1.281
10	0.047	0.078	0.172	3.356	1.806	6.067
15	0.109	0.234	0.547	5.346	3.075	15.523
20	0.187	0.406	1.438	7.397	4.679	30.835
25	0.281	0.687	3.250	9.591	6.798	53.444
30	0.437	1.078	6.625	11.856	9.348	84.958
35	0.609	1.562	12.344	14.411	12.564	127.388
40	0.813	2.203	21.438	17.035	16.718	182.578
45	1.063	4.125	35.312	19.868	21.493	252.862
50	1.344	3.891	55.515	22.839	27.222	340.846
60	2.094	6.266	124.203	29.530	41.767	577.313
70	3.078	9.375	246.656	47.140	84.795	918.209
80	4.328	13.406	451.422	82.422	46.374	1 404.285

operations is wasted on executing a non-compiled code. The programs have been run on a desktop PC with Intel Core 2 Duo (2×3.2 GHz) processor, 3GB of RAM memory and Windows XP SP3 operating system. The source code of the software used for the analysis is available from the authors upon request.

Table 5.1 lists times necessary to complete the task of generating 5 000 random correlation matrices of given dimension. The compiled code is faster as expected and the incremental Cholesky decomposition routine allows the onion method to be the clear winner in this case. The difference between the onion method and the vine method in terms of the required calculation time gets bigger as the dimension increases. We can see a different picture on the Matlab 2007b platform. The vine method is faster than the onion method for lower dimensions of correlation matrices ($d < 44$), but our tests showed that this also depends on the processor used for calculations. We have included the results for the vine method based on the D-vine for reference. The C-vine based method of generating correlation matrices performs better in terms of the execution time by a large margin.

5.5 Conclusions

The main goal of this paper was to study and improve existing methods of generating random correlation matrices. Two of such methods include the onion method of Ghosh and Henderson [2003] and the vine method recently proposed by Joe [2006]. Originally the vine method was based on the so-called D-vine. We extend this methodology to any regular vine by studying the relationship between the multiple correlation and partial correlations on a regular vine. The C-vine exhi-

bits computational advantage for generating random correlation matrices, since the recursive formula (2.1) operates only on partial correlations that are already specified on a vine. It is the only vine with this property. This simplifies the generating algorithm and limits the number of calculations.

We also give a simpler explanation of the onion method in terms of elliptical distributions. The generalization of this method yields a procedure to sample from the set of positive definite correlation matrices with joint densities of correlations proportional to $\det(\mathbf{r})^{\eta-1}$ with $\eta > 0$. This allows the choice of the method suited to the need. The efficiency of the algorithms for generating random correlation matrices depends heavily on the programming language used for implementation. Preferably both methods would be implemented and benchmarked before the final decision is made on the usage of one or another, however the onion method with some heavy optimizations (like incremental Cholesky decomposition) seems to have an edge in this regard.

For the vine method, a particular regular vine should be used if the partial correlations associated with this vine are of main interest (i.e., the sequence of conditioning is most natural for the variables) and they are needed as part of the generation of the random correlation matrix.

CHAPTER 6

Sample-based estimation of correlation ratio with polynomial approximation

> There are in fact two things, science and opinion; the former begets knowledge, the latter ignorance.
>
> *Hippocrates*

6.1 Introduction

Suppose a model is defined as a function $G = G(X_1, X_2, \ldots, X_n)$. The aim of sensitivity analysis is to investigate how much the uncertainty in X_i's, $i = 1, 2, \ldots, n$, or combinations thereof, contributes to the uncertainty in G. In this paper we concentrate on the notion of the so-called *correlation ratio* — a variance based measure.

The correlation ratio (CR) of random variable G with respect to random variable X is defined as

$$\eta^2(G|X) = \frac{\mathrm{Var}(\mathbf{E}(G|X))}{\mathrm{Var}(G)}.$$

Evidently, this is not a correlation coefficient of random variables; it is not symmetric and it is always non-negative.

Thanks largely to the work of McKay [1997] the correlation ratio is becoming recognized as a key notion in global sensitivity analysis. Other authors have

[5]This chapter is based on the publication *Sample-based estimation of correlation ratio with polynomial approximation* by Daniel Lewandowski, Roger M. Cooke and Radboud J. Duintjer Tebbens published in *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, Volume 18, Issue 1, pages 1–17, 2007.

studied this subject as well [e.g. Chan et al., 1997, Ishigami and Homma, 1990, Cooke and Lewandowski, 2001]. Saltelli et al. [2000a] offer an extensive overview of sensitivity analysis methods, including variance-based approaches. Recently the correlation ratio has been applied and compared with other sensitivity measures in [Duintjer Tebbens et al., 2008]. Theoretically, the correlation ratio is an attractive tool for quantifying importance because it represents the fraction of the variance of G that can be attributed to variation of X. However, there is an evident problem with computing it in a simple and accurate manner — estimation of the conditional expectation $\mathbf{E}(G|X)$ is the real challenge. A number of algorithms have been developed to overcome this difficulty, some more successful than others. Instructive among the lesser successful are the methods proposed by Kendall and Stuart [1961] and Sobol' [1993]. The first relies on a user-selected parameter (the number of bins for discretizing the model) which fully controls the value of the estimates. The second leads to very large deviations in the results and possible negative values although some may consider this as a strength of this method as it gives unbiased estimates. In general there is no need to approximate $\mathbf{E}(G|X)$ in order to estimate the correlation ratio (methods like FAST and Sobol' explained in [Saltelli et al., 2000b] do not deal with that at all, for instance). However the regression curve $\mathbf{E}(G|X)$ arises naturally in sensitivity analysis and having that determined is a useful byproduct.

Recognizing the drawbacks of the standard estimation methods, we look for methods which:

1. are based only on samples and do not require additional simulation and/or special simulation methods,
2. give an approximation of $\mathbf{E}(G|X)$ in analytical form,
3. do not require any input from the user, as this could control the result,
4. are generic, ie. not model specific,
5. are easy to implement in computer code,
6. have accuracy at least on par with other known methods.
7. have little computational cost.

The first point really means that we are interested in methods of estimating the correlation ratio from pseudo random or fully random samples only. The Bayesian method of Oakley and O'Hagan [2004], described in section 6.5.1, performs best if the samples for input variables are carefully chosen and therefore needs a special sampling algorithm. However, it also works with pseudo random samples very well, and therefore we include this method in our comparison. Theorems introduced in this paper help develop a new method of estimating correlation ratios complying with this specification. The main objective therefore is to present and compare 3 variants of this new method and decide which one performs best. The best adaptation of the method will be compared with two already known

state-of-the-art methods of estimating the correlation ratio on an example of a multivariate model.

This chapter is organized as follows. Section 6.2 places the correlation ratio into a broader context of global sensitivity measures. Section 6.3 presents a general definition of the correlation ratio and section 6.4 list properties of the correlation ratio. Next, in section 6.5 we describe methods proposed by Oakley and O'Hagan [2004] and Li et al. [2002]. Section 6.6 introduces 3 variations of the new method of estimating the correlation ratio. The performance of this method is investigated in section 6.7 with conclusions and discussion following in section 6.8.

6.2 Global sensitivity measures

The correlation ratio belongs to a family of global quantitative measures of importance of input factors for a given model; it is a variance-based non-parametric method closely related to Sobol' indices [Sobol', 1993, Chan et al., 2000b]. Sobol's method relies on decomposing the model function $G(\mathbf{U})$ into orthogonal summands of increasing dimensionality with zero mean, where $\mathbf{U} = (U_1, U_2, \ldots, U_n)$ is a vector of length n of statistically independent uniform random variables on $[0, 1]$ with realizations $\mathbf{u}$:

$$G(\mathbf{u}) = G_0 + \sum_{i=1}^{n} G_i(u_i) + \sum_{1 \le i < j \le n} G_{ij}(u_i, u_j) + \ldots + G_{1,2,\ldots,n}(u_1, u_2, \ldots, u_n), \quad (6.1)$$

where G_0 denotes the expectation of $G(\mathbf{U})$ and

$$\begin{aligned} G_i(u_i) &= E(G|u_i) - G_0; \\ G_{ij}(u_i, u_j) &= E(G|u_i, u_j) - G_i(u_i) - G_j(u_j) - G_0; \text{ etc.} \end{aligned}$$

Similarly, higher–order terms can be obtained. This is the starting point for the high-dimensional model representations (HDMR), tools for estimating G_i's. HDMR expresses the model output G as a function expansion as given in eq.(6.1). It can be generalized to non-uniform and correlated inputs as it is done in (see [Li et al., 2006, Bedford, 1998]). Li et al. [2002] approximate the HDMR component functions analytically by orthonormal polynomials, polynomial spline functions and ordinary polynomials (formulae exist for determining coefficients of orthonormal polynomials), as well as numerically by using kernel smoothers. They do not, however, consider the problem of overfitting which is evidently possible if the order of the polynomial is too high.

With the assumption of independence of inputs and given eq.(6.1) the variance of G may be written:

$$\begin{aligned} \mathrm{Var}(G(\mathbf{U})) &= \textstyle\sum_{i=1}^{n} Var(G_i(X_i)) + \sum_{1 \le i < j \le n} Var(G_{ij}(X_i, X_j)) + \\ &\quad + \ldots + Var(G_{1,2,\ldots,n}(X_1, X_2, \ldots, X_n)). \end{aligned} \quad (6.2)$$

The Sobol' k-th order *sensitivity index* is defined as

$$S_{i_1,\ldots,i_k} = \frac{\mathrm{Var}(G_{i_1,\ldots,i_k}(U_{i_1}, \ldots, U_{i_k}))}{\mathrm{Var}(G)}.$$

Sobol' indices sum up to unity. The first order Sobol' indices were used already by Pearson [1903].

The role of the correlation ratio in quantifying importance is based on the well-known relation (which does not require $\{U_i\}$ to be independent):

$$\mathrm{Var}(G) = \mathbf{E}(\mathrm{Var}(G|U_i)) + \mathrm{Var}(\mathbf{E}(G|U_i)),$$

If the expected reduction in variance of G with U_i fixed is small, then the variance $\mathrm{Var}(\mathbf{E}(G|U_i))$ is large. Normalizing by $\mathrm{Var}(G)$, $\frac{\mathrm{Var}(\mathbf{E}(G|U_i))}{\mathrm{Var}(G)}$ represents the fraction of the variance of G which is "explained" by U_i. The use of Sobol' indices as a sensitivity measure is then motivated by the fact that they explain *all* the variance, according to eq.(6.2). For a more detailed overview of Sobol' indices see [Chan et al., 2000a]. The following section suggests another motivation of the correlation ratio, not based on variance reduction, but on optimal prediction.

6.3 Definition of correlation ratio

Building on the concept of Sobol' indices, we more generally define for any random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and any subset $\mathbf{X}^{(k)}$ of k components of $\mathbf{X}$, $(1 \leq k \leq n)$:

Definition 6.3.1. *The correlation ratio η^2 of $G = G(\mathbf{X})$ with respect a to random vector $\mathbf{X}^{(k)}$ is*

$$\eta^2\left(G|\mathbf{X}^{(k)}\right) = \frac{\mathrm{Var}\left(\mathbf{E}\left(G|\mathbf{X}^{(k)}\right)\right)}{\mathrm{Var}(G)}. \tag{6.3}$$

The correlation ratio can be motivated in terms of optimal prediction. One may ask for which function $f : \mathbb{R}^k \mapsto \mathbb{R}$ with $\sigma^2_{f(\mathbf{X}^{(k)})} < \infty$ is the correlation $\rho^2(G, f(\mathbf{X}^{(k)}))$ maximal? The answer is given by the generalized result of Cooke and Lewandowski [2001] (similar to a result of Whittle [1992]).

Theorem 6.3.1. *Let $\mathbf{X}^{(k)}$, G and and $f(\mathbf{X}^{(k)})$ have finite variance. Then*

$$\max_f \rho^2(G, f(\mathbf{X}^{(k)})) = \rho^2(G, \mathbf{E}(G|\mathbf{X}^{(k)})) = \frac{\mathrm{Var}(\mathbf{E}(G|\mathbf{X}^{(k)}))}{\mathrm{Var}(G)} = \eta^2(G|\mathbf{X}^{(k)}).$$

Proof. Let $\delta(\mathbf{X}^{(k)})$ be any function with finite variance and write $f(\mathbf{X}^{(k)}) = \mathbf{E}(G|\mathbf{X}^{(k)}) + \delta(\mathbf{X}^{(k)})$. Put $A = \sigma^2_{\mathbf{E}(G|\mathbf{X}^{(k)})}$, $B = \mathrm{Cov}(\mathbf{E}(G|\mathbf{X}^{(k)}), \delta(\mathbf{X}^{(k)})) = \mathrm{Cov}(G, \delta(\mathbf{X}^{(k)}))$, $C = \sigma^2_G$, and $D = \sigma^2_\delta$. Then

$$\begin{aligned}
\rho^2(G, \mathbf{E}(G|\mathbf{X}^{(k)}) + \delta(\mathbf{X}^{(k)})) &= \frac{(A+B)^2}{C(A+D+2B)}, \\
\frac{\sigma^2_{\mathbf{E}(G|\mathbf{X}^{(k)})}}{\sigma^2_G} &= \frac{A}{C}, \\
\frac{(A+B)^2}{C(A+D+2B)} \leq \frac{A}{C} &\iff B^2 \leq AD.
\end{aligned}$$

The latter inequality follows from the Cauchy-Schwarz inequality. ∎

If $k = 1$ then the conditioning set of variables $\mathbf{X}^{(k)}$ contains only one element which we denote by X. If the optimal regression of G on $\mathbf{X}^{(k)} = \{X\}$ is linear, that is, $\mathbf{E}(G|X) = aX + b$, then

$$\begin{aligned}\mathrm{Var}(\mathbf{E}(G|X)) &= \mathrm{Var}(aX+b) = \\ &= \frac{\mathrm{Cov}^2(aX+b, X)}{\mathrm{Var}(X)} = \frac{\mathrm{Cov}^2(\mathbf{E}(G|X), X)}{\mathrm{Var}(X)} = \frac{\mathrm{Cov}^2(G, X)}{\mathrm{Var}(X)},\end{aligned}$$

and eq.(6.3) becomes the product moment correlation squared $\rho^2(G, X)$.

Sobol' indices coincide with the correlation ratio when the explanatory variables are independent uniforms. However, when the variables are not independent, the motivation of Sobol' indices in terms of variance decomposition, as in eq.(6.2) is lost. It suffices to consider $G = X + Y$ with $X = Y$. Then $\eta^2(G|X) = \eta^2(G|Y) = \eta^2(G|(X, Y)) = 1$, and they obviously do not sum to one. The correlation ratio admits a more general motivation in terms of prediction, according to Theorem 6.3.1.

Remark. We know from Theorem 6.3.1 that $\eta^2(G|X) = \rho^2(G, X)$ if the regression curve $E(G|X)$ is linear. Hence the notion of correlation ratio can be used for testing the linearity of the regression. Kendall and Stuart [1961] test the linearity of the regression with statistic

$$k = \eta^2(G|X) - \rho^2(G, X). \tag{6.4}$$

The statistic $0 \leq k \leq 1$, with $k = 0$ if $E(G|X)$ is a linear function of X.

6.4 Properties of correlation ratios

The first lemma is straightforward and uses the linearity property of covariance. We consider a partition of $\mathbf{X}$ into s disjoint subsets $\mathbf{X}^i$ of its components such that $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^s)$ ($\mathbf{X}^i \neq \emptyset$, $i = 1, \ldots, s$; $s \geq 1$). The components of a given subset do not have to be independent. If the $\{\mathbf{X}^i\}$ are independent, then their correlation ratio's explain all of the variance.

Lemma 6.4.1. *Let $G = G(\mathbf{X})$; $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^s)$ ($\mathbf{X}^i \neq \emptyset$, $i = 1, \ldots, s$; $s \geq 1$), then:*

$$Cov\left(G, \sum_{i=1}^{s} E(G|\mathbf{X}^i)\right) = \sum_{i=1}^{s} Var\left(E\left(G|\mathbf{X}^i\right)\right).$$

The next proposition is straightforward, and Proposition 6.4.3 uses Lemma 6.4.1.

Proposition 6.4.2. *Let $g_i : \mathbb{R}^{k_i} \to \mathbb{R}$ where k_i is the length of vector $\mathbf{X}^i$, $i = 1, \ldots, s$. Let $\{\mathbf{X}^i\}_{i=1}^s$ be mutually independent, and let $G = \sum_i^s g_i(\mathbf{X}^i)$ with $\sigma_{g_i}^2 < \infty$, such that $\sigma_G^2 > 0$. Then*

$$\sum_{i=1}^{s} \eta^2(G|\mathbf{X}^i) = 1.$$

Proof.

$$E(G|\mathbf{X}^i) = E\left(\sum_{j=1}^{s} g_j(\mathbf{X}^j)|\mathbf{X}^i\right) = g_i(\mathbf{X}^i) + \sum_{j\neq i} E(g_j(\mathbf{X}^j)),$$

so that $Var(E(G|\mathbf{X}^i)) = Var(g_j(\mathbf{X}^j))$. Since :

$$Var(G) = \sum_{i=1}^{s} Var(g_i(\mathbf{X}^i))$$

we have:

$$\sum_{i=1}^{s} \eta^2(G|X_i) = \frac{\sum_{i=1}^{s} Var(E(G|\mathbf{X}^i))}{Var(G)} = 1.$$

■

The additive form of G is essential. Let $G = X \cdot Y$, $X \perp Y$, $E(X) = E(Y) = 0$. Then $Var(E(G|X)) = Var(X \cdot E(Y)) = 0 = Var(E(G|Y))$. Without additivity we can get only:

Proposition 6.4.3. *Let $G = G(\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^s)$ with $Cov(E(G|\mathbf{X}^i), E(G|\mathbf{X}^j)) = 0$, $i \neq j$; then*

$$\sum_{i=1}^{s} \eta^2(G|\mathbf{X}^i) \leq 1.$$

Proof. Lemma 6.4.1 and the zero covariance assumption imply

$$Cov\left(G, \sum_{i=1}^{s} E(G|\mathbf{X}^i)\right) = \sum_{i=1}^{s} Var\left(E(G|\mathbf{X}^i)\right) = Var\left(\sum_{i=1}^{s} E(G|\mathbf{X}^i)\right). \quad (6.5)$$

On the other hand by the properties of variance

$$\begin{aligned} Var\left(G - \sum_{i=2}^{s} E(G|\mathbf{X}^i)\right) = \\ = \; Var(G) + Var\left(\sum_{i=2}^{s} E(G|\mathbf{X}^i)\right) - 2\,Cov\left(G, \sum_{i=2}^{s} E(G|\mathbf{X}^i)\right) \\ = \; Var(G) - Var\left(\sum_{i=2}^{s} E(G|\mathbf{X}^i)\right) \geq 0. \end{aligned} \quad (6.6)$$

Then, by eq.(6.5) and (6.6) we have:

$$\begin{aligned} \rho\left(E(G|\mathbf{X}^1), G - \sum_{i=2}^{s} E(G|\mathbf{X}^i)\right) = \\ = \; \frac{Cov(E(G|\mathbf{X}^1), G - \sum_{i=2}^{s} E(G|\mathbf{X}^i))}{\sqrt{Var(E(G|\mathbf{X}^1))}\sqrt{Var(G) - Var\left(\sum_{i=2}^{s} E(G|\mathbf{X}^i)\right)}} \\ = \; \frac{\sqrt{Var(E(G|\mathbf{X}^1))}}{\sqrt{Var(G) - Var\left(\sum_{i=2}^{s} E(G|\mathbf{X}^i)\right)}} \leq 1. \end{aligned}$$

Thus

$$Var(E(G|\mathbf{X}^1)) + Var\left(\sum_{i=2}^{s} E(G|\mathbf{X}^i)\right) \leq Var(G).$$

■

Proposition 6.4.4. *If $k = 1$ (thus $\mathbf{X}^{(k)} = (X)$), then*

$$\eta^2(G|X) = \frac{\rho^2(G, X)}{\rho^2(X, E(G|X))}.$$

Proof. Since $Cov(G, X) = Cov(E(G|X), X)$,

$$\begin{aligned}
\rho^2(G, X) &= \frac{Cov(E(G|X), X)}{Var(G)Var(X)} \cdot \frac{Var(E(G|X))}{Var(E(G|X))} \\
&= \frac{Cov(E(G|X), X)}{Var(E(G|X))Var(X)} \cdot \frac{Var(E(G|X))}{Var(G)} \\
&= \rho^2(E(G|X), X) \cdot \rho^2(G, E(G|X)).
\end{aligned}$$

However $\rho^2(G, E(G|X)) = \eta^2(G|X)$. ■

Proposition 6.4.5. *Let $h : \mathbb{R}^k \mapsto \mathbb{R}^m$, $m \leq k$, with $\sigma^2_{h(\mathbf{X}^{(k)})} < \infty$. Then*

$$\eta^2(G|\mathbf{X}^{(k)}) \geq \eta^2(G|h(\mathbf{X}^{(k)})).$$

Proof. Consider quantities $\max_f \rho^2(G, f(\mathbf{X}^{(k)}))$ and $\max_g \rho^2(G, g(h(\mathbf{X}^{(k)})))$. The maximization procedure over all fs maximizes over all possible $g \circ h$ as well. Hence

$$\max_f \rho^2(G|f(\mathbf{X}^{(k)})) \geq \max_g \rho^2(G|g(h(\mathbf{X}^{(k)}))).$$

■

6.5 Standard methods of estimating correlation ratio

State-of-the-art methods for computing the correlation ratio include the Bayesian approach of Oakley and O'Hagan [2004] and State Dependent Parameter (SDP) model by Ratto et al. [2006]. We describe them both briefly here. The HDMR method of Li et al. [2002] stops where we start. It approximates the component functions but does not deal with the prevention of overfitting. It must be noted that a variety of other approaches exist for carrying out this task, like *FAST* [see Saltelli et al., 1999]. We do not consider these in this paper in view of the requirements formulated in section 6.1.

It is assumed from now on that the sample size is m. Symbol $\mathbf{x}_j$ denotes the j-th vector of realizations of $\mathbf{X}$ and $\mathbf{x}_{i,j}$ is the j-th realization of X_i.

6.5.1 Bayesian approach

The first method employs the Bayesian paradigm by emulating G as a Gaussian process whose parameters are assigned hyper prior distributions and updating using model evaluations $G(\mathbf{x}_j)$, $j = 1, \ldots, m$. For further reading on this method please refer to the article of Oakley and O'Hagan [2004].

The biggest advantage of this approach is that it does not require a large number of simulations and therefore it is best suited for applications when computing the model evaluations is rather complicated and time consuming. On the other hand it requires the user to specify many parameters and to implement routines for numerical integration. The sensitivity of this method to various specifications of the input parameters is yet to be determined as there is no study on this subject (the choice of samples for instance).

6.5.2 State Dependent Parameter models

The State Dependent Parameter modelling developed in [Ratto et al., 2006], in turn, can be applied to any Monte Carlo sample and can be seen as one of the *postprocessing* methods, ie. the analysis is done after the creation of the sample. The idea is to extract the signal ($\mathbf{E}(G|X_i)$) from noisy data ($G(X_i)$). In order to prepare simulation data, which does not need to exhibit any *temporal* order, for smoothing with this method one has to sort the values of X_i in an increasing order (with $Y = G(X_i)$ sorted accordingly) and *pretend* that this ordered statistic specifies a time series. The change in Y as X_i changes its value from $x_{i,j}$ to $x_{i,j+1}$ is modelled as a random walk process. The forward filtering algorithm has been coupled with backward recursive smoothing in this case Fixed Interval Smoothing algorithm since the data is available for the whole range and does not come sequentially.

Unfortunately, there is no computer implementation of this method available at the time of this writing. Therefore a full comparison of the new method with the SDP approach is not possible although it clearly has potential.

6.5.3 Sobol' method

Sobol' [1993] introduced a method using Monte–Carlo simulation. Let $\mathbf{X}_{\sim i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. If we can sample $\mathbf{X}'_{\sim i}$ from the conditional distribution $(\mathbf{X}_{\sim i}|X_i)$ independently of $\mathbf{X}_{\sim i}$, and if the evaluation of G is not too expensive, then the following simple algorithm may be applied:

Algorithm 6.5.1 (Estimation of correlation ratio with Sobol's method).

1. Sample $\mathbf{x}$ from $\mathbf{X}$;
2. Compute $G = G(\mathbf{x})$;
3. Sample $\mathbf{x}'_{\sim i}$ from $(\mathbf{X}_{\sim i}|X_i = x_i)$ independent of $\mathbf{X}_{\sim i} = \mathbf{x}_{\sim i}$;
4. Compute $G' = G((x_i, \mathbf{x}'_{\sim i}))$;
5. Store $Z = G \cdot G'$;

6. Repeat.

The average value of Z will approximate $E(E^2(G|X_i))$, from which the estimation of the correlation ratio $\eta^2(G|X_i)$ may be computed as

$$\eta^2(G|X_i) \approx \frac{E(E^2(G|X_i)) - E^2(G)}{\sigma_G^2}.$$

If $\mathbf{X}_{\sim i}$ and X_i are independent, then this algorithm poses no problems. If $\mathbf{X}_{\sim i}$ and X_i are not independent, then it may be difficult to sample from $(\mathbf{X}_{\sim i}|X_i)$. The biggest weakness of this method is the high variance of the estimates, especially for small samples. It is not unusual to obtain negative values of the estimation if the true value is close to 0.

It could be more reasonable to choose one of the *postprocessing* methods of estimating the correlation ratio, i.e. estimate $Var(E(G|X_i))$ based only on a large sample generated before the analysis.

6.5.4 Kendall-Stuart method

Kendall and Stuart [1961] propose a method that might be described as "pedestrian". Let m be the number of samples per variable.

Algorithm 6.5.2 (Estimation of correlation ratio with Kendall-Stuart method).

1. Collect m samples of $(G, \mathbf{X})$;
2. Order the X_i values $x_{i(1)}, ..., x_{i(m)}$ from smallest to largest;
3. Divide the samples into M cells $C_1, ..., C_M$, where C_1 contains the samples with the m_1 smallest X_i values, C_2 contains samples with the m_2 smallest X_i values which are bigger than those in C_1, etc.;
4. Compute $\widehat{G}_j = E(G|X_i \in C_j), j = 1, ..., M$;
5. Compute the unbiased variance of these conditional expectations, weighted by the number of samples, as

$$Var(E(G|X_i)) \approx Var(\widehat{G}) = \sum_{j=1}^{M} \frac{n_j(\widehat{G}_j - E(G))^2}{m-1}.$$

This is an intuitive transliteration of the mathematical definition. The good news is that its badness is illuminating. The problem lies in the choice of M and m_j. If M is sufficiently large, then m_j is either 0 or 1. Take only those C_j's with $m_j = 1$. Then C_j contains exactly one sample, say $(g', \mathbf{x}')$ and $E(G|X_i \in C_j) = g'$ for all j's. Taking the variance of these numbers will simply return the unconditional variance of G. On the other hand, if we take $M = 1$, then all sample values $(g', \mathbf{x}')$ will satisfy $x_i' \in C_1$ and $E(G|X_i \in C_1) = E(G)$, so the variance of the conditional expectation will be zero. Appropriately choosing the size and number of the cells C_j we traverse the values between $Var(G)$ and 0. Variations on the pedestrian method using kernel estimators are discussed in [Kurowicka and Cooke, 2006a], and experience the same issues.

6.6 Polynomial approximation methods

The problems of estimating the correlation ratio are, for the most part, caused by the recurring issue of estimating the regression curve $\mathbf{E}(G|X)$ based on data. There is a great deal of literature on the latter [Draper and Smith, 1998, Kleinbaum et al., 1998]; but we propose a simpler strategy that can be easily implemented.

For simplicity, we restrict attention to the case, where the explanandum X is a one-dimensional random variable rather then a vector.

The method we propose assumes that the regression function is *analytic*, that is it can be approximated as a Taylor expansion, i.e., a polynomial function. Having said that one can immediately observe that Theorem 6.3.1 gives a good instrument for estimating $\mathbf{E}(G|X)$. Intuitively, since the regression curve is a function that maximizes $\rho^2(G, f(X))$ over all possible $f(X)$, then under the above assumption of smoothness we are searching for a polynomial $g_d(X)$ of degree d that maximizes $\rho^2(G, g_d(X))$. Optimization methods can be implemented with the coefficients of the polynomial as independent variables.

6.6.1 Polynomial fit

For fixed d the optimization problem can be formulated as:

$$\text{maximize } \rho^2(G, p_0 + p_1 X + \ldots + p_d X^d) \tag{6.7}$$

Optimization routines are time consuming, however. The following theorem states that equivalent results can be obtained by simply applying the least–squares error method to fit the polynomial.

Theorem 6.6.1. *Let $G = G(\mathbf{X})$ with $\sigma_G^2 < \infty$ and $X \in \mathbf{X}$. Then*

$$\arg\min_f \mathbf{E}(G - f(X))^2 = \mathbf{E}(G|X).$$

Proof. Decompose the variance of $G - f(X_i)$ in order to obtain

$$\mathbf{E}(G - f(X))^2 = \mathrm{Var}(G - f(X)) + \mathbf{E}^2(G - f(X)).$$

Minimizing the right hand side of the above equation implies setting $\mathbf{E}(G) = \mathbf{E}(f(X))$ (hence $\mathbf{E}^2(G - f(X)) = 0$). Express f as

$$f(X) = \mathbf{E}(G|X) + \delta(X),$$

where $\mathbf{E}(\delta(X)) = 0$, and note that

$$\begin{aligned} \mathbf{E}(G \cdot \mathbf{E}(G|X)) &= \mathbf{E}(\mathbf{E}(G|X)^2) \\ \mathbf{E}(\delta(X) \cdot \mathbf{E}(G|X)) &= \mathbf{E}(\mathbf{E}(G\delta(X)|X)) = \mathbf{E}(G \cdot \delta(X)). \end{aligned}$$

Then

$$\mathbf{E}(G - (\mathbf{E}(G|X) + \delta(X)))^2 = \mathbf{E}(G^2 + \delta(X)^2) - \mathbf{E}(\mathbf{E}(G|X)^2)$$

attains its minimum when $\delta(X) = 0$ and hence $f(X) = \mathbf{E}(G|X)$. ■

Henceforth we use the least squares error method to fit a polynomial to data. If one still prefers to apply the optimization problem (6.7) then consider the following. The solution of (6.7) is unique up to positive affine transformation (since the correlation is invariant under positive affine transformations). It is very likely that the approximation $v(x)$ of the regression curve $E(G|X=x)$ obtained in that way will not even pass through the scatter plot of G vs X. On the other hand the least-squares error approach always leads to a regression approximation $v'(x)$ that has this feature. Hence there exists a linear transformation $a\,v(x)+b=v'(x)$, where $a>0$ and b are real constants.

The next proposition gives formulae for calculating a and b in case $E(G|X=x)$ takes the form of a polynomial.

Proposition 6.6.2. *Let* $p=(p_0,p_1,\ldots,p_d)$, $\overline{\mathbf{X}}=(1,X,X^2,\ldots,X^d)$ *and*

$$p^*=\arg\max_p \rho^2\left(G,p\overline{\mathbf{X}}^T\right),\quad p'=\arg\min_p E\left(G-p\overline{\mathbf{X}}^T\right)^2.$$

Let $v(X)=p^*\overline{\mathbf{X}}^T$ *and* $v'(X)=p'\overline{\mathbf{X}}^T$. *Then there exist real constants* $a\neq 0$ *and* b *such that*

$$a\,v(X)+b=v'(X),$$

where

$$\begin{aligned} a &= \frac{Cov(G,v(X))}{Var(v(X))},\\ b &= E(G)-aE(v(X)). \end{aligned}$$

Proof. Assume λ to represent one of the p_i's, $i=1,2,\ldots,d$. Then let $d/d\lambda$ denote the operator of differentiation with respect to one of the coefficients of $v=v(X)$. Function v as a solution of the optimization problem (6.7) satisfies the following equation

$$\frac{d}{d\lambda}\log(\rho^2(G,v))=\frac{2\frac{d}{d\lambda}Cov(G,v)}{Cov(G,v)}-\frac{2\frac{d}{d\lambda}\sigma_v}{\sigma_v}=0. \tag{6.8}$$

Note that $\frac{d}{d\lambda}\sigma_v=\frac{Cov(u,\frac{d}{d\lambda}v)}{\sigma_v}$. Substituting $\frac{d}{d\lambda}\sigma_v$ into (6.8) and simplifying yields

$$\frac{\frac{d}{d\lambda}Cov(G,v)}{Cov(G,v)}=\frac{Cov(v,\frac{d}{d\lambda}v)}{\sigma_v^2}$$

and therefore

$$\frac{\frac{d}{d\lambda}Cov(G,v)}{Cov(v,\frac{d}{d\lambda}v)}=\frac{Cov(G,v)}{\sigma_v^2}=\frac{Cov\left(G,\frac{1}{a}E(G|X)-\frac{b}{a}\right)}{Var\left(\frac{1}{a}E(G|X)-\frac{b}{a}\right)}=a. \tag{6.9}$$

The latter equality follows from Lemma 6.4.1. In order to obtain the formula for b note that

$$E(v)=E\left(\frac{1}{a}E(G|X)-\frac{b}{a}\right)=\frac{1}{a}E(G)-\frac{b}{a}$$

and hence

$$b=E(G)-aE(v).$$

■

The above theorem can be illustrated by applying both the least–squares error and the optimization methods in order to determine v and v'. Simply determine coefficients a and b by formulating a minimization problem of the sum of squared differences between $a\,v(X) + b$ and G as a function of a and b. Then compare them with those obtained by using the formulae given by Proposition 6.6.2.

6.6.2 Prevention of overfitting

Fitting a polynomial to data introduces a problem of overfitting. The challenge is not to fit (in the least-squares error sense) a function that predicts perfectly values of the fitted sample, but a function that will be representative for the whole population from which the samples were drawn. Therefore there is a need for introducing a mechanism to prevent overfitting. Since it has been assumed that the model is a polynomial, the only parameter that can be used for controlling the overfitting is the degree of the polynomial. Once the degree is fixed the coefficients are uniquely determined by applying the least-squares error method.

Ideally, the number of independent samples from the same joint distribution is unlimited. How can we escape from the trap of overfitting then? One can use the algorithm given below:

Algorithm 6.6.1 (Overfitting prevention).

1. Split the sample into test and validations samples.
2. Fit a polynomial of degree d to the test sample,
3. Calculate the test correlation ratio using this polynomial as the regression curve,
4. Calculate correlation ratios for the remaining validation data sets using the same polynomial as the regression curve (build up a distribution of correlation ratios),
5. Check if the correlation ratio for the test sample is significantly higher than the other ones.

Clearly, if the polynomial fit is representative only for the test sample, then its correlation ratio will be in the tail of the distribution of remaining validation correlation ratios and we can reject the null hypothesis that this given polynomial is a good approximation to the regression curve. Otherwise, the correlation ratio of the test sample will be somewhere closer to the median of the distribution and gives no evidence to reject the null hypothesis.

This method can be applied only if evaluating a model is not computationally intensive and a large number of data sets can be produced. If this is not the case a different method can be applied.

In order to experimentally determine the distribution of correlation ratios given only one data set, one may be tempted to use resampling methods. We give an equal importance to the test part (polynomial fit, test correlation ratio) and the validation part (determining the distribution of validation correlation ratios given

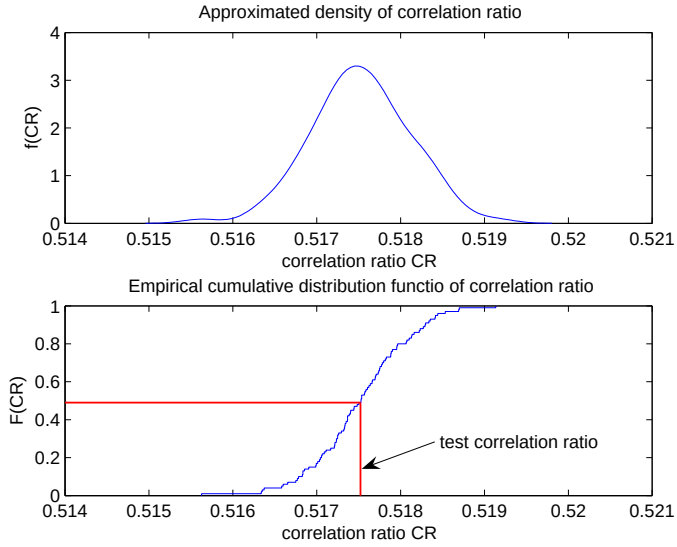


Figure 6.1: *Empirical distribution of correlation ratio obtained using the jackknife method and the value of the original correlation ratio (red line).*

the polynomial fit), thus we split the whole sample into two data sets of the same size and fit a polynomial only to the test part of the sample. Then one can apply various resampling methods (like bootstrap or jackknife) to the entire sample in order to use as much data as possible and to create validation samples needed to determine the distribution of correlation ratios. However this method does not give satisfactory results because the created validation samples contain, among others, samples used to obtain the fit. Experimenting with these resampling methods showed that the test correlation ratio is always equal (jackknife) or close (bootstrap) to the average of the empirical distribution of validation correlation ratios. Figure 6.1 shows one example of such case. Validation samples are created with the jackknife method and the test correlation ratio is exactly equal to the average of the validation CR's.

Therefore we consider three other methods for determining the optimal degree of the polynomial estimation of the regression curve:

Adjusted R^2 This statistic is a rather standard tool used in regression analysis for evaluating impact of additional variables on a model's performance. The multiple correlation R^2 can be computed as the squared correlation $\rho^2(G, g_d(X))$. The adjusted R^2, accounting for the number of parameters in the model, is

$$adj\, R^2 = 1 - (1 - R^2)\frac{n-1}{n-d-1}$$

The adjusted R^2 can decrease if increasing the polynomial degree d is not associated with a sufficient increase in R^2. Choose the degree d maximizing

the adjusted R^2.

Early stopping This idea is based on an approach applied in machine learning models such as neural networks. The sample is split into two subsets: a test sample and a validation sample. A polynomial of degree d is fitted to the test sample and used to estimate the correlation ratio for the validation sample. Choose the lowest d such that the correlation ratio with polynomial of degree d on the validation set is greater than with $d+1$.

Wilcoxon rank sum test The one-sided Wilcoxon rank sum test also compares two data sets of estimates of correlation ratio based on the test sample (data set 1) and the validation sample (data set 2) and tries to detect the shift in their distributions. The null hypothesis is that both distributions are equal. The alternative is that data set 1 is statistically larger than data set 2. The test statistic is the sum of ranks of the test observations among all combined and sorted test and validation observations. Its distribution can be easily tabulated or approximated by the normal distribution [Hodges and Lehmann, 1970].

Our specific application of this test relies on the following reasoning. First split a given sample into two equally sized subsets (T — test sample and V — validation sample), then fit a polynomial of degree d to the test sample. Now divide both T and V into 10 smaller data sets of equal size and calculate the approximate correlation ratios for each of these based on the polynomial fitted on the test sample. In the end 10 values of correlation ratio for the test sample and 10 corresponding values of correlation ratio for the validation sample are obtained. They form two sets that will be compared with the help of the Wilcoxon rank sum test. The sum of the ranks W_T of the test group is expected to be larger than this sum W_V for the validation group. We use the following p-value as an indication of overfitting

$$P(W_T \geq w_T) = p_W,$$

where w_T is the realization of the rank sum of the test sample correlation ratios. Small p-value indicates overfitting. For the calculations presented next, we choose degree d for which the p-value is closest to 0.05 from above.

6.7 Simulations and Results

The performance of all of the variations of the polynomial method introduced in section 6.6.2 is compared in terms of their ability to estimate the true correlation ratio. The search algorithm is restricted to polynomials of degree from 1 to 20, as the fitting algorithms in generally available programs (eg. MATLAB) experience numerical instabilities for degrees greater than 20. The sample sizes that expose sensitivity for overfitting are therefore also relatively small. Of course, if higher degree polynomials can be reliably fitted, the overfitting issues will apply to larger sample sizes. The synthetic benchmark model used for simulations is chosen such

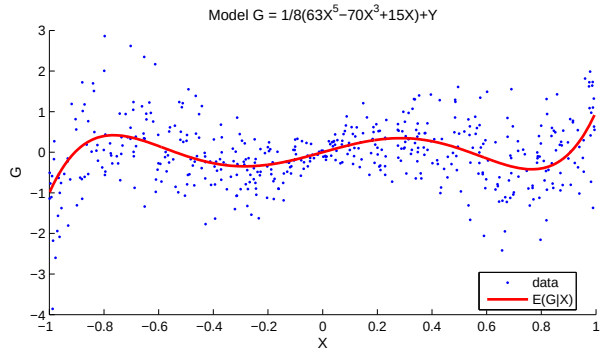


Figure 6.2: *Scatter plot of 500 samples generated given model* $G = \frac{1}{8}(63X^5 - 70X^3 + 15X) + Y$.

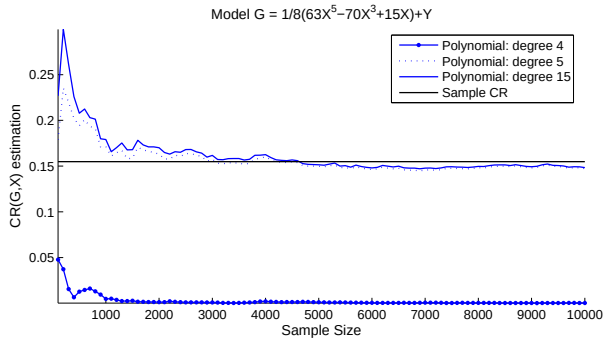
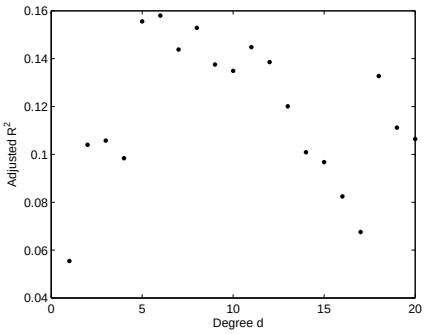


Figure 6.3: *Convergence of the estimation of correlation ratio* $\widehat{\eta}^2(G|X)$ *as a function of sample size* n.

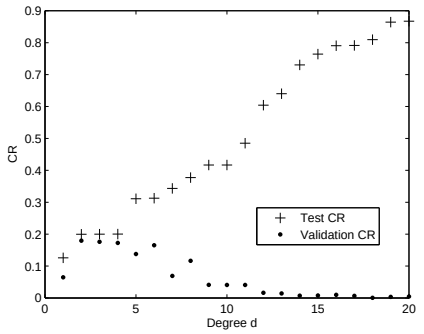
that the true $\eta^2(G|X)$ can be easily calculated analytically (X is the explaining variable and Y is added noise). Let $G = f(X) + Y = \frac{1}{8}(63X^5 - 70X^3 + 15X) + Y$ where $X \sim U[-1, 1]$ and $Y \sim \mathcal{N}(0, \sqrt{|X|})$, $E(Y|X) = 0$. Thus the true regression function $\mathbf{E}(G|X) = f(X)$ is known. This highly non–linear model presented in Figure 6.2 exhibits heteroscedasity in error variance, $\eta^2(G|X) \approx 0.1538$.

6.7.1 Influence of sample size

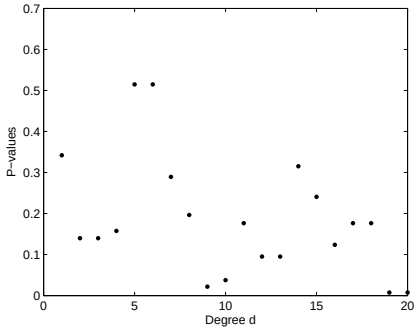
The sample size is a crucial factor in estimating any statistical quantity, therefore we study its influence on the accuracy of the estimations. It can be observed in Figure 6.3 that small sample sizes cause problems in estimating the correlation ratio accurately, as expected. The estimations of the correlation ratio are compared against the sample correlation ratio computed on the whole data set rather than the true η^2 in order to avoid penalizing the estimator for features of the data. Since the regression function is given the sample correlation ratio can be computed as the ratio of the sample variances $\mathrm{Var}(f(X))$ and $\mathrm{Var}(G)$. The po-



(a) Adjusted R^2



(b) Early stopping



(c) Wilcoxon rank sum test

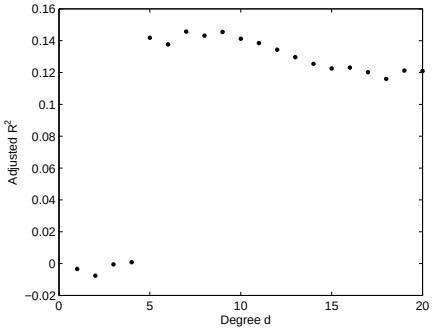
Figure 6.4: *Various statistics vs degree of polynomial approximation for 60 samples.*

lynomial approximation of degree 4 badly underestimates the sample correlation ratio of 0.155. It simply does not exhibit enough variability. On the other hand polynomial approximations of degree 5 (the degree of the model polynomial) and 15 yield good estimates for sample sizes greater than 400, indicating little sensitivity to the polynomial degree once it is at least equal to the degree of the true regression polynomial.

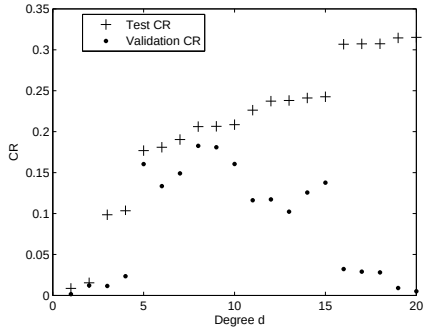
6.7.2 Overfitting

As it has been already mentioned an important issue for the polynomial methods of estimating the correlation ratio is the prevention of overfitting. Figure 6.4 shows a typical picture of what one may expect from the values of the adjusted R^2, the test and the validation CR's and p-values versus the degree of the fitted polynomial approximation for a small sample size, in this case 60. In this situation the adjusted R^2 statistic is rather unstable.

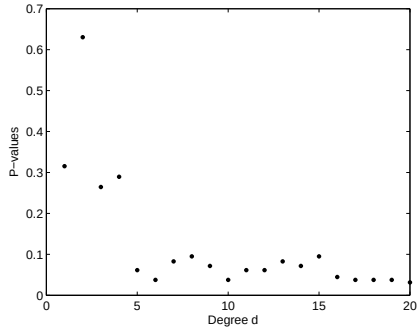
We proposed two other techniques for preventing overfitting designed with this specific issue in mind. Early stopping trains the polynomial approximation



(a) Adjusted R^2



(b) Early stopping



(c) Wilcoxon rank sum test

Figure 6.5: *Various statistics vs degree of polynomial approximation for 200 samples.*

first on test data and then checks its performance on validation data. Figure 6.4b shows values of the estimates of the correlation ratio both on the test data and the validation data against the degree of the polynomial approximation. The validation-set correlation ratio gradually increases as the degree increases and eventually starts decreasing when the degree of the approximation becomes too high. We stop when the correlation ratio on the validation set starts to decrease. This method is more eager to penalize data overfitting by reducing the optimal degree of the polynomial approximation.

The Wilcoxon rank sum test for preventing overfitting is much more *forgiving* in a sense that it rejects the hypothesis of overfitting only after there is a clear evidence to do so. This evidence is the p-value being as close to 0.05 as possible, but not lower. The threshold value (0.05 in our case) should reflect analyst's particular risk attitude. The example we present in Figure 6.4c shows the p-values to be very noisy for this small sample but a general tendency for decreasing value as the degree increases can be observed.

Things become clearer with a larger data set of 200 samples (see Figure 6.5).

There is a clear jump of the adjusted R^2 statistics when the degree of the approximation polynomial changes from 4 to 5. This jump can be explained by the fact that the true model is also a fifth order polynomial in X. The early stopping method also correctly detects the underlying model as the fifth order polynomial. The maximum of the validation-set correlation ratio is attained for degree equal to 5 and gradually decreases when the polynomial degree increases giving some evidence for overfitting. The behavior of the p-value of the Wilcoxon rank sum test is also much more stable than with only 200 samples. The degree with the p-value closest to 0.05 from the top is 5 as well.

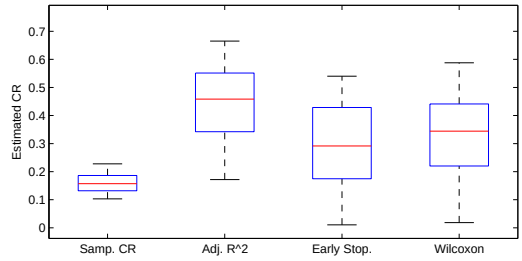
6.7.3 Robustness

The robustness of the estimation methods will be studied given three sample sizes — 60, 200 and 1000 samples. We estimate the statistical fluctuation of the estimation by iterating the estimation process 500 times. One iteration consists of the following steps:

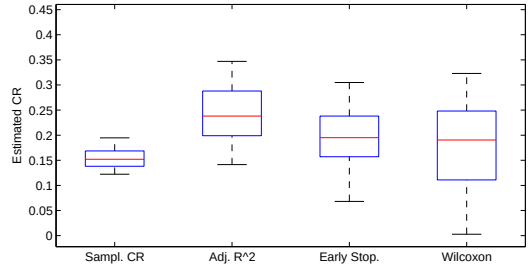
1. Generate n samples of X, Y and compute $G = G(X, Y)$;
2. Fit polynomials of degree 1 to 20 to the whole sample and calculate the adjusted R^2 for each polynomial (Adjusted R^2 method);
3. Fit polynomials of degree 1 to 20 to the first half of the sample and calculate the estimated correlation ratio on the other half of the sample for each polynomial (Early stopping method);
4. Fit polynomials of degree 1 to 20 to the first half of the sample, then split each half into 10 subsamples and calculate the p-value of the Wilcoxon rank sum test statistics for each polynomial (Wilcoxon rank sum test method).

Figure 6.6a shows the box plots of the estimates of correlation ratio calculated based on 60 samples using various polynomial methods presented in this paper. The lower and upper lines of the boxes are the 25^{th} and 75^{th} percentiles of the sample and the whiskers are the 5^{th} and 95^{th} percentiles. The lines in the middle of the box plots show the medians. The first box plot (denoted as *Samp. CR* in Figure 6.6) represents the distribution of the estimates calculated using the true regression function, ie. the error of the estimates occurs only due to statistical fluctuation in samples. The remaining distributions contain variability also due to the model approximation. Selecting an optimal polynomial based on the adjusted R^2 tends to overestimate CR (data overfitting) compared to the early stopping and Wilcoxon methods. The best performing method both in terms of the accuracy and low variability is the early stopping algorithm.

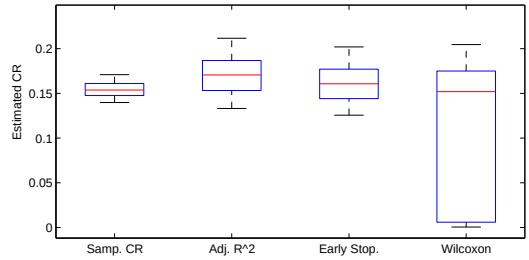
Figure 6.6b shows the box plots of the distribution of estimates given 200 samples. Quick comparison with Figure 6.6a shows that now the variability in the estimates is considerably smaller (maximal standard error of the order of 0.1 compared to 0.16). All methods perform better with larger number of samples providing more accurate estimates. Again, the best version of the polynomial method was early stopping. The Wilcoxon rank sum test starts to exhibit an



(a) 60 samples



(b) 200 samples



(c) 1000 samples

Figure 6.6: *Box plots of the estimates of the correlation ratio.*

undesirable feature — the erratic nature of the p-value causes in some cases to choose 4 or less as the optimal polynomial degree. The value of the correlation ratio is more heavily underestimated then (for example in Figure 6.3), making the box plots look very stretched. Figures 6.6b and 6.6c confirm these observations.

The same model has been used for initial comparison of the Bayesian method with early stopping. The Bayesian method has been implemented in GEM-SA — Gaussian Emulation Machine for Sensitivity Analysis software and we use it in the analysis. 50 sets of samples were generated with 100 samples of X, Y and G per set. The estimates were converted to percentages and compared in this form in Table 6.1. The Bayesian method underestimated the value of $\eta^2(G|X)$

Table 6.1: *Estimates of the correlation ratio - Bayesian and Early stopping methods*

Method	Mean	RMSE
Bayesian	3.78	1.71
Early stop	18.97	10.11

(15.38) with estimates tightly concentrated around value 3.78. Increasing the number of samples to 400 (maximum supported by GEM-SA) did not cure this problem. This suggests that the Bayesian method may have problems with non-normal models and should be further explored. The early stopping algorithm on the other hand produced a more sensible average estimate.

6.7.4 The analytic function of Oakley and O'Hagan

This model for benchmarks has been proposed by Oakley and O'Hagan [2004]. It is a multivariate model with 15 inputs

$$G(\mathbf{X}) = \mathbf{a}_1^T\mathbf{X} + \mathbf{a}_2^T \sin(\mathbf{X}) + \mathbf{a}_3^T \cos(\mathbf{X}) + \mathbf{X}^T\mathbf{M}\mathbf{X}, \tag{6.10}$$

where $\mathbf{X}$ is a vector of independent standard normal random variables. Scalar vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ and matrix $\mathbf{M}$ are chosen such that the importance of the inputs can be classified into 3 categories based on the appropriate values of the correlation ratio. The same model has also been studied in [Ratto et al., 2006].

The full analysis of methods described in section 6.5 is not viable at this moment as the authors of the SDP method could not supply the code with the implementation. Therefore we base our findings on the comments of the authors in [Ratto et al., 2006]. On the other hand, the method of Oakley and O'Hagan [2004] has been implemented in GEM-SA and we use this software in our analysis.

Note that the estimates of the correlation ratio are presented on the percentage scale rather than fractions and all the results are calculated based on percentages.

Oakley and O'Hagan [2004] report that given 250 evaluations of eq.(6.10) at carefully chosen design points for $\mathbf{X}$ the standard deviations for the correlation ratio estimates of $X_1, \ldots, X_5$ is about 0.2, for $X_6, \ldots, X_{10}$ is 0.5 and for $X_{11}, \ldots, X_{15}$ is about 1. Since our method does not require any specific methods of generating the sample we compare it with O'Hagan's method using pseudo random samples produced in MATLAB. This is of course the situation less favorable for the Bayesian method, but it complies with the desiderata declared in section 6.1. The decisive factor when we chose to limit the number of runs to 24 was long execution time of GEM-SA software. Also, out of these 24 runs only 10 distinct vectors of 15 estimates (for each input variable) were returned by GEM-SA. This suggests that the maximum likelihood optimization routine for the hyperparameters of the Bayesian method gets stuck at some fixed points quite often. This may give a misleading picture of the mean and RMSE of the estimates.

Figure 6.7 shows the mean estimates of the correlation ratios for this model based on 24 iterations, 250 samples per variable each. The estimates produced by the Bayesian method are much closer to the true values despite the fact that the

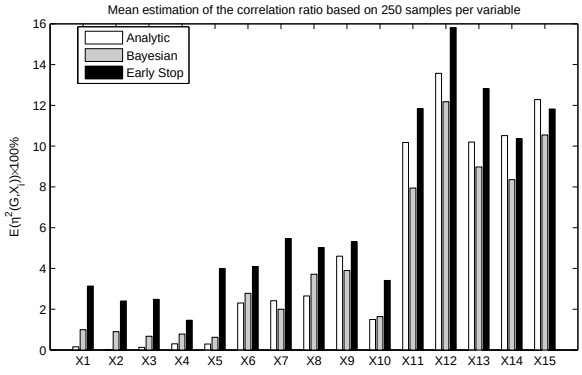


Figure 6.7: *Mean estimates of the correlation ratio based on 24 iterations and 250 samples per variable.*

input sample was not chosen optimally. The early stopping algorithm tends to overestimate the correlation ratios, especially if the true value is close to 0. The power of the prevention of overfitting is limited for a small sample size like this. The RMSE's of the results are also smaller for the Bayesian method (0.5, 1.5 and 2.5 for each of the three groups of input variables respectively) although not on a par with the results reported by Oakley and O'Hagan [2004] if the sample is carefully selected (0.2, 0.5 and 1). In order to achieve a comparable RMSE with the early stopping method the number of samples would have to be increased to about 750 as Table 6.2 shows. This, however, is not enough to have similar mean estimates — for this 1000 samples have to be generated. Overall the early stopping method needs substantially more samples than the Bayesian approach. It will definitely not beat the SDP method either, which seems to perform very well in terms of determining values of the correlation ratios given 1000 samples per variable.

6.8 Conclusions and discussion

There are many ways to quantify sensitivity. We have argued that the correlation ratio $\eta^2(G|X)$ is particularly attractive in this regard, although it cannot always be computed on-the-fly, and may be difficult to compute analytically.

The correlation ratio can be accurately estimated if the regression $\mathbf{E}(G|X)$ of G on X is determined with sufficient accuracy. This paper develops a benchmark for testing candidates for good estimates of $\mathbf{E}(G|X)$. The polynomial method assumes that the underlying model is sufficiently smooth and can be accurately approximated with a polynomial. In order to prevent overfitting we employ three well motivated techniques based on: the adjusted R^2, early stopping algorithm, and Wilcoxon rank sum test. The early stopping method is most resistant to overfitting, has no "tweakable" parameters, is easy to implement and gave the best

Table 6.2: *Mean and standard deviations of the estimates of the correlation ratio obtained with the early stopping method (750 and 1000 samples per variable, 100 iterations).*

	Analytical	Mean		Standard deviation	
		750 samples	1000 samples	750 samples	1000 samples
X_1	0.1560	1.0499	0.7591	0.9011	0.6797
X_2	0.0186	1.0346	0.5218	0.8172	0.4934
X_3	0.1307	1.0311	0.6731	0.7570	0.5236
X_4	0.3045	1.1800	0.9251	1.1149	0.6570
X_5	0.2905	0.9849	0.7772	0.8972	0.6373
X_6	2.3035	2.9734	2.8181	1.2959	0.9932
X_7	2.4151	3.1584	2.9750	1.4744	1.2283
X_8	2.6517	2.8456	3.0997	1.3179	1.3230
X_9	4.6036	5.3172	5.5461	1.7640	1.6890
X_{10}	1.4945	2.0598	2.0152	1.0798	1.1111
X_{11}	10.1823	10.4025	10.7995	2.0375	2.1275
X_{12}	13.5708	13.9139	13.8106	2.3893	2.0873
X_{13}	10.1989	10.0289	10.3519	2.2431	1.9953
X_{14}	10.5169	11.0706	10.4579	2.4762	2.1103
X_{15}	12.2818	12.4564	12.4932	2.3133	2.0299

results. Therefore we used this specific algorithm for further comparison with the Bayesian method. The Bayesian method performed very well on the benchmark model proposed by Oakley and O'Hagan [2004], but experienced difficulties with the model in section 6.7. Without questioning the advantages of Bayesian methods for calculating the correlation ratio, there is a need for a simple generic method that works for a wide variety of models and sample sizes. Polynomial approximations perform decently in this regard, with early stopping as front runner and are very cheap to run when implemented in computer code. Obviously there is a trade-off here between the cost of needing a lot of samples (depends on how expensive the model is to run), and the cost of the algorithm itself. It should be noted that one run of GEM-SA takes 5 minutes to complete one calculation of estimates of $\eta^2(G, X_i)$ for the model described in section 6.7.4 on the current top-of-the-line dual core Intel processor (Intel Core 2 Extreme X6800) with only the option to calculate main effects selected and all the remaining program options set to default.

Polynomial approximation methods can also be extended for estimation of joint effects of 2 or more random variables on the output. One dimensional polynomial functions would simply be replaced by their multidimensional counterparts. Possible future research can look more into the robustness of various methods of estimating the correlation ratio for different models as the choice of benchmark models mattered quite a lot in this study.

CHAPTER 7

Conclusions

> The point of quotations is that one can use another's words to be insulting.
>
> *Amanda Cross*

This thesis aims at approaching the problem of statistical dependence modelling from many different angles to show the complexity of the issue and show ways of dealing with it. Chapter 2 forms the point of reference for the remaining papers incorporated into this thesis. It describes the standard tools for modelling high dimensional data with some parametric families of multidimensional copulae. We also study various dependence concepts and measures expressing interactions between random variables in a quantitative way. Pearson's product moment and Spearman's rank correlations in their unconditional and conditional forms are among the best known concepts of dependence. However, they capture only linear (product moment correlation) or monotonic (rank correlation) dependence between random variables. This often is not satisfactory, as more complicated dependence structures can be observed and these must be properly modelled as well. Tail dependence concepts allow for more accurate modelling of tails of multivariate distributions and this is crucial in applications to financial and insurance markets, where risks are found to be extremes of analyzed distributions. This is the reason for the increasing popularity of tail dependent distributions, like Student's t, Clayton or Gumbel, in actuarial science. We also introduce an entirely new Dirichlet-type copula. It has been constructed without applying Sklar's theorem as it is a special case of the generalized Dirichlet distribution. Unfortunately the correlation structure of this copula is fixed and depends on its dimension only. This limits the number of possible applications significantly. The above mentioned distributions are examples of families of multivariate distributions. However, another way of constructing multivariate distributions is to couple bivariate pieces in a systemized manner, and a tool for this is the vine-copula method.

The dependence vine is build on the concept of dependence tree. Here conditional independence statements existing implicitly in a dependence tree, have been replaced with conditional dependence statements, quantified with conditional rank correlations and modelled with conditional bivariate copulae. Dependence modelling with vines requires a copula for which the conditional and inverse conditional cumulative distribution functions can be efficiently computed, and preferably are given in a closed form. Chapter 3 shows some examples of copulae satisfying this requirement. They are members of a broad class of copulae called generalized diagonal band (GDB) copulae. This class is a product of a very intuitive geometrical method of construction. The whole copula density is generated from the density put on one of the boundaries of the copula domain. Future work should concentrate on studying links between this generating density and the properties of the resulting copula. We also extend and in some places correct the work of Meeuwissen [1993] on a subclass of GDB copula, that is obtainable through mixing of ordinary diagonal band copulae.

Chapter 4 of the thesis describes the use of the DAD algorithm to construct discretized minimally informative copulae with respect to the independent copula given some moment constraints. It extends the minimally informative copula developed by Bedford and Meeuwissen [1997]. Their copula has been constructed with just one constraint, namely $\mathbf{E}[XY]$, where X and Y are uniformly distributed on interval $[-\frac{1}{2}, \frac{1}{2}]$. In order to simplify calculations the copula itself is also defined on a square $[-\frac{1}{2}, \frac{1}{2}]^2$ and its density has the form

$$f(x, y) = \kappa(x)\kappa(y) \exp(\lambda xy).$$

From the set up of this copula we see that it must be centrally symmetric. Since the exponent term is a symmetric function, it follows that the product of the kappa functions is a symmetric function too. Hence κ is an even function on $[-\frac{1}{2}, \frac{1}{2}]$. The D_1AD_2 approach presented in chapter 4 generalizes this copula. Vectors D_1 and D_2 are simply discretized counterparts of the kappa function in the 2-dimensional case. The algorithm for determining these vectors is extremely simple and consists of projecting an initial density for the copula on each margin successively to impose uniform marginals for the final copula density. The resulting minimally informative copula with respect to the independent copula under the given moment constraints that has been fit to the World Bank data produces a good overall fit to the data, and realizes the lowest possible level of information. Frank's copula with the maximum likelihood parameter estimate achieved higher likelihood (at the expense of higher information).

The use of the minimum information principle makes the D_1AD_2 approach attractive for expert elicitation applications. Experts are asked their opinion on expectations of some functions of variables of interest and this is translated into a minimally informative copula density given the assessments. We show an example of running such a procedure in which a 3-dimensional discretized minimally informative copula is being constructed.

The next chapter departures from copula modelling and concentrates on another application of vines — generating random correlation matrices of size $d \times d$ from the joint density of all correlation matrices of the same size. The matrices

can be drawn from the joint density being proportional to a power of the determinant of the correlation matrix. The uniform distribution is a special case. The idea was introduced by Joe [2006] and was based on the D-vine. The method however is not limited to the use of this one type of copula. We argued that the C-vine is less computationally demanding and can successfully applied as well. In fact, we extend the method to be applicable to any regular vine. This brings new applications of this method of generating random correlation matrices. For instance, we can generate correlation matrices conditional on correlation values in an arbitrary tree. The Onion method proves to be very efficient computationally, however in some setups the C-vine method shows better performance. The Onion method has also been extended to allow generating random correlation matrices non-uniformly from the set of semi-positive definite correlation matrices.

An essential step in statistical modelling is sensitivity analysis and chapter 6 is dedicated to this subject. The chapter concentrates on the notion of correlation ratio, a variance based global sensitivity measure. We show some properties of the correlation ratio and its links to other concepts used in sensitivity analysis, namely Sobol' indices, high dimensional model representations (HDMR) and state dependent parameter models (SDP). Calculations of the correlation ratio can be very tedious and quite often analytical solutions do not exist. Therefore we concentrated our efforts on developing a numerical method of estimating this quantity based on samples. We estimate the regression curve via a simple least-squares error fit of a polynomial. However there are two dangers in doing so without any control mechanism. Fitting a polynomial of too low degree may result in a very bad fit, which does not correspond well to the true regression curve. On the other hand, a polynomial of a very high degree exhibits a very good fit to this specific sample, but cannot be seen as a good estimator of the regression for the whole population. Therefore we introduced an overfitting prevention method to overcome this problem. Three different criteria have been tested for detecting the overfitting and the best performing algorithm is based on an early stopping approach. The whole method of estimating the correlation ratio from a sample is very easy to implement and performs well even with moderate sample size.

References

B. Abdous, C. Genest, and B. Rémillard. *Statistical Modeling and Analysis for Complex Data Problems*, chapter Dependence properties of meta-elliptical distributions, pages 1–15. Kluwer, Dordrecht, The Netherlands, 2005.

M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. National Bureau of Standards, US Government Printing Office, Washington, DC., 1964.

S. Andreassen, K. Olesen, F. Jensen, and F. Jensen. HUGIN: a shell for building Bayesian belief universes for expert systems. In *Proceedings of Eleventh International Joint Conference on Artificial Intelligence*, pages 1080–1085, 1989.

K. Baba, R. Shibata, and M. Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian and New Zealand Journal of Statistics*, 46(4):657–664, December 2004.

T. Bedford. Sensitivity indices for (tree)-dependent variables. In K. Chan, S. Tarantola, and F. Campolongo, editors, *SAMO'98, Proceedings of Second International Symposium on Sensitivity Analysis of Model Output*, pages 17–20, 1998.

T. Bedford and R. Cooke. Vines — a new graphical model for dependent random variables. *Annals of Statistics*, 30(4):1031–1068, 2002.

T. Bedford and A. Meeuwissen. Minimally informative distributions with given rank correlation for use in uncertainty analysis. *Journal of Statistical Computation and Simulation*, 57(1–4):143–174, 1997.

J. Bojarski. A new class of band copulas — distributions with uniform marginals. Technical Publication, Institute of Mathematics, Technical University of Zielona Góra, 2001.

J. Borwein, A. Lewis, and R. Nussbaum. Entropy minimization, DAD problems, and doubly stochastic kernels. *Journal of Functional Analysis*, 123:264–307, 1994.

K. Chan, A. Saltelli, and S. Tarantola. Sensitivity analysis of model output: variance-based methods make the difference. In *Proceedings of the 29th conference on Winter simulation*, pages 261–268, 1997.

K. Chan, A. Saltelli, and S. Tarantola. Winding Stairs: A sampling tool to compute sensitivity indices. *Statistics and Computing*, 10:187–196, 2000a.

K. Chan, S. Tarantola, A. Saltelli, and I. Sobol. *Sensitivity Analysis*, chapter Variance-Based Methods, pages 167–197. Wiley, 2000b.

R. Clemen, G. Fisher, and R. Winkler. Assessing dependence: Some experimental results. *Management Science*, 46(8):1100–1115, 2000.

R. Cooke. *Experts in Uncertainty*. Oxford University Press, 1991.

R. Cooke. Markov and entropy properties of tree- and vine-dependent variables, 1997. Proceedings of the ASA Section on Bayesian Statistical Science.

R. Cooke and L. Goossens. Procedures guide for structured expert judgement in accident consequence modelling. *Radiation Protection and Dosimetry*, 90(3): 303–311, 2000.

R. Cooke and D. Lewandowski. Bayesian sensitivity analysis. In Y. Hayakawa, T. Irony, and M. Xin, editors, *System and Bayesian Reliability — Essays in Honor of Professor Richard E. Barlow on His 70^{th} Birthday*, volume 5, pages 315–331. World Scientific, 2001.

R. Cooke and R. Waij. Monte Carlo sampling for generalized knowledge dependence, with application to human reliability. *Risk Analysis*, 6(3):335–343, 1986.

I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.

S. Demarta and A. J. McNeil. The t copula and related copulas. *International Statistical Review*, 73(1):111–129, April 2005.

N. Draper and H. Smith. *Applied regression analysis*. Wiley series in probability and statistics. John Wiley and Sons, Inc., 1998.

R. J. Duintjer Tebbens, K. M. Thompson, M. G. M. Hunink, T. M. Mazzuchi, D. Lewandowski, D. Kurowicka, and R. M. Cooke. Uncertainty and sensitivity analyses of a dynamic economic evaluation model for vaccination programs. *Medical Decision Making*, 28(2):182–200, 2008.

P. Embrechts, A. McNeil, and D. Straumann. *Risk Management: Value at Risk and Beyond*, chapter Correlation and dependence in risk management: Properties and pitfalls, pages 176–223. Cambridge: Cambridge University Press, 2002.

H.-B. Fang, K.-T. Fang, and S. Kotz. The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82(1):1–16, July 2002.

T. Ferguson. A class of bivariate uniform distributions. *Statistical Papers*, 36: 31–40, 1995.

M. Frank. On the simultanous associativity of $f(x, y)$ and $x - y - f(x, y)$. *Aequationes Math.*, 19:194:226, 1979.

C. Genest. Frank's family of bivariate distributions. *Biometrika*, 74:549–555, 1987.

C. Genest and J. MacKay. The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283, November 1986.

C. Genest and L. Rivet. Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, 88:1034–1043, 1993.

S. Ghosh and S. G. Henderson. Behavior of the NORTA method for correlated random vector generation as the dimension increases. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13(3):276–294, July 2003.

R. Gupta, J. Misiewicz, and D. Richards. Infinite sequences with sign–symmetric Liouville distributions. *Probability and Mathematical Statistics*, 16(1):29–44, 1996.

P. Gustafson and L. J. Walker. An extension of the Dirichlet prior for the analysis of longitudinal multinomial data. *Journal of Applied Statistics*, 30:293–310, 2003.

J. Hodges and E. Lehmann. *Basic Concepts of Probability and Statistics*. Holden-Day, 1970.

W. Hoeffding. Masstabinvariante korrelationstheorie. *Schriften des Mathematischen Instituts un des Instituts für Andewandre Mathematic der Universität Berlin*, 5:179–233, 1940.

R. Iman and W. Conover. A distribution-free approach to inducing rank correlation amoung input variables. *Communications in Statistics — Simulation and Computation*, 11(3):311–334, 1982.

T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis for computer models. In *Proceedings of the ISUMA '90 First International Symposium on Uncertianty Modelling and Analysis*, pages 398–403, 1990.

F. Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York, 2001.

H. Joe. *Multivariate Models and Dependence Concepts*. Chapman & Hall, London, 1997.

H. Joe. Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97:2177–2189, 2006.

M. Kendall and A. Stuart. *The Advanced Theory of Statistics — Volume 2, Inference and Relationship*. Charles Griffin and Company Limited, London, 1961.

D. G. Kleinbaum, L. L. Kupper, K. E. Muller, and A. Nizam. *Applied Regression Analysis and Multivariable Methods*. An Alexander Kugushev book. Brooks/Cole Publishing Company, 1998.

S. Kotz and J. R. van Dorp. *Beyond Beta, Other Continuous Families of Distributions with Bounded Support and Applications*. World Scientific Press, Singapore, September 2004.

B. Kraan. *Probabilistic Inversion in Uncertainty Analysis and related topics*. Phd thesis, isbn 90-9015710-7, TU Delft, 2002.

B. Kraan and T. Bedford. Probabilistic inversion of expert judgements in the quantification of model uncertainty. Research Paper No. 2003/12, 2003.

D. Kurowicka and R. Cooke. *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley, 2006a.

D. Kurowicka and R. M. Cooke. Completion problem with partial correlation vines. *Linear Algebra and Its Applications*, 418:188–200, 2006b.

D. Kurowicka, J. Misiewicz, and R. Cooke. *Elliptical Copulae*, pages 209–214. Monte Carlo Simulation. Balkema, Rotterdam, 2001.

J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112–147, 1998.

G. Li, S.-W. Wang, and H. Rabitz. Practical appraoches to construct RS-HDMR component functions. *Journal of Physical Chemistry*, 106(37):8721–8733, 2002.

G. Li, J. Hi, S.-W. Wang, P. G. Georgopoulos, J. Schoendorf, and H. Rabitz. Random sampling-high dimesional model representations (RS-HDMR) and orthogonality of its different order component functions. *Journal of Physical Chemistry*, 110(7):2474–2485, 2006.

D. D. Mari and S. Kotz. *Correlation and Dependence*. Imperial College Press, London, UK, 2001.

M. McKay. Nonparametric variance-based methods of assessing uncertainty importance. *Reliability Engineering and System Safety*, 57:267–279, 1997.

A. M. Meeuwissen. *Dependent Random Variables in Uncertainty Analysis*. PhD thesis, Delft University of Technology, Delft, The Netherlands, 1993.

O. Morales-Napoles, D. Kurowicka, and A. Roelen. Eliciting conditional and unconditional rank correlations from conditional probabilities. *Reliability Engineering & System Safety*, 93:775–777, March 2007.

R. B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer New York, 2 edition, 2007.

R. B. Nelsen. Properties of a one-parameter family of bivariate distributions with specified marginals. *Communications in Statistics, Theory and Methods*, 15: 3277–3285, 1986.

W. Nordhaus. A question of balance: Economic modeling of global warming. To be published by Yale University Press in 2008, 2008.

J. E. Oakley and A. O'Hagan. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society*, 66:751–769, 2004.

K. Pearson. Mathematical contributions to the theory of evolution — on homotyposis in homologous but differentiated organs. *Proceedings of the Royal Society of London*, 71:288–313, 1903.

M. Ratto, A. Saltelli, S. Tarantola, and P. Yound. Improved and accelerated sensitivity analysis using state dependent parameter models. Eur 22251 en, Joint Research Centre, European Commission, 2006.

A. Saltelli, S. Tarantola, and K. Chan. A quantitative, model independent method for global sensitivity analysis of model output. *Technometrics*, 41:39–56, 1999.

A. Saltelli, K. Chan, and E. M. Scott, editors. *Sensitivity Analysis*. Wiley, 2000a.

A. Saltelli, S. Tarantola, and F. Campolongo. Sensitivity analysis as an ingredient of modeling. *Statistical Science*, 15(4):377–395, 2000b.

A. Sklar. Fonctions de répartiton ă n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

M. Smith. Modelling sample selection using Archimedean copulas. *The Econometrics Journal*, 6(1):99–123, June 2003.

I. Sobol'. Sensitivity analysis for nonlinear mathematical models. *Mathematical Modelling and Computational Experiment*, 1:407–414, 1993.

G. G. Venter. Tails of copulas. In *Proceedings of the Casualty Actuarial Society Casualty Actuarial Society*, pages 68–113, 2002.

P. Whittle. *Probability Via Expectation*. Springer Texts in Statistics. Springer, 1992.

G. Yule and M. Kendall. *An introduction to the theory of statistics*. Charles Griffin & Co, Belmont, California., 14 edition, 1965.

APPENDIX A

More examples of GDB copulae

We present here more examples of GDB copulae generated by various distributions defined on the interval $[0, 1]$ and a formulation of the relative information of the GDB copula in terms of its generating function.

A.1 Beta distribution as the generating function

Let the generating function g be in the class of beta distributions.

Definition A.1.1. *Random variable X is beta distributed with parameters s and q (denoted $Beta(q, s)$) if its probability density function has the form*

$$f(x) = \frac{x^{q-1}(1-x)^{s-1}}{B(q,s)}. \tag{A.1}$$

Explicit formula for the parameters of the beta distributions, q and s, as a function of the product moment correlation ρ are needed. We have

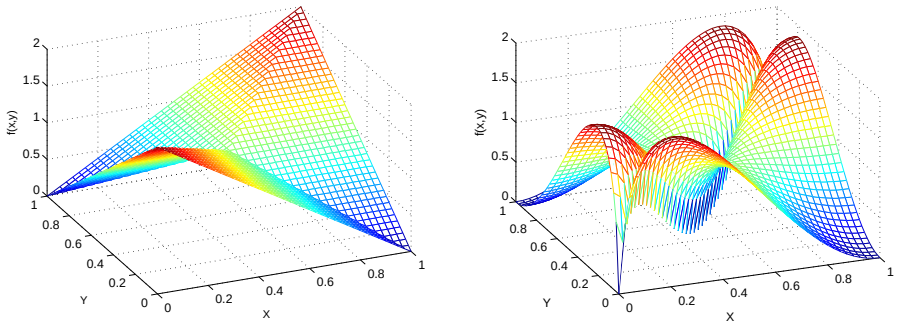
$$\begin{aligned} \mathbf{E}(X^2) &= \frac{q\,(q+1)}{(q+s)\,(1+q+s)}, \\ \mathbf{E}(X^3) &= \frac{q\,(q+1)\,(q+2)}{(q+s)\,(1+q+s)\,(s+q+2)}, \end{aligned}$$

and thus

$$\rho = 1 - 2\,\frac{q\,\left(3\,q\,s + q^2 + 3\,q + 3\,s + 2\right)}{(q+s)\,(1+q+s)\,(s+q+2)}.$$

This problem can be solved analytically and the solution is given below

$$\begin{aligned} s &= s, \\ q &= \frac{1}{3(\rho+1)}\,[H(s,\rho)]^{\frac{1}{3}} + (6s^2 + 6s + \rho + 1)\,[H(s,\rho)]^{-\frac{1}{3}} - s - 1, \end{aligned}$$



(a) $Beta(1, 2)$ **(b)** $Beta(1.5506, 3)$

Figure A.1: *GDB copulae realizing correlation* $\rho = 0.4$ *generated by the beta distributions with different parameters.*

where

$$\begin{aligned} H(s,\rho) \quad = \quad & -(\rho+1)^2\left[27s+81s^2+54s^3\right]- \\ & -(\rho+1)3\sqrt{3}\left[1+18s+99s^2+270s^3+405s^4+324s^5+108s^6+\right. \\ & +\rho\left(3+36s+117s^2+54s^3-243s^4-324s^5-108s^6\right)+ \\ & \left.+\rho^2\left(3+18s+18s^2\right)-\rho^3\right]. \end{aligned}$$

For any $\rho \in (-1, 1)$ there exist an entire family of beta distributions generating a GDB copula with this given correlation. The optimal choice of parameters q and s should allow to generate a copula with minimal relative information.

Remark. By the construction of the GDB copula and the properties of the beta distribution one can notice, that if the beta distribution with parameters $q = a$ and $s = b$ generates a copula with correlation ρ, then the beta distribution with parameters $q = b$ and $s = a$ generates a copula with correlation $-\rho$.

A.2 Distribution based on cosine function as the generating function

Let the generating function $g(x)$ with parameters q and p be given by the formula

$$g(x) = p + 1_{x\in[0,\frac{1}{q}]}(1-p)q\left(\cos(q\pi x)+1\right). \tag{A.2}$$

where $p \in [0, 1]$ and $q \geq 1$. This generating function ensures smoothness of the generated copula along the diagonals, and hence, lower relative information compare to the truncated exponential density function for instance. It can be shown that the relationship between the correlation realized by this copula and

its parameters is given as follows

$$\rho = \frac{1-p}{q^3}\left(q^3 + 2\,\frac{q\left(6-\pi^2\right)}{\pi^2} + 12\,\frac{4-\pi^2}{\pi^4} + 1\right)$$

and solving this formula for q allows to find the analytical solution

$$\begin{aligned} q &= q, \\ p &= 1 - \rho q^3\left(q^3 + 2\,\frac{q\left(6-\pi^2\right)}{\pi^2} + 12\,\frac{4-\pi^2}{\pi^4} + 1\right)^{-1}, \end{aligned}$$

One can notice, that for lower correlations $q = 1$ and for higher correlations $p = 0$ and there is only one value of ρ for which $q = 1$ and $p = 0$ (this is for $\rho \approx 0.4927$). Hence in fact, the generating function (A.2) can be controlled only by one parameter at the time simplifying further calculations.

Assume that $p = 1$ first. Then

$$g(x) = p + (1-p)\left(\cos(\pi x) + 1\right).$$

Employing eq.(3.8) allows to find the bivariate cdf for $x \le y$, $x + y \le 1$ and its derivatives

$$\begin{aligned} F(x,y) &= \frac{x\pi^2 y + (1-p)\sin(y\pi)\sin(x\pi)}{\pi^2}, \\ f(x,y) &= 1 + (1-p)\cos(y\pi)\cos(x\pi), \\ F_{Y|X}(y) &= \frac{y\pi + (1-p)\sin(y\pi)\cos(x\pi)}{\pi}. \end{aligned}$$

Note that the density $f(x,y)$ is given by the same formula everywhere on the unit square, not only for $x \le y$, $x + y \le 1$. The same holds for the conditional cumulative distribution function $F_{Y|X}(y)$.

Now assume $p = 0$ and $q \ge 1$. Then for $y \le x$ and $x + y \le 1$ we have

$$f(x,y) = \begin{cases} q\left(1 + \cos(xp\pi)\cos(y\,q\pi)\right), & \text{if } y \le -x + \frac{1}{q}; \\ \frac{q(1+\cos(xp\pi)\cos(y\,q\pi)+\sin(xq\pi)\sin(y\,q\pi))}{2}, & \text{if } y > -x + \frac{1}{q} \text{ and } y > x - \frac{1}{q}; \\ 0, & \text{elsewhere.} \end{cases}$$

Then the conditional cdf for any $x, y \in [0,1]$ is given as the following

$$F_{Y|X}(y) = \begin{cases} 0, & y < x - 1/q; \\ \frac{yq\pi+\cos(xq\pi)\sin(yq\pi)}{\pi}, & x \le \frac{1}{q} \text{ and } y \le -x + \frac{1}{q}; \\ \frac{\pi(yq-q+1)+\cos(xq\pi)\sin(yq\pi)+\frac{1}{2}\sin(2q\pi)}{\pi}, & x > 1 - \frac{1}{q} \text{ and } y > -x + 2 - \frac{1}{q}; \\ 1, & x < 1 - \frac{1}{q} \text{ and } y > x + \frac{1}{q}; \\ \frac{yq\pi-xq\pi+\pi+\sin(q\pi(y-x))}{2\pi}, & \text{elsewhere.} \end{cases}$$

Unfortunately, the conditional cdf's are not analytically invertible.

A.3 Relative information of the GDB copula in terms of its generating function

The following corollary allows to express the relative information of a GDB copula in terms of its generating density g.

Corollary A.3.1. *The relative information of a GDB copula with density $f(x,y)$ generated by the generating function g with respect to the uniform distribution is*

$$I(f|u) = \int_0^1 g(v)\log(g(v)+g(1))dv - \int_0^1\int_0^1 \frac{\frac{dg(t)}{dt}tg(v)}{g(v)+g(t)}dtdv - \log(2).$$

Proof. Since the GDB copula is symmetric along both diagonals, we can calculate the relative information only for one of the regions bounded by the diagonals and multiply it by 4

$$\begin{aligned} I(f|u) &= 4\int_0^{1/2}\int_y^{1-y} f(x,y)\log(f(x,y))dxdy \\ &= 2\int_0^{1/2}\int_y^{1-y} (g(x+y)+g(x-y))\left[\log\left(g(x+y)+g(x-y)\right)-\log(2)\right]dxdy \\ &= 2\int_0^{1/2}\int_y^{1-y} (g(x+y)+g(x-y))\log\left(g(x+y)+g(x-y)\right)dxdy - \log(2). \end{aligned}$$

The last equality follows from the fact that

$$4\int_0^{1/2}\int_y^{1-y} f(x,y)dxdy = 2\int_0^{1/2}\int_y^{1-y} g(x+y)+g(x-y)dxdy = 1.$$

Let $x = \frac{1}{2}(t+v)$, $y = \frac{1}{2}(-t+v)$. Then the Jacobian is $1/2$ and

$$\begin{aligned} I(f|u) &= \int_0^1\int_0^v (g(v)+g(t))\log\left(g(v)+g(t)\right)dtdv - \log(2) \\ &= I - \log(2). \\ I &= \int_0^1\int_0^v (g(v)+g(t))\log\left(g(v)+g(t)\right)dtdv \\ &= \int_0^1 g(v)\int_0^v \log\left(g(v)+g(t)\right)dtdv + \int_0^1 g(t)\int_t^1 \log\left(g(v)+g(t)\right)dvdt \\ &= \int_0^1 g(v)\int_0^v \log\left(g(v)+g(t)\right)dtdv + \int_0^1 g(v)\int_v^1 \log\left(g(v)+g(t)\right)dtdv \\ &= \int_0^1 g(v)\int_0^1 \log\left(g(v)+g(t)\right)dtdv. \end{aligned}$$

Integrating $\int_0^1 \log\left(g(v)+g(t)\right)dt$ gives

$$\int_0^1 \log\left(g(v)+g(t)\right)dt = log(g(v)+g(t))|_0^1 - \int_0^1 \frac{\frac{dg(t)}{dt}tg(v)}{g(v)+g(t)}dt.$$

Hence

$$I(f|u) = \int_0^1 g(v)\log(g(v)+g(1))dv - \int_0^1\int_0^1 \frac{\frac{dg(t)}{dt}tg(v)}{g(v)+g(t)}dtdv - \log(2).$$

■

APPENDIX B

Mixtures of diagonal band copulae with discontinuous mixing measures

Mixtures of diagonal band copulae described in chapter 3.3 allow for only one point with discrete mass, that is at the origin. We extend this class of mixtures to include discrete mass at any finite set of points in the interval $[-1, 1]$.

B.1 Introduction

Let us define a more general mixing measure first

Definition B.1.1. *Let $0 \leq p \leq 1$. A probability distribution $M(\theta)$, $M(\theta) : [-1, 1] \to [0, 1]$ is called a mixing measure if its derivative with respect to θ consists of an absolutely continuous part $m(\theta) \geq 0$ with $\int_{-1}^{1} m(\theta)\, d\theta = p$ and a discrete part with mass $p_i > 0$ at θ_i, $-1 \leq \theta_{i-1} \leq \theta_i \leq 1$, $\cup_i \{\theta_i\} = \mathcal{A}$ and $\sum_i p_i = 1 - p$. $\mathcal{A}$ may be empty.*

The mixture of diagonal band copulae is defined then as in Definition 3.3.1. For any two mixing measures M_1 and M_2 and any $\lambda \in [0, 1]$, $\lambda M_1 + (1 - \lambda)M_2$ is a mixing measure as well, and

$$c_{\lambda M_1 + (1-\lambda) M_2} = \lambda c_{M_1}(x, y) + (1 - \lambda) c_{M_2}(x, y).$$

The mixing measure $M(\theta)$ can be determined as follows. By Definition 3.3.1 a conditional density $c_M(x, 0)$ is a mixture of diagonal bands, thus $c_M(x, 0) = \int_{-1}^{1} d_\theta(x, 0)\, dM(\theta)$. Let us rewrite it as follows

$$\begin{aligned} c_M(x, 0) &= \int_{-1}^{0} \frac{\mathbf{1}_{\{\theta \in [-x, 0]\}}}{1 + \theta}\, dM(\theta) + \int_{0}^{1} \frac{\mathbf{1}_{\{\theta \in (0, 1-x]\}}}{1 - \theta}\, dM(\theta) \\ &= \int_{-x}^{0} \frac{1}{1 + \theta}\, dM(\theta) + \int_{0}^{1-x} \frac{1}{1 - \theta}\, dM(\theta). \end{aligned} \tag{B.1}$$

Note that

$$\int_{-x}^{0} \frac{1}{1+\theta}\, dM(\theta) \geq 0 \tag{B.2}$$

is a nondecreasing, nonegative function of x, whereas

$$\int_{0}^{1-x} \frac{1}{1-\theta}\, dM(\theta) \geq 0 \tag{B.3}$$

is a nonincreasing, nonnegative function of x. Now we shall decompose density $c_M(x,0)$ into two nonnegative components, $g^+(x)$ and $g^-(x)$ that we could relate to (B.2) and (B.3) respectively. We introduce two functions $g^+(x)$ and $g^-(x)$ differentiable almost everywhere, with the derivatives with respect to x defined in (B.4) and (B.5), and set $c_M(x,0) = g(x) = g^+(x) + g(0) - g^-(x)$.

$$\frac{d}{dx}\, g^+(x) = \max\left\{ \frac{d}{dx}\, g(x), 0 \right\}, \;\; g^+(0) = 0 \tag{B.4}$$

$$\frac{d}{dx}\, g^-(x) = \max\left\{ -\frac{d}{dx}\, g(x), 0 \right\}, \;\; g^-(0) = 0 \tag{B.5}$$

The nondecreasing component $g^+(x)$ of the conditional density $c_M(x,0)$ corresponds to mixing step functions given in (3.10). Similarly, the nonincreasing component $g(0,0) - g^-(x)$ corresponds to mixing step functions given in (3.11).

B.2 Determining the continuous part of the mixing measure

The continuous part $m(\theta)$ of the mixing measure M can be determined by differentiating eq.(B.1)

$$\begin{aligned} \frac{d}{dx}\, g(x) &= \frac{d}{dx}\, g^+(x) + \frac{d}{dx}\, (g(0) - g^-(x)) = \frac{d}{dx}\, g^+(x) - \frac{d}{dx}\, g^-(x) \\ &= \frac{m(-x)}{1-x} - \frac{m(1-x)}{x}. \end{aligned}$$

The last equality emerges from substituting $dM(\theta)$ with $m(\theta)d\theta$ in eq.(B.1) and noticing that $\theta = x$ if $\theta < 0$ and $\theta = 1 - x$ when $\theta > 0$. By the construction of eq.(B.4) and eq.(B.5) there is no $x \in [0,1]$ such that $\frac{d}{dx}\, g^-(x)$ are $\frac{d}{dx}\, g^+(x)$ are both nonzero. This implies that the step functions (3.11) do not contribute to the mixture at point x if $\frac{d}{dx}\, g^+(x) = 0$, thus $m(-x) = 0$. Similar reasoning justifies setting $m(1-x) = 0$ where $\frac{d}{dx}\, g^-(x) = 0$. Eventually we have

$$\begin{aligned} \frac{d}{dx}\, g^+(x) = \frac{m(-x)}{1-x} \quad &\Longrightarrow \quad m(\theta) = (1+\theta)\frac{d}{dx}\, g^+(-\theta), \;\; \theta < 0, \\ -\frac{d}{dx}\, g^-(x) = -\frac{m(1-x)}{x} \quad &\Longrightarrow \quad m(\theta) = (1-\theta)\frac{d}{dx}\, g^-(1-\theta), \;\; \theta > 0. \end{aligned}$$

B.3 Determining the discontinuous part of the mixing measure

The discontinuities of $g(x)$ correspond to the discontinuities of M in the following way. Let x_i, $i = 1, 2, \ldots, n$, be the locations of discontinuities of $g(x)$. Let A and B be two disjoint subsets of the set of indices $i \in \{1, 2, \ldots, n\}$ such that

- if $g(x_i^+) > g(x_i^-)$ then $i \in A$,
- if $g(x_i^+) < g(x_i^-)$ then $i \in B$,

where $g(x_i^+) = \lim_{x \to x_i^+} g(x)$ is the limit from above, and $g(x_i^-) = \lim_{x \to x_i^-} g(x)$ is the limit from below. The following holds then

Proposition B.3.1. *M has a jump of size p_i at $-x_i$ if $i \in A$, or at $1 - x_i$ if $i \in B$.*

$$p_i = \begin{cases} (1 - x_i)\left(g(x_i^+) - g(x_i^-)\right), & \text{if } i \in A, \\ -x_i\left(g(x_i^+) - g(x_i^-)\right), & \text{if } i \in B, \end{cases}$$

Proof. A single diagonal band copula is also a mixture of diagonal band copulae with a discrete mixing measure assigning weight 1 to the parameter of that copula. Hence if we consider a conditional density $d_\theta(x, 0)$ of a diagonal band density $d_\theta(x, y)$, where $\theta > 0$, as a conditional density of a mixture of diagonal bands, then the jump p of the mixing measure M at point θ is 1. In order to reflect the fact that p_1 depends on the difference of the both limits of $d_\theta(x, 0)$ at $x = 1 - \theta$, we assume that

$$p_1 = a\left(d_\theta(x^+, 0) - d_\theta(x^-, 0)\right)$$

where a is a monotonic, real function of x. We determine a with the following calculations

$$1 = a\left(d_\theta(x^+, 0) - d_\theta(x^-, 0)\right) = -\frac{a}{1 - \theta} = -\frac{a}{x}$$

Hence $a = -x$. Similar reasoning holds for determining $a = 1 - x$ when $\theta < 0$. ■

Combining both the information on the continuous and the discontinuous part of the mixing measure full expressions for $g^+(x)$ and $g^-(x)$ can be determined

$$\begin{aligned} g^+(x) &= \int_0^x \frac{d}{ds} g^+(s)\, ds + \sum_{i \in A,\, x_i \le x} \left(g(x_i^+) - g(x_i^-)\right), \\ g^-(x) &= \int_0^x \frac{d}{ds} g^-(s)\, ds - \sum_{i \in B,\, x_i \le x} \left(g(x_i^+) - g(x_i^-)\right). \end{aligned}$$

B.4 Formulation of the theorem

We already know that mixtures of diagonal band copulae are in the class of GDB copulae. In this chapter we show what conditions have to be imposed on the generating density g of the GDB copula to allow this copula to be represented as a mixture of diagonal band copulae. We do this by showing that for a generating

density g satisfying the conditions specified in Theorem B.4.2 below, there exist a mixing measure $M(\theta)$, such that

$$g(u) = \int_{-1}^{1} d_\theta(u, 0)\, dM(\theta).$$

It follows from the fact that if the conditional density is a mixture of diagonal bands, than the entire density over the unit square is a mixture of diagonal bands as well. We introduce the following lemma first

Lemma B.4.1. *For any bounded function $g : [a, b] \to \mathbb{R}$ with finite number n of discontinuities at x_i, $i = 1, 2, \ldots, n$,*

$$\sum_{i=1}^{n} (b - x_i) \left(g(x_i^+) - g(x_i^-) \right) = \int_a^b \sum_{x_i \le x} \left(g(x_i^+) - g(x_i^-) \right) dx. \tag{B.6}$$

Proof. The right hand side of eq.(B.6) can be expressed as

$$\int_a^b \sum_{x_i \le x} \left(g(x_i^+) - g(x_i^-) \right) dx = \sum_{i=1}^{n} \int_{x_i}^b \left(g(x_i^+) - g(x_i^-) \right) dx = \sum_{i=1}^{n} (b - x_i) \left(g(x_i^+) - g(x_i^-) \right).$$

■

We formulate the main theorem.

Theorem B.4.2. *Let $c(x, y)$ be the density of a generalized diagonal band copula generated with generating density $g(u)$, $u \in [0, 1]$. If g is bounded on $[0, 1]$, has finite number of discontinuities and*

$$g(0) - g^-(1) \ge 0 \tag{B.7}$$

then $c(x, y)$ is a density of a mixture of diagonal bands.

Proof. Start with constructing the mixing measure $M(\theta)$

$$\begin{aligned}
\int_{-1}^{1} m(\theta)\, d\theta + \sum_{i=1}^{n} p_i &= \int_0^1 m(-x)\, dx + \int_0^1 m(1 - x)\, dx + \sum_{i=1}^{n} p_i \\
&= \int_0^1 (1 - x) \frac{d}{dx} g^+(x)\, dx + \int_0^1 x \frac{d}{dx} g^-(x)\, dx + \\
&\quad + \sum_{i \in A} (1 - x_i) \left(g(x_i^+) - g(x_i^-) \right) - \sum_{i \in B} x_i \left(g(x_i^+) - g(x_i^-) \right) \\
&= \int_0^1 \frac{d}{dx} g^+(x)\, dx - \int_0^1 x \left(\frac{d}{dx} g^+(x) - \frac{d}{dx} g^-(x) \right) dx + \\
&\quad + \sum_{i \in A} \left(g(x_i^+) - g(x_i^-) \right) - \sum_{i=1}^{n} x_i \left(g(x_i^+) - g(x_i^-) \right) \\
&= g^+(1) - \left(\int_0^1 x \frac{d}{dx} g(x)\, dx + \sum_{i=1}^{n} x_i \left(g(x_i^+) - g(x_i^-) \right) \right)
\end{aligned}$$

Solving $\int_0^1 x \frac{d}{dx} g(x)\, dx$ by parts gives

$$\begin{aligned}
&\int_0^1 x \frac{d}{dx} g(x)\, dx = \\
&= \left\{ \begin{array}{ll} u = x & v = g(x) - \sum_{x_i \le x} \left(g(x_i^+) - g(x_i^-)\right) - G(0) \\ v' = \frac{d}{dx} g(x) & u' = 1 \end{array} \right\} = \\
&= g(1) - \sum_{i=1}^{n} \left(g(x_i^+) - g(x_i^-)\right) - g(0) - \\
&\quad - \int_0^1 \left[g(x) - \sum_{x_i \le x} \left(g(x_i^+) - g(x_i^-)\right) - g(0) \right] dx \\
&= g(1) - 1 - \sum_{i=1}^{n} \left(g(x_i^+) - g(x_i^-)\right) + \int_0^1 \sum_{x_i \le x} \left(g(x_i^+) - g(x_i^-)\right) dx.
\end{aligned}$$

By Lemma B.4.1

$$-\sum_{i=1}^{n} \left(g(x_i^+) - g(x_i^-)\right) + \int_0^1 \sum_{x_i \le x} \left(g(x_i^+) - g(x_i^-)\right) dx = -\sum_{i=1}^{n} x_i \left(g(x_i^+) - g(x_i^-)\right).$$

Hence

$$\int_{-1}^{1} m(\theta)\, d\theta + \sum_{i=1}^{n} p_i = g^+(x) - g(1) + 1.$$

However $g(1) = g^+(1) + g(0) - g^-(1)$. Thus

$$\int_{-1}^{1} m(\theta)\, d\theta + \sum_{i=1}^{n} p_i = g^+(1) + 1 - g^+(1) - g(0) + g^-(1) = 1 - g(0) + g^-(1).$$

By the assumption $g(0) - g^-(1)$ is nonnegative and we have $\int_{-1}^{1} m(\theta)\, d\theta + \sum_{i=1}^{n} p_i \le 1$. The weight $m(0)$ assigned to the uniform density is determined by the fact that $M(\theta)$ must be a mixing measure, thus

$$m(0) + \int_{-1}^{1} m(\theta)\, d\theta + \sum_{i=1}^{n} p_i = 1$$

■

We call $g^-(1) \ge 0$ the total decrement of function g.

APPENDIX C

Computer source code for generating random correlation matrices

We list the source code of MATLAB scripts used to generate random correlation matrices uniformly from the set of semi-positive definite correlation matrices with the onion and the vine methods.

Listing C.1: *Vine method with C-vine*

```
function y = GenerateCorMatrixCVine(d)

% GENERATECORMATRIXCVINE generates a correlation matrix
%    of size d x d with the vine method (C-vine used).
%
% 'd' - [in] Dimension of the generated correlation matrix.
% 'y' - [out] Correlation matrix of dimension d x d.

% Initialization speeds up calculations
y = eye(d);

% row = 1
alp = 1+(d-2)/2;
y(1,2:d) = 2*betarnd(alp,alp,1,d-1)-1;
prr(1,:) = y(1,:);

% row > 1
for m = 2:d-1
    alp = 1+(d-1-m)/2;
    prr(m,m+1:d) = 2*betarnd(alp,alp,1,d-m)-1;
    for i = m+1:d
        tem = prr(m,i);
        for k = m-1:-1:1
            tem = prr(k,m)*prr(k,i) + ...
                tem*sqrt((1-prr(k,m)*prr(k,m))*(1-prr(k,i)*prr(k,i)));
        end
        y(m,i) = tem;
    end
end
y = y+y'-eye(d);
```

Listing C.2: *Vine method with D-vine*

```
function y = GenerateCorMatrixDVine(d)

% GENERATECORMATRIXDVINE generates a correlation matrix
%    of size d x d with the vine method (D-vine used).
%
% 'd' - [in] Dimension of the generated correlation matrix.
% 'y' - [out] Correlation matrix of dimension d x d.

% Initialization
y = eye(d);
alp = d/2;

% First off-diagonal
prr = 2*betarnd(alp,alp,1,d-1)-1;
for i = 1:d-1
    y(i,i+1) = prr(i);
    y(i+1,i) = prr(i);
end

% Remaining off-diagonals
for m = 2:d-1
    alp = alp - 0.5;
    prr = 2*betarnd(alp,alp,1,d-m)-1;
    for i = 1:d-m
        y(i,i+m) = PartCorr2Corr(y,i,i+m,prr(i));
        y(i+m,i) = y(i,i+m);
    end
end

% Helper functions for GenerateCorMatrixDVine

function y = PartCorr2Corr(mat, jstart, jend, prr)

% PARTCORR2CORR calculates the product moment correlation based on
% already filled positions in the correlation matrix and the corresponding
% partial correlation.
%
% 'mat' - [in] Partially generated correlation matrix.
% 'jstart' - [in] Row index of the computed correlation.
% 'jend' - [in] Column index of the computed correlation.
% 'prr' - [in] Value of the partial correlation with conditioned set
%           [jstart, jend].

nrow = jend - jstart - 1;
a = zeros(nrow,nrow);
b = zeros(nrow,2);
for i = jstart+1:jend-1
    ii = i - jstart;
    for j = jstart+1:jend-1
        jj = j - jstart;
        a(ii,jj) = mat(i,j);
    end
    b(ii,1) = mat(i,jstart);
    b(ii,2) = mat(i,jend);
end

x = a\b;

tem11 = 0;
for ii = 1:nrow
    tem11 = tem11 + x(ii,1)*mat(ii + jstart,jstart);
end

tem13 = 0;
for ii = 1:nrow
    tem13 = tem13 + x(ii,2)*mat(ii + jstart,jstart);
end
```

```
68
69  tem33 = 0;
70  for ii = 1:nrow
71      tem33 = tem33 + x(ii,2)*mat(ii + jstart,jend);
72  end
73
74  y = tem13 + prr*sqrt((1-tem11)*(1-tem33));
```

Listing C.3: *Onion method*

```
1   function y = GenerateCorMatrixOnion(d)
2
3   % GENERATECORMATRIXONION generates a correlation matrix
4   %    of size d x d with the onion method.
5   %
6   % 'd' - [in] Dimension of the generated correlation matrix.
7   % 'y' - [out] Correlation matrix of dimension d x d.
8
9   y = eye(d, d); % initialize
10
11  % row = 1
12  b = sqrt(betarnd(1/2,d/2,1,1));
13  u = 2*unidrnd(2,1,1)-3;
14  q = b*u;
15  y(1, 2) = q;
16  y(2, 1) = q;
17
18  % row > 2
19  c = eye(d,d);
20  for k = 2:d-1
21      c(1:k,1:k) = IncrementalChol(y(1:k,1:k),c(1:k-1,1:k-1));
22      b = sqrt(betarnd(k/2,(d-k+1)/2,1,1));
23      u = GenerateSphereUnif(k);
24      q = c(1:k,1:k)*b*u;
25      y(1:k,k+1) = q;
26      y(k+1,1:k) = q';
27  end
28
29  % Helper functions for GenerateCorMatrixOnion
30
31  function y = GenerateSphereUnif(n)
32
33  % GENERATESPHEREUNIF generates 1 sample of n-dimensional
34  %    uniform distribution on a sphere in R^n.
35  %
36  % 'n' - [in] Dimension of uniform distribution on a sphere to sample
37  %              from.
38  % 'y' - [out] Vector of length n containing 1 sample of n-dimensional
39  %               uniform distribution on a sphere in R^n.
40
41  N = normrnd(0,1,n,1);
42  y = N./sqrt(sum(N.^2));
43
44
45  function y = IncrementalChol(m, c)
46
47  % INCREMENTALCHOL computes the Cholesky decomposition incrementally
48  %    when new 'q' is generated and appended to 'y'.
49  %
50  % 'm' - [in] Leading principal minor of dimension k x k of the
51  %              correlation matrix.
52  % 'c' - [in] Cholesky decomposition of the leading principal minor
53  %              of dimension k-1 x k-1 of the correlation matrix.
54  % 'y' - [out] Cholesky decomposition of the leading principal minor
55  %               of dimension k x k of the correlation matrix.
56
57  k = size(m,1);
58
```

```
59  for i = 1:k-1
60      tem = 0;
61      for j = 1:i-1
62          tem = tem+c(i,j)*c(k,j);
63      end
64      if (abs(m(k,i)-tem)>1.e-5)
65          c(k,i) = (m(k,i)-tem)/c(i,i);
66      else
67          c(k,i) = 0;
68          c(i,k) = 0;
69      end
70
71      tem = 0;
72      for j = 1:k-1
73          temp = tem+c(k,j)*c(k,j);
74      end
75      if (m(k,k)-tem<=0)
76          c(k,k) = 0;
77      else
78          c(k,k) = sqrt(m(k,k)-tem);
79      end
80  end
```

The code has been optimized for speed with the help of the built-in profiling tool of MATLAB.

Listing C.3 uses the idea of prof. Harry Joe (personal communication) for computing the Cholesky decomposition incrementally. It means that a lot of computational time is saved since we do not perform the full Cholesky decomposition of matrix y, but save it in matrix c and append new row as new q is generated (line 21 of the listing).

Summary

High Dimensional Dependence
Copulae, Sensitivity, Sampling

Daniel Lewandowski

Uncertainty analysis has definitely past its infant times. Whether it is a regulatory obligation, a desire to optimize processes of all kinds, or simply a curiosity, uncertainty analysis allows dealing with random in nature phenomena within a well developed framework. It is no more a question of simple statistical analysis, but rather a matter of full scale high-dimensional modelling, where dependencies between variables are among the most important aspects. For already quite some time industries operating with hazardous materials are subject to very demanding probabilistic risk assessment regulations. Entities like nuclear power plants, chemical factories or airliners often include departments responsible for constant monitoring of risk factors. These, in turn, may exhibit high correlations between each other. Therefore it is not only important to model the marginal distributions of variables in question; even more crucially the dependence structures must be captured to reflect the interactions as these may alter final results significantly. Actuarial sciences make use of high dimensional copulae for maximizing the profit and minimizing risks and current guidelines often recommend usage of tail dependent copulae for modelling assets. Currently copulae are among most popular methods of modelling dependent random variables and most likely they will preserve their position as such in the future.

The usage of copulae has been simplified over the years of development of software tools for uncertainty analysis. A good example in this regard is the software developed at the Delft University of Technology called UNICORN. It implements various copulae with properties that should satisfy many users. Future years should bring even more advanced software solutions with tools allowing efficient specification of complex dependence structures in a matter of minutes and fast sampling to obtain results at site. Recently there has been a lot of effort devoted to the development of graphical representations of dependence structures, like dependence vines or continuous Bayesian belief networks. Especially the

last concept is currently actively developed at the Department of Mathematics of Delft University of Technology.

Although this study centers the bulk of the work on copulae, we have also broadened the perspective with departures to the field of sensitivity analysis, expert judgement and studies on correlation matrices. Chapter 2 forms the point of reference for the remaining papers incorporated into this thesis. It describes the standard tools for modelling high dimensional data with some parametric families of multidimensional copulae. We also study various dependence concepts and measures expressing interactions between random variables in a quantitative way. On the other hand, chapter 5 departures from copula modelling and concentrates on another application of vines - generating random correlation matrices of size $d \times d$ from the joint density of all correlation matrices of the same size. The matrices can be drawn from a joint density being proportional to a power of the determinant of the correlation matrix. The uniform distribution is a special case. The idea was introduced by Joe [2006] and was based on the D-vine. The method however is not limited to the use of this one type of copula. We argued that the C-vine is less computationally demanding and can successfully applied as well. In fact, we extend the method to be applicable to any regular vine. This brings new applications of this method of generating random correlation matrices. For instance, we can generate correlation matrices conditional on correlation values in an arbitrary tree. The Onion method proves to be very efficient computationally, however in some setups the C-vine method shows better performance. The Onion method has also been extended to allow generating random correlation matrices non-uniformly from the set of semi–positive definite correlation matrices.

An essential step in probabilistic risk analysis is sensitivity analysis and chapter 6 is dedicated to this subject. The chapter concentrates on the notion of correlation ratio, a variance based global sensitivity measure. Therefore we concentrated our efforts on developing a numerical method of estimating this quantity based on samples. We estimate the regression curve via a simple least-squares error fit of a polynomial. However there are two dangers in doing so without any control mechanism. Fitting a polynomial of too low degree may result in a very bad fit, which does not correspond well to the true regression curve. On the other hand, a polynomial of a very high degree exhibits a very good fit to this specific sample, but cannot be seen as a good estimator of the regression for the whole population. Therefore we introduced an overfitting prevention method to overcome this problem. Three different criteria have been tested for detecting the overfitting and the best performing algorithm is based on an early stopping approach. The whole method of estimating the correlation ratio from a sample is very easy to implement and performs well even with moderate sample size.

The subject of multidimensional statistical dependence modelling turned out be far more complex than initial views of the author on this issue. For him this work has probably brought more questions than it answered - feeling scientists should be familiar with. Future research is therefore well motivated and this thesis may not be the last word of the author on this story yet.

Samenvatting

Hoog-Dimensionale Afhankelijkheden
Copula's, Gevoeligheden, Trekkingen

Daniel Lewandowski

De onzekerheidsanalyse is ongetwijfeld uit zijn kinderschoenen gegroeid. Of het nu vanwege een wettelijke verplichting is, of vanwege de wens om processen van allerlei te optimaliseren, of gewoon vanwege nieuwsgierigheid, de onzekerheidsanalyse staat het toe om met natuurlijke fenomenen om te gaan in een goed ontwikkeld raamwerk. Het gaat steeds minder om een eenvoudige statistische analyse, maar meer om groot-schalig modelleren in meerdere dimensies waarbij de afhankelijkheid tussen groot-heden één van de belangrijkste aspecten is. Reeds lang zijn industriën die werken met gevaarlijke stoffen onderhavig aan weten regelgeving die een probabilistische risico-analyse voorschrijven. Organisaties en bedrijven zoals kerncentrales, chemische fabrieken of luchtvaartmaatschappijen hebben speciale afdelingen die de risico-factoren continu in de gaten houden. Deze factoren kunnen een hoge onderlinge correlatie vertonen. Het is daarom niet alleen belangrijk om de marginale kansdichtheden van deze stochasten te bepalen; ook de afhankelijkheidsstructuur moet bepaald worden, omdat de inherente interactie tussen de stochasten het resultaat van de analyse sterk kan beïnvloeden. De actuariële wetenschap maakt gebruik van hoger dimensionele copula's voor het maximaliseren van de winst en voor het minimaliseren van de risico's. Huidige voorschriften raden vaak aan om staart-afhankelijke copula's toe te passen bij het modelleren van verliezen of effecten. Op dit moment zijn copula's één van de meest populaire methoden om afhankelijke stochasten te modelleren en waarschijnlijk zullen zij dit ook blijven in de toekomst.

De toepassing van copula's is de afgelopen jaren sterk vereenvoudigd door de ontwikkeling van programmatuur voor het uitvoeren van onzekerheidsanalyses. Een goed voorbeeld hiervan is het programma UNICORN dat aan de Technische Universiteit Delft is ontwikkeld. In deze software zijn verschillende copula's geïmplementeerd met eigenschappen die de meeste gebruikers tevreden zouden moeten stellen. In de komende jaren zullen steeds meer geavanceerde software-

applicaties het mogelijk maken om complexe afhankelijkheidsstructuren in slechts enkele minuten te definieren en door te rekenen. Recentelijk is er veel inspanning geleverd in het grafisch weergeven van dit soort afhankelijkheidsstructuren. Voorbeelden hiervan zijn zogenaamde "vines" en "continuous Bayesian belief networks". Met name aan deze laatste wordt actief gewerkt aan de wiskundeafdeling van de Technische Universiteit van Delft.

Alhoewel het grootste deel van deze studie op copula's is geconcentreerd, hebben we onze grenzen ook verlegd met uitstapjes naar de gevoeligheidsanalyse, het gebruik van expert meningen en naar de studie van correlatie matrices. Hoofdstuk 2 vormt het beginpunt voor de verdere artikelen die aan de basis liggen van dit proefschrift. Hierin worden de standaardmethoden beschreven voor het modelleren van hoger dimensionale gegevens met een aantal parametrische families van copula's. We bestuderen ook verschillende concepten van afhankelijkheid tussen stochasten en manieren om interacties tussen deze stochasten kwantitatief te beschrijven. Daartegenover stappen we in hoofdstuk 5 af van de copula's en concentreren we ons op een andere toepassing van vines: het genereren van willekeurige correlatiematrices van grootte $d \times d$ vanuit de gezamelijke kansdichtheid over alle correlatiematrices van deze grootte. De matrices kunnen getrokken worden uit de gezamelijke kansdichtheid die proportioneel is aan de macht van de determinant van de correlatiematrix. De uniforme verdeling is hiervan een bijzonder geval. Dit idee komt van Joe [2006] en was gebaseerd op de zogenaamde D-vine. De methode is echter niet beperkt tot deze ene copula. Wij tonen aan dat de C-vine ook goed toegepast kan worden en dat deze bovendien minder rekenintensief is. We laten ook zien dat de methode uitgebreid kan worden naar elke reguliere vine. Dit creëert nieuwe mogelijkheden om deze methode toe te passen bij het trekken van willekeurige correlatiematrices. Dit maakt het bijvoorbeeld mogelijk om correlatiematrices te genereren conditioneel op correlatiewaarden in een willekeurige boom. De zogenaamde 'Onion' methode blijkt rekentechnisch erg efficient te zijn, maar in sommige opstellingen laat de C-vine een betere prestatie zien. Deze Onion methode is ook uitgebreid om het mogelijk te maken om willekeurige correlatiematrices op een niet-uniforme manier uit een verzameling van semi-positief definiete matrices te trekken.

Een essentiële stap in een probabilistische risicoanalyse is een gevoeligheidsanalyse en hoofdstuk 6 gaat over dit onderwerp. Dit hoofdstuk bespreekt de notie van de correlatie-ratio die een globale maat van gevoeligheid is op basis van de variantie. We concentreren onze inspanning op de ontwikkeling van numerieke methoden voor het schatten van deze ratio op basis van trekkingen. We schatten de regressiecurve via een eenvoudige kleinste kwadraten methode voor het fitten van polynomen. Er zijn echter twee valkuilen als we dit zonder een controlemechanisme doen. Het fitten van een polynoom van een te lage graad zal resulteren in een slechte fit die niet goed overeenkomt met de ware regressiecurve. Aan de andere kant zal een polynoom van een te hoge graad weliswaar goed fitten, maar kan deze niet als een goede schatting voor de hele populatie beschouwd worden. Hiervoor introduceren we een procedure voor het voorkomen van dit probleem dat bekend staat als "overfitting". Drie verschillende criteria voor het toetsen op overfitting zijn getest en het criterium met de beste resultaten is gebaseerd

op een aanpak van vroegtijdig stoppen. De hele methode voor het schatten van de correlatie-ratio van trekkingen is heel eenvoudig en werkt goed, zelfs met een beperkte hoeveelheid trekkingen.

Het onderwerp van hoger dimensionale statistische afhankelijkheid is veel complexer gebleken dan de auteur in eerste instantie dacht. Voor hem heeft dit werk waarschijnlijk meer vragen dan antwoorden opgeleverd. Een gevoel waar vele wetenschappers zich ongetwijfeld in zullen herkennen. Verder onderzoek is daarom zeker wenselijk en dit proefschrift kan wel eens niet het laatste woord van de auteur over dit onderwerp zijn.

Curriculum Vitae

Daniel Lewandowski was born in Kożuchów, Poland, on May 20, 1977. After finishing the Technical High School in Zielona Góra, Poland, with the specialization general electronics, he begun his studies in mathematics at the University of Zielona Góra. After two and half years of studies in Poland, he moved to the Netherlands to continue the studies in the Risk and Environmental Modelling programme offered by the Delft University of Technology. He graduated in 2003 both at the Delft University of Technology and the University of Zielona Góra. In the following year he worked on two projects at the University of Strathclyde in Glasgow, UK and the George Washington University, Washington D.C., USA. In 2004 he went on to become a PhD student (in Dutch: *Assistent in Opleiding* or AIO) with the Faculty of Electrical Engineering, Mathematics and Computers Systems at the Delft University of Technology.