Klein, Ingo

Working Paper

Copulabasierte Ordnung für Paare ordinalskalierter Merkmale

Diskussionspapier, No. 38/2001

Provided in Cooperation with:
Friedrich-Alexander-University Erlangen-Nuremberg, Chair of Statistics and Econometrics

Suggested Citation: Klein, Ingo (2001) : Copulabasierte Ordnung für Paare ordinalskalierter Merkmale, Diskussionspapier, No. 38/2001, Friedrich-Alexander-Universität Erlangen-Nürnburg, Lehrstuhl für Statistik und Ökonometrie, Nürnberg

This Version is available at:
https://hdl.handle.net/10419/29579

Friedrich-Alexander-Universität Erlangen-Nürnberg
Wirtschafts-und Sozialwissenschaftliche Fakultät

Diskussionspapier
38 / 2001

Copulabasierte Ordnung für Paare ordinalskalierter Merkmale

Ingo Klein

Lehrstuhl für Statistik und Ökonometrie
Lehrstuhl für Statistik und empirische Wirtschaftsforschung
Lange Gasse 20 · D-90403 Nürnberg

Copulabasierte Ordnung für Paare ordinalskalierter Merkmale

Ingo Klein
Lehrstuhl für Statistik und Ökonometrie
Universität Erlangen–Nürnberg
Lange Gasse 20
D–90403 Nürnberg
Germany
E-mail: ingo.klein@wiso.uni-erlangen.de

Abstract

Characterizations of measures of concordance depend on an ordering of bivariate distributions with fixed marginals or use the concept of a copula to define a ordering without fixed but with continuous marginals. Ordinal variables with a fixed number of categories have discrete bivariate distributions. But for discrete distributions copulas are not unique. Therefore, Scarsini (1984) propose an ordering for discrete distributions. But it is not possible to check whether a measure of concordance hold this ordering. Following Scarsini we consider for discrete distributions a special copula and define an ordering based on this special copula. We give formulas for the traditional measures of concordance (like Kendall's τ and Spearman's ρ) which hold the ordering and are only slight modifications of the well-known formulas.

1 Einleitung

Zumeist werden in der Literatur Axiomatiken für bivariate Abhängigkeitsmaße vorgeschlagen, die auf einer Ordnung fußen, mit der sich nur bivariate Verteilungen mit identischen Randverteilungen vergleichen lassen (siehe die Übersicht in Klein (2000)). Es gibt einige Ansätze, die diese Restriktion aufheben (siehe die Übersicht in Averous et al. (1999)). Diese Ansätze verzichten zwar auf die Forderung identischer Randverteilungen, gehen aber von stetigen Randverteilungen aus. Einige dieser Vorschläge basieren auf dem Konzept der Copula (siehe z.B. Schweizer & Wolff (1981)), worunter eine bivariate Verteilungsfunktion auf dem Einheitsquadrat verstanden wird. Der Übergang von einer bivariaten Verteilungen zu ihrer Copula "rechnet" den Einfluß der Randverteilungen aus der gemeinsamen Verteilung heraus. Man könnte von einer "Randverteilungsstandardisierung" sprechen. Die bekannten Definitionsformeln für die wichtigsten Konkordanzmaße bei Vorliegen stetiger bivariater Verteilungen (wie z.B. Kendalls τ und Spearmans ρ) halten die copulabasierte Ordnung ein. Die Copula ist aber nur eindeutig festgelegt, wenn es sich um eine stetige bivariate Verteilung handelt. Eine solche Annahme ist für ordinalskalierte Merkmale mit einer fixen, zumeist kleinen Zahl möglicher Merkmalsausprägungen unrealistisch, da dann große Mengen von Bindungen unvermeidbar sind. Scarsini (1984) erweitert deshalb die copulabasierte Ordnung auf diskrete bivariate Verteilungen, für die es wegen der fehlenden Eindeutigkeit unendlich viele Copulas geben kann. Deshalb ist diese Erweiterung auch nicht operational. In Anlehnung an Scarsini wird eine spezifische Copula diskutiert, die die Form einer bivariaten, stückweise linear approximierenden Verteilungsfunktion auf dem Einheitsquadrat besitzt. Mit dieser spezifischen Copula läßt sich eine spezifische Konkordanzordnung einführen, die einen Vergleich von bivariaten Häufigkeitsmatrizen mit unterschiedlichen Anzahlen von Merkmalsausprägungen und/oder unterschiedlichen Randverteilungen erlaubt. Die üblichen Formeln für die bekannten Konkordanzmaße bei Vorliegen von Ties halten diese Ordnung nicht ein und erlauben somit keinen Konkordanzvergleich bei unterschiedlichen Randverteilungen. Wendet man aber die entsprechenden Formeln dieser Maße für stetige bivariate Verteilungen auf die spezifische Copula an, so ergeben sich modifizierte neue Berechnungsformeln, die im Regelfall einfacher zu berechnen sind als die "Klassischen". Die so modifizierten Maße halten naturgemäß die copulabasierte Ordnung ein und stehen, wie eingehend gezeigt wird, in einem engen Zusammenhang zu den "klassischen" Konkordanzmaßen bei Vorliegen von Ties.

2 Copula: Bivariate Verteilungsfunktion auf dem Einheitsquadrat

Da eine Copula im wesentlichen die Eigenschaften einer bivariaten Verteilungsfunktion besitzt, ist die Quasi-Monotonie von zentraler Bedeutung:

Definition 2.1 *Eine Funktion $\phi : \mathbb{R}^2 \to \mathbb{R}$ heißt quasi-monoton, wenn*

$$\phi(x, y) - \phi(x', y) - \phi(x, y') + \phi(x, y) \geq 0$$

für alle $x \geq x'$, $y \geq y'$ gilt.

Eine Copula ist eine auf das Einheitsquadrat restringierte stetige bivariate Verteilungsfunktion, deren Randverteilungen Rechteckverteilungen über dem Intervall $[0, 1]$ sind. D.h. durch den Übergang zu Copulas findet eine Elimination des Einflusses der Randverteilungen statt.

Definition 2.2 *Eine Copula ist eine Funktion* $C : [0, 1] \times [0, 1] \to [0, 1]$ *mit:*

1. C *ist quasi-monoton,*
2. $C(u, 0) = C(0, v)$, $C(u, 1) = u$, $C(1, v) = v$ *für alle* $u, v \in [0, 1]$.

Die folgenden wichtigen Eigenschaften zeichnen eine Copula aus (siehe z.B. Scarsini (1984), Schweizer & Wolff (1981)):

1. C ist stetig auf $[0, 1] \times [0, 1]$.
2. Sei $F_{X,Y}$ eine bivariate Verteilungsfunktion mit den Randverteilungsfunktionen F_X und F_Y, dann existiert eine Copula C mit

   $$F_{X,Y}(x, y) = C_{X,Y}(F_X(x), F_Y(y)) \text{ für alle } x, y \in \mathbb{R}.$$

   Sind F_X und $_Y$ stetig, dann ist $C_{X,Y}$ eindeutig. Ansonsten ist die Eindeutigkeit nur auf dem Bildbereich von F_X und F_Y gegeben.
3. Sind F_X und F_Y stetig, dann ist $C_{X,Y}$ die bivariate Verteilung von $U = F_X(X)$, $V = F_Y(Y)$.
4. Die Copula $C^0_{X,Y}(u, v) = uv$ für $u, v \in [0, 1]$ gehört zu der bivariaten Verteilung $F^0_{X,Y}(x, y) = F_X(x)F_Y(y)$ für $x, y \in \mathbb{R}$, die sich bei Unabhängigkeit von X und Y einstellt.
5. Jede Copula läßt sich durch die zwei Copula $C^-(u, v) = \max(0, u + v - 1)$ und $C^+ = \min(u, v)$ entsprechend

   $$C^-(u, v) \leq C(u, v) \leq C^+(u, v)$$

   alle $u, v \in [0, 1]$ begrenzen. Betrachtet man zu einer bivariaten Verteilung $F_{X,Y}$ mit den Randverteilungen F_X und F_Y die bivariaten Verteilungen

   $$\begin{aligned} F^-_{X,Y}(x, y) &= C^-(F_X(x), F_Y(x)) = \max(0, F_X(x) + F_Y(y) - 1) \\ F^+_{X,Y}(x, y) &= \min(F_X(x), F_Y(y)) \end{aligned}$$

   für $x, y \in \mathbb{R}$, dann besitzen $F^-_{X,Y}$ und $F^+_{X,Y}$ dieselben Randverteilungen wie $F_{X,Y}$ und sind im Sinne der positiven Quadrantenordnung von Lehmann (1966) die "mininale" bzw. "maximale" bivariate Verteilung bei gegebenen Randverteilungen. Im folgenden werden diese Verteilungsfunktionen Fréchet-Hoeffding-Grenzen genannt.

6. Wenn g streng monoton zunehmend (abnehmend) auf dem Bildbereich von F_X ist und $Y = g(X)$ gilt, dann ist $C_{X,Y} = C^+$ ($C_{X,Y} = C^-$.

7. Sind g bzw. h auf dem Bildbereich von F_X bzw. F_Y streng monoton zunehmend, dann ist $C_{g(X),h(Y)} = C_{X,Y}$. D.h. eine Copula ist invariant bezüglich streng monoton zunehmenden Transformationen der Argumente. Damit ist eine Copula als "Verteilungsfunktion" für ordinalskalierte Merkmale prädestiniert.

8. Streng monoton abnehmende Transformationen g bzw. h bewirken eine "Spiegelung" der Copula in dem folgenden Sinne:

$$\begin{aligned} C_{g(X),Y}(u,v) &= v - C_{X,Y}(1-u,v), \\ C_{X,h(Y)}(u,v) &= u - C_{X,Y}(u,1-v), \\ C_{g(X),h(Y)}(u,v) &= u + v - 1 + C_{X,Y}(1-u,1-v) \end{aligned}$$

   für $u, v \in [0,1]$.

Eine Copula ist insofern eine "standardisierte" bivariate Verteilung, als der Einfluß der Randverteilungen F_X und F_Y eliminiert wird. Damit eignen sich Copula, um eine Abhängigkeitsordnung einzuführen, die nicht die Fixierung der Randverteilungen voraussetzt.

3 Copulaordnung stetiger bivariater Verteilungsfunktionen

Lehmann (1966) betrachtete die Eigenschaft der positiven Quadrantenabhängigkeit zweier Zufallsvariablen, womit sich die positive Quadrantenordnung auf der Menge der bivariaten Verteilungsfunktionen mit identischen Randverteilungen einführen läßt.

Definition 3.1 *Seien F und F' bivariate Verteilungsfunktionen mit denselben Randverteilugen F_X und F_Y, dann heißt F positiver quadrantenabhängig als F' (kurz: $F \succeq^q F'$), wenn*

$$F(x,y) \geq F'(x,y) \text{ für alle } x, y \in \mathbb{R}$$

gilt.

Streng genommen wird nur definiert, ob eine bivariate Verteilung nicht negativer quadrantenabhängig ist als eine andere. In diesem Sinne soll im folgenden stets "positiver abhängig als" verstanden werden. $\succeq^q$ definiert die stochastische Dominanz von bivariaten Verteilungen.
Die positive Quadrantenabhängigkeit wird von Scarsini (1984) auf einen Ordnung verallgemeinert, die bivariate Verteilungen mit unterschiedlichen Randverteilungen bezüglich der positiven Abhängigkeit zu vergleichen erlaubt. Sie basiert auf den zu den Verteilungen gehörenden Copula, die den Einfluß der unterschiedlichen Randverteilungen eliminieren können.

Definition 3.2 *Seien $F_{X,Y}$ und $F_{X',Y'}$ stetige bivariate Verteilungsfunktionen mit den Copula C und C'. Dann heißt $F_{X,Y}$ positiver copulaabhängig als $F_{X',Y'}$ (kurz: $F_{X,Y} \succeq^c F_{X',Y'}$), wenn*

$$C_{X,Y}(u,v) \geq C_{X',Y'}(u,v) \text{ für alle } u,v \in [0,1]$$

gilt.

Wenn $F_{X,Y}$ und $F_{X',Y'}$ dieselben Randverteilungen besitzen und $F_{X,Y}$ positiver quadrantenabhängig ist als $F_{X',Y'}$, dann ist

$$F_{X,Y}(x,y) = C_{X,Y}(F_X(x), F_Y(y)) \geq C_{X',Y'}(F_{X'}(x), F_{Y'}(y)) = F_{X',Y'}(x,y)$$

für alle $x, y \in \mathbb{R}$. D.h. bei Fixierung der Randverteilungen geht die copulabasierte Ordnung in die positive Quadrantenordnung über.

4 Bivariate Abhängigkeitsmaße (Konkordanzmaße)

4.1 Axiomatische Anforderungen

Scarsini (1984) gibt - ganz ähnlich wie Klein (2000) - für ein Maß der monotonen Abhängigkeit (im folgenden nur noch Konkordanzmaß genannt) einen Katalog elementarer Anforderungen an. In Erweiterung zu Klein wird die Beschränkung des Vergleichs von bivariaten Verteilungen mit identischen Randverteilungen nicht benötigt und die copulabasierte Ordnung verwendet. Dafür geht Scarsini zunächst nur von bivariaten Verteilungen mit stetigen Randverteilungen aus, die im Falle ordinalskalierter Merkmale mit fixierter Anzahl möglicher Merkmalsausprägungen nicht sinnvoll sind, da eine Vielzahl von Bindungen vorliegen. Wir geben eine kurze, nicht formale Definition eines Konkordanzmaßes nach Scarsini (1984).

Definition 4.1 *Ein Maß ist ein Konkordanzmaß, wenn*

1. *es für jede stetige bivariate Verteilungsfunktion definiert ist,*
2. *es symmetrisch ist, d.h. für $F_{X,Y}$ und $F_{Y,X}$ denselben Wert annimmt,*
3. *es nur Werte zwischen -1 und $+1$ annimmt,*
4. *es im Falle der Unabhängigkeit, den Wert 0 annimmt,*
5. *sich die Werte für $F_{X,Y}$ und $F_{-X,-Y}$ nur durch das Vorzeichen unterscheiden,*
6. *es über eine Stetigkeitseigenschaft bezüglich der schwachen Konvergenz von Folgen bivariater Verteilungen verfügt und*
7. *es die copulabasierte Ordnung einhält.*

Scarsini zeigt u.a., daß ein Konkordanzmaß den Wert -1 ($+1$) annimmt, wenn Y fast sicher eine streng monoton abnehmende (zunehmende) Funktion von X ist.

4.2 Konkordanzmaße mit Scorefunktion

Scarsini (1984) betrachtet eine allgemeine Klasse von Konkordanzmaßen, die die soeben genannten Anforderungen erfüllen. Die Elemente dieser Klasse hängen von einer Scorefunktion a ab, die auf $[-1/2, 1/2]$ eine beschränkte und ungerade Funktion ist. a determiniert via

$$I(F_{X,Y}) = \int_0^1 \int_0^1 ka(u-1/2)a(v-1/2)dC_{X,Y}$$

mit der Normalisierungskonstante

$$k = \left(\int_0^1 a^2(u-1/2)du \right)^{-1}$$

ein Konkordanzmaß.

Beispiel 4.1 *Setzt man $a(u) = u$ für $u \in [-1/2, 1/2]$, dann erhält man den Rangkorrelationskoeffizienten ρ von Spearman*

$$\rho(F_{X,Y}) = 12 \int_0^1 \int_0^1 (u-1/2)(v-1/2)dC_{X,Y}(u,v).$$

Integriert man die einzelnen Summanden des Integranden - soweit möglich aus - so erhält man die Darstellung

$$\rho(F_{X,Y}) = 12 \int_0^1 \int_0^1 uvdC_{X,Y}(u,v) - 3.$$

Partielle Integration führt schließlich zu einer dritten Darstellung von Spearmans ρ als

$$\rho(F_{X,Y}) = 12 \int_0^1 \int_0^1 (C_{X,Y}(u,v) - uv)dudv.$$

Setzt man die über $[0,1]$ rechteckverteilten Zufallsvariablen $U = F_X(X)$ und $Y = F_Y(Y)$ ein, so ergibt sich aus der zweiten und dritten Darstellung vonb Spearmans ρ mittels Variablentransformation die bekannteren Formeln

$$\rho(F_{X,Y}) = 12 \int \int (F_{X,Y}(x,y) - F_X(x)F_Y(y))f_X(x)f_Y(y)dxdy$$

und

$$\rho(F_{X,Y}) = 12 \int \int F_X(x)F_Y(y)f_{X,Y}(x,y)dxdy - 3.$$

Beispiel 4.2 *Wählt man $a(u) = sign(u)$ für $u \in [-1/2, 1/2]$, dann ergibt sich die Quadrantenstatistik q von Blomquist, die im folgenden nicht weiter diskutiert werden soll.*

4.3 Konkordanzmaße ohne Scorefunktion

Nicht Spezialfall der auf einer Scorefunktion beruhenden Klasse von Konkordanzmaßen sind z.B. Kendalls τ, das Maß G von Gini und eine Erweiterung eines von Vogel & Wiede (1994) vorgeschlagenen Maßes. Für Kendalls τ und das Maß von Gini (per Analogieschluß) zeigt Scarsini, daß sie Konkordanzmaße im Sinne der Definition sind. Da von stetigen Verteilungen ausgegangen wird, treten keine Ties auf, so daß Kendalls τ und das γ-Maß von Goodman & Kruskal identisch sind.

Beispiel 4.3 *Ausgangspunkt für das Kendallsche Konkordanzmaß ist die sog. Konkordanzwahrscheinlichkeit*

$$\begin{aligned} F^c &= \int\int P(X > x, Y > y) f_{X,Y}(x,y)dxdy \\ &= \int\int (1 - P(X \leq x) - P(Y \leq y) + P(X \leq x, Y \leq y)) f_{X,Y}(x,y)dxdy \\ &= 1 - \int F_X(x) f_X(x) - \int F_Y(y) f_Y(y)dy + \int\int F_{X,Y}(x,y) f_{x,y} dxdy. \end{aligned}$$

Wegen $U = F_X(X)$ und $V = F_Y(Y)$ rechteckverteilt über dem Intervall $[0,1]$ ist $\int F_X(x) f_X(x)dx = \int F_Y(y) f_Y(y)dy = 1/2$, so daß

$$F^c = \int\int F_{X,Y}(x,y) f_{X,Y}(x,y)dxdy$$

ist. Für die sog. Diskordanzwahrscheinlichkeit gilt analog

$$\begin{aligned} F^d &= \int\int P(X < x, Y > y) f_{X,Y}(x,y)dxdy \\ &= \int\int (P(X < x) - P(X < x, Y \leq y)) f_{X,Y}(x,y)dxdy \\ &= 1/2 - F^c. \end{aligned}$$

D.h. $F^c + F^d = 1/2$. Damit ist das Kendallsche τ

$$\tau(F_{X,Y}) = \frac{F^c - F_d}{F^c + F^d} = \frac{2F^c - 1/2}{1/2} = 4F^c - 1$$

oder

$$\tau(F_{X,Y}) = 4 \int\int F_{X,Y}(x,y) f_{X,Y}(x,y)dxdy - 1.$$

Beachtet man $F_{X,Y}(x,y) = C_{X,Y}(F_X(x), F_Y(y)$, und setzt man wiederum die über $[0,1]$ rechteckverteilten Zufallsvariablen $U = F_X(X)$ und $V = F_Y(Y)$ ein, so folgt mit

$$\tau(F_{X,Y}) = 4 \int_0^1 \int_0^1 C_{X,Y}(u,v) dC_{X,Y}(u,v) - 1$$

für $u, v \in [0,1]$ eine Darstellung von Kendalls τ mittels Copula.

Beispiel 4.4 *Ein anderes Maß, das Scarsini betrachtet stammt von Gini und lautet*

$$G = 2\int_0^1 \int_0^1 (|1-u-v| - |u-v|) dC_{X,Y}(u,v).$$

Beispiel 4.5 *Vogel & Wiede (1994) haben ein Konkordanzmaß für diskrete bivariate Verteilungen vorgeschlagen, das randverteilungsnormiert ist, d.h. die Extremwerte* -1 *und* $+1$ *für die Fréchet-Hoeffding-Grenzen annimmt und damit von den Randverteilungen abhängt. Das Maß von Vogel & Wiede geht von einem Abstandskonzept zwischen empirischen Verteilungen aus und berücksichtigt insbesondere die Abstände zwischen einer bivariaten Verteilung und den zugehörigen Fréchet-Hoeffding-Grenzen bzw. der bivariaten Verteilung, die sich bei Unabhängigkeit einstellt. Überträgt man den Ansatz von Vogel & Wiede auf bivariate stetige Verteilungen, dann bietet sich an, den Anstand zweier stetiger bivariater Verteilungen über den Abstand der zugehörigen Copula zu messen, z.B. in Anlehnung an Vogel & Wiede als*

$$d(F_{X,Y}, F_{X',Y'}) = d(C_{X,Y}, C_{X',Y'}) = \int_0^1 \int_0^1 |C_{X,Y}(u,v) - C_{X',Y'}(u,v)| du dv.$$

Dann definiert

$$\eta(F_{X,Y}) = \begin{cases} 1 - d(C_{X,Y}, C^+_{X,Y})/d(C^0_{X,Y}, C^+_{X,Y}) & \text{für } d(C_{X,Y}, C^+_{X,Y}) \le d(C^0, C^+_{X,Y}) \\ d(C_{X,Y}, C^-_{X,Y})/d(C^0_{X,Y}, C^-_{X,Y}) - 1 & \text{für } d(C_{X,Y}, C^+_{X,Y}) > d(C^0_{X,Y}, C^+_{X,Y}) \end{cases}$$

ein Maß, von dem noch zu zeigen ist, daß es die Anforderungen von an ein Konkordanzmaß im Sinne von Scarsini erfüllt. Für fixierte Randverteilungen sind die Anforderungen in Klein (2000) überprüft worden.

5 Copulaordnung bivariater diskreter Verteilungen

Der Bildbereich von Verteilungsfunktionen diskreter Zufallsvariablen besteht nur aus einer höchstens abzählbaren Teilmenge des Intervalls $[0,1]$. Copula sind jedoch auf dem ganzen Intervall $[0,1]$ definiert, so daß zunächst eine Subcopula als "Copula" auf abzählbaren Teilmengen eingeführt wird.

Definition 5.1 *Seien* A *und* B *Teilmengen von* $[0,1]$ *die jeweils* 0 *und* 1 *enthalten. Eine Subcopula ist eine Funktion* $C^* : A \times B \to [0,1]$ *mit:*

1. C^* *ist quasi-monoton,*
2. $C^*(u,1) = u$ *für alle* $u \in B$ *und* $C^*(1,v) = v$ *für alle* $v \in A$.

Wir betrachten speziell bivariate diskrete Verteilungen zweier qualitativ-geordneter Merkmale U bzw. V, die die möglichen Ausprägungen

$$u_1 < u_2 < \ldots u_k \text{ bzw } v_1 < v_2 < \ldots < v_l.$$

besitzen können. Es sollen in einer Erhebung n paarweise Beobachtungen (x_p, y_p), $p = 1,2,\ldots,n$ vorliegen, wobei es für alle $p = 1,2,\ldots,n$ genau ein $i \in \{1,2,\ldots,k\}$ und ein

$j \in \{1, 2, \ldots, l\}$ gibt, so daß $x_p = u_i$ und $y_p = v_j$ gilt. (x_p, y_p), $p = 1, 2, \ldots, n$ heißen Merkmalswerte. Für $i = 1, 2, \ldots, k$ und $j = 1, 2, \ldots, l$ bezeichnen wir mit $n_{U,V}(i, j)$ die Anzahl der Beobachtungspaare mit den Merkmalsausprägungen (u_i, v_j). Die relative Häufigkeit, mit der (u_i, v_j) auftritt wird mit $f_{U,V}(i, j)$ bezeichnet. Die relativen Randhäufigkeiten sind

$$\begin{aligned} f_U(i) &= \sum_{j=1}^{l} f_{ij} = n_U(i)/n = f_{i.}, \\ f_V(j) &= \sum_{i=1}^{k} f_{ij} = n_V(j)/n = f_{.j}. \end{aligned}$$

für $i = 1, 2, \ldots, k$ und $j = 1, 2, \ldots, k$. Summiert man die relativen Häufigkeiten entsprechend der Ordnung von U und V auf, so ergeben sich die kumulierten relativen Häufigkeiten

$$F_{U,V}(i, j) = \sum_{p \le i, q \le j} f_{UV}(p, q) = F_{ij}$$

mit den kumulierten relativen Randhäufigkeiten

$$\begin{aligned} F_U(i) &= \sum_{p=1}^{i} \sum_{j=1}^{l} f_{U,V}(p, j) = F_{U,V}(i, l) = F_{il}, \\ F_V(j) &= \sum_{i=1}^{k} \sum_{q=1}^{j} f_{U,V}(i, q) = F_{U,V}(k, j) = F_{kj} \end{aligned}$$

für $i = 1, 2, \ldots, k$ und $j = 1, 2, \ldots, l$. Es wird im folgenden sowohl die längere Notation mit den Merkmalen U und V als Index und (i, j) als Argument als auch die Kurznotation mit ij als Index verwendet. $\mathbf{F}_{U,V}$ (oder kurz $\mathbf{F}$) bezeichnet im folgenden die $(k \times l)$-Tabelle (oder Matrix) der kumulierten relativen Häufigkeiten $(F_{ij})_{i=1,2,\ldots,k, j=1,2,\ldots,l}$. Damit es sich um "echte" $(k \times l)$-Tabellen handelt, soll angenommen werden, daß $f_U(k) \neq 0$ und $f_V(l) \neq 0$ sind.

Beispiel 5.1 *Sei die $(k \times l)$-Matrix $\mathbf{F}$ von kumulierten relativen Häufigkeiten gegeben. Definiere mittels der kumulierten relativen Randverteilungen von die Mengen $A = \{0, F_{1,l}, \ldots, F_{k,l}\}$ und $B = \{0, F_{k,1}, \ldots, F_{k,l}\}$. Dann ist $C^*(u, v) = F_{i,j}$ für $u = F_{i,l}$ und $v = F_{l,j}$ eine Subcopula.*

Subcopulas lassen sich auf vielfältige Weise zu Copulas auf der Menge $[0, 1] \times [0, 1]$ erweitern. Scarsini (1984) führt eine Konkordanzordnung zwischen diskreten bivariaten Verteilungen mit Hilfe der erweiterten Copula ein. Wir wollen die Definition nur für den Spezialfall zweier qualitativ-geordneter Merkmale mit endlich vielen möglichen Ausprägungen angeben. Zunächst sind Mengen "minimaler" bzw. "maximaler" Copula einzuführen.

Bezeichnung 5.1 *Es sei die $(k \times l)$-Matrix $\mathbf{F}$ der kumulierten relativen Häufigkeiten gegeben. Sei weiterhin $\mathcal{C}_{\mathbf{F}}$ die Menge aller Copula mit $C(F_{il}, F_{kj}) = F_{ij}$ für $i = 1, 2, \ldots, k$*

und $j = 1, 2, \ldots, l$. Wir betrachten die Teilmenge $\mathcal{C}_{\mathbf{F}}^-$, deren Elemente bez. der stochastischen Dominanz in dem folgenden Sinne minimal sind: Für alle $C \in \mathcal{C}_{\mathbf{F}} \setminus \mathcal{C}_{\mathbf{F}}^-$ und $C^- \in \mathcal{C}^-$ existieren $u, v \in [0, 1]$ mit

$$C(u, v) > C^-(u, v).$$

In einem ähnlichen Sinne kann die Teilmenge $\mathcal{C}_{\mathbf{F}}^+$ "maximaler" Copula ausgezeichnet werden. Für alle $C \in \mathcal{C}_{\mathbf{F}} \setminus \mathcal{C}_{\mathbf{F}}^+$ und $C^+ \in \mathcal{C}_{\mathbf{F}}^+$ existieren $u, v \in [0, 1]$ mit

$$C(u, v) < C^+(u, v).$$

Scarsini (1984) führt eine copulabasierte Ordnung der monotonen Abhängigkeit ein, in dem er verlangt, daß die "minimalen" Copula der Verteilung mit der stärkeren positiven Abhängigkeit die "maximalen" Copula der anderen Verteilung dominiert.

Definition 5.2 *Seien $\mathbf{F}$ und $\mathbf{F}'$ nicht notwendig gleich dimensionierte Matrizen kumulierter relativer Häufigkeiten. Dann heißt $\mathbf{F}$ positiver copulaabhängig als $\mathbf{F}'$ (kurz: $\mathbf{F} \succeq \mathbf{F}'$), wenn*

$$C_{\mathbf{F}}^-(u, v) \geq C_{\mathbf{F}'}^+(u, v) \quad \textit{für alle } u, v \in [0, 1]$$

und alle $C_{\mathbf{F}}^- \in \mathcal{C}_{\mathbf{F}}^-$ und $C_{\mathbf{F}'}^+ \in \mathcal{C}_{\mathbf{F}'}^+$ gilt.

Da für stetige bivariate Verteilungen die Copula eindeutig sind, reduziert sich in diesem Falle diese Definition auf die obige Eigenschaft der stochastischen Dominanz der Copula. Mit dieser Definition sind prinzipiell positive Abhängigkeiten zwischen Merkmalen mit unterschiedlichen Randverteilungen und unterschiedlicher Anzahl möglicher Merkmalsausprägungen vergleichbar. Leider ist die Definition nicht operabel, da das geforderte Dominanzkriterium nicht für alle Elemente der jeweiligen Mengen minimaler und maximaler Copula überprüft werden kann. Im folgenden soll deshalb eine spezifische Copula herausgegriffen werden.

6 Spezifische Copula C^s

Für eine $(k \times l)$-Matrix $\mathbf{F}$ kumulierter relativer Häufigkeiten ist die Subcopula auf der Menge $A \times B$ mit $A = \{0, F_{1,l}, \ldots, F_{k,l}\}$ und $B = \{0, F_{k,1}, \ldots, F_{k,l}\}$ definiert. Die zugehörigen Funktionswerte F_{ij} sollen nun linear verbunden werden, d.h. es wird

$$\begin{aligned} C_{\mathbf{F}}^s(u, v) &= F_{i-1,j-1} + \frac{F_{i-1,j} - F_{i-1,j-1}}{f_{.j}}(v - F_{k,j-1}) \\ &\quad + \frac{F_{i,j-1} - F_{i-1,j-1}}{f_{i.}}(u - F_{i-1,l}) + \frac{f_{ij}}{f_{i.} f_{.j}}(u - F_{i-1,l})(v - F_{k,j-1}) \end{aligned}$$

für $F_{i-1,l} < u \leq F_{i,l}$, $F_{k,j-1} < v \leq F_{k,j}$ und $0 \leq u, v \leq 1$. eingeführt, wobei $F_{0,0} = F_{0,i} = F_{j,0} = 0$ für $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$ gesetzt wird.
Dann ist offensichtlich

$$F_{i,j} = C_{\mathbf{F}}^s(F_{i,l}, F_{l,j})$$

für $i = 1, 2, \ldots, k$ und $j = 1, 2, \ldots, l$. Durch die lineare Interpolation entsteht offensichtlich eine auf $[0,1] \times [0,1]$ quasi-konvexe Funktion mit

$$C_{\mathbf{F}}(u, 1) = C_{\mathbf{F}}(u, F_{kl}) = u$$

und

$$C_{\mathbf{F}}(1, v) = C_{\mathbf{F}}(F_{kl}, v) = v$$

für $u, v \in [0,1]$, so daß das folgende Lemma gilt.

Lemma 6.1 *$C^s_{\mathbf{F}}$ ist eine Copula.*

Die zu dieser spezifischen Copula gehörenden Häufigkeitsdichte ist

$$f_{ij}/(f_{i.}f_{.j}) \text{ für } F_{i-1,l} < u \leq F_{il},\ F_{k,j-1} < v \leq F_{k,j}$$

und $0 \leq u, v \leq 1$. Sie ist eine stückweise konstante Funktion, so daß die Häufigkeitsmasse f_{ij} gleichmäßig auf das Quadrat $(F_{i-1,l}, F_{i,l}] \times (F_{k,j-1}, F_{k,j}]$ aufgeteilt wird. Diese Aufteilung mutet auf den ersten Blick willkürlich an. Sie basiert aber auf einer Annahme die im Sinne der Entropiemaximierung die geringste benötigte Information voraussetzt.
Statt eine copulabasierte Ordnung über alle Elemente der Mengen "minimaler" bzw. "maximaler" Copula einzuführen, werden für zwei Matrizen kumulierter Häufigkeiten nur die spezifischen Copula verglichen, um zu wissen, ob zwischen ihnen eine Ordnung der monotonen Abhängigkeit besteht.

Definition 6.1 *Seien $\mathbf{F}$ und $\mathbf{F}'$ nicht notwendig gleich dimensionierte Matrizen kumulierter relativer Häufigkeiten. Dann heißt $\mathbf{F}$ positiver copulaabhängig als $\mathbf{F}'$ (kurz: $\mathbf{F} \succeq^s \mathbf{F}'$), wenn*

$$C^s_{\mathbf{F}}(u, v) \geq C^s_{\mathbf{F}'}(u, v) \text{ für alle } u, v \in [0,1]$$

gilt.

Auch aus dieser Ordnung folgt bei fixierter Randverteilung offensichtlich die positive Quadrantenordnung.

7 Ausgewählte Abhängigkeitsmaße

Es stellt sich die Frage, ob die "klassischen" Konkordanzmaße die Ordnung $\succeq^s$ erhalten. Wenn diese Maße durch die Copula stetiger bivariate Verteilungen definiert sind, wie dies z.B. für Spearmans ρ und Kendalls τ der Fall ist, so kann diese Definition auf die Copula C^s angewendet werden. Diese Maßzahlen stimmen nicht mit den "klassischen" Varianten überein, da diese von punktförmigen Häufigkeitsmassen und keiner gleichförmigen Verteilung der Häufigkeitsmasse über Teilquadraten ausgehen.

7.1 Kendalls τ

Die klassische Variante von Kendalls τ für eine $(k \times l)$-Matrix $\mathbf{F}$ kumulierter relativer Häufigkeiten lautet bekanntlich

$$\tau^d(\mathbf{F}) = \frac{N_c - N_d}{\binom{n}{2}}$$

mit N_c (N_d) als Anzahl der konkordanten (diskordanten) Paare. Für diese Anzahlen gilt (siehe Klein (2000)):

$$\begin{aligned} N_c &= n^2 \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij}^c f_{ij} \\ N_d &= n^2 \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij}^d f_{ij} \end{aligned}$$

mit

$$\begin{aligned} f_{ij}^c &= 1 - F_{kj} - F_{il} + F_{ij} \\ f_{ij}^d &= F_{k,j-1} - F_{i,j-1} \end{aligned}$$

für $i = 1, 2, \ldots, k$ und $j = 1, 2, \ldots, l$ und $F_{i,j} = 0$ für $i = 0$ oder $j = 0$. Definiert man $F_c = N_c/n$ und $F_d = N_d/n$, dann ist

$$\tau^d(\mathbf{F}) = \frac{n(F_c - F_d)}{n(n-1)/2} = \frac{F_c - F_d}{(n-1)/2}$$

eine Funktion der kumulierten relativen Häufigkeiten und von n. Wir schreiben deshalb im folgenden $\tau^d(\mathbf{F}, n)$.
Wie erwähnt lautet Kendalls τ für eine stetige bivariate Verteilung F mit der Copula C_F:

$$\tau = 4 \int_0^1 \int_0^1 C_F(u, v) dC_F(u, v) - 1.$$

Diese Formel läßt sich auch auf die spezifische Copula C^s anwenden, wodurch sich eine weitere Formel für Kendalls Maß ergibt, die ebenfalls eine Funktion von F_{ij}, $i = 1, 2, \ldots, k$ und $j = 1, 2, \ldots, l$ ist. Diese Variante wird im folgenden mit τ^s bezeichnet.

Theorem 7.1 *Sei* $\mathbf{F} = (F_{ij})_{i=1,2,\ldots,k, j=1,2,\ldots,l}$ *eine bivariate Matrix kumulierter relativer Häufigkeiten. Dann gilt:*

$$\tau^s(\mathbf{F}) = 2 \sum_{i=1}^{k} \sum_{j=1}^{l} \left(F_{i-1,j} + F_{i,j-1} + 1/2 f_{ij} \right) f_{ij} - 1.$$

mit $F_{ij} = 0$ *für* $i = 0$ *oder* $j = 0$.

Beweis: Es ist

$$dC^s_{\mathbf{F}}(u,v) = f_{ij}/f_{i.}f_{.j}dudv \text{ für } F_{i-1,l} < u \leq F_{il} \text{ und } F_{k,j-1} < v \leq F_{kj}$$

und $i = 1, 2, \ldots, k$ und $j = 1, 2, \ldots, l$. Damit folgt

$$\tau^s(\mathbf{F}) = 4\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{f_{ij}}{f_{i.}f_{.j}}\int_{F_{i-1,l}}^{F_{il}}\int_{F_{k,j-1}}^{F_{kj}} C^s_{\mathbf{F}}(u,v)dudv - 1.$$

Dabei ist

$$\int_{F_{i-1,l}}^{F_{il}}\int_{F_{k,j-1}}^{F_{kj}} C^s_{\mathbf{F}}(u,v)dudv = A_{ij} + B_{ij} + C_{ij} + D_{ij}$$

mit

$$\begin{aligned} A_{ij} &= \Bigg(F_{i-1,j-1} - \frac{F_{i-1,j} - F_{i-1,j-1}}{f_{.j}}F_{k,j-1} \\ &\quad - \frac{F_{i,j-1} - F_{i-1,j-1}}{f_{i.}}F_{i-1,l} + \frac{f_{ij}}{f_{i.}f_{.j}}F_{i-1,l}F_{k,j-1}\Bigg) \\ &\quad f_{i.}f_{.j} \end{aligned}$$

wegen $\int_{F_{i-1,l}}^{F_{il}}\int_{F_{k,j-1}}^{F_{kj}} dudv = f_{ij}/(f_{i.}f_{.j}$. B_{ij} ist durch

$$\begin{aligned} B_{ij} &= \left(\frac{F_{i-1,j} - F_{i-1,j-1}}{f_{.j}} - \frac{f_{ij}}{f_{i.}f_{.j}}F_{i-1,l}\right)\int_{F_{i-1,l}}^{F_{il}}\int_{F_{k,j-1}}^{F_{kj}} vdudv \\ &= \left(\frac{F_{i-1,j} - F_{i-1,j-1}}{f_{.j}} - \frac{f_{ij}}{f_{i.}f_{.j}}F_{i-1,l}\right) f_{i.}f_{.j}(F_{k,j} + F_{k,j-1})/2 \end{aligned}$$

gegeben. Analog ist

$$C_{ij} = \left(\frac{F_{i,j-1} - F_{i-1,j-1}}{f_{i.}} - \frac{f_{ij}}{f_{i.}f_{.j}}F_{k,j-1}\right) f_{i.}f_{.j}(F_{i,l} + F_{i-1,l})/2.$$

Schließlich folgt für D_{ij}:

$$\begin{aligned} D_{ij} &= \frac{f_{ij}}{f_{i.}f_{.j}}\int_{F_{i-1,l}}^{F_{il}}\int_{F_{k,j-1}}^{F_{kj}} uvdudv \\ &= (F_{i,l} + F_{i,l-1})/2 \cdot (F_{k,j} + F_{k,j-1})/2. \end{aligned}$$

Faßt man in der Summe $A_{ij} + B_{ij} + C_{ij} + D_{ij}$ Terme, die $F_{i-1,j-1}$ enthalten, Terme, die $F_{i-1,j}$ enthalten, Terme, die $F_{i,j-1}$ enthalten und Terme, die f_{ij} enthalten zusammen, so erhält man

$$\int_{F_{i-1,l}}^{F_{il}}\int_{F_{k,j-1}}^{F_{kj}} C^s_F(u,v)dudv = 1/2\,(F_{i-1,j} + F_{i,j-1} + 1/2f_{ij})\,f_{i.}f_{.j}.$$

Damit ist

$$\tau^s(\mathbf{F}) = 2\sum_{i=1}^{k}\sum_{j=1}^{l}(F_{i-1,j} + F_{i,j-1} + 1/2f_{ij})\,f_{ij} - 1. \quad \square$$

Zu beachten ist, daß τ^s nur von kumulierten relativen Häufigkeiten abhängt und damit unabhängig von n ist.
τ^d und τ^s unterscheiden sich nur durch den Faktor $(n-1)/n$, wie der nächste Satz zeigen wird. Vorbereitend für den Beweis dieses Satzes sind zwei Lemmata hilfreich:

Lemma 7.1 *Sei* $\mathbf{F} = (F_{ij})_{i=1,2,\ldots,k,j=1,2,\ldots,l}$ *eine* $(k\times l)$*-Matrix kumulierter relativer Häufigkeiten. Dann gelten*

$$\sum_{j=1}^{l}(F_{kj} + F_{k,j-1})f_{.j} = 1$$

und

$$\sum_{i=1}^{k}(F_{il} + F_{i-1,l})f_{i.} = 1.$$

Beweis: Es ist

$$\begin{aligned} 1 &= 2\int_0^1 udu = 2\sum_{j=1}^{l}\int_{F_{k,j-1}}^{F_{kj}} udu \\ &= 2\sum_{j=1}^{l} 1/2(F_{kj}^2 - F_{k,j-1}^2) = \sum_{j=1}^{l}(F_{kj} + F_{k,j-1})f_{.j}. \end{aligned}$$

Die zweite Beziehung kann analog gezeigt werden. □

Lemma 7.2 *Sei* $\mathbf{F} = (F_{ij})_{i=1,2,\ldots,k,j=1,2,\ldots,l}$ *eine* $(k\times l)$*-Matrix kumulierter relativer Häufigkeiten. Dann gilt:*

$$\sum_{i=1}^{k} F_{i-1,l}f_{i.} - \sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{q=1}^{j} f_{i,q}f_{ij} + 1/2\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}^2 = 1/2.$$

Beweis: Zu zeigen ist, daß

$$\sum_{i=1}^{k} F_{i,l}f_{i.} - \sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{q=1}^{j} f_{i,q}f_{ij} + 1/2\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}^2 - 1/2$$

ist. Setzt man

$$\sum_{i=1}^{k} 1/2(F_{il} + F_{i-1,l})f_{i.} = 1/2$$

ein, so ist zu zeigen

$$D = \sum_{i=1}^{k} F_{i,l}f_{i.} - \sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{q=1}^{j} f_{i,q}f_{ij} + 1/2\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}^2 - \sum_{i=1}^{k} 1/2(F_{il} + F_{i-1,l})f_{i.} = 0.$$

Faßt man den ersten und letzten Summanden zusammen, so ist

$$D = 1/2 \sum_{i=1}^{k} f_{i.}^2 - \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{q=1}^{j} f_{iq} f_{ij} + 1/2 \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} f_{ij}.$$

Mit $f_{i.}^2 = \sum_{j=1} \sum_{q=1}^{j} f_{iq} f_{ij}$ und

$$\sum_{j=1}^{l} \sum_{q=1}^{l} f_{iq} f_{ij} = 2 \sum_{j=1}^{l} \sum_{q=1}^{j} f_{iq} f_{ij} - \sum_{j=1}^{l} f_{ij}^2$$

folgt $D = 0$. $\square$
Damit läßt sich der Satz beweisen, der die beiden Varianten von Kendalls τ verbindet.

Theorem 7.2 *Sei* $\mathbf{F} = (F_{ij})_{i=1,2,\ldots,k, j=1,2,\ldots,l}$ *eine* $(k \times l)$*-Matrix kumulierter relativer Häufigkeiten. Dann gilt:*

$$\tau^s(\mathbf{F}) = \frac{n-1}{n} \tau^d(\mathbf{F}, n)$$

Beweis: Wegen

$$F_{i,j-1} = F_{k,j-1} - f_{ij}^d$$

und

$$F_{i-1,j} = F_{ij} - \sum_{p=1}^{i} f_{pj} = f_{ij}^c + F_{kj} + F_{il} - 1 - \sum_{q=1}^{j} f_{iq}$$

ist

$$F_{i,j-1} + F_{i-1,j} = f_{ij}^c - f_{ij}^d + F_{kj} + F_{k,j-1} + F_{il} - 1 - \sum_{q=1}^{j} f_{iq}$$

für $i = 1, 2, \ldots, k$ und $j = 1, 2, \ldots, l$. Setzt man dies in $\tau^s(F)$ ein, so ergibt sich

$$\begin{aligned} \tau^s(\mathbf{F}) &= 2 \sum_{i=1}^{k} \sum_{j=1}^{l} \left(f_{ij}^c - f_{ij}^d + F_{kj} + F_{k,j-1} + F_{il} - 1 - \sum_{q=1}^{j} f_{iq} + 1/2 f_{ij} \right) f_{ij} - 1 \\ &= n^2 \sum_{i=1}^{k} \sum_{j=1}^{l} \left(f_{ij}^c - f_{ij}^d \right) f_{ij} / (n^2/2) + 2D \end{aligned}$$

mit

$$D = \sum_{i=1}^{k} \sum_{j=1}^{l} (F_{kj} + F_{k,j-1}) f_{.j} + \sum_{i=1}^{k} \sum_{j=1}^{l} (F_{il} - \sum_{q=1}^{j} f_{iq} + 1/2 f_{ij}) f_{ij} - 3/2.$$

Nach den vorstehenden Lemmata hat der erste Summand den Wert 1 und der zweite den Wert 0.5, so daß insgesamt $D = 0$ ist. Mit

$$N_c - N_d = n^2 \sum_{i=1}^{k} \sum_{j=1}^{l} \left(f_{ij}^c - f_{ij}^d \right) f_{ij}$$

und $\binom{n}{2}/(n^2/2) = (n-1)/n$ ist schließlich

$$\tau^s(\mathbf{F}) = \frac{n-1}{n} \tau^d(\mathbf{F}, n) \ \square$$

Liegen keine Ties vor, so ist wegen $(n-1)/n < 1$ $|\tau^s(\mathbf{F})|$ immer kleiner als $|\tau^d(\mathbf{F}, n)|$. Mit wachsendem n vesrchwindet die Differenz zwischen den beiden Varianten von Kendalls τ. Scarsini (1984) hat die vorstehende Beziehung zwischen τ^d und τ^s für den Spezialfall gezeigt, daß keine Ties vorliegen. Es sind dann $u_1 < u_2 < \ldots < u_n$ und $v_1 < v_2 < \ldots < v_n$ geordnete Merkmalswerte, wobei als Beobachtungspaare

$$(u_1, v_{\pi(1)}), (u_2, v_{\pi(2)}, \ldots, (u_n, v_{(\pi(n)})$$

auftreten. π bezeichnet eine Permutation auf der Menge von $\{1, 2, \ldots, n\}$. Die relativen kumulierten Randhäufigkeiten sind in diesem Fall

$$F_{k,j} = j/n \ \text{ und } \ F_{i,l} = i/n$$

und es folgt:

$$f_{i,\pi(i)} \neq 0 \Longrightarrow f_{iq} = 0, f_{pj} = 0 \ \text{ für } p \neq i \text{ und } q \neq \pi(i)$$

für $i = 1, 2, \ldots, k$ und $j = 1, 2, \ldots, l$. D.h. in der Kontingenztabelle der relativen Häufigkeiten ist in jeder Zeile und in jeder Spalte genau ein Element von Null verschieden.
Bekanntlich nimmt τ^d im Falle perfekten monotonen Zusammenhangs nur dann die Grenzen -1 bzw. $+1$ an, wenn keine Bindungen vorliegen. Selbst in diesem Falle kann für endliches n τ^s die genannten Grenzen nicht annehmen, da $(n-1)/n < 1$ ist.
τ^s hält per Konstruktion die Copulaordnung $\succeq^s$ ein. Bei fixiertem n hält damit offensichtlich auch τ^d diese Copulaordnung ein. Wie bereits erwähnt hängt τ^d explizit von n ab, so daß es eine Matrix $\mathbf{F}$ kumulierter relativer Häufigkeiten und $n < n'$ geben kann mit $\tau^d(\mathbf{F}, n) > \tau^d(\mathbf{F}, n')$, wie das folgende Beispiel zeigt:

Beispiel 7.1 *Sei*

$$\mathbf{F} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1 \end{pmatrix}.$$

Dann sind $\tau^d(\mathbf{F}, 2) = 1$ *und* $\tau^d(\mathbf{F}, 100) = 0.5051$. $\tau^s(\mathbf{F}) = \lim_{n\to\infty} \tau^d(\mathbf{F}, n) = 0.5$.

Interessanter ist die Fragestellung, ob es $\mathbf{F}$, $\mathbf{F}'$ und n, n' gibt, so daß $\mathbf{F} \succeq^s \mathbf{F}'$ gilt und damit $\tau^s(\mathbf{F}) \geq \tau^s(\mathbf{F}')$, aber $\tau^d(\mathbf{F}, n) < \tau^d(\mathbf{F}', n')$ sind.

Beispiel 7.2 *Seien*

$$\mathbf{F} = \begin{pmatrix} 0.1 & 0.41 \\ 0.1 & 0.39 \end{pmatrix} \ \textit{und} \ \mathbf{F} = \begin{pmatrix} 0.1 & 0.411 \\ 0.1 & 0.389 \end{pmatrix}.$$

Dann sind $\mathbf{F} \succeq \mathbf{F}'$ *und* $\tau^s(\mathbf{F}) = -0.004 > -0.0044 = \tau^s(\mathbf{F}')$. *Es ist aber* $\tau^d(\mathbf{F}, 2) = -0.008 < \tau^d(\mathbf{F}', 10) = -0.0049$, *so daß* τ^d *die Copulaordnung nicht einhält.*

7.2 Spearmans ρ

Wenn keine Bindungen vorliegen gibt es eine eindeutig festgelegte Formel für den Rangkorrelationskoeffizienten von Spearman (siehe z.B. Scarsini (1984) oder Büning & Trenkler (1994)):

$$\rho^d(\mathbf{F}) = \frac{1}{n(n^2-1)} \left(12 \sum_{i=1}^{n} i\pi(i) - 3n(n+1)^2 \right).$$

Im Falle von Bindungen ist prinzipiell jede Erweiterung dieser Formel denkbar, die im Spezialfall des Fehlens von Bindungen in die vorstehende Form übergeht.
Eine Möglichkeit der Verallgemeinerung ist die Anwendnung des Bravais-Pearson-Korrelationskoeffizienten auf die kumulierten relativen Randhäufigkeiten (siehe Heiler & Michels (1994), Klein (2000)):

$$\rho^d(\mathbf{F}) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(F_{il} - \overline{F}_{.l})(F_{kj} - \overline{F}_{k.})f_{ij}}{\sqrt{\sum_{i=1}^{k}(F_{il} - \overline{F}_{.l})^2 f_{i.} \sum_{j=1}^{l}(F_{kj} - \overline{F}_{k.})^2 f_{.j}}}$$

mit $\overline{F}_{.l} = \sum_{i=1}^{k} F_{il} f_{i.}$ und $\overline{F}_{k.} = \sum_{j=1}^{l} F_{kj} f_{.j}$. Gehören zu $\mathbf{F}$ die geordnet-kategorialen Merkmale U und V, kann ρ^d auch für die Merkmale $-U$ und V berechnet werden. Leider führt dies nicht nur zu einem Vorzeichenwechsel von τ^d, sondern der Absolutbetrag von τ^d verändert sich. Dieser Nachteil kann durch die Behandlung von Ties mittels Midranks gemildert werden.
Zu sog. Midranks gelangt man über die Transformation

$$\overline{F}_{il} = 1/2(F_{il} + F_{i-1,l} + 1/n)$$

bzw.

$$\overline{F}_{kj} = 1/2(F_{kj} + F_{k,j-1} + 1/n)$$

für $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$. Als mögliche Erweiterung von ρ^d im Falle von Bindungen kann der Bravais-Pearson-Korrelationskoeffizient auf die Midranks angewendet werden. Beachtet man, daß wegen Lemma 7.1 die Mittelwerte der Midranks durch

$$\sum_{i=1}^{k} \overline{F}_{il} f_{i.} = 1/2 \sum_{i=1}^{k} (F_{il} + F_{i,l-1}) f_{i.} + 1/(2n) = 1/2 + 1/(2n)$$

und

$$\sum_{j=1}^{l} \overline{F}_{kj} f_{.j} = 1/2 \sum_{j=1}^{l} (F_{kj} + F_{k,j-1}) f_{.j} + 1/(2n) = 1/2 + 1/(2n)$$

gegeben sind und

$$\sum_{i=1}^{k}\sum_{j=1}^{l} \overline{F}_{il}\overline{F}_{kj} f_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{l} ((F_{il} + F_{i-1,l})/2)((F_{kj} + F_{k,j-1})/2) f_{ij} + 1/(2n) + 1/(4n^2)$$

bzw.

$$\sum_{i=1}^{k} \overline{F}_{il}^2 f_{i.} = \sum_{i=1}^{k} ((F_{il} + F_{i-1,l})/2)^2 f_{i.} + 1/(2n) + 1/(4n^2)$$

und

$$\sum_{j=1}^{l} \overline{F}_{kj}^2 f_{.j} = \sum_{j=1}^{l} ((F_{kj} + F_{k,j-1})/2)^2 f_{.j} + 1/(2n) + 1/(4n^2)$$

sind, so ist der Bravais-Pearson-Korrelationskoeffizient der Midranks

$$\rho_m^d(\mathbf{F}) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}((F_{il} + F_{i-1,l})/2)((F_{kj} + F_{k,j-1})/2)f_{ij} - 1/4)}{\sqrt{\sum_{i=1}^{k}(((F_{il} + F_{i-1,l})/2)^2 f_{i.} - 1/4))(\sum_{j=1}^{l}((F_{kj} + F_{k,j-1})/2)^2 f_{.j} - 1/4))}}.$$

Diese Form ist bei fixierten Randverteilungen nur via f_{ij} von F_{ij} abhängig, so daß ähnlich wie in Klein (2000) für die Formel ρ^d nachgewiesen werden kann, daß auch ρ_m^d die positive Quadrantenordnung einhält.
Wie erwähnt lautet die Formel für den Rangkorrelationskoeffizienten von Spearman für stetige bivariate Verteilungen F

$$\rho(F) = 12 \int_0^1 \int_0^1 uvdC_F^s(u,v) - 3.$$

Wendet man diese Formel auf die Copula $C_{\mathbf{F}}^s$ an, so ergibt sich nach wenigen Rechenschritten

$$\rho^s(\mathbf{F}) = 12 \sum_{i=1}^{k}\sum_{j=1}^{l}((F_{il} + F_{i-1,l})/2)((F_{kj} + F_{k,j-1})/2)f_{ij} - 3.$$

Diese Formel hält per Konstruktion die copulabasierte Ordnung und damit die positive Quadrantenordnung ein. ρ_m^d und ρ^s unterscheiden sich lediglich durch den Nenner. Ist

$$\sum_{i=1}^{k}((F_{il} + F_{i-1,l})/2)f_{i.} - 1/4) = \sum_{j=1}^{l}((F_{kj} + F_{k,j-1})/2)f_{.j} - 1/4 = 1/12,$$

so gehen die beiden Formeln ineinander über. Exakt gilt die folgende offensichtliche Beziehung:

Theorem 7.3 *Es ist*

$$\rho^s(\mathbf{F}) = \rho_m^d(\mathbf{F}) \frac{\sqrt{\sum_{i=1}^{k}(\overline{F}_{il} - 1/2 - 1/(2n))^2 f_{i.} \sum_{j=1}^{l}(\overline{F}_{kj} - 1/2 - 1/2n)^2 f_{.j}}}{1/12}.$$

Beide Maße unterscheiden sich also nur in einem Quotienten, der ausschließlich von den kumulierten Randhäufigkeiten abhängt. Bei geeigneter Wahl ist mithin klar, daß es bivariate kumulierte Häufigkeitsmatrizen $\mathbf{F}$ und $\mathbf{F}'$ geben kann, so daß $\mathbf{F} \preceq^s \mathbf{F}'$ und mithin $\rho^s(\mathbf{F}) \geq \rho^s(\mathbf{F}'$ sind und gleichzeitig $\rho_m^d(\mathbf{F}) \leq \rho_m^d(\mathbf{F}'$ ist, womit ρ_m^d die copulabasierte Ordnung nicht einhält.
Im Falle fehlender Ties kann die etwas unhandliche Beziehung zwischen ρ^s und ρ_m^d weitgehend aufgelöst werden. Das folgende Ergebnis wurde auf andere Weise bereits von Scarsini (1984) nachgewiesen:

Korollar 7.1 *Wenn keine Ties vorliegen, gilt*

$$\rho^s(\mathbf{F}) = (n^2-1)/n^2 \rho^d(\mathbf{F}) = (n^2-1)/n^2 \rho_m^d(\mathbf{F}).$$

Beweis: Zunächst ist

$$\begin{aligned}
\rho^s(\mathbf{F}) &= 12 \sum_{i=1}^{n} \frac{1}{2}\left(\frac{i}{n} + \frac{i-1}{n}\right) \frac{1}{2}\left(\frac{\pi(i)}{n} + \frac{\pi(i)-1}{n}\right) f_{i,\pi(i)} - 3 \\
&= \frac{3}{n^3} \sum_{i=1}^{n} (2i-1)(2\pi(i)-1) - 3 \\
&= \frac{1}{n^3}\left(12 \sum_{i=1}^{n} i\pi(i) - 6n(n+1) + 3n - 3n^3\right) \\
&= \frac{1}{n^3}\left(\sum_{i=1}^{n} i\pi(i) - 3n(n^2+2n+1)\right) \\
&= \frac{1}{n^3}\left(\sum_{i=1}^{n} i\pi(i) - 3n(n+1)^2\right).
\end{aligned}$$

Um eine analoge Darstellung von $\rho_m^d(\mathbf{F})$ zu erhalten, sind

$$\overline{F}_{.l} = \sum_{i=1}^{n}\sum_{j=1}^{n} F_{il} f_{ij} = \sum_{i=1}^{n} F_{il} f_{i.} = \sum_{i=1}^{n} i/n 1/n = (n+1)/(2n)$$

und analog $\overline{F}_{k.} = (n+1)/(2n)$ einzusetzen. Weiterhin ist

$$\sum_{i=1}^{n} F_{il}^2 f_{i.} = 1/n^3 \sum_{i=1}^{n} i^2 = (n+1)(2n+1)/(6n^2).$$

Mithin ist auch $\sum_{j=1}^{n} F_{kj}^2 f_{.j} = (n+1)(2n+1)/(6n^2)$. Damit ist

$$\begin{aligned}
\rho^d(F) &= \frac{\sum_{i=1}^{n} i\pi(i)/n^2 - (n+1)^2/(4n^2)}{(n+1)(2n+1)/(6n^2) - (n+1)^2/(4n^2)} \\
&= \frac{1}{n(n^2-1)}\left(12 \sum_{i=1}^{n} i\pi(i) - 3n(n+1)^2\right) = \frac{n^2}{n^2-1}\rho^s(F).
\end{aligned}$$

Die Beziehung zwischen ρ^s und ρ_m^d ergibt sich einfacher, da sich beide Formeln nur im Nenner unterscheiden. Ohne Ties ist aber

$$\begin{aligned}\sum_{i=1}^{k}((F_{il}+F_{i-1,l}+1/n)/2-(1/2+1/(2n)))^2 1/n &= \sum_{i=1}^{n}((2i-1+1)/(2n)\\ &\quad -(n+1)/(2n))^2 1/n\\ &= 1/n^3\sum_{i=1}^{n}(i-(n+1)/2)^2\\ &= (n-1)n(n+1)/(12n^3)\\ &= (n^2-1)/(12n^2).\end{aligned}$$

Für die letzte Umformung siehe z.B. Büning & Trenkler (1994), S. 234. Derselbe Ausdruck ergibt sich auch für $\sum_{j=1}^{l}((F_{kj}+F_{k,j-1}+1/n)/2-(1/2+1/(2n)))^2 1/n$, so daß insgesamt

$$\rho^s(\mathbf{F}) = (n^2-1)/n^2\rho_m^d(\mathbf{F})$$

ist. □

Mit $n \to \infty$ verschwindet die Differenz zwischen $\rho^s(\mathbf{F})$ und $\rho_m^d(\mathbf{F})$, wenn keine Bindungen vorliegen. Ähnlich wie im Falle des Maßes von Kendall kann es n bzw. n' und bivariate kumulierte Häufigkeitsmatrizen geben, so daß ρ_m^d auch im Falle fehlender Ties die copulabasierte Ordnung nicht einhält.

7.3 Beziehung zwischen Kendalls τ^s und Spearmans ρ_m^d

Nelson (1999) hat die Beziehung der Wertebereiche von Kendalls τ und Spearmans ρ für stetige bivariate Verteilungen untersucht. Die dabei gefundenen Ergebnisse lassen sich unmittelbar auf τ^s und ρ^s übertagen, obwohl diese explizit für diskrete bivariate Verteilungsfunktionen bzw. im Falle zweier komparativer Merkmale nur für bivariate kumulierte Häufigkeitsmatrizen definiert sind. In Anlehnung an Nelson (1999), S. 141ff. gelten:

$$-1 \le 3\tau^s(\mathbf{F}) - 2\rho^s(\mathbf{F}) \le 1$$

und

$$\frac{3\tau^s(\mathbf{F})-1}{2} \le \rho^s(\mathbf{F}) \le \frac{1+2\tau^s(\mathbf{F})-\tau^s(\mathbf{F})^2}{2}$$

für $\tau^s(\mathbf{F}) \ge 0$ bzw.

$$\frac{-1+2\tau^s(\mathbf{F})+\tau^s(\mathbf{F})^2}{2} \le \rho^s(\mathbf{F}) \le \frac{3\tau^s(\mathbf{F})+1}{2}$$

für $\tau^s(\mathbf{F}) \le 0$.

8 Zusammenfassung und Ausblick

Konkordanzordnungen werden zumeist für stetige bivariate Verteilungen angegeben und beziehen sich häufig auf die Situation fixierter Randverteilungen. Unter Verwendung des Konzeptes der für stetige bivariate Verteilungen eindeutig festgelegten Copula kann eine Konkordanzordnung eingeführt werden, die auf die Annahme fixierter Randverteilungen verzichtet. Bivariate diskrete Verteilungen fixieren lediglich die Copula auf diskreten Werten. Durch eine lineare Interpolation dieser Subcopula entsteht eine spezifische Copula. Von Konkordanzmaßen wird gefordert, daß sie zumindest die auf dieser spezifischen Copula basierenden Konkordanzordnung einhalten. Berechnet man die klassischen Konkordanzmaße für diese spezifische Copula, so ergeben sich neue Formeln, die im Falle von Kendalls τ und Spearmans ρ nur leicht von den bekannten Formeln anweichen, aber im Gegensatz zu diesen die copulabasierte Ordnung einhalten.
Neben Kendalls τ und Spearmans ρ lassen sich weitere Konkordanzmaße betrachten, deren (theoretische) Formeln für bivariate stetige Verteilungen und deren Abhängigkeit von der Copula bekannt sind. Dies betrifft z.B. Ginis Konkoradanzmaß, Bplomqvists q oder das von Schweitzer & Wolf (1981) genannte Maß. Ebenso ist die oben angegeben theoretische Version für das Maß von Vogel & Wiede (1994) diskutierbar. Dies soll weiteren Arbeiten vorbehalten bleiben.

Literatur

1. Averous, J. & Dortet-Bernadet, J.-L. (2000). LTD and RTI dependence orderings. *Canadian Journal of Statistics* **28**, 151-157.

2. Büning, H. & Trenkler, G. (1994). *Nichtparametrische statistische Methoden.* de Gruyter, Berlin.

3. Heiler, S. & Michels, P. (1994). *Deskriptive und Explorative Datenanalyse.* Oldenbourg-Verlag, München.

4. Klein, I. (2000). Bivariate Abhängigkeitsmessung für ordinalskalierte Merkmale. *Diskussionspapier der Lehrstühle für Statistik der Universität Erlangen-Nürnberg* **28/2000**.

5. Nelson, R.B. (1999). *An Introduction to Copulas.* Springer, New York.

6. Scarsini, M. (1984). On measures of concordance. *Stochastica* **8**, 201-218.

7. Schweizer, B. & Wolff, E.F. (1981). On nonparametric measures of dependence for random variables. *Annals of Statistics* **9**, 879-885.

8. Tchen, A.H. (1980). Inequalities for distributions with given marginals. *Annals of Probability* **8**, 814-827.

9. Vogel, F. & Wiede, T. (1994). Ein neues Zusammenhangsmaß für ordinalskalierte Merkmale. *Jahrbücher für Nationalökonomie und Statistik* **213**, 1-30.