Some measures of local and global dependence

Karl Ove Hufthammer

Master's thesis in statistics
Mathematical statistics

Department of Mathematics
University of Bergen

30th November 2005

Acknowledgements

I would like to thank my supervisor, professor Dag Tjøstheim, for help and support throughout the writing of this thesis.

The rest of the staff has also been extremely helpful, and they have taught me almost everything I know (or have forgotten) about probability and statistics.

Thanks also to my fellow students at Kroepeliens for making my time here an enjoyable period of my life. The waffles would never have been the same without you!

Finally, I would like to thank my family and friends – for being there when I needed you.

Bergen, 30th November 2005
Karl Ove Hufthammer

Contents

1 Introduction 7

1.1 Continuous numbering 8

1.2 Software used 8

2 Measures of global dependence 9

2.1 Concepts and definitions 9

2.1.1 Measures of concordance 12

2.2 Correlation and regression 12

2.2.1 Regression and correlation 13

2.2.2 Correlation in mixtures 17

2.2.3 Rank correlation 18

2.3 Kendall's tau 20

2.4 Problems with correlation 21

2.4.1 Zero correlation does not imply independence 21

2.4.2 Correlation range not attainable 22

2.5 Problems with rank-based measures of association 27

2.6 Measures of concordance 27

2.7 Other measures of dependence 28

2.8 Summary and conclusions 28

3 Measures of local dependence 30

3.1 Introduction 30

3.2 Properties of measures of local dependence 31

3.3 Correlation curves 31

3.3.1 A simple generalisation 32

3.3.2 The multiple correlation curve 32

3.3.3 Basic properties of the correlation curve 33

3.3.4 Estimating the correlation curve 34

3.3.5 Correlation curves and transformations 38
3.4 The local dependence function 41
3.4.1 Cross-product ratios 41
3.4.2 Defining the local dependence function 43
3.4.3 Properties of the local dependence function 48
3.4.4 Ordering of dependence 48
3.4.5 Constant local dependence 49
3.4.6 Dependence maps 50
3.4.7 Local dependence in transformations 62
3.4.8 Local dependence in mixtures 62
3.5 Regional dependence 64
3.6 Local dependence function of Bairamov and Kotz 66
3.7 Summary and conclusions 67

4 Copulas 69
4.1 Introduction 69
4.2 Defining distribution functions 70
4.3 Defining copulas 72
4.4 Examples 75
4.5 Local dependence and copulas 78
4.6 Summary and conclusions 80

5 Graphical methods 81
5.1 Chi-plots 81
5.2 Kendall plots 85
5.3 Summary and conclusions 88

6 Summary and future research 89

Notation and definitions

Here is a short list of some of the more frequently used notation used in this thesis. Any other nonstandard notation used will be explained in the text.

Symbol	Description
$X, Y, Z, W, \ldots$	random variables
$x, y, z, w, \ldots$	real numbers or nonrandom variables
$f, g, f_1, f_2, f_X, \ldots$	marginal probability density functions
$F, G, F_1, F_2, F_X, \ldots$	corresponding cumulative distribution functions
h and H	multivariate density and distribution functions
$\mathbb{P}(A)$	probability of the event A
$\mathbb{E}(X)$	mean of X
$\mathrm{Cov}(X, Y)$	covariance between X and Y
$\mathrm{corr}(X, Y)$ or ρ_{XY}	correlation between X and Y
$\mathrm{Var}(X)$ or σ_X^2	variance of X
$\mathrm{SD}(X)$ or σ_X	standard deviation of X
$\mathbb{I}$	unit interval, $[-1, 1]$
$\mathbb{R}$ and $\overline{\mathbb{R}}$	sets of real numbers, $(-\infty, \infty)$ and $[-\infty, \infty]$
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_y^2, \rho)$	bivariate normal distribution with means μ_X and μ_2, variances σ_X^2 and σ_Y^2 and correlation ρ
U and V	random variables uniformly distribution on $\mathbb{I}$
$\mathrm{Dom}\, X$	domain of X
$\mathrm{Ran}\, X$	range of X

We will, for instance, let f_X denote the probability density function of the random variable X; but we will frequently omit any subscripts when it is clear from context which variable is intended. We may also use f, g, h and other lowercase letters as general functions. Again, the meaning will be clear from context.

1 Introduction

Dependence between random variables is a much studied topic in probability and statistics, and it is the subject of this thesis. We will look at various measures of the strength and direction dependence, both from a theoretical and empirical point of view.

Some of the measures characterise 'overall dependence', and these are discussed in chapter 2. But the dependence between variables often *varies* over their support, and the 'local dependence' is therefore of special interest. We will look closely at two approaches to quantifying this dependence, and we will examine the properties and problems of the resulting functions in chapter 3.

The global and local measures only capture *some* of the dependence in the distributions, but in chapter 4 we will look at a function that describes the entire dependence between two or more variables. This function – the copula – does not depend on the marginal distributions, and is thus a pure dependence concept. Moreover, some of our earlier measures can be expressed as transformations of this copula.

Lastly, we will look at two types of graphical displays that may be of help in determining *if* there is a dependence between two variables, and, possibly, to infer which type of dependence there is.

1.1 Continuous numbering

Note that, to make the text easier to follow, we use a continuous numbering of theorems, lemmas, definitions and examples. This means that, for instance, example 3.4.2 can be followed by definition 3.4.3, which is followed by theorem 3.4.4, etc. Equations are numbered by sections.

1.2 Software used

All computer calculations and simulations were done on the statistical computing software package R, version 2.1.1 or 2.2.0, running on the SuSE Linux 9.2 operating system on an AMD Athlon$^{\text{TM}}$ XP 3200+ or an AMD Athlon$^{\text{TM}}$ XP 1600+ computer. See R development core team (2005). All source code in the thesis is written for this software package.

2
Measures of global dependence

2.1 Concepts and definitions

A variable Y is said to be completely dependent on X if there exists a function f such that $Y = f(X)$ with probability one. If Y is completely dependent on X and X is completely dependent on Y, we say that the two variables are (mutually) completely dependent. The other extreme is of course independence between the variables.

We wish to have a *measure of dependence* – a real-valued function that measures the *degree* of dependence between the two (or more) variables. There are a few 'natural' properties such a measure of dependence should have (Rényi 1959):

1. The measure, say h, should be defined for any pair (X, Y) of nonconstant variables.
2. $h(X, Y) = h(Y, X)$. We say that the measure is *symmetrical*.
3. $0 \leq h(X, Y) \leq 1$.
4. $h(X, Y) = 0$ if and only if X and Y are independent.
5. $h(X, Y) = 1$ if Y is completely dependent of X or X is completely dependent on Y.

6. For all one-to-one functions (injective functions) α and β, $h\big(\alpha(X), \beta(Y)\big) = h(X, Y)$. We say that the measure is *transformation invariant*.
7. If (X, Y) are bivariate normal (see section 2.2.1 on page 13), $h(X, Y)$ should be equal to the correlation coefficient (see equation 2.2 on page 12).

See also page 170 of Nelsen (1999) for references to discussions of and modifications of these properties. And note that most of them are easily extended to more than two variables (see below for a few examples). Now let us also discuss a few suggested properties of our own:

8. The range of h should be defined on an interval or a ratio scale.
9. The measure should *not* be symmetrical.
10. The measure should use natural (physical) units for ease of interpretation. In other words, it should *not* be transformation invariant.
11. The measure should be comparable across different data sets and variables.

Of course, no measure of dependence can satisfy *all* of these properties, since, for instance, property 2 and 9 are contradictory.

Range on an interval scale

A range being defined on an interval scale means that an increase in value from x to $x + \Delta$ is equivalent (by some meaningful definition of 'equivalent') to an increase in value from y to $y + \Delta$. A ratio scale has in addition a *meaningful* zero, so that, for instance, a value of $2x$ means 'twice as much association' as a value of x.

Basically, we here require that the interior points of the range of h, not only the endpoints 0 and 1, have a natural interpretation.

Symmetry or nonsymmetry

Symmetry was in property 2 on the preceding page (and in Rényi 1959) only defined when measuring the association between *two* variables (though the variables could be vectors). But it is easily extended to more variables, by requiring that $h(X_1, \ldots, X_n) = h(X_{k_1}, \ldots, X_{k_n})$ (where $k_1, \ldots, k_n$ is a permutation of the numbers 1 to n). When there is no clear 'cause and effect' or 'input and output' relationship, we may prefer a symmetrical measure of association.

Now consider the case where X is a random variable which may take both positive and negative values, and let $Y = X^2$. Knowledge of the value of X completely determines the value of Y, and $\mathrm{Var}(Y \mid X) = 0$. But knowledge of the value of Y does *not* uniquely determine the value of X, since X can be either $\sqrt{Y}$ or $-\sqrt{Y}$, i.e. $\mathrm{Var}(X \mid Y) \neq 0$. In other words, Y is completely dependent on X but X is not completely dependent on Y. We may here prefer a non-symmetrical measure of association.

Transformation invariance

A *transformation invariant* measure of association is a measure which is invariant to (a subset A of) all injective transformations of the variables. More precisely, if $\boldsymbol{X}$ is a vector of variables (either random variables or observations), g is an arbitrary function in A and h is the measure of association, we should have $h\big(g(\boldsymbol{X})\big) = h(\boldsymbol{X})$.

We may be satisfied letting A contain only a subset of all possible transformations. We may for example take A to be a set of linear functions, or a set of (strictly) increasing functions. If a measure is invariant to strictly increasing transformations, we call it a *scale-invariant* measure. We will in chapter 4 on page 69 see that the measures of dependence that are scale-invariant are exactly those measures that are dependent (in a mathematical sense) only on a functional of the multivariate distribution called the 'copula'.

In practical situations a measure of association which uses natural units could be useful. Consider the example $Y = \beta X + \epsilon$, where β is the measure of association and X, Y and ϵ are random variables. This measure of association is obviously not invariant to (even linear) transformations of X (β has to change for Y to have the same distribution), but is useful nontheless, for example in predicting Y given X. (And it can easily have a physical interpretation.) We will in this thesis mostly limit our discussion to measures of associations which are invariant to (increasing) linear transformations or which are scale-invariant.

Comparability

Measures of dependence are often used to compare the level of dependence among different data sets, or to compare various models for the same data set. We may, for instance, have observations of variables Y, X_1 and X_2, and wish to investigate whether there is a greater level of dependence between Y and X_1 than between Y and X_2.

We may also wish that the measure can be used on both continuous and on discrete variables; however, we will in this thesis mostly limit our discussion to absolutely continuous distributions and measures of dependence on these (but will note when extensions to other distributions exist).

Finally, observe that different measures can be useful in different situations, and which ones we use may depend on which features of the dependence we are interested in. We will now look closer at a measure of dependence called 'correlation'.

2.1.1 Measures of concordance

When looking at two random variables, we are often not only interested in the *degree* of dependence, but also in its *direction* – whether 'large' values of one variable is associated with 'large' values of the other (positive dependence), or with 'small' values (negative dependence). We will later give a proper definition of this notion (called concordance), but let us first look at a few frequently used measures of (directional) dependence. The most commonly encountered of these is called the *correlation*:

2.2 Correlation and regression

Let X and Y be two random variables with existing second-order moments. The covariance function is then defined as

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}\Big[\big(X-\mathbb{E}(X)\big)\big(Y-\mathbb{E}(Y)\big)\Big]\\ &= \mathbb{E}(XY)-\mathbb{E}(X)\,\mathbb{E}(Y).\end{aligned} \tag{2.1}$$

We see that the the covariance will be positive (and 'large') when large (small) values of X (that is, values greater than the mean) is associated with large (small) values of Y with high probability; and the covariance will be negative when large (small) values of X is associated with small (large) values of Y. We can now easily make the covariance invariant to positive linear transformations: Assuming non-zero variances, we define the correlation coefficient $\rho_{X,Y}$ as

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\operatorname{SD}(X)\operatorname{SD}(Y)}. \tag{2.2}$$

If one of the variances is zero, $\rho_{X,Y}$ is defined to be zero.

It can be shown (see, for example, Casella and Berger 2001, pages 172–173) that $-1 \leq \rho \leq 1$ and that $|\rho| = 1$ if and only if Y is almost surely a linear transformation of X. If X and Y are independent, the covariance (and the correlation) must be zero:

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) \\ &= \mathbb{E}(X)\,\mathbb{E}(Y) - \mathbb{E}(X)\,\mathbb{E}(Y) = 0.\end{aligned}$$

The converse is not true; see section 2.4 on page 21 for a counterexample. Let now $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate distribution, and denote $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ by S_{XY} (and let S_{XX} and S_{YY} have corresponding definitions). We can estimate the correlation coefficient by the *sample correlation coefficient*

$$R_{X,Y} = \hat{\rho}_{X,Y} = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}. \tag{2.3}$$

For information on bias and consistency, see the end of the following subsection.

2.2.1 Regression and correlation

Consider a pair of random variables (X, Y) with a bivariate distribution, where X has the marginal mean μ_X and variance $\sigma_X^2 \neq 0$, and $Y = \alpha + \beta X + \epsilon$, where ϵ is independent of X and has zero mean and variance $\sigma_\epsilon^2 \neq 0$. We have:

$$\begin{aligned}\mu_X &= \mathbb{E}(X) \\ \mu_Y &= \mathbb{E}(Y) = \alpha + \beta\,\mathbb{E}(X) = \alpha + \beta\mu_X \\ \sigma_X^2 &= \operatorname{Var}(X) \\ \sigma_Y^2 &= \operatorname{Var}(Y) = \operatorname{Var}(\alpha + \beta X + \epsilon) \\ &= \beta^2 \operatorname{Var}(X) + \operatorname{Var}(\epsilon) \\ &= \beta^2\sigma_X^2 + \sigma_\epsilon^2 \\ \rho_{X,Y} &= \frac{\operatorname{Cov}(X, \alpha + \beta X + \epsilon)}{\operatorname{SD}(X)\ \operatorname{SD}(\alpha + \beta X + \epsilon)} = \frac{\beta\ \operatorname{Cov}(X,X)}{\sigma_X\sigma_Y} \\ &= \beta\frac{\sigma_X^2}{\sigma_X\sigma_Y} = \beta\frac{\sigma_X}{\sigma_Y}\end{aligned} \tag{2.4}$$

Note that the squared correlation can be written:

$$\rho_{X,Y}^2 = \frac{\operatorname{Var}\big(\mathbb{E}\,(Y \mid X)\big)}{\operatorname{Var}(Y)} = \frac{\operatorname{Var}(\alpha + \beta X)}{\operatorname{Var}(\alpha + \beta X + \epsilon)}. \tag{2.5}$$

The squared correlation can thus be viewed as the proportion of variance of Y 'explained' by the linear association with X. Often the additional requirement that ϵ and X are normally distributed is imposed, and we write $(X, Y) \sim \mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$; see Bickel and Doksum (2001, pages 497–502).

2.2.1.1 Estimating the regression coefficients

Now consider the general linear case where we have a random sample $X_1, \ldots, X_n$ from a distribution (or $X_i = x_i$ are fixed numbers), and let $Y_i = \alpha + \beta X_i + \epsilon_i$, where ϵ_i are i.i.d. random variables with zero mean and finite variances, $i = 1, \ldots, n$. The least-squares estimators of α and β are:

$$\begin{aligned} B &= \hat{\beta} = \frac{S_{XY}}{S_{XX}} \\ A &= \hat{\alpha} = \bar{Y} - B\bar{X}. \end{aligned} \tag{2.6}$$

These are unbiased estimators, and are also equal to the conditional maximum likelihood estimators when the ϵ's are normally distributed. Let us also introduce the notation $\hat{Y}_i = A + BX_i$ for the predicted Y's.

It is easily shown that the sample correlation coefficient $R_{X,Y}$ defined in equation 2.3 on the previous page can be written

$$R_{X,Y} = B \frac{\sqrt{\frac{S_{XX}}{n-1}}}{\sqrt{\frac{S_{YY}}{n-1}}} = B \frac{S_X}{S_Y}. \tag{2.7}$$

Note the similarity to ρ_{XY} in equation 2.4 on the preceding page. The parameters β, σ_X and σ_Y have all been replaced by their (unbiased) estimators.

When the X_i's are fixed numbers and not random variables, we define the *coefficient of determination* using the formula of the square of the sample correlation coefficient, equation 2.3 on the previous page, with the X_i's replaced by x_i. We will denote this statistics by $R^2_{x,Y}$,

Like the square of the population correlation, the square of the sample correlation can also be seen as the proportion of (sample) variance of Y 'explained' by the linear association with X,

$$R^2_{X,Y} = \frac{\sum (\hat{Y}_i - \bar{\hat{Y}})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}, \tag{2.8}$$

or as the proportionate reduction in variation (sample variance) by the linear regression on X,

$$R^2_{X,Y} = \frac{\sum (Y_i - \bar{Y})^2 - \sum (Y_i - \hat{Y})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum (Y_i - \hat{Y})^2}{\sum (Y_i - \bar{Y})^2}. \tag{2.9}$$

These formulas have natural generalisations to multiple linear regression, $Y = \alpha + \sum_{i=1}^{k} \beta_i X_i + \epsilon$. Consult any book on this topic for further details. But see Kvålseth (1985) and Anderson-Sprecher (1994) for information on problems with interpretation of, and definition of, $R^2_{X,Y}$ in nonlinear models, when transformations are used and/or when model comparison of different models are of interest.

Also note that there are many other interpretations of $R^2_{X,Y}$. See Rodgers and Nicewander (1988), and Rovine and von Eye (1997) for a non-random sample of some of these.

2.2.1.2 Consistency of estimates

The statistics $R_{X,Y}$ and $R^2_{X,Y}$ are *biased* (but consistent) estimators of $\rho_{X,Y}$ and $\rho^2_{X,Y}$, respectively. And do note that we need to sample from the bivariate distribution to get valid (that is, consistent) estimates; when we sample from a conditional distribution, say $Y \mid X$, and let $X_i = x_i$ be fixed (chosen) numbers, the estimators ($R_{x,Y}$ and $R^2_{x,Y}$) are usually *not* consistent, and can have arbitrary bias. Consider an example where (X, Y) has a bivariate $\mathcal{N}(0, 0, 1, 1, \rho = 0.7)$ distribution and we have a sample of one million observations, but where we have fixed the x_i observations so that half of them have a value of $-\lambda$ and the other half the value λ, and we have sampled Y from the conditional distribution. The slope is here equal to the correlation: $\beta = \rho = 0.7$. Here is the result of one simulation where we try to estimate the slope and the correlation (the output is slightly edited for readability):

```
> n = 10^6 # Number of observations
> rho = .7 # Correlation
> lambda = 2 # End points
>
> m = n/2
> x1 = rep( −lambda, m )
> x2 = rep( lambda, m )
> x = c( x1, x2 ) # Let half of them have the value −lambda, the other half lambda.
> y = rnorm(n, rho∗x, sqrt(1−rho^2)) # Generate y observations from
> # the conditional Y | X=x distribution.
>
```

```
> lm( y~x )$coefficients

 (Intercept) x
0.0004948295 0.7007136920

> cor(x,y)

0.8908783
```

This shows that the estimate of α, the intercept, is 0.00, the estimate of $beta$, the slope, is 0.70 and the estimate of the correlation is 0.89. More runs of the same program gave approximately the same estimates. (Estimated standard error of the statistic $R_{x,Y}$ was 0.0002, computed from 100 simulations.)

While the estimate of β is very good (and the variance of the least-squares estimator is σ_Y^2/S_{xx}, so putting half of the x_i observations on each endpoint thus gives us the estimator with the lowest variance for the given range), the estimate of the correlation is severely biased. Reducing λ to 0.2 gives us the estimates 0.70 and 0.19 for β and ρ, respectively. It can also be shown that the mean of $R_{x,Y}$ is an increasing function of λ (see references at the end of this section).

Note that we need not have the x_i values split into two groups to observe severe bias. Here is the result of a simulation with one million observations, where the x_i values had a (frequency) distribution similar to that of a standard normal distribution, but with each observation scaled by a constant, c (before generating the corresponding y_i observations from the conditional $Y \mid X = x$ distribution):

c	$\hat{\beta}$	$R_{x,Y}$
1.0	0.70	0.70
10.0	0.70	0.99
0.1	0.69	0.10

(Repeating the simulation several times gave approximately the same numbers.)

These results show that the sample correlation $R_{x,Y}$ (or its square) should *not* be used to estimate the population correlation (or its square), unless care is taken to ensure that the distribution of the nonrandom x_i's mimics that of X, or at least that they have approximately the same variance; for it can be shown that the mean of $R_{x,Y}$ is largely a function of the ratio of the variance of X and the (sample) variance of the x_i's. And, indeed, for $\lambda = 1$, the sample correlation in the earlier two-split example is (consistently) estimated to approximately 0.70.

For more information on the distribution of $R_{x,Y}$ under nonrandom sampling, see Warren (1971) (and references therein), section 3.2 in van Belle (2002) and the discussion on the book's Web site.

2.2.2 Correlation in mixtures

We are now interested in expressing the correlation in a mixture of two distributions as a function of the correlations in each distribution.

Let (X, Y) have the (bivariate) distribution H_1 with probability p and H_2 with probability $q = 1 - p$. In other words, let $(X, Y) = Z(X_1, Y_1) + (1 - Z)(X_2, Y_2)$, where (X_1, Y_1) has distribution H_1, (X_2, Y_2) has distribution H_2, and Z is 1 with probability p and 0 with probability $q = 1 - p$.

We use the following notation:

$$\begin{aligned}
&\mathbb{E}(X \mid Z = 1) = \mu_{X_1}, && \mathbb{E}(X \mid Z = 0) = \mu_{X_2}, && \mathbb{E}(Y \mid Z = 1) = \mu_{Y_1},\\
&\mathbb{E}(Y \mid Z = 0) = \mu_{Y_2}, && \mathrm{SD}(X \mid Z = 1) = \sigma_{X_1}, && \mathrm{SD}(X \mid Z = 0) = \sigma_{X_2},\\
&\mathrm{SD}(Y \mid Z = 1) = \sigma_{Y_1}, && \mathrm{SD}(Y \mid Z = 0) = \sigma_{Y_2}, && \mathrm{corr}(X, Y \mid Z = z) = \rho_z.
\end{aligned}$$

To express the correlation between X and Y, we can use the the two variance and covariance identities

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\big(\mathrm{Var}(X \mid Z)\big) + \mathrm{Var}\big(\mathbb{E}(X \mid Z)\big) \quad \text{and}\\
\mathrm{Cov}(X, Y) &= \mathbb{E}\big(\mathrm{Cov}(X, Y \mid Z)\big) + \mathrm{Cov}\big(\mathbb{E}(X \mid Z), \mathbb{E}(Y \mid Z)\big).
\end{aligned}$$

Proof of the first identity can be found in Casella and Berger (2001, pages 167–168). The proof of the second identity is similar. We now have

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}\big(\mathrm{Cov}(X, Y \mid Z)\big) + \mathrm{Cov}\big(\mathbb{E}(X \mid Z), \mathbb{E}(Y \mid Z)\big)\\
&= p\,\mathrm{Cov}(X, Y \mid Z = 1) + q\,\mathrm{Cov}(X, Y \mid Z = 0)\\
&\quad + \mathrm{Cov}\big(Z\mu_{X_1} + (1 - Z)\mu_{X_2}, Z\mu_{Y_1} + (1 - Z)\mu_{Y_2}\big)\\
&= p\rho_1\sigma_{X_1}\sigma_{Y_1} + q\rho_2\sigma_{X_2}\sigma_{Y_2} + \mathrm{Cov}\big(Z(\mu_{X_1} - \mu_{X_2}), Z(\mu_{Y_1} - \mu_{Y_2})\big)\\
&= p\rho_1\sigma_{X_1}\sigma_{Y_1} + q\rho_2\sigma_{X_2}\sigma_{Y_2} + pq(\mu_{X_1} - \mu_{X_2})(\mu_{Y_1} - \mu_{Y_2}).
\end{aligned}$$

And of course,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Cov}(X, X) = p\sigma^2_{X_1} + q\sigma^2_{X_2} + pq(\mu_{X_1} - \mu_{X_2})^2,\\
\mathrm{Var}(Y) &= \mathrm{Cov}(Y, Y) = p\sigma^2_{Y_1} + q\sigma^2_{Y_2} + pq(\mu_{Y_1} - \mu_{Y_2})^2.
\end{aligned}$$

This gives us the correlation

$$\rho_{X,Y} = \frac{p\rho_1\sigma_{X_1}\sigma_{Y_1} + q\rho_2\sigma_{X_2}\sigma_{Y_2} + pq(\mu_{X_1} - \mu_{X_2})(\mu_{Y_1} - \mu_{Y_2})}{\sqrt{\left(p\sigma^2_{X_1} + q\sigma^2_{X_2} + pq(\mu_{X_1} - \mu_{X_2})^2\right)\left(p\sigma^2_{Y_1} + q\sigma^2_{Y_2} + pq(\mu_{Y_1} - \mu_{Y_2})^2\right)}}. \quad (2.10)$$

When the means of the X_i's are equal, the means of the Y_i's are equal and all the variances are equal, this expression is simplified to

$$\rho_{X,Y} = p\rho_1 + q\rho_2, \quad (2.11)$$

a weighted mean of the two correlations.

Now consider a mixture of two distributions (populations) with positive correlation and parameters

$$\mu_{X_1} = 0, \mu_{Y_1} = 4, \mu_{X_2} = 3, \mu_{Y_2} = 0,$$
$$\sigma_{X_1} = \sigma_{X_2} = \sigma_{X_2} = \sigma_{Y_2} = 1,$$
$$p = q = \tfrac{1}{2} \text{ and } \rho_1 = \rho_2 = \tfrac{1}{2}.$$

We get a correlation of

$$\rho_{X,Y} = -\sqrt{\frac{5}{13}} \approx -0.62.$$

We see the perhaps surprising result that – even though the correlations in the subpopulations H_1 and H_2 are positive (both equal to one half) – the correlation in the mixture population is negative. Figure 2.1 on the facing page illustrates why this is so.

2.2.3 Rank correlation

Using transformed variables is one alternative to calculating the correlation directly. This may, for instance, be desirable if the association between X and Y is thought to be non-linear. Since the correlation measures the degree of *linear* association, transforming the variables to a scale of measurement where the association is believed to be approximately linear may be fruitful.

It is elementary to verify that when X and Y are continuous variables with distribution functions F and G, respectively, $U = F(X)$ and $V = G(Y)$ are uniformly distributed on $\mathbb{I} = [0, 1]$; see Casella and Berger (2001, pages 54–55) for details. We will now look at one, perhaps extreme, form of 'transformed correlation', called

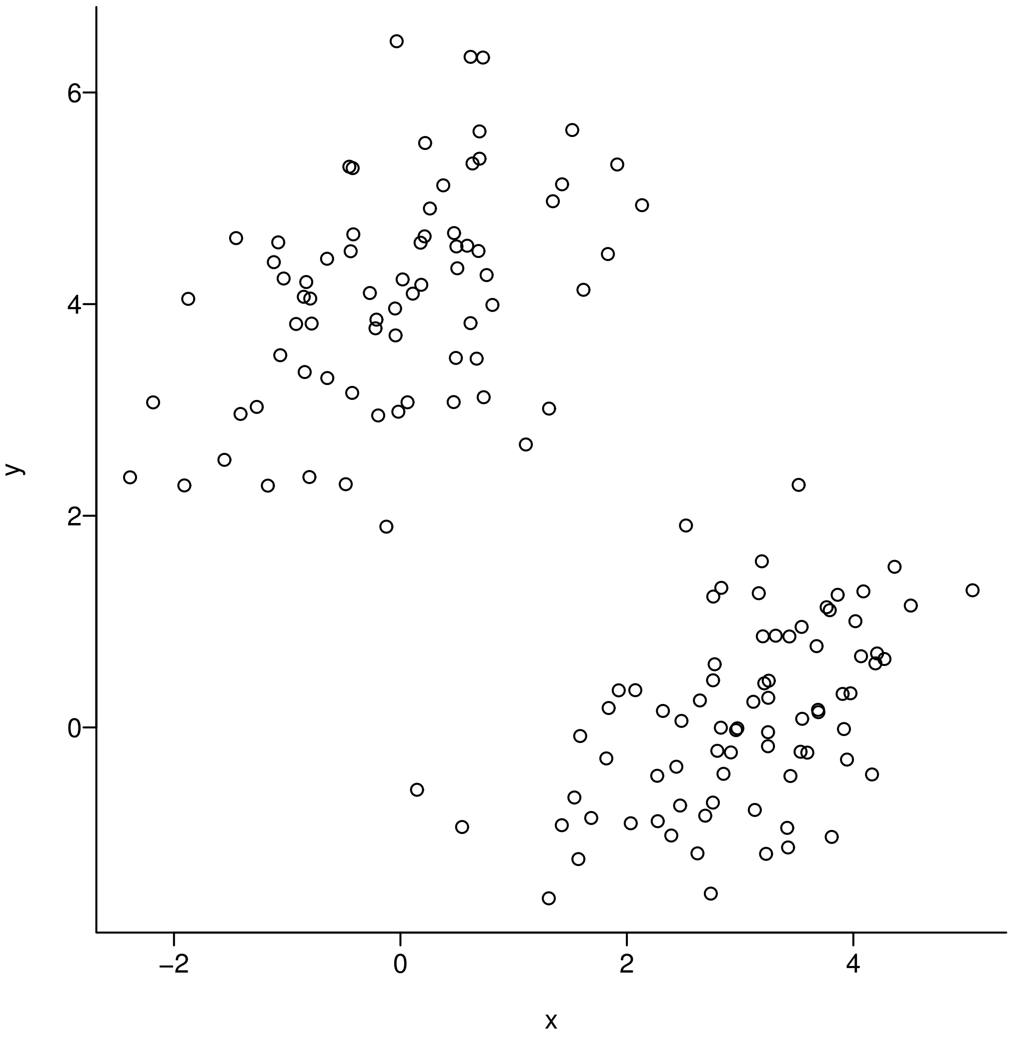


Figure 2.1: Scatterplot of 150 observations from an even mixture of two bivariate normal distributions, both with correlation $\frac{1}{2}$, with the first distribution having means $\mu_{X_1} = 0$ and $\mu_{Y_1} = 4$, the second $\mu_{X_2} = 3$ and $\mu_{Y_2} = 0$, and all marginal distributions having unit variance. The correlation in the mixture distribution is $\rho \approx -0.62$, and the estimated correlation is $r \approx -0.67$. We note that even though the correlations in each subpopulation are positive, the overall correlation is negative. Equation 2.10 on the facing page shows why results like this one can occur.

rank correlation, *Spearman's rank order correlation* or just *Spearman's rho*. It is defined as the correlation between U and V:

$$\begin{aligned}\rho_S(X,Y) = \operatorname{corr}\bigl(F(X),G(Y)\bigr) &= \operatorname{corr}(U,V) \\ &= \frac{\mathbb{E}(UV) - \mathbb{E}(U)\,\mathbb{E}(V)}{\mathrm{SD}(U)\ \mathrm{SD}(V)} \\ &= \frac{\mathbb{E}(UV) - \frac{1}{4}}{\frac{1}{12}} \\ &= 12\,\mathbb{E}(UV) - 3. \qquad (2.12)\end{aligned}$$

The sample rank correlation is calculated by replacing each value by its rank i, and then calculating the usual sample correlation on these ranks. Naturally, both the population and sample rank correlation share the same range as ordinary correlation, $[-1,1]$.

Rank correlation is a measure of *monotone* association, and is invariant to all increasing transformations of the original data. It is a very robust measure of association, meaning that its estimator is not affected much by a few observations, unlike in ordinary correlation, where *one* observation can completely determine the value of the estimated correlation.

2.3 Kendall's tau

Another popular rank-based measure of global association is Kendall's tau. It is based on the idea that two variables are positively dependent if large (small) values of one variable tend to occur with large (small) values of the other variable.

We say that two observations (x_1,y_1) and (x_2,y_2) from a random variable (X,Y) are *concordant* if and only if

$$q = (x_1 - x_2)(y_1 - y_2) > 0. \qquad (2.13)$$

If the inequality is changed to 'strictly less than', we say that the variables are *discordant*. Now define $Q = (X_1 - X_2)(Y_1 - Y_2)$, where (X_1,Y_1) and (X_2,Y_2) are two independent samples from the (X,Y) distribution. Kendall's tau, denoted by τ, is a measure of concordance, and we define it as

$$\begin{aligned}\tau &= \mathbb{P}(Q > 0) - \mathbb{P}(Q < 0) \\ &= 2 \cdot \mathbb{P}(Q > 0) - 1. \qquad (2.14)\end{aligned}$$

(The last equality is true when the variables are continuous.)

In observations from a random sample of pairs of variables, τ is usually estimated by

$$\hat{\tau} = \frac{\text{the number of concordant pairs} - \text{the number of discordant pairs}}{\text{the number of pairs}}. \tag{2.15}$$

Just like Spearman's rho, Kendall's tau is obviously invariant to strictly increasing transformations, and both τ and its estimate take values in the interval $[-1, 1]$. For $n > 10$, $\hat{\tau}$ is, for most purposes, well approximated by a normal distribution. See Mari and Kotz (2001) and references therein for details and other properties.

Finally, we note that Spearman's rho can also be seen as a measure of concordance; it is the probability of concordance minus the probability of discordance for (X, Y) and (X', Y'), where X' and Y' has the same marginal distributions as X and Y, respectively, but are otherwise completely independent of (X, Y), and of each other. In other words,

$$\rho_S(X, Y) = \mathbb{P}\big((X - X')(Y - Y') > 0\big) - \mathbb{P}\big((X - X')(Y - Y') < 0\big). \tag{2.16}$$

See Nelsen (1999, pages 134–136) for further details. There also exists several important relationships between Spearman's rho and Kendall's tau. One of them is the inequality $-1 \leq 3\tau - 2\rho_S \leq 1$. The proof, along with other inequalities, can be found in the book cited above, pages 141–146.

2.4 Problems with correlation

There are mainly two serious problems with correlation as a measure of association: 1) zero correlation does not imply independence, and 2) the range of correlation, $[-1, 1]$, is not attainable for all (pairs of marginal) distributions.

2.4.1 Zero correlation does not imply independence

While it is true that independent variables have zero correlation, the converse is not true, as the following example shows.

Let X be any symmetric random variable with mean 0 and existing third moment, and let $Y = X^2$. We have $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) = \mathbb{E}(X^3) = 0$. This is an extreme example, where we have *complete* association (one variable is a function of the other), but the correlation is still zero.

The reason this happens is, of course, that the positive association for $X \geq 0$ and the negative association for $X < 0$ 'cancel each other out'. To illustrate this, look at the correlation between $W = |X|$ and $Y = W^2 = X^2$, where X has a standard normal distribution. The distribution of W is called the *standard folded normal distribution*.

The covariance is now $\text{Cov}(W, W^2) = \mathbb{E}(W^3) - \mathbb{E}(W)\,\mathbb{E}(W^2)$. Straightforward integration gives us the needed terms

$$\begin{aligned}\mathbb{E}(W^3) &= \sqrt{\frac{8}{\pi}},\\ \mathbb{E}(W) &= \sqrt{\frac{2}{\pi}} \text{ and}\\ \mathbb{E}(W^2) &= 1,\end{aligned}$$

so the covariance is $\sqrt{\frac{2}{\pi}}$. Dividing by the standard deviations,

$$\begin{aligned}\text{SD}(W) &= \sqrt{1 - \frac{2}{\pi}} \text{ and}\\ \text{SD}(Y) &= 2,\end{aligned}$$

gives us the correlation, $\rho_{W,Y} = \frac{1}{\sqrt{\pi - 2}} \approx 0.94$. Similarly, the correlation between $-W$ and $(-W)^2 = W^2$ is $-\frac{1}{\sqrt{\pi-2}} \approx -0.94$.

This shows that conditional on X being non-negative, we have strong *positive* correlation between X and X^2, and conditional on X being negative, we have strong *negative* correlation. In other words, the 'local monotone association' varies over the support of X. We will in the next chapter give examples of measures of local dependence which quantifies and formalises this notation of 'local monotone association'.

2.4.2 Correlation range not attainable

As previously mentioned, the correlation will always lie between -1 and 1. But this range may not always be attainable. Before looking closer at this, though, let us first look at a basic property of bivariate distribution functions: It is well known that if (X, Y) has the distribution function H, then H has an upper and a lower bound:

$$H_-(x,y) \leq H(x,y) \leq H_+(x,y). \tag{2.17}$$

Let F and G be the marginal distributions of X and Y. The bounds are then

$$\begin{aligned} H_-(x,y) &= \max\bigl(F(x)+G(y)-1,0\bigr) \quad \text{and} \\ H_+(x,y) &= \min\bigl(F(x),G(y)\bigr). \end{aligned} \tag{2.18}$$

The proof of the right inequality in inequality 2.17 on the preceding page is almost trivial, since $\mathbb{P}(X \leq x, Y \leq y)$ is never greater than $\mathbb{P}(X \leq x)$ or $\mathbb{P}(Y \leq y)$; thus, it is never greater than the minimum of these these two marginal probabilities. The proof of the left inequality is also simple: $\mathbb{P}(X \leq x, Y \leq y) = 1 - \mathbb{P}(X > x \text{ or } Y > y) \geq 1 - \bigl(\mathbb{P}(X > x) + \mathbb{P}(Y > y)\bigr) = F(x) + G(y) - 1$. And since a probability must always be non-negative, the result follows.

We note that the bounds H_- and H_+, called Fréchet bounds, are themselves distribution functions. Let U have a uniform distribution on $\mathbb{I}$. The upper bound $H_+(x,y)$ is now the distribution function of $(X,Y) = \bigl(F^-(U), G^-(U)\bigr)$, and the lower bound $H_-(x,y)$ is the distribution function of $(X',Y') = \bigl(F^-(U), G^-(1-U)\bigr)$, where F^- is the *generalised inverse*, $F^-(u) = \inf\{x \mid F(x) \geq u\}$ (and similar for G^-). See Joe (1997, pages 58–59) for details and proof.

We will later see extensions of these bounds to higher dimensions.

There are several (generalised) expressions for covariance listed in (Mari and Kotz 2001, pages 151–152). One of the more useful is

$$\operatorname{Cov}(X,Y) = \iint \bigl(H(x,y) - F(x)G(y)\bigr)\, \mathrm{d}x\, \mathrm{d}y. \tag{2.19}$$

Using this and the formula for the correlation (equation 2.2 on page 12), we see that two marginal distributions attain their minimum and maximum correlations when their joint distribution is H_- and H_+, respectively.

Example 2.4.1

> When F is the $\mathcal{N}(\mu_X, \sigma_X^2)$ distribution and G is the $\mathcal{N}(\mu_Y, \sigma_Y^2)$ distribution, we obtain the highest correlation when $X = F^-(U)$ and $Y = G^-(U) = G^-\bigl(F(X)\bigr)$ (U being uniformly distributed on $\mathbb{I}$), that is, when X has the $\mathcal{N}(\mu_X, \sigma_X^2)$ distribution and Y can be written $Y = \mu_Y + \sigma_Y \frac{X-\mu_x}{\sigma_X}$.
>
> Similarly, we attain the lowest correlation when $X = F^-(U)$ and $Y = G^-(1-U) = G^-\bigl(1-F(X)\bigr)$, that is, when X has the $\mathcal{N}(\mu_X, \sigma_X^2)$ distribution and Y can written $Y = \mu_Y - \sigma_Y \frac{X-\mu_x}{\sigma_X}$.
>
> Obviously, the highest and lowest attainable correlations are here 1 and -1, respectively.

Let us now look at an example where the bounds are tighter than 1 and -1:

Example 2.4.2

If the two variables X and Y have lognormal distributions, the lower bound is not attainable, since that would involve Y being written as $Y = -aX + b$ for positive a, which is not possible, since both X and Y are non-negative. But we can also find closed-form expressions for the bounds. It can be shown (de Veaux 1976, cited in Shih and Huang 1992) that the maximum and minimum possible correlation between two lognormal variables whose logarithms have a bivariate normal distribution $\mathcal{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ is

$$\max\operatorname{corr}(X,Y) = \frac{e^{\sigma_X\sigma_Y} - 1}{\sqrt{(e^{\sigma_X^2} - 1)(e^{\sigma_Y^2} - 1)}} \quad \text{and} \tag{2.20}$$

$$\min\operatorname{corr}(X,Y) = \frac{e^{-\sigma_X\sigma_Y} - 1}{\sqrt{(e^{\sigma_X^2} - 1)(e^{\sigma_Y^2} - 1)}}. \tag{2.21}$$

For standard lognormal variables, these expressions reduce to give a possible correlation range of $[-1/e, 1]$. Here, the upper bound corresponds to $Y = X$. When we let one of σ_X or σ_Y increase towards infinity (and let the other stay constant), both the lower and the upper bound converge to zero.

Note that the above bounds are independent of ρ. And also note that the restriction on the logarithms having a bivariate normal distribution can be removed, as it *includes* the case of complete dependence, $Y = s(X)$ (with probability 1) for some monotone function s, namely $s(X) = G^-\big(F(X)\big)$ (maximum) and $s(X) = G^-\big(1 - F(X)\big)$, where F and G are the distribution functions of X and Y, respectively.

Correlation estimates using equation 2.3 on page 13 are *not* restricted by these bounds (consider a sample of size 2), but for large samples they will usually lie inside the range (asymptotically almost surely).

This could in fact be used to estimate the bounds when analytical expressions are difficult to obtain. We can use computers to quickly simulate many (perhaps a few hundred thousand) uniform variables, insert these into the expressions for maximum and minimum association and then estimate the resulting correlation.

When the marginal distributions are unknown, estimating the quantile functions (the inverse of the distribution functions) and using these seem to give good results, even for moderately many observations (how many depends on the distributions). One variant on this method is computing the estimated correlation of the sorted sample $(X_{(i)}, Y_{(i)})$ or $(X_{(i)}, Y_{(n-i+1)})$, $i = 1, \ldots, n$, as detailed in Shih and Huang (1992).

Here is one example of this last method. Assume that we want to estimate the maximum and the minimum correlation possible between a standard normal variable and a standard lognormal variable, and, furthermore, that we are able to simulate as many observations as we require from these distributions. We can easily write a program to estimate the extremal correlations, based on, for example, one hundred thousand observations (though as few as one hundred observations usually give reasonable estimates in this case). The output has been slightly edited for readability:

```
> n = 10^5 # One hundred thousand observations.
> x = rnorm(n) # From the standard normal distribution.
> y = rlnorm(n) # From the standard lognormal distribution.

> cor( sort(x), sort(y) ) # Estimated maximum correlation.
0.771875

> cor( sort(x), sort(y, decreasing=TRUE) ) # Estimated minimum correlation.
−0.7691114
```

It looks like the maximum and minimum correlation are approximately ±0.77. Note that, in this simulation, the original observations were taken from independent variables, but this method of estimating the extremal correlations will work just as well when the variables are dependent (even highly dependent), as long as the observation *pairs* are taken from a random sample.

Also observe that we can calculate the exact values fairly easy for these distributions. It is not difficult to show that we achieve the maximum correlation when X is any standard normal distribution and Y can be written $Y = e^X$:

Let F and G be the distribution functions of X and Y, respectively, and let U be a variable uniformly distributed on $\mathbb{I} = [-1, 1]$. X and Y can now be defined as $X = F^{-}(U)$ and $Y = G^{-}(U)$, respectively. It follows from basic properties of the standard lognormal distribution that Y can also be written $Y = e^{X'}$, where X' is a standard normal variable; in other words, X' is a variable with the same distribution as X, namely F. We have

$$\begin{aligned} G(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(e^{X'} \leq y) \\ &= \mathbb{P}(X' \leq \ln y) = F(\ln y). \end{aligned}$$

Furthermore,

$$U = G(Y) = F(\ln Y),$$

so

$$\ln Y = F^{-}(U),$$

and

$$Y = e^{F^{-}(U)} = e^{X}.$$

We have the maximum covariance

$$\begin{aligned}
\max \operatorname{Cov}(X, Y) &= \mathbb{E}(Xe^{X}) - \mathbb{E}(X)\,\mathbb{E}(e^{X}) = \mathbb{E}(Xe^{X}) - 0 \cdot \mathbb{E}(e^{x}) \\
&= \mathbb{E}(Xe^{x}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} xe^{x}e^{-\frac{1}{2}x^2} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} xe^{-\frac{1}{2}(x^2-2x)} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} xe^{-\frac{1}{2}(x-1)^2+\frac{1}{2}} \\
&= e^{\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} xe^{-\frac{1}{2}(x-1)^2} \\
&= e^{\frac{1}{2}}.
\end{aligned}$$

The last line follows from recognising the integral as the mean of a $\mathcal{N}(1, 1)$ variable. Now, using the variances of X and e^X, 1 and $e^2 - e$, respectively[1], we can calculate the correlation:

$$\begin{aligned}
\rho_{X,Y} &= \frac{\operatorname{Cov}(X, Y)}{\operatorname{SD}(X)\operatorname{SD}(Y)} \\
&= \frac{e^{\frac{1}{2}}}{\sqrt{1 \cdot (e^2 - e)}} \\
&= \frac{1}{\sqrt{e-1}} \approx 0.763.
\end{aligned}$$

Similarly, the minimum correlation is $-\frac{1}{\sqrt{e-1}} \approx -0.763$, and we see that the estimates were very close to the exact values.

[1] The formula for the variance of a lognormal variable can be found in Casella and Berger 2001, page 109, or in almost any basic statistics books.

2.5 Problems with rank-based measures of association

If we measure association between variables by using rank-based measures of association, both Spearman's rho and Kendall's tau have the problem that a value of zero for the given statistic does not imply independence. All the examples mentioned for correlation also holds for these measures.

However, the problem of reduced range for some distributions is no longer present. Since Spearman's rho is just ordinary correlation applied on the probability-integral transformed variables, the bounds reduce to $\operatorname{corr}(U, 1-U) = -1$ and $\operatorname{corr}(U, U) = 1$. And since Kendall's tau is also invariant to transformations (of the distributions, for the population measure, and of the observed values, for the statistic), it shares the same range, $[-1, 1]$.

2.6 Measures of concordance

We have previously looked at Spearman's rho and Kendall's tau – two 'measures of concordance'. But we have not specified any properties such a measure of concordance should satisfy. Joe (1997, page 136) lists the following 'desirable properties':

1. The measure, say k, should be defined for any pair (X, Y) of continuous variables.
2. $k(X, Y) = k(Y, X)$.
3. $-1 \leq k(X, Y) \leq 1$, with $k(X, X) = 1$ and $k(X, -X) = -1$.
4. $k(X, Y) = 0$ if X and Y are independent.
5. $k(-X, Y) = k(X, -Y) = -k(X, Y)$.
6. For all increasing functions α and β, $k\big(\alpha(X), \beta(Y)\big) = k(X, Y)$.

We will use this as our definition of a *measure of concordance* (or *concordance measure*). Note that Joe (1997, page 136) presented these properties in a different notation, based on the concept of copulas. A copula is, basically, a function that completely characterises the dependence between variables while being invariant to strictly increasing transformations of the marginals; but we will leave the details and definitions to chapter 4 on page 69.

Joe also added two other copula-based properties to the above list. These two properties relate the concept of concordance and the concept of a copula, and are not of much interest without having a fully developed theory of copulas. Consequently, we omit the properties from our definition. The reader may consult the work cited above for more information.

It should be clear from the definition that both Spearman's rho and Kendall's tau are measures of concordance, while correlation is not. And if we were to strengthen property 4 on the previous page to require equivalence between independence and zero concordance, none of the measures would satisfy all the properties.

2.7 Other measures of dependence

We have so far only lightly touched on the subject of dependence measures; there exists many other measures, some of them frequently used, that we have *not* looked at. Surveys of these can be found in books such as Mari and Kotz (2001), Nelsen (1999) and Joe (1997). See also the article Lehmann (1966).

2.8 Summary and conclusions

We have in this chapter looked at common measures of global dependence, and have defined some desirable properties such measures should have. We have examined three measures that are frequently used, and frequently used as 'measures of dependence'. We have looked at how these can, and can *not*, be estimated, and we have explored some interpretations of these measures, and of their sample counterparts.

The measure of dependence in mixtures of distributions, and in the relationship between the dependence in the mixture and in each subpopulation, is of special interest. We have therefore calculated an equation showing how the correlation in a mixture can be expressed as a function of the correlation in two subdistributions. This result will of course also hold analogously for Spearman's rho, which is related to ordinary correlation.

The relationship between the dependence in a mixture and the dependence in its constituents is just one of several examples we have discussed that shows how a single 'global measure of dependence' has serious problems capturing the dependence between (two) variables. What we may need is a 'measure of local dependence’, which can be allowed to vary over the range of values the variables take. We will now, in the subsequent chapters, look at how such a measure can reasonably be defined, what properties it should possess and a few examples of such measures.

3
Measures of local dependence

3.1 Introduction

As we have already seen, sometimes global measures of dependence do not contain enough information on the *nature* of association. Here is one additional example:

We clearly have a positive association between the age and the height of human beings: Older people – adults – are generally higher (on average) than younger people – babies. Of course, we do not have complete association, since people of a certain age vary much in height, and vice versa; but we expect the strength of association (based on a suitable global measure of dependence) to be somewhat high.

However, the strength of association is not constant: For humans aged 0–10, the level of association will be high[1], but for humans aged 40–50 (the same number of years), it will be low, or even nil.

In the previous chapter, we looked at an example (section 2.4.1 on page 21), $Y = X^2$, where we had high positive correlation in one area (positive X), high

[1] The exact value, and what we consider a 'high' level of association, is of course a function of the measure of dependence we use. We should choose a suitable measure based on the natural properties of what we are measuring (for example, not use a measure based on a linear model of association if we do not have good *reason* to believe the association to be linear) and the sampling methodology used. The definition of 'high' and 'low' levels of association may also depend on the context.

negative correlation in a different area (negative X), but an overall correlation of zero. Of course, in this case, we had Y being completely dependent on X, so any reasonable measure of local dependence should show complete dependence. But we may easily extend this example to one where we do not have complete dependence, for example $Y = X^2 + \epsilon$ where ϵ is 'noise' – perhaps a standard normally distributed variable.

We clearly need a 'local' measure of association, or, in other words, a *measure of local dependence*.

3.2 Properties of measures of local dependence

A 'good' local measure of association *should* preferably possess the same properties as defined in section 2.1 on page 9 for global measures of association, with the exception that property 4 is changed to only require independence to *imply* zero local dependence, and not necessarily be implied by it. But, in addition, the measure must be allowed to vary inside the support of the variables. In other words, the measure should be a function of both the random variables and of mathematical variables (real numbers). We may write this as $h = h_{XY}(x, y)$, where h is the measure of local dependence, but we will usually leave the dependence on the random variables implicit, and omit the subscript. We also note that any global measure of dependence will, of course, also be a (constant) local measure of dependence.

Measures of local dependence may also be measures of 'local concordance'; that is, they may possess the properties of section 2.6 on page 27. And, in fact, the two measures we will examine in this chapter are both 'measures of local concordance' – in the sense that they include information on the sign of the dependence (loosely: is Y locally an *increasing* or a decreasing function of X?). But they do not have all the properties a real measure of concordance should have: One of the measures is not symmetrical, and none of them are scale-invariant.

3.3 Correlation curves

Let us again look at the usual correlation in the linear case from section 2.2.1 on page 13, which, as shown in equation 2.4 on page 13, can be written

$$\rho = \beta \frac{\sigma_X}{\sigma_Y} = \frac{\beta\sigma_X}{\sqrt{(\beta\sigma_X)^2 + \sigma_\epsilon^2}}. \tag{3.1}$$

Bjerve and Doksum (1993) suggested a general local measure of association, the *correlation curve*, based on localising ρ by conditioning on X:

$$\rho(x) = \frac{\beta(x)\sigma_X}{\sqrt{\left(\beta(x)\sigma_X\right)^2 + \sigma_\epsilon^2(x)}} \tag{3.2}$$

where

$$\begin{aligned} \mu(x) &= \mathbb{E}\left(Y \mid X = x\right) \\ \beta(x) &= \mu'(x) \\ \sigma_\epsilon^2(x) &= \mathrm{Var}(Y \mid X = x) \end{aligned}$$

Whilst equation 3.1 on the previous page is based on the linear case, we here only require that X is a continuous random variable, $\mu(x)$ is continuously differentiable and all variances are finite. Y can be either discrete or continuous, or a mixture.

Note that the correlation curve does not require linear association or homoscedasticity (that $\mathrm{Var}(Y \mid X = x)$ is constant for all x). And *in* the linear, homoscedastic case (see section 2.2.1 on page 13), $\beta(x) = \beta$ (constant slope) and $\sigma_\epsilon^2(x) = \sigma_\epsilon^2$ (constant residual variance), so $\rho(x) = \rho$ for all x. In other words, we have constant local correlation.

3.3.1 A simple generalisation

Of course, any measure of location and scale can be used in defining a correlation curve. Following Bjerve and Doksum (1993), let $m(x)$ and $\tau(x)$ be measures of location and scale of $Y \mid X = x$, and let τ_X be a measure of scale of X. We assume, among other things, that the two measures of scale are of the same type (see the cited article for further details). Now define the generalised correlation curve by

$$\rho(x) = \rho_{X,Y}(x) = \frac{m'(x)\tau_X}{\sqrt{\left(m'(x)\tau_X\right)^2 + \tau^2(x)}}. \tag{3.3}$$

3.3.2 The multiple correlation curve

Blyth (1994) tried to generalise the correlation curve to the case of multiple covariates by localising the multiple correlation coefficient from the linear homoscedastic model,

$$Y = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X} + \epsilon,$$

where $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, $\mathbb{E}(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2$, and $\boldsymbol{X}$ and ϵ are independent. The multiple correlation coefficient is here defined to be

$$\rho^2 = \frac{\mathrm{Var}\big(\mathbb{E}\,(Y \mid \boldsymbol{X})\big)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(\alpha + \boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}})}{\mathrm{Var}(\alpha + \boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}} + \epsilon)} = \frac{\boldsymbol{\beta}\boldsymbol{\Sigma}\boldsymbol{\beta}^{\mathrm{T}}}{\boldsymbol{\beta}\boldsymbol{\Sigma}\boldsymbol{\beta}^{\mathrm{T}} + \sigma_\epsilon^2}. \tag{3.4}$$

Replacing the regression coefficients and residual variance with their conditional analogues,

$$\begin{aligned} \sigma_\epsilon^2(\boldsymbol{x}) &= \mathrm{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) \quad \text{and} \\ \beta_i(x) &= \frac{\partial}{\partial x_i}\,\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}), \end{aligned}$$

we get the multiple correlation curve

$$\rho^2(x) = \frac{\boldsymbol{\beta}(\boldsymbol{x})\boldsymbol{\Sigma}\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x})}{\boldsymbol{\beta}(\boldsymbol{x})\boldsymbol{\Sigma}\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x}) + \sigma_\epsilon^2(\boldsymbol{x})}. \tag{3.5}$$

It is not clear if this is a good measure of local association, and I see no convincing arguments why it should be.

3.3.3 Basic properties of the correlation curve

Here are some basic properties of the correlation curve, taken from Bjerve and Doksum (1993) and Doksum *et al.* (1994).

A well-defined range We easily see from the formula for the general correlation curve, equation 3.3 on the preceding page, that $-1 \leq \rho(x) \leq 1$, with equality holding if and only if $\sigma_\epsilon^2(x) = \mathrm{Var}(Y \mid X = x) = 0$, the case where X completely determines or 'explains' Y *locally*.

Independent variables and zero correlation When X and Y are independent, $m(x)$ is constant, so $m'(x) = 0$, and $\rho(x) = 0$ for all x (assuming that $\tau(x) \neq 0$).

Function of standardised slope It is easy to see that $\rho^2(x)$ is an increasing function of the standardised regression slope $\sigma_x\beta(x)/\sigma(x)$.

Invariance and equivariance Let X^* and Y^* be linear transformations of X and Y, respectively. We have $\rho_{X^*,Y^*}(x^*) = \pm\rho_{X,Y}(x)$. See Bjerve and Doksum (1993) for proof and further details.

Y* as a function of *X $Y = g(X)$ implies that the scale measure $\tau(x)$ is zero for all x, and we easily see from equation 3.3 on page 32 that $\rho(x) = \pm 1$ (where we define $0/0 = 1$).

The converse, that $\rho(x) = \pm 1$ for all x implies that Y is almost surely a function of X, is also true *if* $\tau(x) = 0$ implies that $Y = g(x)$ for some g. (This property does hold for $\tau(x) = \mathrm{Var}(Y \mid X = x)$.)

Regression dependence If $\mathbb{P}(Y \leq y \mid X = x)$ is nonincreasing in x, we say that Y is *positively regression dependent* on X, and this implies that $\rho(x) \geq 0$. See Lehmann (1966) for more information on regression dependence.

Symmetry and nonsymmetry The correlation curve is usually not symmetric; that is, $y = \rho_{X,Y}(x)$ and $x = \rho_{Y,X}(y)$ do not describe the same curve. Bjerve and Doksum (1993) suggested a simple ad hoc extension:

$$\rho^*(x,y) = \begin{cases} \mathrm{sign}\big(\rho_{X,Y}(x)\big)\sqrt{\rho_{X,Y}(x)\rho_{Y,X}(y)} & \mathrm{sign}\big(\rho_{X,Y}(x)\big) = \mathrm{sign}\big(\rho_{Y,X}(y)\big), \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

They offered no arguments why this last, symmertical measure should be a reasonable measure, and I can find no arguments either. We will therefore dismiss it from our consideration of possible measures of local dependence.

3.3.4 Estimating the correlation curve

There are several methods we can use to estimate the correlation curve, and the *best* method depends on the data set studied. We will here take a brief look at two estimation methods presented in Doksum *et al.* (1994), and which seem to work well for a wide range of possible distributions. A more detailed description of these, and references to theoretic studies of their properties, can be found in the cited article. See also Blyth (1994) for a possible estimation method for grouped data.

3.3.4.1 Neighbourhood estimates

Neighbourhood estimates are natural estimates of mean, $\mu(x)$, scale, $\sigma_\epsilon(x)$ and the derivative of the mean, $\beta(x)$, based on subsets of data pairs near x. Let k and x be fixed numbers, and let $I_k(x)$ be the set of indices of the x_i's closest to x, but with an equal number, $k/2$, of x_i's on either side. If $x = x_i$ for some i's, replace k by k

minus the number of such x_i's. For k odd, replace k by $k-1$. Now we can estimate $\mu_p(x) = \mathbb{E}(Y^p \mid X = x)$ by the average of all y_i^p, $i \in I_k(x)$, and $\beta(x)$ by

$$\hat{\beta}(x) = \frac{\overline{y}^+(x) - \overline{y}^-(x)}{\overline{x}^+(x) - \overline{x}^-(x)}, \tag{3.7}$$

where $\overline{y}^+(x)$ is the average of all the y_i with indices in $I_k(x)$ and lying strictly to the right of x (and similar for the other values).

The correlation curve can now be estimated by replacing each function in equation 3.2 on page 32 by its neighbourhood estimate. For estimating σ_X, we can use any reasonable estimate (for example the sample standard deviation used in equation 2.7 on page 14).

For details on conditions of consistency, on asymptotic confidence intervals and on choice of k, consult Doksum *et al.* (1994).

Here is my implementation of neighbourhood estimates:

```
# x = x vector. Must be sorted in increasing order.
# y = y vector.
# x0 = the point the estimates should be evaluated at.
# k = smoothing parameter. Use (approximately) this many points to to calculate each estimate.
estimate.neigh = function( x, y, x0 = mean(x), k = .3∗length(x)^(6/7) )
{
  n = length(x)
  k = k − sum(x==x0) # Reduce k by 1 for each value of x equal to x0.
  r = floor(k/2) # Use approx. kn/2 values to the left (and right) of x0.
  mid.min = which.min( x < x0 ) # Index of first x equal to or greater than x0.
  mid.max = which.max( x > x0 ) − 1 # Index of last x equal to x0 ...
  if( all( x != x0 ) ) mid.max = mid.min # ... or of first x greater than x0 if none are equal.
  if ( r > (mid.min−1) ) r = mid.min − 1 # Decrease value of r at boundaries, if necessary.
  if ( r > n − mid.max ) r = n − mid.max
  ys = y[ (mid.min − r):(mid.max + r) ] # The y values used.
  n = length(ys) # Now calculate and return estimates of mu(x),
                                                # sigma^2(x) and beta(x):
  mu = sum( ys ) / k
  mu2 = sum( ys^2 ) / k
  sig2 = max( mu2 − mu^2, 0 )
  xx = sum( x[ (mid.max+1):(mid.max+r) ] ) − sum( x[ (mid.min−r):(mid.min−1) ] )
  beta = ( sum( y[ (mid.max+1):(mid.max+r) ] ) − sum( y[ (mid.min−r):(mid.min−1) ] ) ) / xx
  rho = sd(x) ∗ beta / sqrt( sd(x)^2 ∗ beta^2 + sig2 )

  list( "mu" = mu, "sig2" = sig2, "beta" = beta, "rho" = rho )
}
```

3.3.4.2 Kernel estimates

Kernel estimates are based on a similar idea to neighbourhood estimates, but instead of using a fixed number of points on each side of x, we use a fixed window with varying number of points, and weigh the points used according to their distance to x.

Specifically, assume the x_i's are sorted, so $x_1 \leq \cdots \leq x_n$. We define the function $\hat{\mu}_{p,k}(x)$ as an estimate of $\mu_{p,k}(x) = \frac{\partial}{\partial x^{k-1}} \mathbb{E}(Y^p \mid X = x)$:

$$\hat{\mu}_{p,k}(x) = \sum_{i=1}^{n} \left[\frac{1}{b_{p,k}^{k}} \int_{s_{i-1}}^{s_i} w_{p,k}\left(\frac{x-u}{b_{p,k}} \right) \mathrm{d}u \, Y_i^p \right], \tag{3.8}$$

where $w_{p,k}$ are bounded, two times differentiable functions (*kernels*) with finite support, $s_i = (x_i + x_{i+1})/2$, $i = 1, \ldots, n-1$, $s_0 = x_1$, $s_n = x_n$ and (p,k) is equal to $(1,1)$, $(2,1)$ or $(1,2)$. We call $b_{p,k} = b_{p,k}^{(n)}$ the *bandwidths*, and, for ease of notation, we leave their dependence on n implicit. Basically, these determine how smooth our estimated function will be. Higher values make the estimate smoother (decreases the variance), but also increases the bias.

If we now let $W_{p,k}(u)$ be the integral of $w_{p,k}(t)$ from $-\infty$ to u, $\hat{\mu}_{p,k}(x)$ is easily seen to be equal to

$$\hat{\mu}_{p,k}(x) = -\sum_{i=1}^{n} \left[\frac{1}{b_{p,k}^{k-1}} \left(W_{p,k}\left(\frac{x-x_i}{b_{p,k}} \right) - W_{p,k}\left(\frac{x-x_{i-1}}{b_{p,k}} \right) \right) Y_i^p \right]. \tag{3.9}$$

When we can evaluate the integrals analytically, this latter formula is easy and fast to use in computer implementations. And, as before, we estimate the correlation curve by replacing each function in equation 3.2 on page 32 by its estimate from equation 3.9.

The estimates used are the Gasser-Müller estimates introduced in Gasser and Müller (1979) and further studied in Gasser and Müller (1984), and were based on the assumption of constant conditional variance. In these two articles consistency and asymptotic properties of the estimates are shown. These results were later proved to hold (with slightly stricter assumptions) for the hetereoscedastic model in Doksum *et al.* (1994). The details can be found in the article, but we note one sufficient statement for (pointwise) weak consistency:

$$\max_{1 \leq i \leq n} \mathbb{E}(Y^4 \mid X = x_i) \leq B < \infty, \tag{3.10}$$

as $n \to \infty$, where B does not depend on the sample size, and

$$\begin{aligned} b^{(n)} &\to 0 \quad \text{and} \\ n(b^{(n)})^3 &\to \infty \end{aligned} \tag{3.11}$$

(where $b^{(n)}$ are the three bandwidths, $b^{(n)}_{1,1}, b^{(n)}_{1,2}$ and $b^{(n)}_{2,1}$.

Here is my implementation of these kernel estimates:

```
# x = x vector. Must be sorted in increasing order.
# y = y vector.
# x0 = the point the estimates should be evaluated at.
# b = smoothing parameter (bandwidth).
estimate.gasser = function( x, y, x0 = mean(x), b )
{
  n = length(x)
  Wp = function( u ) # The integral of w_p1(t) from −1 to u, p = 1, ..., 2.
    {
      (−u^3/4 + 3∗u/4 + 1/2)∗(abs(u) <1) + (u > 1)
    }
  W12 = function( u ) # The integral of w_12(t) from −1 to u.
    {
      (15∗u^4/16 + 15/16 − 15∗u^2/8)∗(abs(u) <1)
    }
  s = numeric(n+1) # Note: Vectors in R start at index 1, not 0.
  s[1] = x[1] # s_0
  s[n+1] = x[n] # s_n
  s[2:n] = ( x[1:(n−1)] + x[2:n] ) / 2 # s_1, ... s_{n−1} = ...

  sup = Wp( (x0 − s[2:(n+1)]) / b ) − Wp( (x0 − s[1:n]) / b ) # Now calculate and return estimates:

  mu = − sum( sup ∗ y )
  mu2 = − sum( sup ∗ y^2 )
  sig2 = max( mu2 − mu^2, 0 )
  beta = − sum( (W12( (x0 − s[2:(n+1)]) / b ) − W12( (x0 − s[1:n]) / b )) ∗ y ) / b
  rho = sd(x) ∗ beta / sqrt( sd(x)^2 ∗ beta^2 + sig2 )
  list( "mu" = mu, "sig2" = sig2, "beta" = beta, "rho"= rho )
}
```

This implementation is just meant as an illustration of the algorithm, and is not optimised in any way. It uses the same bandwidths for the three estimates, but a better approach is having different bandwidths and bandwidths that varies with x. See Doksum *et al.* (1994) for a possible choice of better-performing bandwidths.

However, the kernels used in the implementation are the optimal kernels – the kernels which minimise the asymptotic mean square error. These can be shown (Gasser *et al.* 1985) to be equal to

$$
\begin{aligned}
w_{k,1} &= \frac{3}{4}(1-t^2), |t| \leq 1, \quad k = 1,2, \quad \text{and} \\
w_{1,2} &= \frac{15}{4}(t^3 - t), |t| \leq 1.
\end{aligned}
$$

Figure 3.1 on the next page shows Gasser-Müller estimates for the four different quantities, based on 150 observations from the model $Y = X^3 - X^2 - 25X + \epsilon(X)$, where $\epsilon \sim \mathcal{N}(\mu = 0, \sigma^2 = X^4)$ and X has a uniform distribution on $[-6, 6]$.

It is easy to show that the correlation curve here converge to $\sqrt{\frac{108}{109}} \approx 0,995$ as $|x|$ tends to infinity, and that it is equal to -1 when x is 0 (because the conditional variance $\sigma_\epsilon(x)$ is then 0).

Simple limit calculations show that *if* the ratio of the conditional mean and the conditional variance converges to plus or minus infinity, that is, $\lim_{|x|\to\infty} \frac{\beta(x)}{\sigma_\epsilon^2(x)} = \pm\infty$, the correlation curve will always converge to ± 1. And if it converges to zero, the correlation curve will also converge to 0. Finally, if it converges to a non-zero, finite constant, the correlation curve will also converge to a (different) constant.

Increasing the bandwidths makes the estimated functions smoother, but, at the same time, it increases the bias. The figure suggests that the estimators perform less well at the boundaries, and this impression is confirmed by performing estimates on other distributions. The problem was noted in the original articles (Gasser and Müller 1979, 1984; Gasser *et al.* 1985), where the authors proposed using special boundary kernels at the boundaries, and investigated their properties. Using boundary kernels can improve the estimates greatly, but we will not look at them in this thesis.

Note that there also exist other good estimators of the functions we are interested in, such as local polynomial kernel estimators. Any standard text on kernel smoothing or local regression, such as Wand and Jones (1995), has more information on this. And see Wilcox (2005) for a comparison of various estimators of the conditional variance.

3.3.5 Correlation curves and transformations

While the correlation curve is invariant under linear transformations (except that the sign may change), it is not invariant under other monotone transformations. Here

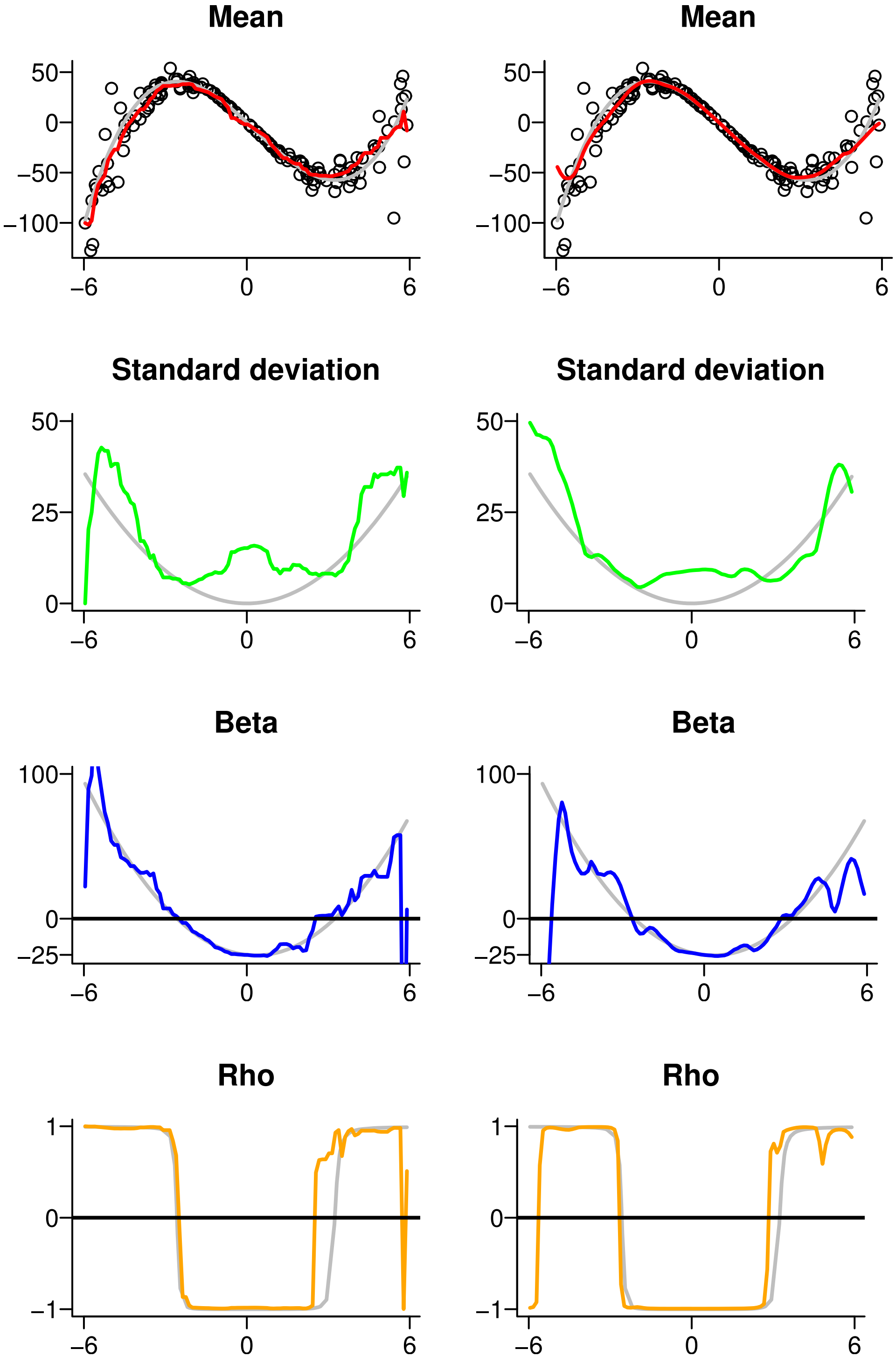


Figure 3.1: Neighbourhood (left) and Gasser-Müller (right) estimates of the conditional mean, the conditional variance, the derivative of the conditional mean, and the correlation curve. The grey lines show the real values of the various functions. The estimates are based on 150 observations from the model $Y = X^3 - X^2 - 25X + \epsilon(X)$, where $\epsilon \sim \mathcal{N}(\mu = 0, \sigma^2 = X^4)$ and X has a uniform distribution on $[-6, 6]$.

is one example: Let $Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ and let $W = a + bZ + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Then the correlation (curve) is

$$\rho_{Z,W}(z) = \rho_{Z,W} = \frac{\beta\sigma_Z}{\sqrt{(\beta\sigma_Z)^2 + \sigma_\epsilon^2}}, \tag{3.12}$$

that is, constant. Let us now look at $X = e^Z$ and $Y = e^W$. Conditional on X, Y has a lognormal distribution, with mean

$$\begin{aligned}\mu(x) = \mathbb{E}(Y \mid X = x) &= e^{a + b\ln x + \frac{1}{2}\sigma_\epsilon^2} \\ &= x^b e^{a + \frac{1}{2}\sigma_\epsilon^2}.\end{aligned}$$

The slope is

$$\beta(x) = \mu'(x) = bx^{b-1}e^{a+\frac{1}{2}\sigma_\epsilon^2},$$

and the conditional variance is

$$\mathrm{Var}(Y \mid X = x) = e^{2(a + b\ln x + \frac{1}{2}\sigma_\epsilon^2)}(e^{\sigma_\epsilon^2} - 1).$$

The variable X also has a lognormal distribution, with variance

$$\mathrm{Var}(X) = \mathrm{Var}(e^Z) = e^{\mu_x + \frac{1}{2}\sigma_X^2}.$$

These equations follow from basic properties of the lognormal distribution (see, for instance, Casella and Berger 2001, page 109). Inserting all of them into equation 3.2 on page 32 and doing some straightforward algebra, we get the correlation curve

$$\begin{aligned}\rho_{X,Y}(x) &= \mathrm{sign}(b)\sqrt{\frac{b^2 x^{2(b-1)} e^{2a + \mu_x + \sigma_\epsilon^2 + \frac{1}{2}\sigma_x^2}}{x^{2(b-1)} e^{2a + \sigma_\epsilon^2}\left(b^2 e^{\mu_x + \frac{1}{2}\sigma_x^2} + x^2(e^{\sigma_\epsilon^2} - 1)\right)}} \\ &= \mathrm{sign}(b)\sqrt{\frac{b^2 e^{\mu_x + \frac{1}{2}\sigma_x^2}}{b^2 e^{\mu_x + \frac{1}{2}\sigma_x^2} + x^2(e^{\sigma_\epsilon^2} - 1)}} \qquad (3.13)\end{aligned}$$

We see that, although the correlation curve between Z and W is constant, the correlation curve between $X = e^Z$ and $Y = e^W$ varies strongly with x. It may therefore be difficult to look at the square of correlation curves as a *general* measures of the 'proportion of local variability explained' by the regression on x, which was the interpretation suggested by Doksum *et al.* (1994).

This interpretation is still reasonable, though, with some restrictions. In order for the correlation curve to have a simple interpretation, we need to postulate that Y is related to X in the following manner:

$$Y = m(X) + \tau(X)\epsilon. \tag{3.14}$$

Here, m and τ are two general continuous functions (m being continously differentiable) and ϵ is the 'noise variable' – a random variable independent of X and with zero mean. We see that in the previous example Y and X were not related in this manner (the 'noise' was in fact a multiplicative factor, not an additive term).

Under this additive model, what we now mean when we say that the correlation curve is $\rho(x_0)$ at $X = x_0$ is that if we extended the relationship between Y and X at $X = x_0$ to hold (linearly) for other values of X,

$$Y = m'(x_0)X + \tau(x_0)\epsilon, \tag{3.15}$$

$\rho^2(x_0)$ can be interpreted in the same way as $R^2_{X,Y}$ in section 2.2.1.1 on page 14 is. But do note that $\rho^2(x)$, unlike $R^2_{X,Y}$, is *not* a symmetrical measure.

In some cases, we can transform the two variables to achieve an additive (nonlinear) regression model, but this is in general not possible. And if we do not know what the (functional) relationship between the variables is, but still want to estimate a correlation curve based on data alone, transforming the variables might not be *feasible*, even if it is *possible* in theory.

3.4 The local dependence function

Holland and Wang (1987a) suggested a completely different measure of local dependence. But before introducing it, let us briefly look at the concept of cross-product ratios.

3.4.1 Cross-product ratios

A two-dimensional contingency table $[P_{ij}]$ describes a joint probability function for two discrete random variables, say X and Y. All association between X and Y is contained in the local cross-product ratios

$$\alpha_{i,j} = \frac{P_{i,j}\, P_{i+1,j+1}}{P_{i,j+1}\, P_{i+1,j}}, \tag{3.16}$$

for all i and j where the probabilities exist. Together with the marginal distributions, $\mathbb{P}(X = x)$ and $\mathbb{P}(Y = y)$, these *uniquely* determine the contingency table – the joint distribution. See, for instance, Goodman (1969) and references therein for further details.

The cross-product ratios, often called odds ratios (see Nelsen 1999, page 79, for an explanation why), are much used as measures of dependence for discrete variables. Consider a simple crosstabulation of 'high' and 'low' values of two variables:

	low	high
high	c	d
low	a	b

The numbers a, b, c and d are either the number of observations in each category or the frequency (the number divided by the sum $n = a + b + c + d$). The cross-product ratio is $\alpha = ad/bc$. Under independence, we would expect the joint frequencies to be approximately equal to the product of the corresponding marginal frequencies, or, equivalently, the number of observations to be approximately equal to n times this product.

When $\alpha = 1$, we have $ad = bc$, and the 'expected' number of observations under independence is equal to the observed frequencies. For example, for a:

$$\begin{aligned}
\hat{a} &= n \frac{a+b}{n} \frac{a+c}{n} \\
&= \frac{(a+b)(a+c)}{n} \\
&= \frac{a^2 + ab + ac + bc}{a+b+c+d} \\
&= \frac{a^2 + ab + ac + ad}{a+b+c+d} \\
&= \frac{a(a+b+c+d)}{a+b+c+d} \\
&= a.
\end{aligned}$$

And when 'high' observations of one variable occur together with 'high' observations of the other and, at the same time, 'low' observations occur together with 'low' observations, that is, we have relatively many observations in the 'high, high' and 'low, low' cells, α will be greater than 1, and we say that we have *positive dependence*. When the observations are mostly placed in the 'high, low' and 'low, high' cells,

α will be between 0 and 1, and we have *negative dependence*. We may also look at the *log odds* function, $\theta = \ln \alpha$, which, of course, will be negative for negative dependence, positive for positive dependence and zero for independence. Note that we can replace the relative frequencies with probabilities to get a population measure of dependence for contingency tables (the 'approximate' statements above will then hold exactly).

Finally, note that the (logarithm of the) cross product ratios in a contingency table will remain unchanged when we multiply any row or column with a constant. We say that the ratios are *invariant to marginal replacements*, or that the log odds is a *margin-free* property of a distribution.

Holland and Wang (1987a) introduced a continuous analogue of this measure of dependence, which they called the *local dependence function*:

3.4.2 Defining the local dependence function

Consider (X, Y) with the joint density $f(x, y)$ defined on a (possible infinite) cartesian product set, K. Partition K into a fine rectangular grid, and let $R_{x,y}$ denote the rectangle containing the point (x, y) and having sides of length Δx and Δy. We have

$$P_{x,y} = \mathbb{P}\big((X, Y) \in R_{x,y}\big) \approx f(x, y)\,\Delta x \Delta y, \tag{3.17}$$

where f is the joint density function of (X, Y).

For each rectangle in the grid, pick one pair (x, y) contained in that rectangle. Based on all these pairs, construct a contingency table with the elements $P_{x,y}$. Now consider the four cells (i, k), (i, l), (j, k) and (j, l) in K, with $i < j$ and $k < l$:

$$\begin{array}{cc} (j,k) & (j,l) \\ (i,k) & (i,l) \end{array}$$

The cross-product ratio is

$$\begin{aligned} \alpha\big((i,k),(j,l)\big) &= \frac{P_{i,k}P_{j,l}}{P_{i,l}P_{j,k}} \\ &\approx \frac{f(i,k)\Delta i \Delta k \cdot f(j,l)\Delta j \Delta l}{f(i,l)\Delta i \Delta l \cdot f(j,k)\Delta j \Delta k} \\ &= \frac{f(i,k)\,f(j,l)}{f(i,l)\,f(j,k)}. \end{aligned}$$

Let $\theta\big((i,k),(j,l)\big) = \ln \alpha\big((i,k),(j,l)\big)$. Now let us look at θ near the point (x,y):

$$\frac{1}{\Delta x \Delta y}\theta\big((x,y),(x+\Delta x, y+\Delta y)\big) \tag{3.18}$$

Letting Δx and Δy go towards zero and taking limits, we obtain

$$\begin{aligned}
\gamma(x,y) &= \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{\theta\big((x,y),(x+\Delta x, y+\Delta y)\big)}{\Delta x \Delta y} \\
&= \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \Bigg[\frac{\ln f(x,y) + \ln f(x+\Delta x, y+\Delta y)}{\Delta x \Delta y} \\
&\qquad - \frac{\ln f(x, y+\Delta y) + \ln f(x+\Delta x, y)}{\Delta x \Delta y} \Bigg] \\
&= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \left[\lim_{\Delta y \to 0} \frac{1}{\Delta y} \big(\ln f(x+\Delta x, y+\Delta y) - \ln f(x+\Delta x, y) \big) \right] \\
&\quad - \lim_{\Delta x \to 0} \frac{1}{\Delta x} \left[\lim_{\Delta y \to 0} \frac{1}{\Delta y} \big(\ln f(x, y+\Delta y) - \ln f(x,y) \big) \right] \\
&= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \left[\frac{\partial}{\partial y} \ln f(x+\Delta x, y) - \frac{\partial}{\partial y} \ln f(x,y) \right] \\
&= \frac{\partial^2}{\partial x \, \partial y} \ln f(x,y),
\end{aligned} \tag{3.19}$$

which we will call the *local dependence function*. We assume both partial derivatives are continuous (which implies that they are equal).

Example 3.4.1: Bivariate normal distribution

> Let us look at the nondegenerate bivariate normal distribution $(X,Y) \sim \mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. It has the following density and local dependence function:
>
> $$\begin{aligned}
f(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\Bigg[-\frac{1}{2\,(1-\rho^2)} \Bigg(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 \\
&\qquad - 2\rho \left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \Bigg) \Bigg]. \\
\frac{\partial}{\partial y} \ln f(x,y) &= -\frac{1}{2\,(1-\rho^2)} \left[-2\rho \left(\frac{x-\mu_X}{\sigma_X\sigma_Y}\right) + 2\left(\frac{y-\mu_Y}{\sigma_Y}\right) \right] \\
\gamma(x,y) &= \frac{\partial^2}{\partial x \, \partial y} \ln f(x,y) = \frac{\rho}{1-\rho^2} \frac{1}{\sigma_X\sigma_Y}.
\end{aligned} \tag{3.20}$$
>
> In other words, the (nondegenerate) bivariate normal distribution has constant

local dependence (compare with section 3.3 on page 31, where we showed that it has a constant local correlation curve). We can now easily invert equation (3.20) to express the global correlation ρ as a function of the local correlation:

$$\rho(\gamma) = \frac{-1 \pm \sqrt{1 + 4\left(\gamma\sigma_X\sigma_Y\right)^2}}{2\gamma\sigma_X\sigma_Y} \tag{3.21}$$

We will in section 3.4.5 on page 49 see a complete characterisation of all bivariate distributions with constant local dependence.

Let us now look at an example where we do *not* have constant local dependence, but where the dependence is a function of only one of the variables:

Example 3.4.2: Not constant dependence

Let $X \sim \mathcal{N}(0, \sigma_X^2)$, and let $Y = X^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. This is similar to the example in section 2.4.1 on page 21, but we have added some noise so that the (bivariate) distribution has a density and an associated local dependence function.

A measure of local dependence should *intuitively* be negative for negative values of x and positive for positive values of x, be increasing in $|x|$ and be decreasing in σ_ϵ. And a quick calculation shows that this is true for the local dependence function:

First, note that $Y \mid X = x \sim \mathcal{N}(x^2, \sigma_\epsilon^2)$. We have:

$$\begin{aligned}
\gamma(x, y) &= \frac{\partial^2}{\partial x\, \partial y} \ln f_{X,Y}(x, y) \\
&= \frac{\partial^2}{\partial x\, \partial y} \ln \left[f_X(x) f_{Y|X}(y \mid x) \right] \\
&= \frac{\partial^2}{\partial x\, \partial y} \ln f_{Y|X}(y \mid x) \\
&= \frac{\partial^2}{\partial x\, \partial y} \ln \frac{1}{\sqrt{2\pi}\,\sigma_\epsilon} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(y - x^2)^2 \right] \\
&= \frac{\partial^2}{\partial x\, \partial y} \frac{-1}{2\sigma_\epsilon^2}(y - x^2)^2 = \frac{\partial}{\partial x} \frac{-2}{2\sigma_\epsilon^2}(y - x^2) \\
&= \frac{2x}{\sigma_\epsilon^2}.
\end{aligned} \tag{3.22}$$

Note that we have not used that X is normal in the derivation; the expression will be valid regardless of the distribution of X (but its domain will vary, and will always be equal to the support of X).

As the calculations in the previous example showed, the local dependence function only depends on the conditional distribution of Y given X or of X given Y. In other words, it is invariant to marginal replacements, just like its discrete analogue.

3.4.2.1 Another way of deriving the local dependence function

Jones (1996) suggested a completely different way of deriving the local dependence function, based on localising the correlation coefficient. Define the convenience function

$$M_{x_0,y_0}(A,B) = \mathbb{E}\big(w_{x_0,y_0}(X,Y)AB)\big) - \frac{\mathbb{E}\big(w_{x_0,y_0}(X,Y)A\big)\,\mathbb{E}\big(w_{x_0,y_0}(X,Y)B\big)}{\mathbb{E}\big(w_{x_0,y_0}(X,Y)\big)}, \quad (3.23)$$

where A and B are two random variables. One 'natural' way of localising the correlation is

$$\psi(x_0,y_0) = \frac{M_{x_0,y_0}(X,Y)}{\sqrt{M_{x_0,y_0}(X,X)M_{x_0,y_0}(Y,Y)}}. \quad (3.24)$$

When $w_{x_0,y_0}(x,y) = 1$ for all (x,y), we get the usual correlation coefficient $\rho_{X,Y}$. A different wx_0,y_0, which localises the correlation around (x_0,y_0), is a bivariate density:

$$w_{x_0,y_0}(x,y) = \frac{1}{h_1h_2}K^*\left(\frac{x_0-x}{h_1},\frac{y_0-y}{h_2}\right). \quad (3.25)$$

We see that the function K^* is also a density, or kernel, and we will call h_1 and h_2 the *bandwidths*. Jones suggested using a product density of identical symmetrical marginal distributions. We will now follow Jones's derivation of the local dependence function, but with some added details to make the exposition clearer.

Let $f = f(x,y)$ be the density of (X,Y). We are now interested in finding a measure of local dependence based on equation 3.24, and we first define

$$g_{ij}(x_0,y_0) = \iint u^i v^j K(u)K(v)f(x_0-h_1u,y_0-h_2v)\,\mathrm{d}u\,\mathrm{d}v. \quad (3.26)$$

Using standard Taylor approximation, we have

$$\begin{aligned} f(x_0-h_1u,y_0-h_2v) &= f(x_0,y_0) - uh_1f^x(x_0,y_0) - vh_2f^y(x_0,y_0) \\ &\quad + \tfrac{1}{2}u^2h_1^2f^{xx}(x_0,y_0) + \tfrac{1}{2}v^2h_1^2f^{yy}(x_0,y_0) + uvh_1h_2f^{xy}(x_0,y_0) + o(h_1^2+h_2^2), \end{aligned}$$

where $f^{xy} = \frac{\partial^2 f}{\partial x \partial y}$, and similar for the other functions. When $h_1, h_2 \to 0$, equation 3.26 on the facing page becomes

$$
\begin{gathered}
g_{00} \sim f, \quad g_{10} \sim -h_1 s_2 f^x, \\
g_{01} \sim -h_2 s_2 f^y, \quad g_{11} \sim h_1 h_2 s_2^2 f^{xy}, \\
g_{20} \sim s_2 f, \quad g_{02} \sim h_2 s_2 f,
\end{gathered}
$$

where $s_2 = \int u^2 K(u)\,\mathrm{d}u$ (and we leave the dependence of (x_0, y_0) implicit). The first term of equation 3.23 on the preceding page contains

$$
\begin{aligned}
\mathbb{E}\big(w_{x_0,y_0}(X,Y)XY\big) &= \iint xy \frac{1}{h_1 h_2} K\bigg(\frac{x_0 - x}{h_1}\bigg) K\bigg(\frac{y_0 - y}{h_1}\bigg)\,\mathrm{d}x\,\mathrm{d}y \\
&= \iint (x_0 - h_1 u)(y_0 - h_2 v) \frac{1}{h_1 h_2} K(u)K(v)\,\mathrm{d}u\,\mathrm{d}v,
\end{aligned}
$$

which, using the g_{ij} notation, can be written

$$
x_0 y_0 g_{00} - h_1 y_0 g_{10} - h_2 x_0 g_{01} + h_1 h_2 g_{11}.
$$

Similarly, we can write

$$
\frac{\mathbb{E}\big[w_{x_0,y_0}(X,Y)X\big]\,\mathbb{E}\big[w_{x_0,y_0}(X,Y)Y\big]}{\mathbb{E}\big[w_{x_0,y_0}(X,Y)\big]}
$$

as

$$
\frac{\big(x_0 g_{00} - h_1 g_{10}\big)\big(y_0 g_{00} - h_2 g_{01}\big)}{g_{00}},
$$

so the numerator of equation 3.24 on the facing page is

$$
\begin{aligned}
h_1 h_2 \big(g_{11} - \frac{g_{01} g_{10}}{g_{00}}\big) &\sim h_1 h_2 \left(h_1 h_2 s_2^2 f^{xy} - \frac{1}{f} h_1 h_2 s_2^2 f^x f^y\right) \\
&= h_1^2 h_2^2 s_2^2 \left(f^{xy} - \frac{f^x f^y}{f}\right),
\end{aligned}
$$

and the denominator is the square root of the product of

$$
x_0^2 g_{00} - 2 h_1 x_0 g_{10} - h_1^2 g_{20} - \frac{(x_0 g_{00} - h_1 g_{10})^2}{g_{00}} = h_1^2 \left(g_{20} - \frac{g_{10}^2}{g_{00}}\right) \sim h_1^2 s_2 f
$$

and $h_2^2 s_2 f$. This gives us

$$
h_1 h_2 s_2 \frac{1}{f}\left(f^{xy} - \frac{f^x f^y}{f}\right),
$$

where all the functions are evaluated at (x_0, y_0). Renormalising, by dividing by the constant $h_1 h_2 s_2$, gives us

$$\frac{1}{f}\left(f^{xy} - \frac{f^x f^y}{f}\right)\bigg|_{(x_0,y_0)} = \frac{\partial^2}{\partial x\,\partial y} \ln f(x,y)\bigg|_{(x,y)=(x_0,y_0)} = \gamma(x_0, y_0). \tag{3.27}$$

3.4.3 Properties of the local dependence function

Here are some properties of the local dependence function, noted (and proved) by Holland and Wang (1987a):

Independence and zero local dependence From equation 3.19 on page 44, we immediately see that $\gamma(x,y) = 0$ for all (x,y) if and only if X and Y are independent.

A margin-free measure The local dependence is a function only of the conditional distribution of X given Y, or of Y given X. We say that γ is a *margin-free* measure of dependence. The proof is obvious.

Unique joint distribution For any integrable function $\gamma(x,y)$ defined over $K = (a,b) \times (c,d)$, and for any continuous densities $f(x)$ and $g(y)$ defined on (a,b) and (c,d), respectively, there is a unique joint density $f(x,y)$ on K with local dependence $\gamma(x,y)$ and marginal distributions $f(x)$ and $g(y)$.

Sankaran and Gupta (2004) also showed that the local dependence function along with the conditional means, $\mathbb{E}(Y \mid X = x)$ and $\mathbb{E}(X \mid Y = y)$, characterised at least certain bivariate distributions, but did not supply a proof of the general case.

3.4.4 Ordering of dependence

The local dependence function can sometimes also be used to compare the overall dependence between different distributions. Holland and Wang (1987a) suggested the following dependence ordering:

Definition 3.4.3: Positive dependence

> Let f and g be two bivariate densities on the same support K, and with the same marginal distributions. We say that g is more positively dependent than f if and only if
>
> $$\gamma_f(x,y) \le \gamma_g(x,y) \quad \text{for all } (x,y) \in K.$$

We will denote this by $f \preceq g$. We will also write $(X, Y) \preceq (W, Z)$ if and only if $f \preceq g$, where f and g are the joint density functions of (X, Y) and (W, Z), respectively.

If the inequality is strict for at least one (x, y), we say that g is strictly more positively dependent. And if the two densities have different marginals, we can transform them in order to compare their dependence.

Theorem 3.4.4: The ordering $\preceq$ is a partial ordering

The ordering $\preceq$ is a partial ordering, that is

1. $f \preceq f$ (reflectivity),
2. $f \preceq g$ and $g \preceq h$ implies $f \preceq h$ (transitivity) and
3. $f \preceq g$ and $g \preceq f$ implies $f = g$ (antisymmetry).

See the cited article for further details.

3.4.5 Constant local dependence

Recall that in example 3.4.1 on page 44, we showed that the bivariate normal distribution has constant local dependence. Jones (1998) found that the distributions having constant local dependence θ are precisely the distributions that have the joint density

$$f(x, y) = a(x; \theta)\, b(y; \theta)\, e^{\theta x y}, \tag{3.28}$$

where $a(x; \theta)$ and $b(y; \theta)$ are arbitrary functions (only restricted so that $f(x, y)$ is a real density). We now easily see that the conditional density of Y given $X = x$ is proportional to

$$b(y; \theta)\, e^{\theta x y}.$$

More details and one additional example of a distribution with constant local dependence can be found in Jones (1998). With the exception of this example (and the bivariate normal distribution), Jones did not find any other distributions with the property of constant local dependence in the literature; and while it is possible to create new distributions based on equation 3.28, 'it is unclear why such distributions might be useful apart from having constant dependence' (Jones 1998).

3.4.6 Dependence maps

The local dependence function is a function of both x and of y, and it is difficult to graph and to interpret. Its values have no natural units, and are affected by scaling. One example is equation 3.20 on page 44, where we showed that the local dependence function is a decreasing function of the standard deviations of X and Y in the bivariate case.

Jones and Koch (2003) suggested plotting a contour plot of (an estimate of) $\sigma_X \sigma_Y \gamma$ instead of a contour plot of γ. They also noted that other (robust) measures of scales could be used. But the main part of their paper was devoted to a type of plot they named *dependence maps*.

A dependence map is a graphical display of an estimate of the local dependence function on a two-dimensional grid over the values of X and Y we are interested in. At each grid point, a colour representation of the value of the local dependence function is shown. The colours are chosen so that it is easy to see if the local dependence is positive or negative, and to compare the strength of the dependence in various areas. Values where $\hat{f}$ is near zero are not shown, for two reasons: The local dependence is of little interest at these points, and its estimate is very unreliable here.

One alternative, which Jones and Koch (2003) argued for, is to only use three colours, to indicate values that are statistically significantly positive, statistically significantly negative or not statically significantly different from 0 (for a user-chosen significance level). Jones and Koch used a permutation test to calculate values were statistically significant, and the reader is referred to their article for details.

In contrast to Jones and Koch (2003), we will not use any significance tests in our dependence maps; instead, we will use a (small) palette of colours to indicate varying levels of local dependence. The colours used are based on the recommendations of (Cleveland 1994, pages 230-233). But we note that a similar approach *could* be used for the significance-based procedure: We could use different colours for values that are significant at various levels (for example, at the 0.01, 0.05, 0.10, 0.2 and 0.4 level). The result of this procedure would be a dependence map that is similar to our dependence map, but where the colour 'axis' has been scaled nonlinearly.

Finally, we note that we can also construct three-dimensional dependence maps, where the third dimension shows the estimated local dependence. See section 3.4.6.2 on page 59 for a short discussion, and an example, of such 3D dependence maps.

3.4.6.1 Estimating the local dependence

To estimate the local dependence function, we can use equation 3.24 on page 46 and replace each expectation with the corresponding sample average. However, Jones (1996) used a different (and incorrect) denominator. The corrected estimator, which appears (in a different notation) in Jones and Koch (2003), is:

$$\hat{\gamma}(x_0, y_0) = \frac{\hat{g}_{11}(x_0, y_0) - \frac{\hat{g}_{01}(x_0, y_0)\hat{g}_{10}(x_0, y_0)}{\hat{g}_{00}(x_0, y_0)}}{h_1^2 h_2^2 s_2^2 \hat{g}_{00}(x_0, y_0)}, \tag{3.29}$$

where

$$\hat{g}_{ij}(x_0, y_0) = \frac{1}{n} \sum_{k=0}^{n} X_k^i Y_k^j \frac{1}{h_1} K\left(\frac{X_k - x_0}{h_1}\right) \frac{1}{h_2} K\left(\frac{Y_k - y_0}{h_2}\right) \tag{3.30}$$

and s_2 is defined as before: $s_2 = \int u^2 K(u)\, \mathrm{d}u$. The estimator can also be derived by fitting the bilinear form $a + bx + cy + dxy$ to the log density *locally*. Details can be found in Jones and Koch (2003).

Jones and Koch (2003) also suggested using the following rule of thumb for bandwidth selection:

$$h_i = \frac{\sigma_i}{n^{\frac{1}{6}}} \left(\frac{2\sqrt{\pi} \int K^2(u)\, \mathrm{d}u}{\int u^2 K(u)}\right)^{\frac{1}{3}} \frac{(1 - \rho^2)^{\frac{5}{12}}}{(1 + \rho^2/2)^{\frac{1}{6}}}, \quad i = 1, 2. \tag{3.31}$$

Since the variances and the correlation are not usually known, they can be estimated using sample variances and sample correlation.

Larger bandwidths give smoother dependence maps (less variance) but larger bias, while smaller bandwidths give dependence maps that fluctuates more, with 'spots' of the opposite colour frequently occurring. Simulations seem to indicate that using slightly higher bandwidths than the ones suggested by the rule of thumb often works well, especially when it is reasonable to expect that the local dependence is a smooth function, without fast fluctuations.

Here is my implementation of the local dependence estimator and of the dependence map:

```
# x = x vector.
# y = y vector.
# points = number of grid points the local dependence should be estimated at.
# levels = (maximum) number of different levels used in the dependence maps.
# plot.zero = plot values close to zero as white if TRUE. If FALSE (the default),
# all values will be either magenta (positive) or cyan (negative).
# bw.scale = scale the automatically selected bandwidths by this vector.
# quant = do not plot estimate where the density is less than this quantile of
# the discretised density.
locest = function( x, y, points = 50, levels = 7, plot.zero = FALSE, bwscale = c(1,1), quant = .6 )
{
  require("lattice") # Needed for plotting the dependence map.

  n = length(x)

  levels = 2∗floor(levels/2) # Use an even number of levels.
  if( plot.zero ) levels = levels + 1 # Or an odd number of levels, so 0 = white.

  xmi = min(x) # Range of x and y, for use later.
  xma = max(x)
  ymi = min(y)
  yma = max(y)

  k = function(u) # Kernel function.
    {
      ifelse(u^2<1,(15/16)∗(1−u^2)^2,0)
    }

  kh = function(h, x) # Scaled kernel function.
    {
      k(x/h)/h
    }

  h = function(i) # Calculate automatic bandwidth.
    {
      if(i == 1)
        s = sd(x) else s = sd(y)
      r = cor(x,y)
      s∗n^(−1/6)∗(10∗sqrt(pi))^(1/3)∗(1−r^2)^(5/12)/(1+r^2/2)^(1/6)
    }

  h1 = h(1)∗bwscale[1] # Calculate bandwidths, and scale them if necessary.
  h2 = h(2)∗bwscale[2]
  k2 = 1/7

  gx = seq(xmi, xma, length = points) # Define the grid points where ...
  gy = seq(ymi, yma, length = points) # ... estimates will be calculated.

  # Now calculate the vectors used in the various estimates:
  xy = x ∗ y
  xx = outer(gx,x,"−") # Matrix of (x_0 − x_i) for all grid points x_0 and all x values x_i.
  yy = outer(gy,y,"−") # Similar for y values.
```

```
kh1 = kh(h1,xx)
kh2 = kh(h2,yy)
fxy = (kh1 %*% (xy*t(kh2))) / n # Sum the product of X_k^i Y_k^j and the two kernels ...
f1 = (kh1 %*% t(kh2)) / n # ... over all data pairs, and divide by n.
fx = (kh1 %*% (x*t(kh2))) / n # Do this for all the four functions (fxy, f1, fx, fy) ...
fy = (kh1 %*% (y*t(kh2))) / n # ... needed in the estimate.

gam = (fxy - fx*fy/f1)/ # Finally estimate the local dependence function.
      ((h1*h2*k2)^2 * f1)

cutoff = quantile(f1,quant) # Determine cutoff value.
remv = f1 < cutoff
gam[remv] = NA # Mark all values where the density is lower than the cutoff as NA.
gam = gam * sd(x) *sd(y) # Scale/normalise the estimated local dependence function.

xp <- yp <- x
for(k in 1:n) # Determine which grid point correspond to each data pair.
  {
    xp[k] = which.min(gx <= x[k])
    yp[k] = which.min(gy <= y[k])
  }

pcol = remv[cbind(xp,yp)] # For each data pair, determine whether the estimated ...
                          # ... local dependence should be plotted or not.

pcol = ifelse(pcol, "white", "black") # Determine colour of datapoints in scatterplot,
                                      # depending on the background colour of the
                                      # corresponding grid point. Note: When a datapoint
                                      # falls on a boundary, parts of it may not be visible.


# Determine the highest observed local dependence (in absolute value).
# This will be used to select the colours in the dependence map so that
# zero dependence will be in the middle of the colour spectrum.
tp = max(abs(c(min(gam[!remv & !is.na(gam)]),max(gam[!remv & !is.na(gam)]))))

# Finally plot the dependence map, with all the data points overlaid.
l = levelplot(gam~x*y, expand.grid( x = gx, y = gy),
              panel = function(...) {
                panel.fill(col = "black")
                panel.levelplot(...)
                panel.xyplot(x,y, col=pcol)
              }, at = seq(-tp, tp, length = levels + 1),
              col.regions = cm.colors(levels),
              )
  print(l)
}
```

Figure 3.3 on page 56 shows the average of 1000 dependence maps based on the same model as in figure 3.2 on the facing page (see the caption for calculation details). The map is an estimate of the expected value of the dependence map under this model.

Recall example 3.4.2 on page 45, where we looked at the local dependence between X and $Y = X^2 + \epsilon$, where ϵ was independent of X. Figure 3.2 on the facing page shows a dependence map calculated with the above algorithm, with X and ϵ both having standard normal distributions. The dependence map captures the negative dependence for negative X and the positive dependence for positive X. It also seems to *indicate* that the local dependence may increase with $|X|$, but additional simulations of the same model shows that this feature is often not visible on the dependence map.

This is also true for simulations for a number of other models that I have tried (not mentioned in this thesis). Whilst the dependence map usually manages to estimate the sign of the local dependence, its size is much more difficult to estimate, even with very many (thousands of) observations. For estimating the sign, a sample size of 150 generally seems to suffice, except for distributions with complex dependence, that is, a rapidly fluctuating local dependence sign.

Note that the value (colour) at each grid point should not be interpreted as a measure of local dependence at the point, but as a measure of local dependence in an larger area around the point. Since it is often natural to assume that the local dependence function is a continuous function, this interpretation seems reasonable.

In the two earlier examples we looked at, the local dependence function was either constant (example 3.4.1 on page 44) or depended on only *one* of the two variables (example 3.4.2 on page 45). Let us therefore now look at an example where the local dependence changes as a function of both variables:

Example 3.4.5: Local dependence in Cauchy distribution

> Let (X, Y) have a bivariate Cauchy distribution (this was also used as an example in Holland and Wang 1987a),
>
> $$\begin{aligned} X &= \frac{Z_1}{W} \quad \text{and} \\ Y &= \frac{Z_2}{W}, \end{aligned}$$

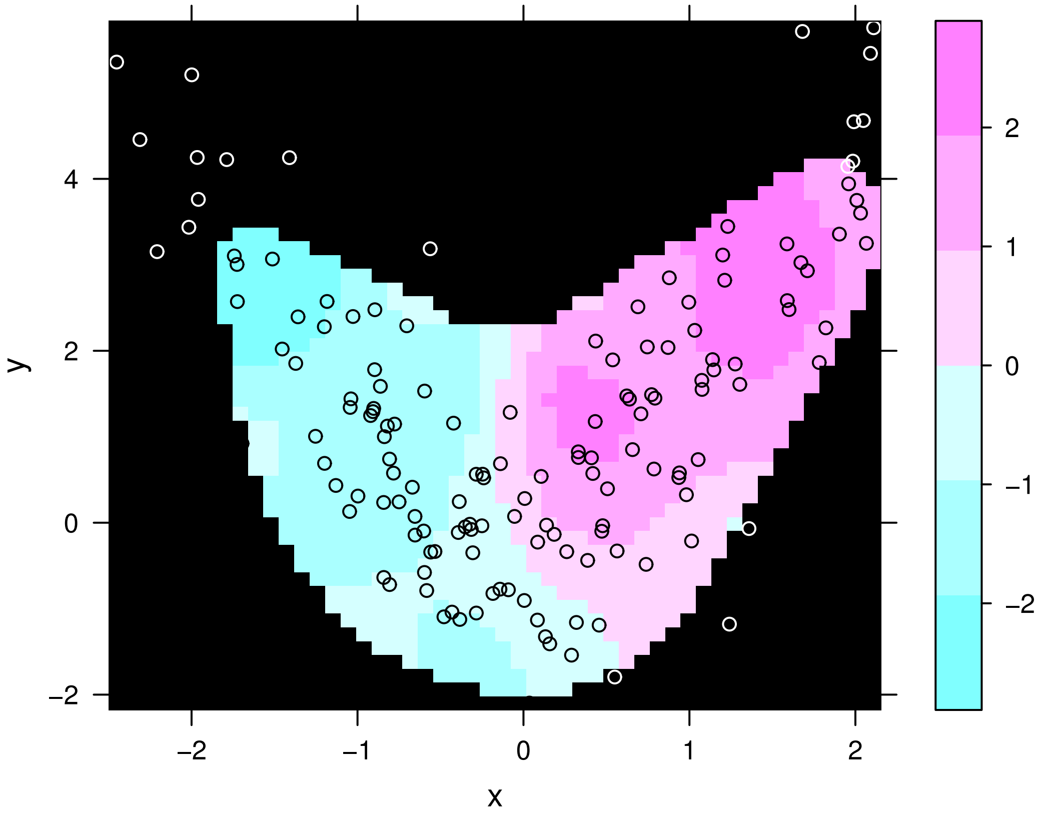


Figure 3.2: Dependence map of 150 observations from the distribution of (X, Y), where $X \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$ and $Y = X^2 + \epsilon$, $\epsilon \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$ and independent of X. (This is a special case of example 3.4.2 on page 45.) Magenta-coloured and cyan-coloured areas indicate (estimated) positive and negative local dependence, respectively, and black areas indicate low-density areas (where we do not have reliable estimates of the local dependence). The black and white circles are the 150 observations. The dependence map captures the negative dependence for negative X and the positive dependence for positive X. It also *indicates* that the local dependence may increase with $|X|$, but additional simulations of the same model shows that this feature is often not visible on the dependence map.

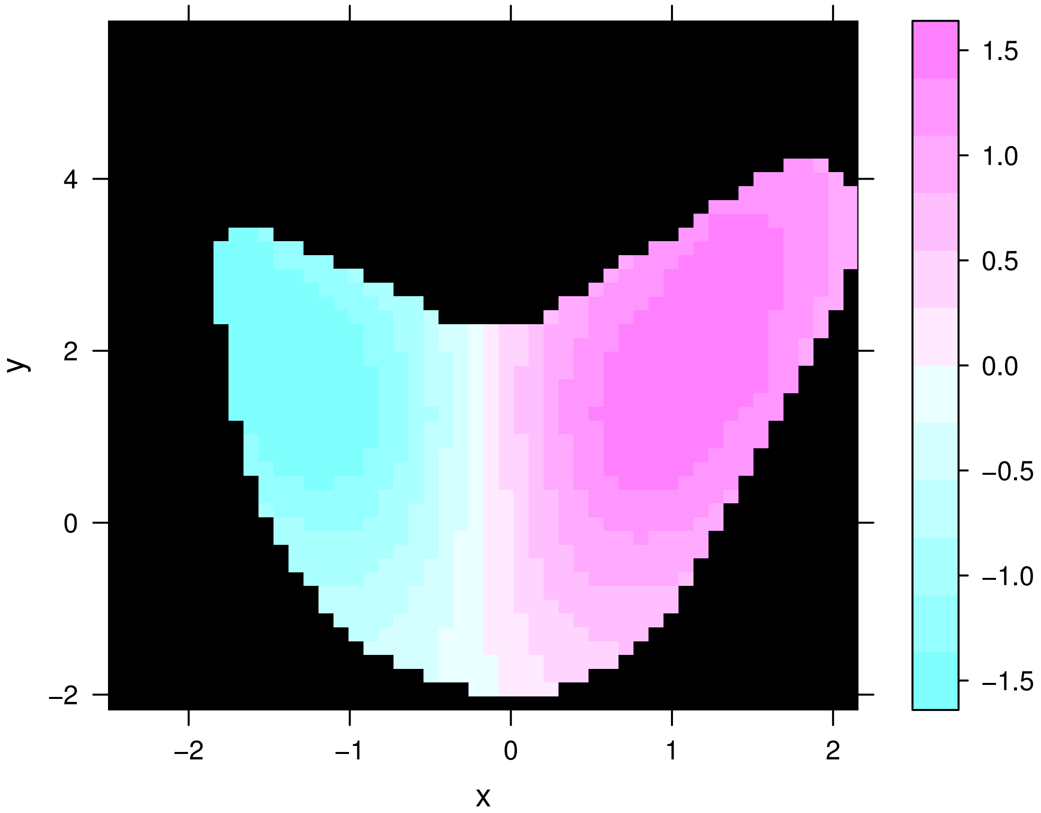


Figure 3.3: Average dependence map of 150 observations from the distribution of (X, Y), where $X \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$ and $Y = X^2 + \epsilon$, $\epsilon \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$ and independent of X (see figure 3.2 on the preceding page). The map was made by creating 1000 dependence maps from this model, before calculating the average value for each map point. The black area was made identical to the one in figure 3.2, to aid in comparing the two maps; however, do note that the number of colours used has been doubled, and the range the colours represent has been changed. This was done to increase the 'depth' resolution of the dependence map. We see that the average dependence map captures the increasing (in absolute value) local dependence for high x's, but has some problems near the boundary – the edge of the map. This is to be expected.

where Z_1, Z_2 and W are independent, Z_1 and Z_2 are standard normal and W^2 has a chi-square distribution with one degree of freedom. It can be shown (Dunnett and Sobel 1954) that the density of (X, Y) is

$$f(x,y) = c(x^2 + y^2 + c)^{-\frac{3}{2}}. \tag{3.32}$$

Now we can easily calculate the local dependence function

$$\begin{aligned}\gamma_g(x,y) &= \frac{\partial^2}{\partial x\,\partial y} \ln f(x,y) \\ &= \frac{\partial^2}{\partial x\,\partial y}\left[\ln c - \frac{3}{2}\ln(x^2 + y^2 + c)\right] \\ &= \frac{\partial}{\partial x}\left[-\frac{3y}{x^2 + y^2 + c}\right] \\ &= \frac{6xy}{(x^2 + y^2 + c)^2}. \end{aligned} \tag{3.33}$$

We see that the local dependence function changes sign in the origin; it is positive in the first and third quadrant, and negative in the second and fourth quadrant.

Because of its heavy tails, the Cauchy distribution frequently gives us values that are extremely large compared to the bulk of observations (for example, the central 90% of observations). This makes it much more difficult to use scatterplots and dependence maps to judge the dependence present in a data set (since most of the data is squeezed into a tiny area on the plot). Figure 3.4 on the next page illustrates this with a scatterplot from the bivariate Cauchy distribution.

One alternative is to transform the data, for example by a probability integral transformation, where we look at (observations from) $U = F(X)$ and $V = F(Y)$, where F is the cumulative distribution functions of X (and of Y), instead of looking at (observations from) X and Y directly. We will in section 4.5 on page 78 show that the *sign* of the local dependence function at each point will be the same at the transformed points. For this example, this means that, since the local dependence function of (X, Y) changes sign at the medians (0.5 quantiles) of the variables, the local dependence function of the transformed variables (U, V) will change sign at the point $(0.5, 0.5)$.

In the bivariate Cauchy distribution, we do of course *know* the (univariate) cumulative distributions, or can at least compute it numerically for each value, but in most applications the marginal distributions will be as unknown to us as the

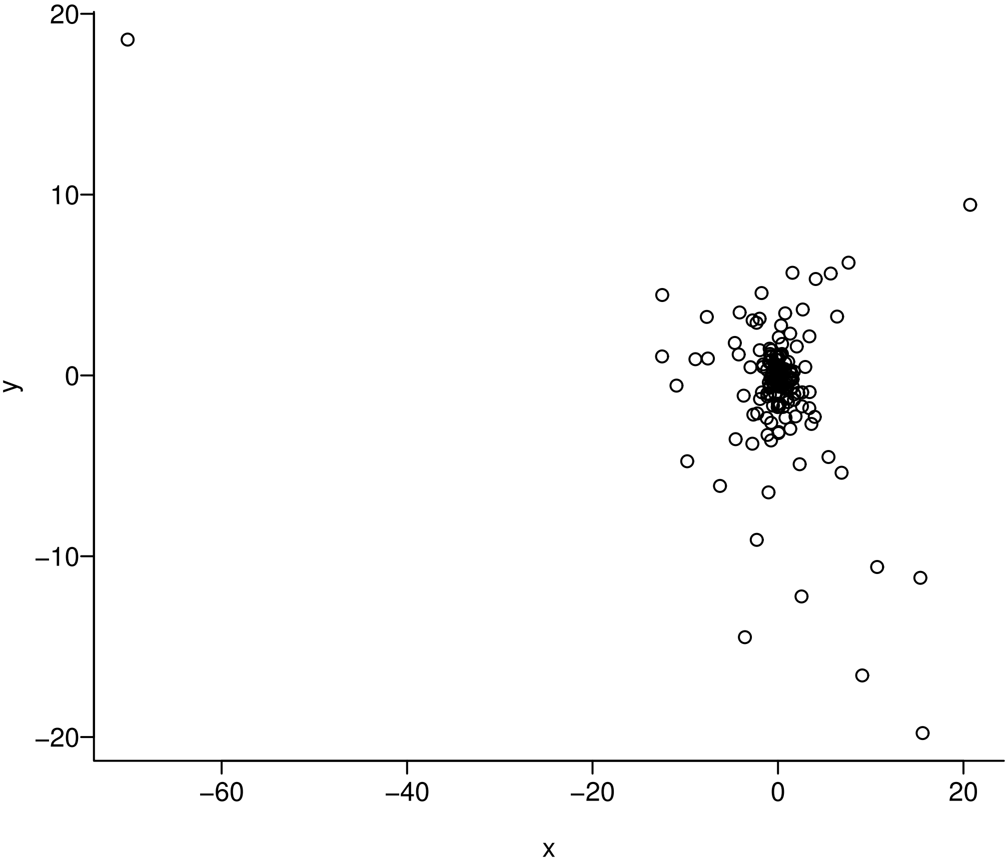


Figure 3.4: Scatterplot of 150 observations from the bivariate Cauchy distribution in example 3.4.5 on page 54. The bulk of the observations are clustered in a small area, a few observations are scattered some distance off, while one observation (top right) is very far away from the rest. Most of the plot is empty space, and all this makes it difficult to see how observations inside the main cluster are distributed, for example, if there is dependence among them.

joint distributions (and, consequently, as the local dependence functions). However, we can use very simple and easy-to-compute estimates of the transformed variables. Let $x_{(1)}, \dots, x_{(n)}$ be the sorted x_i values. We now replace each $x_{(i)}$ by it transformed rank value (see, for instance, Cleveland 1994, pages 136–137):

$$u_i = \frac{i - 0.5}{n}. \tag{3.34}$$

For example, the middle of the $x_{(i)}$ values (assuming n is odd) will get a u value of .5. We define the v values similarly, based on $y_{(i)}$ values. A plot (scatterplot or dependence plot) of the u values against the v values (both reordered so that the pairs (u_i, v_j) match the original $(x_{(i)}, y_{(j)})$ pairs) will now be similar to a plot using real probability integral transformed values; see 5.1b and figure 5.1c on page 82 for a comparison for scatterplots.

Figure 3.5 on the following page shows a dependence map based on 150 transformed observations from the bivariate Cauchy distribution. We see that the sign of the local dependence differ in odd-numbered and even-numbered quadrants, but exactly where the change occurs is not clear (from looking at the map). Further simulations indicate that increasing the bandwidths tend to make the estimates (of the sign) somewhat better.

3.4.6.2 3D dependence maps

As we have already mention, it is possible to draw tree-dimensional plots of the local dependence function. Figure 3.6 on page 61 shows an example. This is similar to a dependence map, but the colour 'axis' has been drawn as a real vertical axis. Seeing the *3D dependence map*, perhaps from various angles, may help in visualising how the dependence varies over the support (it is, for example, easier to judge how *fast* the dependence changes by looking at the slopes), but the map provides little additional *quantitative* information not found in the usual two-dimensional dependence map.

Also, parts of the 3D dependence map ('valleys') may be obscured by other parts ('hills'), so the two-dimensional map may be the preferred graphical display of the local dependence function.

But, of course, a combination of *both* a scatterplot, a density estimate contour plot, a 2D and a 3D dependence map, and other graphical displays (see, for instance, chapter 5 on page 81), along with various numerical measures, will be the a convenient approach to exploring the dependence in bivariate distributions.

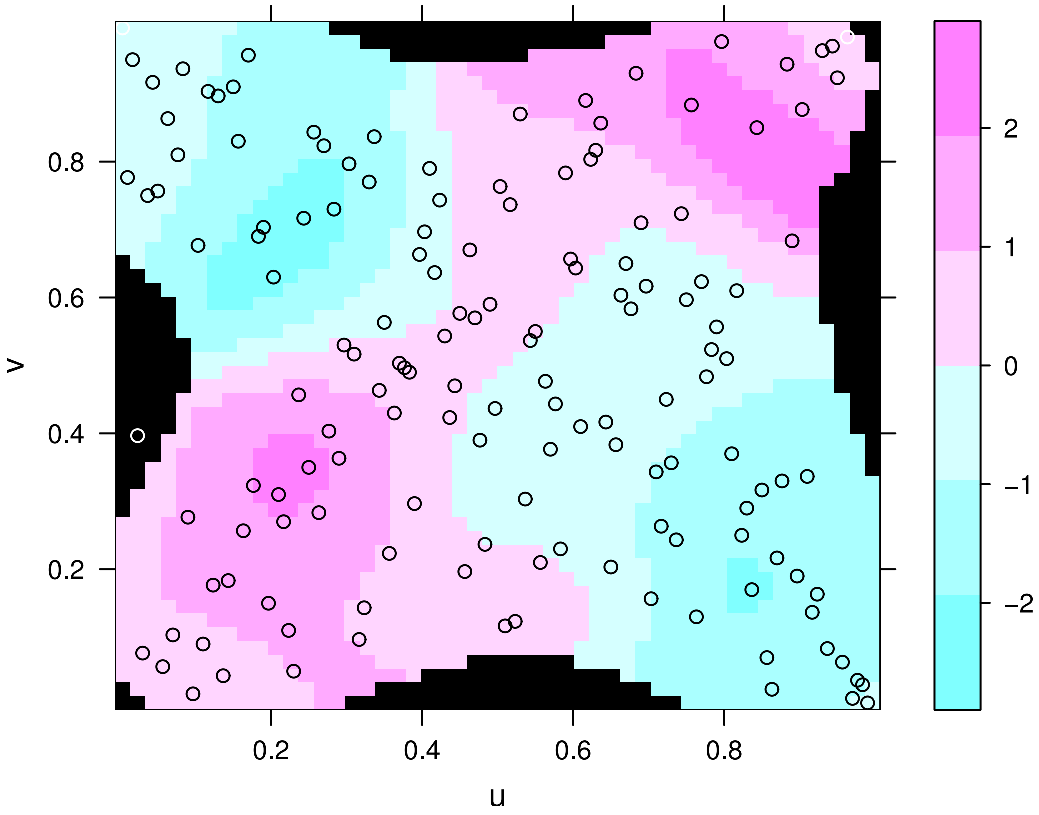


Figure 3.5: Dependence map of 150 transformed observations from the bivariate Cauchy distribution in example 3.4.5 on page 54. The data set is exactly the same as the one in figure 3.4 on page 58, but the observations have been transformed using the rank transformation in equation 3.34 on the preceding page, so that, for instance, the observations (u, v) near $(0.5, 0.5)$ will correspond to pairs of values (x, y) where x is approximately the median of the x_i's and y is approximately the median of y_i's. The dependence map has been calculated using the default options, except for the argument 'quant', which has been set to 0.1, to ensure that most of the map is filled with estimates (even where the estimates are mainly based on a few observations). As before, magenta-coloured and cyan-coloured areas indicate (estimated) positive and negative local dependence, respectively, and black areas indicate low-density areas (where we do not have reliable estimates of the local dependence). The black and white circles are the 150 observations. We see that the sign of the local dependence differ in odd-numbered and even-numbered quadrants, but exactly where the change occurs is not clear (from looking at the map). Increasing the bandwidths tend to make the estimates (of the sign) somewhat better.

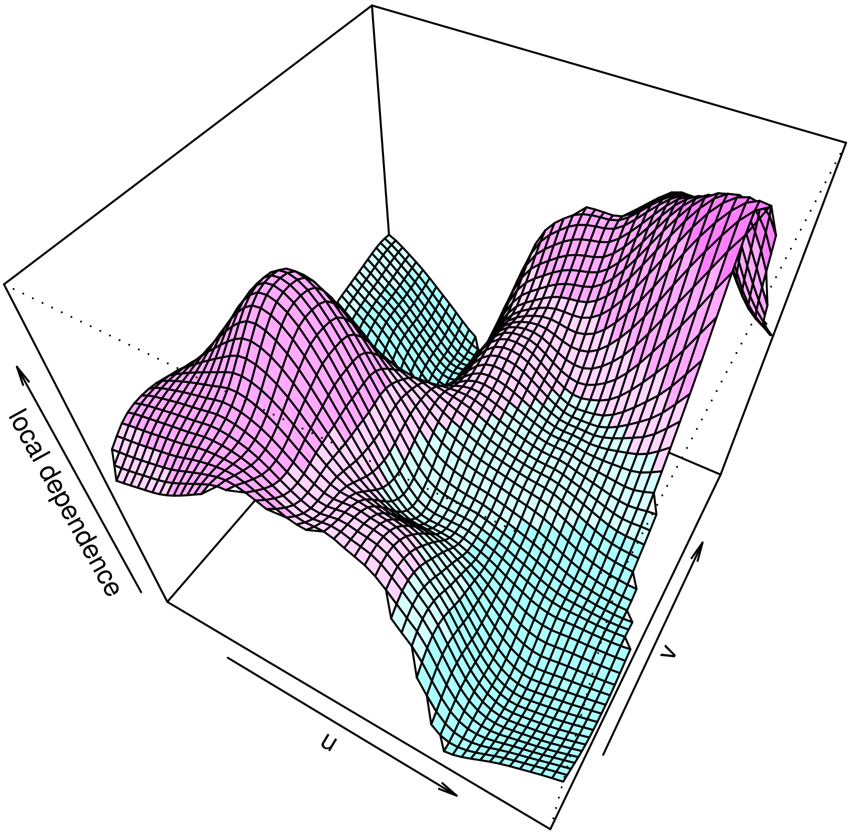


Figure 3.6: Three-dimensional dependence map of 150 transformed observations from the bivariate Cauchy distribution in example 3.4.5 on page 54. The data set is exactly the same as the one in figure 3.4 on page 58, but the observations have been transformed using the rank transformation in equation 3.34 on page 59, so that, for instance, the observations (u, v) near $(0.5, 0.5)$ will correspond to pairs of values (x, y) where x is approximately the median of the x_i's and y is approximately the median of y_i's. The 3D dependence map shows the same information as the ordinary dependence map in figure 3.5 on the facing page, but gives a more immediate impression of how the local dependence changes over the support. And as before, magenta-coloured and cyan-coloured areas indicate (estimated) positive and negative local dependence, respectively. The colours used are the same six as in the two-dimensional dependence map, but their numerical values, that is, where on the vertical axis the change from one colour to another occurs, are not exactly identical (except for the change between magenta and cyan). The actual colours used here are less important, anyway, since we can see *directly* how the estimated local dependence changes.

3.4.7 Local dependence in transformations

We will now look at the local dependence function in linear transformations. Let (Z_1, Z_2) be two variables with density f_{Z_1,Z_2} and local dependence γ_{Z_1,Z_2}, and let $X = a_1 + b_1 Z_1$ and $Y = a_2 + b_2 Z_2$. It is now trivial to show that the density of (X, Y) is

$$f_{X,Y}(x, y) = \frac{1}{b_1 b_2} f_{Z_1,Z_2}\left(\frac{z_1 - a_1}{b_1}, \frac{z_2 - a_2}{b_2}\right).$$

It follows that the local dependence function of (X, Y) is

$$\begin{aligned}\gamma_{X,Y}(x, y) &= \frac{\partial^2}{\partial x\,\partial y} \ln \frac{1}{b_1 b_2} f_{Z_1,Z_2}\left(\frac{x - a_1}{b_1}, \frac{y - a_2}{b_2}\right) \\ &= \frac{\partial^2}{\partial x\,\partial y} \ln f_{Z_1,Z_2}\left(\frac{x - a_1}{b_1}, \frac{y - a_2}{b_2}\right) \\ &= \frac{1}{b_1 b_2} \gamma_{Z_1,Z_2}\left(\frac{x - a_1}{b_1}, \frac{y - a_2}{b_2}\right). \qquad (3.35)\end{aligned}$$

We see that the local dependence function scales as the product of the reciprocals of the scale parameters b_1 and b_2. And observe that example 3.4.1 on page 44, where we looked at the local dependence in the bivariate normal distribution, is a special case of this result.

Holland and Wang (1987a) also noted a generalisation of this result to one-to-one transformations. If $S = c_1(X)$ and $T = c_2(Y)$, where c_1 and c_2 are one-to-one and differentiable functions, we have (with similar notation as before):

$$\gamma_{S,T}(s, t) = \left[\frac{\mathrm{d}}{\mathrm{d}s} c_1^{-1}(s)\right] \left[\frac{\mathrm{d}}{\mathrm{d}t} c_2^{-1}(t)\right] \gamma_{X,Y}\big(c_1^{-1}(s), c_2^{-1}(t)\big). \qquad (3.36)$$

The proof can not be found in the article, but it is similar to the one for linear transformations, and straightforward.

3.4.8 Local dependence in mixtures

Consider a mixture of distributions; that is, let (X, Y) have the density

$$g(x, y) = \sum_{i=1}^{n} p_i f_i(x, y), \qquad (3.37)$$

where $f_i(x, y)$ are bivariate density functions (with existing local dependence functions) and p_i are probabilities that sum to 1. We will now try to express the local dependence of the mixture distribution as a function of the local dependence in the subdistributions. To make the notation simpler, we denote $g(x, y)$ by g, $\frac{\partial^2}{\partial x \partial y} f_i(x, y)$ by f_i^{XY}, $\frac{\partial}{\partial x} f_i(x, y)$ by f_i^X, and $\frac{\partial}{\partial y} f_i(x, y)$ by f_i^Y. The local dependence function of (X, Y) can now be written

$$\begin{aligned}
\gamma_g(x, y) &= \frac{\partial^2}{\partial x \partial y} \ln g(x, y) \\
&= \frac{\partial}{\partial x} \frac{1}{g} \sum_{i=1}^{n} p_i f_i^Y \\
&= -\frac{1}{g^2} \left(\frac{\partial}{\partial x} g \right) \sum_{i=1}^{n} p_i f_i^Y + \frac{1}{g} \sum_{i=1}^{n} p_i f_i^{XY} \\
&= \frac{1}{g} \left[\sum_{i=1}^{n} p_i f_i^{XY} - \frac{1}{g} \left(\sum_{i=1}^{n} p_i f_i^X \right) \left(\sum_{i=1}^{n} p_i f_i^Y \right) \right] \\
&= \frac{1}{\sum_{i=1}^{n} p_i f_i} \left[\sum_{i=1}^{n} p_i f_i^{XY} - \frac{1}{\sum_{i=1}^{n} p_i f_i} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j f_i^X f_j^Y \right]. \qquad (3.38)
\end{aligned}$$

We see that the local dependence function in a mixture is *not* a simple function of the local dependence function of the subdistributions. But if we assume that $f_i(x, y)$ is zero in a two-dimensional open set A_k for all $i \neq k$, we have that for (x, y) in A_k, equation 3.38 is simplified to

$$\begin{aligned}
\gamma_g(x, y) &= \frac{1}{\sum_{i=1}^{n} p_i f_i} \left[\sum_{i=1}^{n} p_i f_i^{XY} - \frac{1}{\sum_{i=1}^{n} p_i f_i} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j f_i^X f_j^Y \right] \\
&= \frac{1}{p_k f_k} \left[p_k f_k^{XY} - \frac{1}{p_k f_k} p_k^2 f_k^X f_k^Y \right] \\
&= \frac{1}{f_k} \left[f_k^{XY} - \frac{1}{f_k} f_k^X f_k^Y \right] \\
&= \frac{\partial^2}{\partial x \partial y} \ln f_k(x, y) \\
&= \gamma_{f_k}(x, y). \qquad (3.39)
\end{aligned}$$

In other words, the local dependence function of the mixture reduces to the local dependence function of the one non-zero density function (f_k) for all (x, y) in A_k. When the densities $f_i, i \neq k$ and their (first-order and second-order mixed) derivatives are ‘close’ to zero, that is, the subdistributions are well separated, the same result holds approximately.

Mixtures of distributions where we in each subdistribution has monotone association (either all positive or all negative association) is one place where the local dependence function is useful. Using dependence maps, we can discover *areas* of positive or negative local dependence. Recall the example in section 2.2.2 on page 17, where we had a mixture of two bivariate distributions, both with positive correlation, but where the correlation in the mixture was negative. Figure 3.7 on the facing page shows a dependence map for this distribution (this is the same data set as in figure 2.1 on page 19). The map shows that we have positive local dependence inside both point clouds.

This example is one where the result in equation 3.39 on the preceding page holds approximately. Numerical calculations show that the local dependence in most areas is approximately $\frac{2}{3}$ (with several decimals correct); it is approximately equal to the local dependence in each subpopulation (see equation 3.20 on page 44), except for a narrow, but infinite, strip between the two subpopulations.

3.5 Regional dependence

Some of the ‘local dependence’ we may encounter in distributions come from the *structural form* of their support, and is really a different type of dependence than the ones we have looked at earlier in this chapter. We note that two variables are never independent if their support is not a cartesian product set. (Remember that the local dependence function is not defined in this case.) Here is one simple example:

Example 3.5.1: Structural dependence

Let (X, Y) have a bivariate uniform distribution on the unit disc, that is,

$$f_{X,Y}(x, y) = \pi, \quad x^2 + y^2 < 1.$$

The marginal densites of X and Y are

$$f_X(x) = f_Y(x) = 2\pi\sqrt{1 - x^2}, \quad 0 < x < 1,$$

and we see that $f_{X,Y}(x, y) = 0 \neq f_X(x)f_Y(y)$ outside the support of (X, Y) but inside the support of X and of Y.

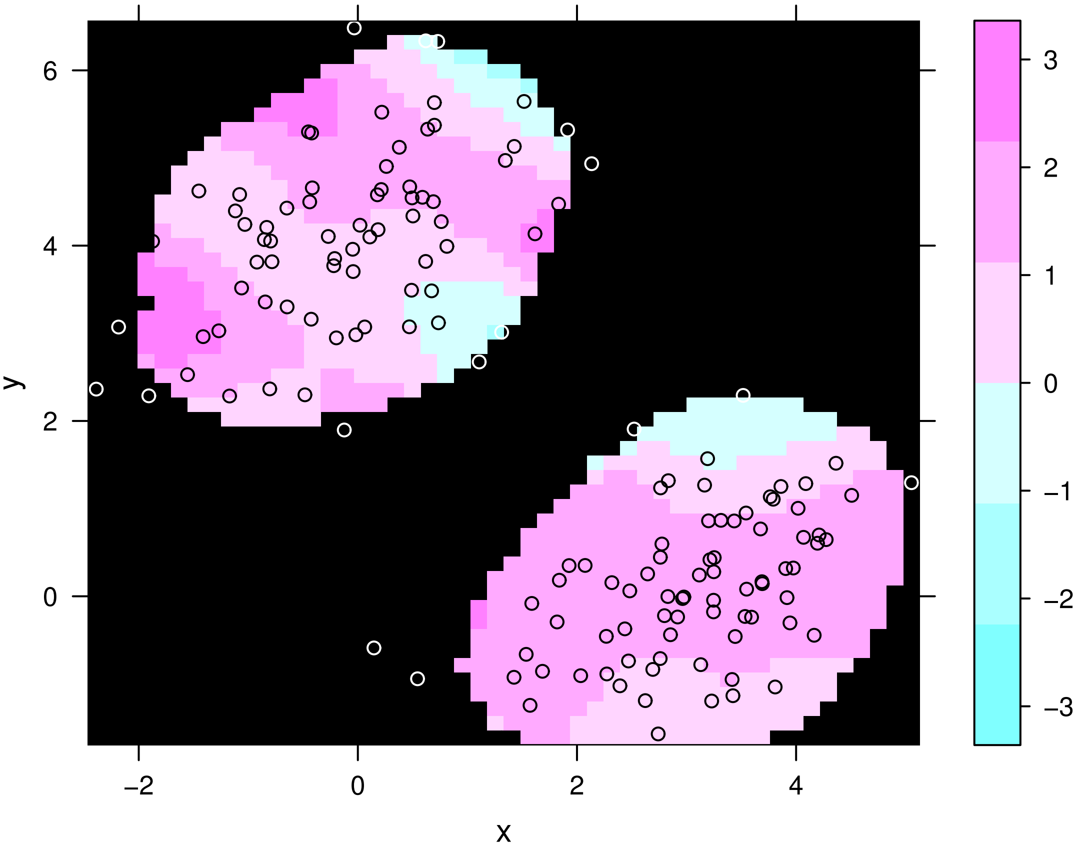


Figure 3.7: Dependence map of 150 observations from an even mixture of two bivariate normal distributions, both with correlation $\frac{1}{2}$. This is the same data set as in figure 2.1 on page 19, and the dependence map has been calculated with default options. Magenta-coloured and cyan-coloured areas indicate (estimated) positive and negative local dependence, respectively, and black areas indicate low-density areas (where we do not have reliable estimates of the local dependence). The black and white circles are the 150 observations. The dependence map indicates that we have positive local dependence (almost) everywhere (that is, inside the two point clouds), even though the overall correlation is negative. Calculating the dependence map on different realisations of the same model gives generally the same results (but with the smaller cyan-coloured spots appearing in different places).

In this example, we saw that the two random variables were not independent; however, they do have several of the properties of independent variables, such as constant conditional mean, $\mathbb{E}(Y \mid X = x) = \mathbb{E}(Y)$ and $\mathbb{E}(X \mid Y = y) = \mathbb{E}(X)$, zero correlation and zero 'local dependence'. (The local dependence function is not defined here, but if we use the function in equation 3.19 on page 44 on the joint density, we get the value 0.)

Based on these observations, Holland and Wang (1987b) introduced a concept of quasi-independence for continuous variables (a similar concept did exist for contingency tables). Two continuous variables X and Y are said to be *quasi-independent* if and only if their joint density $f(x, y)$ can be written as a product of functions of either x or y alone: $f(x, y) = a(x)b(y)$ for all (x, y) in the support of (X, Y). We see that the variables in example 3.5.1 on page 64 are quasi-independent.

Holland and Wang also introduced a *quasi-independent projection* of a density $h = h(x, y)$ of two variables, X and Y (typically defined on a support which is not a cartesian product set). The projection is a new density (on the same same support) that can be written as $f(x, y) = a(x)b(y)$, and where the marginal densities are identical to the original marginal densities. See the article for further information.

A third idea introduced in the article was a (global) measure of regional dependence, for 'quantifying the degree of dependence due to region' in the same article. Again, the reader is referred to the article for further details.

3.6 Local dependence function of Bairamov and Kotz

Kotz and Nadarajah (2003) suggested a measure of local dependence based on conditional moments (originally developed in a different article):

$$H(x, y) = \frac{\mathbb{E}\Big[\big(X - \mathbb{E}\,(X \mid Y = y)\big)\big(Y - \mathbb{E}\,(Y \mid X = x)\big)\Big]}{\sqrt{\mathbb{E}\Big[\big(X - \mathbb{E}\,(X \mid Y = y)\big)^2\Big]\,\mathbb{E}\Big[\big(Y - \mathbb{E}\,(Y \mid X = X)\big)^2\Big]}}. \tag{3.40}$$

Following the naming in Mari and Kotz (2001, pages 175–176), we will call $H(x, y)$ the *local dependence function of Bairamov and Kotz*. We only include it in this thesis for completion. Its properties can be found in the cited article, and will not be reproduced here.

But we do note one property that may be of interest: In contrast to the correlation curve and the local dependence function, the local dependence function of Bairamov and Kotz is *not* constant for the bivariate normal distribution.

3.7 Summary and conclusions

In this chapter, we have looked at the concept of a 'measure of local dependence', and we have defined some properties such measures should possess. We have looked at two candidate measure of local dependence – the correlation curve and the local dependence function.

For the correlation curve, we have examined its properties, possible generalisations and two methods of estimating the curve, with computer implementations. One of the more important features of a good measure of dependence is its interpretability and its generality; in other words, can we use this measure to say something meaningful about the dependence, and for a wide variety of distributions? It turns out that the correlation curve is only meaningful as a measure of local dependence in a subset of models – generalised additive models (with a possible heteroscedastic error term). Transformations of variables from such a model also suffer from lack of an easy interpretation, which limits the correlation curve use as a general measure of local dependence.

One other criticism is the nonsymmetrical form of the curve – the correlation curve is a function of only one variable. This may be appropriate in a pure 'input–output' (perhaps 'dose–response') model of physical interactions, but in general, I see this as shortcoming of the measure. For example, the correlation curve for the transformed distribution in example 3.4.5 on page 54 (the Cauchy distribution) will show zero dependence for both $\rho(x)$ and $\rho(y)$.

The other measure we have looked at is the local dependence function, which has the benefit of being symmetrical, in the sense that it is a function of two variables, and can be used to study the dependence in two-dimensional areas. It has (at least) two possible derivations and interpretations: It can be seen as the continuous analogue of the cross product ratios in a contingency table, or it can be seen as a localisation of the usual correlation function. The exposition in the original proofs of these two derivations were somewhat lacking in detail, and the descriptions in this thesis should hopefully make the proofs easier to read and understand.

The first derivation was based on the cross product ratios, and these are a much used measure of dependence in the discrete case, perhaps because of their nice properties: The cross product ratios along with the marginal distributions completely characterise the bivariate distribution, and they are invariant to marginal replacements. They can, therefore, in one sense, be seen as a pure measure of dependence.

The two properties are shared, or, in fact, inherited, by the local dependence function: The local dependence functions along with the marginal distributions completely characterise the bivariate distribution, and, for a given bivariate distribution, we can replace one of the marginals (leaving the conditional distribution unchanged) without affecting the local dependence function.

We have also looked at a few other properties of the local dependence function, and have examined how it can be estimated. Using this, we have developed an effective computer implementation of the estimate. This implementation is very fast, due to its use of the generalised outer product function and matrix based operations to calculate the estimates. The program can also plot dependence maps, two-dimensional graphical displays of the local dependence function, and these maps are of special interest in that they make it possible to *easily* see how the strength of the local dependence varies over the support.

Using several very simple examples, we have explored how the local dependence behaves, and how good the dependence map manages to capture this behaviour. Our conclusion is that the estimates of the sign of the local dependence are usually good, but estimates of its numerical values are more problematic. But we note that we have not found a natural interpretation of the numerical values either. But at least the local dependence changes in a predictable way for transformed variables, and transformations may be useful, especially for distributions with heavy tails.

The estimates also seem to have some problems near the boundaries of the support, and this is a well-known problem with these kinds of kernel estimates. They will also have problems at other points of discontinuity (of any of the estimated functions); see, for instance, (Wand and Jones 1995, section 2.11). It may be possible to modify the estimates to use special 'boundary kernels'. Some more work on automatic bandwidth selection also needs to be done.

In the chapter on measures of global assciation, we looked at the dependence in a mixture of two distributions, and saw that some common measures of dependence were often not appropriate. In the current chapter, we have looked at the local dependence function in mixtures, and shown that it is *not* a simple function of the local dependence in each subdistribution; however, when the distributions are well-separated, it is approximately reduced to the local dependence in the subdistributions.

Finally, we have looked at the concept of regional or structural dependence, which is a different kind of local dependence, and we have very briefly looked the definition of an alternative measure of local dependence, the local dependence function of Bairamov and Kotz.

4
Copulas

4.1 Introduction

Let us look at some elementary properties of a vector-valued random variable, $\boldsymbol{X} = (X_1, \ldots, X_n)$. The distribution (function) of $\boldsymbol{X}$ tells us which values the variable can assume, and how often it on average assumes them; in other words, it tells us $\mathbb{P}(\boldsymbol{X} \in \boldsymbol{A})$ for arbitrary $\boldsymbol{A}$. It also gives us *complete* information on the association between subsets of $\boldsymbol{X}$; that is, it tells us the value of

$$\mathbb{P}\left((X_{i_1}, \ldots, X_{i_k}) \in \boldsymbol{A} \mid \left(X_{j_1}, \ldots, X_{j_l}\right) \in \boldsymbol{B}\right) \quad \text{for all } \boldsymbol{A}, \boldsymbol{B}.$$

The marginal distributions of X_1 to X_n do not determine their joint distribution, but their joint distribution does determine their marginal distribution. We say that the distribution of $\boldsymbol{X}$ contains *more* information on the dependence between the variables than the marginal distributions do (these contain no information, except for some restrictions on which values $\boldsymbol{X}$ can take). In fact, as noted above, it contains all the information.

Motivated by this, we will now look at probability-based measures of association (among variables $X_1, \ldots, X_n$) that together with the marginal distributions uniquely determine their joint distribution. We will call such functions *dependence-defining*.

If d is such a function, this means that the joint distribution function H can be written $H(x,y) = d(F,G,x,y)$ for marginal distribution functions F and G. The joint distribution function is of course itself such a function, but we are mainly interested in a 'smaller' function – one which says as little as possible about marginal distributions, while still defining the joint distribution *together* with the marginals.

If we have two dependence-defining functions, f and g, we say that $f \preceq g$ if and only if f can be expressed as a function of g. If $f \preceq g$ for all dependence-defining functions g, we say that f is a smallest dependence-defining function. In this chapter, we will look at such a function, called the copula. The information in this chapter is chiefly taken from Nelsen (1999), but we will present it in a form which hopefully makes the various definitions used more intuitively understandable.

It is not difficult to show that when a variable X has a continuous distribution function F, then $F(X)$ has a uniform distribution. When $X_1, \ldots, X_n$ have distribution functions $F_1, \ldots, F_n$, respectively, we say that the joint distribution function of $F_1(X_1), \ldots, F_n(X_n)$ is the copula of $X_1, \ldots, X_n$. We will later define necessary and sufficient statements for an arbitrary function to be a copula.

4.2 Defining distribution functions

A formal definition of a distribution function is needed when working with copulas, and we will use a slightly more general definition of than the one usually used in probability theory (we do not, for example, require left-continuity). All results proved using our definition will therefore be valid for 'ordinary' distribution functions too. But let us first look at some motivation. Consider a two-dimensional random variable (X,Y). Let B be a rectangle (a *2-box*) with lower-left corner (x_1, y_1) and upper-left corner (x_2, y_2) (the coordinates need not be finite). A distribution function for (X,Y), $H(x,y) = \mathbb{P}(X \leq x, Y \leq y)$ should satisfy $\mathbb{P}(B) = H(x_2,y_2) - H(x_2,y_1) - H(x_1,y_2) + H(x_1,y_1) \geq 0$ for all rectangles B. We say that the *H-area of B* is nonnegative. Note that a result in probability theory says that a probability measure defined on rectangles can be uniquely extended to other sets, so it suffices to look at probabilities of such sets. This concept of H-areas can be extended to higher dimensions, where they are called H-volumes:

Definition 4.2.1: *H*-volume

Let $S_1, \ldots, S_n$ be nonempty subsets of $\overline{\mathbb{R}}$, and let H be a real-valued function in n variables with $\operatorname{Dom} H = S_1 \times \cdots \times S_n$. Let $B = [\boldsymbol{a}, \boldsymbol{b}]$ be an n-box with vertices in $\operatorname{Dom} H$. The H-volume of B is defined to be

$$V_H(B) = \sum \operatorname{sgn}(\boldsymbol{c}) H(\boldsymbol{c}), \tag{4.1}$$

where the sum is taken over all vertices $\boldsymbol{c}$ of B, and $\operatorname{sgn}(\boldsymbol{c})$ is

$$\operatorname{sgn}(\boldsymbol{c}) = \begin{cases} 1 & \text{if } c_k = a_k \text{ for av even number of } k\text{'s}, \\ -1 & \text{if } c_k = a_k \text{ for av odd number of } k\text{'s}. \end{cases} \tag{4.2}$$

For example, the H-volume of the 4-box $B = [x_{11}, x_{12}] \times [x_{21}, x_{22}] \times [x_{31}, x_{32}] \times [x_{41}, x_{42}]$ is

$$\begin{aligned} V_H(B) = {} & H(x_{12}, x_{22}, x_{32}, x_{42}) - H(x_{12}, x_{22}, x_{32}, x_{41}) - H(x_{12}, x_{22}, x_{31}, x_{42}) \\ & + H(x_{12}, x_{22}, x_{31}, x_{41}) - H(x_{12}, x_{21}, x_{32}, x_{42}) + H(x_{12}, x_{21}, x_{32}, x_{41}) \\ & + H(x_{12}, x_{21}, x_{31}, x_{42}) - H(x_{12}, x_{21}, x_{31}, x_{41}) - H(x_{11}, x_{22}, x_{32}, x_{42}) \\ & + H(x_{11}, x_{22}, x_{32}, x_{41}) + H(x_{11}, x_{22}, x_{31}, x_{42}) - H(x_{11}, x_{22}, x_{31}, x_{41}) \\ & + H(x_{11}, x_{21}, x_{32}, x_{42}) - H(x_{11}, x_{21}, x_{32}, x_{41}) - H(x_{11}, x_{21}, x_{31}, x_{42}) \\ & + H(x_{11}, x_{21}, x_{31}, x_{41}). \end{aligned} \tag{4.3}$$

A distribution function should naturally be nondecreasing in each argument. But in defining copulas, we will actually need a slightly different property:

Definition 4.2.2: *n*-increasing functions

A real-valued function H in n variables is said to be *n-increasing* if and only if $V_H(B) \geq 0$ for all n-boxes B with vertices in $\operatorname{Dom} H$.

Note that for $n \geq 2$, that H is n-increasing neither implies or is implied by the property that H is nondecreasing in each variable. But see lemma 4.2.4 on the next page for an extra condition which does make H nondecreasing in each variable.

A distribution function H should also have the property that

$$H(x_1, \dots, x_{i-1}, -\infty, x_{i+1}, \dots, x_n) = 0 \quad \text{for all } i \text{ and all } x_j\text{'s}. \tag{4.4}$$

More generally, we can define grounded functions:

Definition 4.2.3: Margins and grounded functions

Suppose that the domain of a real-valued function H in n variables is given by $\operatorname{Dom} H = S_1 \times \cdots \times S_n$, where each S_k has a least element a_k. We say that H

is grounded if and only if $H(\boldsymbol{t}) = 0$ for all $\boldsymbol{t}$ in Dom H such that $t_k = a_k$ for at least one k.

If, in addition, each S_k is nonempty and has a greatest element, b_k, we say that H has margins, and the one-dimensional margins (simply called margins) of H are the functions H_k with Dom $H_k = S_k$ and

$$H_k(x) = H(b_1, \dots, b_{k-1}, x, b_{k+1}, \dots, b_n) \quad \text{for } x \text{ in } S_k. \tag{4.5}$$

The k-dimensional margins are called k-margins, and are defined in the obvious way.

Lemma 4.2.4: *n*-increasing functions nondecreasing in each argument

If $S_1, \dots, S_n$ are nonempty subsets of $\overline{\mathbb{R}}$ and H is a grounded n-increasing function with domain $S_1 \times \cdots \times S_n$, then H is nondecreasing in each argument.

Definition 4.2.5: Distribution functions

An n-dimensional distribution function is a function H with domain $\overline{\mathbb{R}}^n$ such that

1. H is n-increasing,
2. $H(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$ in $\overline{\mathbb{R}}^n$ such that $x_k = -\infty$ for at least one k, and $H(\infty, \dots, \infty) = 1$.

From this, it follows that distribution functions are grounded.

4.3 Defining copulas

Definition 4.3.1: Copula

An n-dimensional copula (an n-copula) is a grounded, n-increasing function C with domain $\mathbb{I}^n$, and with margins C_k, $k = 1, \dots, n$ that satisfy $C_k(u) = u$ for all u in $\mathbb{I}$.

An equivalent definition is:

Definition 4.3.2: Copula – alternative definition

A n-dimensional copula (an n-copula) is a function C from $\mathbb{I}^n$ to $\mathbb{I}$ which satisfy these two properties:

1. For every $\boldsymbol{u}$ in $\mathbb{I}^n$

$$C(\boldsymbol{u}) = 0 \text{ if at least one coordinate of } \boldsymbol{u} \text{ is } 0, \text{ and} \tag{4.6a}$$

$$C(\boldsymbol{u}) = u_k \text{ if all coordinates of } \boldsymbol{u} \text{ except } u_k \text{ are } 1. \tag{4.6b}$$

2. For every $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{I}^n$ such that $\boldsymbol{a} \leq \boldsymbol{b}$,

$$V_C\big([\boldsymbol{a}, \boldsymbol{b}]\big) \geq 0. \tag{4.6c}$$

It can be shown that for any n-copula C, $n \geq 3$, each k-margin of C is a k-copula, $2 \leq k < n$.

The following lemma will be used to show that couplas are uniformly continuous. Its proof is somewhat complicated; see (Nelsen 1999, page 39) and the reference therein.

Lemma 4.3.3

Let $S_1, \ldots, S_n$ be nonempty subsets of $\overline{\mathbb{R}}$, and let H be a grounded n-increasing function with domain $\operatorname{Dom} H = S_1 \times \cdots \times S_n$ and margins $H_1, \ldots, H_n$. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ be points in $\operatorname{Dom} H$. Then

$$|H(\boldsymbol{x}) - H(\boldsymbol{y})| \leq \sum_{k=1}^{n} |H_k(x_k) - H_k(y_k)|. \tag{4.7}$$

It now follows directly from this lemma that copulas a uniformly continuous:

Theorem 4.3.4: Copulas are uniformly continuous

Let C be an n-copula. Then for every $\boldsymbol{u}$ and $\boldsymbol{v}$ in $\mathbb{I}^n$,

$$|C(\boldsymbol{u}) - C(\boldsymbol{v})| \leq \sum_{k=1}^{n} |u_k - v_k|. \tag{4.8}$$

We now have a very important result that ties copulas to distribution functions:

Theorem 4.3.5: Sklar's theorem (1959)

Let H be a joint distribution function with margins $F_1, \ldots, F_n$. Then there exists an n-copula C such that for all $x_1, \ldots, x_n$ in $\overline{\mathbb{R}}$,

$$H(x_1, \ldots, x_n) = C\big(F_1(x_1), \ldots, F_n(x_n)\big). \tag{4.9}$$

> If $F_1, \ldots, F_n$ are continuous, then C is unique. Otherwise C is uniquely determined on $\operatorname{Ran} F_1 \times \cdots \times \operatorname{Ran} F_n$. Conversely, if C is a copula and $F_1, \ldots, F_n$ are distribution functions, then the function H is a joint distribution function with margins $F_1, \ldots, F_n$.

When $F_1, \ldots, F_n$ are continuous, C is simply the function which maps $\big(F_1, \ldots, F_n\big)$ to $H(x_1, \ldots, x_n)$. When they are not continuous, it can be shown that C can be extended to the closure of $\operatorname{Ran} F_1 \times \cdots \times \operatorname{Ran} F_n$ by continuity, and then to $\mathbb{I}^n$ by multilinear interpolation. The details, and proof that C really is a copula, can be found in Nelsen (1999, page 41) and in the references listed there.

When the distribution function of a random variable is not strictly increasing, it has no inverse, but we can define 'quasi-inverses':

Definition 4.3.6: Quasi-inverse functions

> Let F be a distribution function. A *quasi-inverse* of F is any function F^- with domain $\mathbb{I}$ and range $\overline{\mathbb{R}}$ such that
>
> 1. for u in $\operatorname{Ran} F$, $x = F^-(u)$ can be any number such that $F(x) = u$;
>
> 2. for u not in $\operatorname{Ran} F$, $F^-(u) = \inf\{x \mid F(x) \geq u\}$.
>
> Note that we have already used this concept, in example 2.4.2 on page 24.

It is immediately clear that for quasi-inverses $F_1^-, \ldots, F_n^-$ of distribution functions $F_1, \ldots, F_n$ and $\boldsymbol{u}$ in $\mathbb{I}^n$,

$$C(u_1, \ldots, u_n) = H\big(F_1^-(u_1), \ldots, F_n^-(u_n)\big), \tag{4.10}$$

where H and C are as in theorem 4.3.5 on the previous page.

One important property of copulas are their invariance to increasing transformations. We will, for ease of notation, show this for two variables, but the general result, and its proof, is completely analogous. Similar results exist for decreasing transformations and mixed transformations (where, for example, one transformation is increasing and the other decreasing).

Theorem 4.3.7: Copulas are invariant under increasing transformations

> Let Z and W be to continuous random variables with copula $C_{Z,W}$ and distribution functions F_Z and G_W, respectively, and let $X = a(Z)$ and $Y = b(W)$ for increasing functions a and b. Furthermore, let $C_{X,Y}$ denote the copula of (X, Y), and let F_X and G_Y be the distribution functions of X and Y, respectively. We

have $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}\big(a(Z) \leq x\big) = \mathbb{P}\big(Z \leq a^-(x)\big) = F_Z\big(a^-(x)\big)$, and it now follows that

$$\begin{aligned} C_{X,Y}\big(F_X(x), G_Y(y)\big) &= \mathbb{P}\big(a(Z) \leq x, b(W) \leq y\big) \\ &= \mathbb{P}\big(Z \leq a^-(x), W \leq b^-(y)\big) \\ &= C_{Z,W}\Big(F_Z\big(a^-(x)\big), G_W\big(\leq b^-(y)\big)\Big) \\ &= C_{Z,W}\big(F_X(x), G_Y(y)\big). \end{aligned}$$

In section 3.3.5 on page 38, we showed that the correlation curve for (e^X, e^Y), where (X, Y) had a bivariate normal distribution, had a completely different form than the correlation curve for (X, Y) (constant correlation curve); and in section 3.4.7 on page 62 we saw the the local dependence function also changed, albeit in a predictable and 'nice' manner. But the copula will stay the same for both these pairs of variables.

4.4 Examples

We will now look at some simple examples of copulas.

Example 4.4.1: The independence copula

Let H be the joint distribution function of $(X_1, \ldots, X_n)$ and $F_1, \ldots, F_n$ be their marginal distribution functions. Then the X_i's are independent if and only if $H(x_1, \ldots, x_n) = F_1(x_1) \cdots F_n(x_n)$ for all $(x_1, \ldots, x_n)$; that is, they are independent if and only if their copula C is $C(u_1, \ldots, u_n) = u_1 \cdots u_n$. We will denote this copula by $\Pi = \Pi_n$.

There also exists two copulas for 'complete dependence'. Recall section 2.4.2 on page 22, where we showed that for each pair of marginal distributions, there exists an upper and a lower bound on their joint distribution function. The bounds, which also charachterised the dependence, involved only the expressions $F(x)$ and $G(y)$, and can thus be reformulated in terms of copulas. They can also be generalised to multivariate distributions.

Following the notation in Nelsen (1999, page 42), we have

$$\begin{aligned} W^n(\boldsymbol{u}) &= \max(u_1 + \cdots + u_n - n + 1, 0) \quad \text{and} \\ M^n(\boldsymbol{u}) &= \min(u_1, \ldots, u_n), \end{aligned}$$

and these are bounds of the multivariate distributions or, equivalently, the copulas:

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

For $n = 2$, both the upper and lower bounds were distribution functions, and we easily see that they have uniform marginals; in other words, the bounds are themselves copulas.

For n greater than two, the upper bound M^n is still a copula (it corresponds to a distribution with n 'copies' of one variable), but the lower bound is *never* a copula. The bound is still tight, though, in the sense that for any $n \geq 3$ and any fixed $\boldsymbol{u} \in \mathbb{I}^n$, there exists an n-copula C such that $C(\boldsymbol{u}) = W^n(\boldsymbol{u})$.

Since the copula is a distribution function, it is sometimes of interest to use its 'density':

Example 4.4.2: Bivariate normal copula

We will now look at the copula of the bivariate normal distribution. Let (X, Y) have this distribution, with marginals $\mathcal{N}(\mu_X, \sigma_X^2)$ and $\mathcal{N}(\mu_Y, \sigma_Y^2)$ and with joint distribution function H. Let $\boldsymbol{\Phi}$ be the joint distribution function of a standard normal bivariate distribution with the same correlation as (X, Y); that is, $H(x, y) = \boldsymbol{\Phi}\left(\frac{x-\mu_X}{\sigma_X}, \frac{y-\mu_Y}{\sigma_Y}\right)$. Let $\boldsymbol{\phi}$ be the corresponding density, and let Φ and ϕ be the standard normal distribution function and density, respectively. We define

$$\begin{aligned} u = F(x) &= \Phi\Big(\frac{x-\mu_X}{\sigma_X}\Big) \quad \text{and} \\ v = G(y) &= \Phi\Big(\frac{y-\mu_Y}{\sigma_Y}\Big). \end{aligned}$$

This gives us

$$\begin{aligned} C(u, v) &= H(F^{-1}(u), G^{-1}(v)) \\ &= \boldsymbol{\Phi}\Big(\frac{F^{-1}(u)-\mu_X}{\sigma_X}, \frac{G^{-1}(v)-\mu_Y}{\sigma_Y}\Big) \\ &= \boldsymbol{\Phi}\big(\Phi^{-1}(u), \Phi^{-1}(v)\big). \qquad (4.11) \end{aligned}$$

Using elementary rules of differential calculus, we get the copula density

$$c(u, v) = \frac{\boldsymbol{\phi}\big(\Phi^{-1}(u), \Phi^{-1}(v)\big)}{\phi\big(\Phi^{-1}(u)\big)\phi\big(\Phi^{-1}(v)\big)}. \qquad (4.12)$$

Note that this copula does not depend on the means and variances of X and Y; it is location and scale-invariant. This would also follow directly from theorem 4.3.7 on page 74 – the copulas' invariance under increasing functions.

We have earlier looked at Spearman's rho (section 2.2.3 on page 18) and Kendall's tau (section 2.3 on page 20), and observed that these were invariant to strictly increasing transformations of the variables. It is therefore natural to try to express them as functions of the underlying copula. The following result is proved in (Nelsen 1999, page 135):

Example 4.4.3: Spearman's rho as a copula concept

In equation 2.12 on page 20 we showed that Spearman's rho for (X, Y) could be written

$$\rho_S(X,Y) = \rho_S = 12\,\mathbb{E}(UV) - 3,$$

where U and V were the probability integral transformed variables of X and Y: $U = F(X)$ and $V = G(Y)$. Since the distribution function of (U, V) is the copula of (X, Y), we have

$$\begin{aligned}\rho_S &= 12 \iint_{\mathbb{I}^2} uv \,\mathrm{d}C(u,v) - 3 \\ &= 12 \iint_{\mathbb{I}^2} C(u,v)\,\mathrm{d}u\,\mathrm{d}v - 3.\end{aligned}$$

Remembering that the mean of a uniform distribution on $\mathbb{I}$ is 0.5, we see that this can also be written

$$\rho_S = 12 \iint_{\mathbb{I}^2} \big[C(u,v) - uv\big]\,\mathrm{d}u\,\mathrm{d}v.$$

Or, in words, Spearman's rho for (X, Y) is (a constant times) the 'average distance' between the copula of (X, Y) and the independence copula.

A similar result can also be shown for Kendall's tau (modified proof from Nelsen 1999, pages 127–128):

Example 4.4.4: Kendall's tau as a copula concept

Let (X, Y) and (X', Y') be two independent samples from a continuous bivariate distribution, with marginals F and G, and copula C. We have

$$\begin{aligned}\tau &= \mathbb{P}\big((X - X')(Y - Y') \geq 0\big) - \mathbb{P}\big((X - X')(Y - Y') < 0\big) \\ &= 2 \cdot \mathbb{P}\big((X - X')(Y - Y') \geq 0\big) - 1 \\ &= 2 \cdot \big[\mathbb{P}(X \geq X', Y \geq Y') + \mathbb{P}(X < X', Y < Y')\big] - 1.\end{aligned}$$

These probabilities can be evaluated using the copula:

$$\begin{aligned}
\mathbb{P}(X \geq X', Y \geq Y') &= \mathbb{P}(X' \leq X, Y' \leq Y) \\
&= \iint_{\mathbb{R}^2} \mathbb{P}(X' \leq x, Y' \leq y) \, \mathrm{d}C\big(F(x), G(y)\big) \\
&= \iint_{\mathbb{R}^2} C\big(F(x), G(y)\big) \, \mathrm{d}C\big(F(x), G(y)\big) \\
&= \iint_{\mathbb{I}^2} C(u,v) \, \mathrm{d}C(u,v).
\end{aligned}$$

Similarly, the same holds for $\mathbb{P}(X < X', Y < Y')$, and we get the result

$$\tau = 4 \iint_{\mathbb{I}^2} C(u,v) \, \mathrm{d}C(u,v) - 1. \tag{4.13}$$

Kendall's tau can thus be seen as a function of expected value of the copula itself, inserted two uniform variables on $\mathbb{I}$ having the copula as their joint distribution: $\tau = \mathbb{E}\big(C(U,V)\big) - 1$. Contrast this with Spearman's rho, which could be seen as a function of the expected value of the *product* of the variables: $\rho_S = 12\,\mathbb{E}(UV) - 3$.

4.5 Local dependence and copulas

The local dependence function is not a copula concept, but, nevertheless, it may be interesting to try to express it as a function of the copula *and* the marginal densities. Recall that we have

$$\begin{aligned}
C(u,v) &= \mathbb{P}\big(F(X) \leq u, G(Y) \leq v)\big) = \mathbb{P}\big(X \leq F^{-}(u), Y \leq G^{-}(v)\big) \\
&= H\big(F^{-}(u), G^{-}(v)\big),
\end{aligned}$$

where H is the joint distribution function of (X,Y). Let h and c be the density functions corresponding to H and C, respectively. We use the notation

$$\begin{aligned}
u = F(x), &\quad v = G(y) \\
c^u(u,v) &= \frac{\partial}{\partial u} c(u,v) \\
c^v(u,v) &= \frac{\partial}{\partial v} c(u,v) \\
c^{uv}(u,v) &= \frac{\partial^2}{\partial u \, \partial v} c(u,v).
\end{aligned}$$

The local dependence function is

$$\begin{aligned}
&= \frac{\partial^2}{\partial x\,\partial y} \ln \frac{\partial^2}{\partial x\,\partial y} H(x,y) \\
&= \frac{\partial^2}{\partial x\,\partial y} \ln \frac{\partial^2}{\partial x\,\partial y} C\bigl(F(x), G(y)\bigr) \\
&= \frac{\partial^2}{\partial x\,\partial y} \ln c\bigl(F(x), G(y)\bigr) f(x) g(y) \\
&= \frac{\partial}{\partial x} \frac{1}{c\bigl(F(x), G(y)\bigr) f(x) g(y)} \left[c^v(u,v) f(x) g^2(y) + c(u,v) f(x) g'(y)\right] \\
&= \frac{\partial}{\partial x} \frac{1}{c(u,v) g(y)} \left[c^v(u,v) g^2(y) + c(u,v) g'(y)\right] \\
&= \frac{1}{c(u,v) g(y)} \left[c^{uv}(u,v) f(x) g^2(y) + c^u(u,v) f(x) g'(y)\right] \\
&\quad - \frac{c^u(u,v) f(x) g(y)}{c^2(u,v) g^2(y)} \left[c^v(u,v) g^2(y) + c(u,v) g'(y)\right] \\
&= \frac{1}{c(u,v)} \Biggl[c^{uv}(u,v) f(x) g(y) + c^u f(x) g'(y) \frac{1}{g(y)} \\
&\quad - c^u(u,v) c^v(u,v) f(x) g(y) \frac{1}{c(u,v)} + c^u f(x) g'(y) \frac{1}{g(y)}\Biggr] \\
&= \frac{f(x)g(y)}{c^2(u,v)} \Bigl[c^{uv}(u,v) c(u,v) - c^u(u,v) c^v(u,v)\Bigr]. \qquad (4.14)
\end{aligned}$$

We recognise the factor multiplied by $f(x)g(y)$ as the local dependence function of (U, V). We now have a formula we can use to calculate the local dependence function based on the copula and the marginal distributions. For a fixed marginal distribution, we can easily see the effect different copulas will have on the local dependence. Note that we could also have calculated the result indirectly, using equation 3.36 on page 62.

We can also remove the (functional) dependence on the marginal distributions by 'copulising' the bivariate distribution: Instead of looking at the local dependence when X and Y have the values x and y, respectively, we look at the local dependence when X and Y are at their q_X and q_Y quantiles. In other words, we examine the local dependence of $F(X)$ and $G(Y)$ at q_X and q_Y. (See example 3.4.5 on page 54, where we did just that.)

It is now elementary to verify that the copula of a copula is the original copula, so the local dependence in equation 4.14 on the previous page is reduced to

$$\gamma(u,v) = \frac{1}{c^2(u,v)}\Big[c^{uv}(u,v)c(u,v) - c^u(u,v)c^v(u,v)\Big]. \tag{4.15}$$

This could of course also be calculated directly from equation 3.19 on page 44.

4.6 Summary and conclusions

In this chapter, we have looked at copulas – multivariate probability distribution functions with uniform marginals. These do together with the marginal distributions completely characterise the multivariate distributions, and are invariant to any increasing transformations of the original variables. They manage to capture exactly *what* the extra information the multivariate distribution functions contain that is missing from the univariate marginals, while at the same time being independent (in one sense of the word) of marginal distributions. We may therefore rightfully call them pure dependence functions, and it makes sense that any measure of (only) dependence should only depend on the copula.

We have looked at how we can define the copulas in a mathematical manner, and how we can interpret them from a statistical perspective. In an earlier chapter, we looked at bounds on the bivariate distributions, and on how these relate to (complete) dependence. These results were extended to multivariate distributions, and related to copulas. We could also relate two measures of global dependence that we had looked earlier to copulas, and we saw how this lead to new interpretations of their meanings.

Finally, we tried to 'copulise' the local dependence function, to make it a copula concept. We also observed that the local dependence of a copula could be seen as a measure of local dependence at the quantiles of the original distribution.

The local dependence function and the copula has two very different invariance properties: while the copula is invariant to increasing transformations of the marginals, the local dependence function is invariant to *replacements* of the marginals. This means that we can replace one marginal distribution with another, leaving the conditional distribution, and hence the local dependence function, intact. This *will*, on the other hand, (usually) change the copula. The local dependence function may thus be completely unaffected by marginal replacements which change *global* measures of dependence, such as Spearman's rho or Kendall's tau, completely.

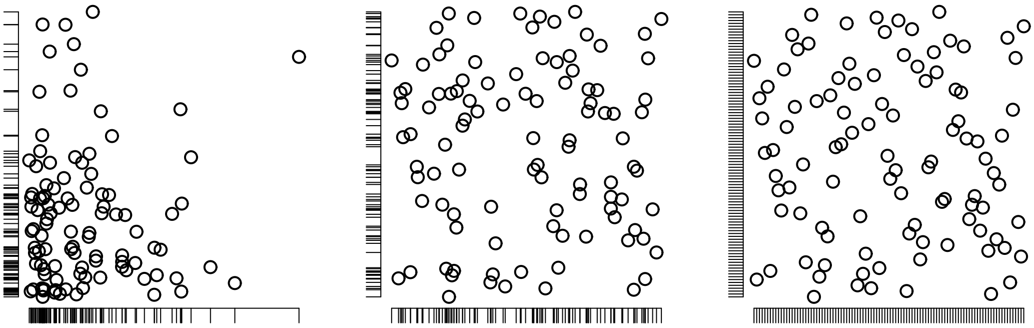
5 Graphical methods

Graphical displays of distribution give us qualitative and some quantitative information about the association between variables. We might for instance plot the mean of $Y \mid X = x$ as a function of X. And when we have observations of random variables, it is common to use scatterplots of pairs of variables. But the marginal distributions of the variables can make it difficult to assess the degree of dependence.

Figure 5.1 on the next page shows an example, based on Genest and Boies (2003). We have 100 pairs of independent exponentially distributed variables, but the (lack of) association is not apparent from a normal scatterplot. If we transform both variables using their probability integral transformations, to make the random variables marginally uniform, the association is easier to see. When the marginal distributions are unknown, we can estimate the transformed variables by replacing all variables with their ranks, perhaps using equation 3.34 on page 59.

5.1 Chi-plots

Fisher and Switzer (1985) have proposed a graphical method for investigating the association between two random variables. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random



(a) Untransformed variables.

(b) Variables transformed with distribution function.

(c) Rank-transformed variables.

Figure 5.1: Scatterplot of transformations of 100 pairs of independent exponential random variables, with marginal distributions.

sample from a bivariate (usually continuous) distribution, and define

$$
\begin{aligned}
H_i &= \frac{1}{n-1}\#\left\{j \neq i : X_j \leq X_i, Y_j \leq Y_i\right\}, \\
F_i &= \frac{1}{n-1}\#\left\{j \neq i : X_j \leq X_i\right\} \text{ and} \\
G_i &= \frac{1}{n-1}\#\left\{j \neq i : Y_j \leq Y_i\right\}.
\end{aligned}
\tag{5.1}
$$

Assuming independence, H_i will approximately factor into F_iG_i, and the difference $H_i - F_iG_i$ can be seen as a measure of the degree of association. The difference has an expected value of zero, and we note that, conditional on (X_i, Y_i), it can be written as a sum of independent variables, and thus it will be asymptotically standard normal when we divide by the estimated standard deviation. The resulting variable is

$$k_i = \frac{\sqrt{n}\,(H_i - F_iG_i)}{\sqrt{F_i(1-F_i)G_i(1-G_i)}}. \tag{5.2}$$

The variable k_i is not defined for those (at most four) i such that $F_i(1-F_i)G_i(1-G_i) = 0$. In the following, we will, to be consistent with Fisher and Switzer (1985), use $\chi_i = k_i/\sqrt{n}$ as the variable of interest.

The numerator, $H_i - F_iG_i$, can be seen as the cross-product difference of the relative counts in a 2×2 table made by partitioning the plane into quadrants centred at (X_i, Y_i). And χ_i is also equal to the sample correlation coefficient between dichotomised variables (X'_{ij} equal 1 when $X_j \leq X_i$ and 0 otherwise, and similarly for Y'_{ij}).

We now define a measure of the distance between (X_i, Y_i) and the centre of the data set:

$$\begin{aligned} \lambda_i &= 4\, s_i \max\left[\left(F_i - \tfrac{1}{2}\right)^2, \left(G_i - \tfrac{1}{2}\right)^2\right] \\ s_i &= \operatorname{sign}\left[\left(F_i - \tfrac{1}{2}\right)\left(G_i - \tfrac{1}{2}\right)\right] \end{aligned} \tag{5.3}$$

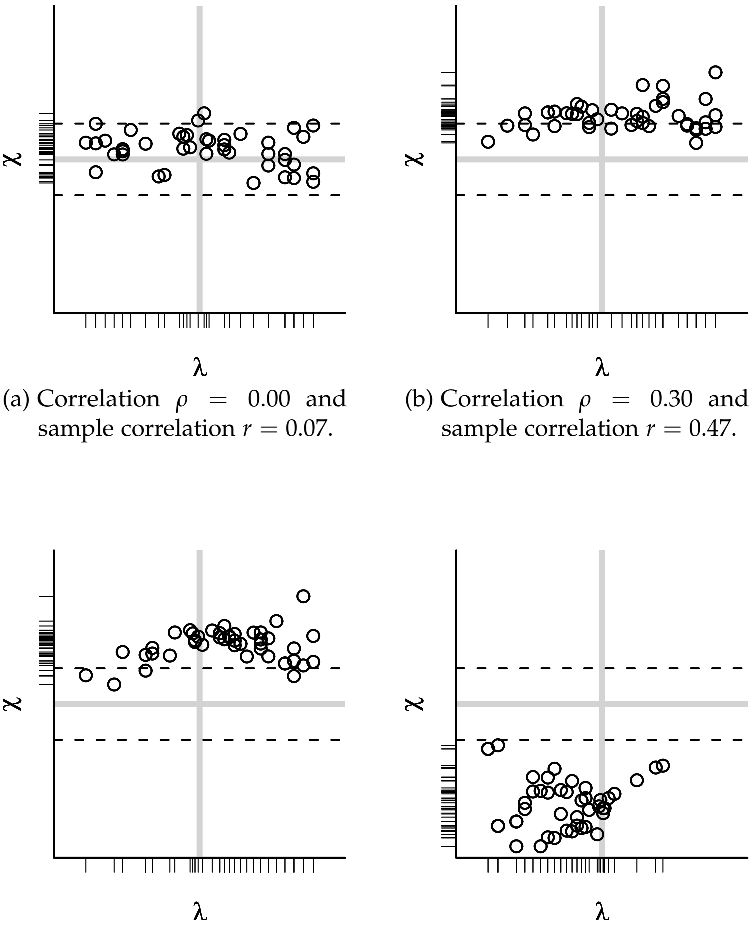
(Other measures of distance than λ_i are possible, and Fisher and Switzer (1985) list several.)

A scatterplot of observations from (λ_i, χ_i) is called a *χ-plot*. We exclude the (at most eight) points where $|\chi_i| \geq 4(1/(n-1) - \frac{1}{2})^2$, since the asymptotic normal approximation will be inappropriate here. Note that both λ_i and χ_i lie in the interval $[-1, 1]$.

To help us distinguish between natural scatter and real dependence, we superimpose horizontal control lines at $\chi = \pm c_p/\sqrt{n}$. We choose c_p so that on average approximately $100p\%$ of the pairs would lie between the guidelines when X and Y are really independent. Since the χ-values are not independent even when X and Y are (see details in Fisher and Switzer 1985), the c_p values are most easily found by Monte Carlo methods. Fisher and Switzer (2001) list these approximate values for $p = 0.90, 0.95$ and 0.99, respectively: 1.54, 1.78 and 2.18. My simulations support these results. The lines will also remain approximately valid even when one variable is nonstochastic.

Figure 5.2 on the next page shows three examples from bivariate normal distributions. Of course, since χ-plots are functions of the ranks of the data, they can not help us distinguish between linear association (which we have here) and other forms of monotone association. But they can be helpful in assessing whether there are monotone association or a more complex form of association. Let us look at a few cases, based on examples in Fisher and Switzer (2001) and in Fisher and Switzer (1985).

Consider an even mixture of a $\mathcal{N}(-1, -1, 1, 1, -.6)$ and a $\mathcal{N}(1, 1, 1, 1, .6)$ distribution. Figure 5.3 on page 85 shows a typical scatterplot, a rank scatterplots and a χ-plot from this distribution. It is based on a random sample of 150 pairs of variables from the first distribution and 150 from the second. Since the domains of the distributions overlap, it is difficult to judge the association from the normal scatterplot, and a rank scatterplot gives the impression that the variables are independent. But from the χ-plot, we clearly see that there is some sort of association, as the points are grouped into two 'lobes'. This pattern is typical of mixtures



(a) Correlation $\rho = 0.00$ and sample correlation $r = 0.07$.

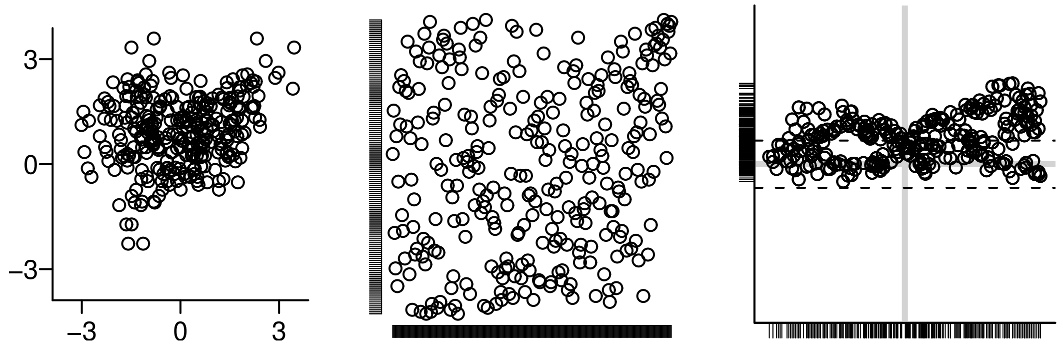
(b) Correlation $\rho = 0.30$ and sample correlation $r = 0.47$.

(c) Correlation $\rho = 0.70$ and sample correlation $r = 0.69$.

(d) Correlation $\rho = -0.95$ and sample correlation $r = -0.95$.

Figure 5.2: Four χ-plots showing samples of 50 pairs from bivariate normal distributions with means 0, standard deviations 1 and various levels of correlation.

of distributions with opposite monotone association. However, it is usually not apparent unless we have at least 100 observations from the mixture distribution.



(a) Scatterplot.

(b) Rank scatterplot.

(c) χ-plot.

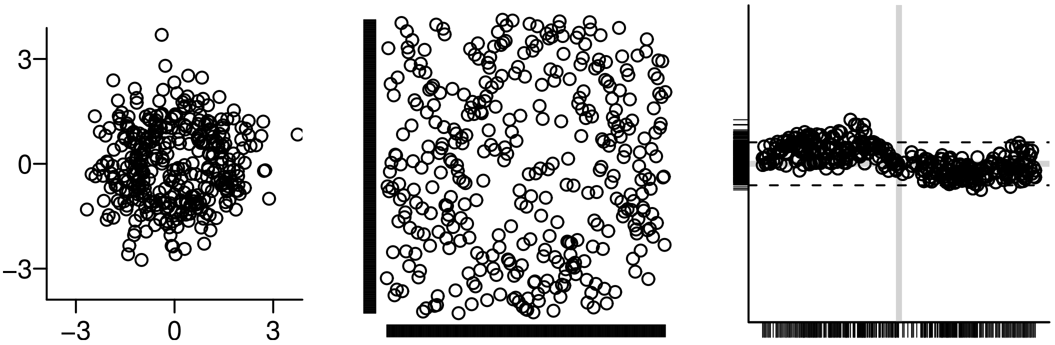
Figure 5.3: Scatterplot, rank scatterplot and χ-plot of observations from 300 pairs from an even mixture of two bivariate normal distributions. For the χ-plot, the resolution of the vertical axis near $\chi = 0$ has been increased by using the monotone transformation $\chi = \sin(\frac{1}{2}\pi\chi')$ (where χ' are the original χ-values).

We get a completely different χ-plot from distributions with holes. Figure 5.4 on the next page shows an example, where we have 500 observations from two independent standard normal variables, where approximately 75% of the pairs inside a circle of radius 1 centred in the origin has been removed. More precisely, the observations are observations from (X, Y), where X can written $X = 1_{\{X^2+Y^2>1\}}X' + I_A(1 - 1_{\{X^2+Y^2>1\}})X'$, where $1_{\{X^2+Y^2>1\}}$ is 1 when $X^2 + Y^2 > 1$ and 0 when $X^2 + Y^2 \leq 1$, I_A is a random variable that is 1 with probability 0.25 and 0 with probability 0.75, and X' and Y' are independent standard normal variables. Y is similary defined.

There are no monotone association (and both the correlation and Spearman's correlation are less than 0.05), but the χ-plot clearly shows that the variables are not independent. Note that if the hole was not centred in the middle of the distribution, we would (naturally) get patterns more similar to the ones in figure 5.2 on the facing page.

5.2 Kendall plots

Recall that when (X, Y) has a distribution with margins F and G, the distribution function of $F(X), G(Y)$ is the copula of (X, Y). This means that the transformed



(a) Scatterplot.

(b) Rank scatterplot.

(c) χ-plot.

Figure 5.4: Scatterplot, rank scatterplot and χ-plot of observation from 500 pairs of independent standard normal variables, where 75% of the observations inside the unit circle have been removed. Again the resolution of the vertical axis near $\chi = 0$ has been increased for the χ-plot, by using the same transformation as in figure 5.3 on the preceding page.

variables in figure 5.1b on page 82 are observations from this copula, and contain all the information on the association between X and Y.

Based on this observation, Genest and Boies (2003) have proposed a graphical measure of association between two variables, similar to the χ-plot and based on the same concept as Q-Q plots (see Wilk and Gnanadesikan 1968).

In Q-Q plots, we plot (sample) quantiles from one distribution against (sample) quantiles from a different distribution in order to compare the two distributions. When the two distributions belong to the same location-scale family, the points will lie on a straight diagonal line (approximately so if sample quantiles are used). And if they belong to the exact same distribution, the straight line will have an orientation of 45% degrees (when the same scale is used on both axes). Any deviation from this can be used to judge the way the distributions differ, such as different location (for example, mean or median), different scale (for example, variance), heavier or lighter tails, or a more complex form of difference (see, for instance, Cleveland 1994, 1993). Instead of (theoretical) quantiles, we can also use (expected values of the) order statistics, and this gives similar results.

Kendall plots (or K-plots) are based on a similar approach, using expected values. We have (observations from) a random sample from a bivariate distribution. First we define H_i as in equation 5.1 on page 82:

$$H_i = \frac{1}{n-1} \# \left\{ j \neq i : X_j \leq X_i, Y_j \leq Y_i \right\}.$$

Then we plot the expected values of the order statistics of the H_i against the observed values under the hypothesis of independence. More specifically: First order H_i to get $H_{(1)} \leq \cdots \leq H_{(n)}$, and compute $w_{n,i}$, the expected value of the i^{th} order statistic $H_{(i)}$ in a random sample of size n. Then plot the pairs $(w_{n,i}, H_{(i)})$. For convenience, Genest and Boies (2003) use the *asymptotic* null distribution of the order statistics.

It is well-known (see, for instance, Casella and Berger (2001), page 229) that the expected value of an order statistic from a continuous distribution with distribution function K_0 and density k_0 is

$$w_{n,i} = n\binom{n-1}{i-1}\int_0^1 w\, k_0(w) K_0(w)^{i-1}(1 - K_0(w))^{n-i}\,\mathrm{d}w. \tag{5.4}$$

Genest and Rivest (1993) has shown that under mild regularity conditions, the empirical distribution function K_n of $H_1, \ldots, H_n$ is a $\sqrt{n}$-consistent estimator of

$$K(w) = \mathbb{P}(H(X,Y) \leq w), \quad 0 \leq w \leq 1,$$

where H is the joint distribution function of (X, Y). Note that K depends only on the copula of the bivariate distribution:

$$\begin{aligned} K(w) &= \mathbb{P}(H(X,Y) \leq w) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1\{H(x,y) \leq w\}\,\mathrm{d}H(x,y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1\{C(F(x),G(y)) \leq w\}\,\mathrm{d}C(F(x),G(y)) \\ &= \int_0^1\int_0^1 1\{C(u,v) \leq w\}\,\mathrm{d}C(u,v) \\ &= \mathbb{P}(C(U,V) \leq w), \end{aligned} \tag{5.5}$$

where U and V are independent uniformly distributed variables on $\mathbb{I}$. It is now easy to show that when X and Y are independent (that is, $H = F \times G$, or $C(U,V) = UV$), we have

$$K(w) = K_0(w) = \mathbb{P}(UV \leq w) = w - w\ln(w), \quad 0 \leq w \leq 1. \tag{5.6}$$

Inserting this into equation 5.4, we can calculate $w_{n,i}$, and plot $H_{(i)}$ against these values. The plot should preferably be based on at least 50 observations. Figure 5.5 on the following page shows a K-plot for independent variables.

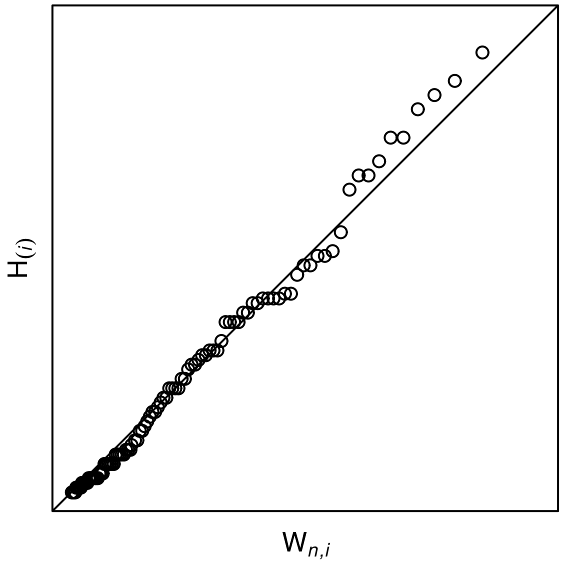


Figure 5.5: K-plot of 100 pairs of observation from a bivariate standard normal distribution with independent components.

For further details (and graphs), see the cited article.

5.3 Summary and conclusions

In this chapter, we have briefly looked at two graphical displays intended to help us discover dependence structure in bivariate distributions. It is not easy to determine the exact form of dependence from these displays, whence they are more useful as diagnostic tools to dermine *if* there is a dependence.

Just like correlation curves and the local dependence function, they can be used to discover if we have dependence of the 'positive' or 'negative' type, but unlike these, they can also be used to determine if there are *structural* dependence, for example, if there are holes or 'almost holes' in the distributions.

6
Summary and future research

In this thesis, we have looked at various measures of global and local dependence. Not surprisingly, the measures of global dependence turned out to be of limited use when the dependence was not of a monotone type. We also found, as is well-known, problems with the interpretation of one of the most frequently used measures of dependence – the correlation.

It might be expected that measures of local dependence would improve the situation, and, indeed, the two measures we looked at manage to capture more of the dependence structure in bivariate distributions. With correlation curves, we can see how the dependence varies over the support of one variable; however, the correlation curve is intrinsically a nonsymmetric measure, it only applies to some very restricted bivariate distributions and it can not always help us discover monotone dependence that is a function of where we are in a two-dimensional area.

The local dependence function, on the other hand, is a function of two coordinates, and *does* help us here. Calculations and simulations indicate that it, and its estimate, manage to capture (at least the 'sign' part of) the two-dimensional dependence. Using two- and three-dimensional graphical displays of the local dependence function, we have found it easy to see how the dependence changes when we move over the support of the bivariate distributions. But despite its merits compared to the correlation curve and the global measures, the local dependence function has its problems: we have not found any natural or 'reasonable' interpretation of its numerical value (except its sign). And, the function does not capture structural dependence.

One such function that *does* capture these type of dependence, and, in fact, any type of dependence one could think of, is the copula. It is a pure 'dependence function', and several of the measures we have looked at can be reformulated as (non-invertible) transformations of the copula. Its rich dependence-characterising property does not come without a price, though: the copula itself is difficult to interpret as a measure of how strong the dependence is (locally).

Our last subject of interest in this thesis, two graphical displays, does not help us with *any* of the above-mentioned problems. It does, however, give us a possibility to infer *if* there is a dependence between two random variables, based on observations. The displays may also indicate possible dependence models (copulas).

This thesis have only covered a tiny subset of all dependence measures, and there are many avenues for future research. The subject of copulas and their application enjoys much current interest, but it seems that simple measures of *local dependence* has been somewhat neglected. For the local dependence function, several improvements seem possible:

- Better automatic bandwidth selection.
- Other estimators, perhaps taking advantage of the fact that the measure is a function only of a *conditional* distribution.
- Modifications of the current estimator to improve the behaviour at boundaries (using boundary kernels?).
- Generalisations to more than two variables. Of course, we can look at pairwise local dependences between the variables, but variables may be pairwise independent while still having a rich multivariate dependence structure. Are there any natural generalisations of the local dependence functions? If so, what information does the sign of the generalised local dependence function indicate?
- And last, but not least, what are possible interpretations of the value of the local dependence function? Does there exist other derivations of the function thay may help in the interpretation?

The thesis ends here.

Bibliography

Anderson-Sprecher R. (1994). 'Model comparisons and R^2'. *The American Statistician*, **volume 48**, number 2, pages 113–117.

van Belle G. (2002). *Statistical Rules of Thumb*. John Wiley & Sons. ISBN 0-471-40227-3.

Bickel P.J. and Doksum K.A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*, volume 1. Prentice-Hall, second edition. ISBN 0-13-850363-X.

Bjerve S. and Doksum K. (1993). 'Correlation curves: Measures of association as functions of covariate values'. *The Annals of Statistics*, **volume 21**, number 2, pages 890–902.

Blyth S.J. (1994). 'Measuring local association: An introduction to the correlation curve'. *Sociological Methodology*, **volume 24**, pages 171–197.

Casella G. and Berger R.L. (2001). *Statistical Inference*. Duxbury, second edition. ISBN 0-534-24312-6.

Cleveland W. (1994). *The Elements of Graphing Data*. Hobart Press, Summit, New Jersey, U.S.A. ISBN 0-9634884-1-4.

Cleveland W.S. (1993). *Visualizing Data*. Hobart Press, Summit, New Jersey, U.S.A. ISBN 0-9634884-0-6.

Doksum K., Blyth S., Bradlow E., Meng X.L. and Zhao H. (1994). 'Correlation curves as local measures of variance explainted by regression'. *Journal of the American Statistical Association*, **volume 89**, number 426, pages 571–582.

Dunnett C.W. and Sobel M. (1954). 'A bivariate generalization of student's t-distribution, with tables for certain special cases'. *Biometrika*, **volume 41**, number 1/2.

Fisher N.I. and Switzer P. (1985). 'Chi-plots for assessing dependence'. *Biometrika*, **volume 72**, pages 253–265.

Fisher N.I. and Switzer P. (2001). 'Graphical assessment of dependence: Is a picture worth 100 tests?' *American Statistician*, **volume 55**, number 3, pages 233–239.

Gasser T. and Müller H. (1979). 'Kernel estimation of regression functions'. In Gasser T. and Rosenblatt M. (editors), 'Smoothing techniques for curve estimation', Number 757 in Lecture notes in mathematics, pages 23–68. Springer-Verlag. ISBN 0-387-09706-6.

Gasser T. and Müller H.G. (1984). 'Estimating regression functions and their derivatives by the kernel method'. *Scandinavian Journal of Statistics*, **volume 11**, pages 171–185.

Gasser T., Müller H.G. and Mammitzsch V. (1985). 'Kernels for nonparametric curve estimation'. *Journal of the Royal Statistical Society, Series B*, **volume 47**, number 2, pages 238–252.

Genest C. and Boies J. (2003). 'Detecting dependence with Kendall plots'. *American Statistician*, **volume 57**, pages 275–284.

Genest C. and Rivest L. (1993). 'Statistical-inference procedures for bivariate archimedean copulas'. *Journal of the American Statistical Association*, **volume 88**, number 423, pages 1034–1043.

Goodman L. (1969). 'How to ransack social mobility tables and other kinds of cross-classification tables'. *American Journal of Sociology*, **volume 75**, number 1, pages 1–40.

Holland P.W. and Wang Y.J. (1987a). 'Dependence functions for continous bivariates densities'. *Communications in Statistics: Theory and Methods*, **volume 16**, number 3, pages 863–876.

Holland P.W. and Wang Y.J. (1987b). 'Regional dependence for continuous bivariate densities'. *Communications in Statistics: Theory and Methods*, **volume 16**, number 1, pages 193–206.

Joe H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall. ISBN 0-412-07331-5.

Jones M. (1996). 'The local dependence function'. *Biometrika*, **volume 83**, number 4, pages 899–904.

Jones M. (1998). 'Constant local dependence'. *Journal of Multivariate Analysis*, **volume 64**, number 2, pages 148–155.

Jones M. and Koch I. (2003). 'Dependence maps: Local dependence in practice'. *Statistics and Computing*, **volume 13**, pages 241–255.

Kotz S. and Nadarajah S. (2003). 'Local dependence functions for the elliptically symmetric distributions'. *Sankhyā*, **volume 65**, number 1, pages 207–223.

Kvålseth T.O. (1985). 'Cautionary note about R^2'. *The American Statistician*, **volume 39**, number 4, pages 279–285.

Lehmann E.L. (1966). 'Some concepts of dependence'. *The Annals of Mathematical Statistics*, **volume 37**, number 5, pages 1137–1153.

Mari D.D. and Kotz S. (2001). *Correlation and dependence*. Imperial College Press. ISBN 1-86094-264-4.

Nelsen R.B. (1999). *An Introduction to Copulas*. Number 139 in Lecture Notes in Statistics. Springer-Verlag. ISBN 0-387-98623-5.

R development core team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Rényi A. (1959). 'On measures of dependence'. *Acta mathematica Academiae Scientiarum Hungaricae*, **volume 10**, pages 441–451.

Rodgers J. and Nicewander W. (1988). '13 ways to look at the correlation-coefficient'. *American Statistician*, **volume 42**, number 1.

Rovine M.J. and von Eye A. (1997). 'A 14th way to look at a correlation coefficient: Correlation as the proportion of matches'. *American Statistician*, **volume 51**, number 1, pages 42–46.

Sankaran P. and Gupta R. (2004). 'Characterizations using local dependence function'. *Communications in Statistics: Theory and Methods*, **volume 33**, number 12, pages 2959–2974.

Shih W.J. and Huang W.M. (1992). 'Evaluating correlation with proper bounds'. *Biometrics*, **volume 48**, number 4, pages 1207–1213.

de Veaux R.D. (1976). 'Tight upper and lower bounds for correlation of bivariate distributions arising in air pollution modeling'. *Technical Report 5*, Department of Statistics, Stanford University.

Wand M. and Jones M. (1995). *Kernel Smoothing*. Number 60 in Monographs on Statistics and Applied Probability. Chapman & Hall. ISBN 0-412-55270-1.

Warren W. (1971). 'Correlation or regression: Bias or precision'. *Applied Statistics*, **volume 20**, number 2, pages 148–164.

Wilcox R.R. (2005). 'Estimating the conditional variance of y, given x, in a simple regression model'. *Journal of Applied Statistics*, **volume 32**, number 5, pages 495–502.

Wilk M.B. and Gnanadesikan R. (1968). 'Probability plotting methods for analysis of data'. *Biometrika*, **volume 55**, number 1, pages 1–17.