“Slepian-Wolf” 编码及相关性研究

黄冬勃

2015 年 12 月 1 日

目录

Chapter 1

信源压缩方法学习

在需要对信息进行实时传输的应用场景下，结合实时性以及节能性，设计合理的信源编码方案，则必须对计算复杂度和信息压缩率综合权衡。实时性要求是在给定一段时间范围以及一定空间范围内，传感节点采集到并传输的信息要保证其独立性与完整性。若在需要利用中继辅助传输的网络中，中继节点在对数据包进行联合编解码时，必须保证在 sink 端进行解码时，能够分离出每个独立数据包，并匹配其相应的时间与空间信息。计算复杂度主要影响信息采集传输的实时性，且低复杂度计算方法对于减少能耗也有贡献；信息压缩率旨在减少通讯数据量，减少节点能耗。为了得到尽可能大的信息压缩率，可以利用单节点参数变化的时间相关性以及相邻节点同一时刻参数的空间相关性，在采样或处理过程中，保证一定失真度的前提下，尽可能多减少信息的冗余度，即获得尽可能大的信息压缩率。综上考虑，为了使能量受限的传感器节点的使用寿命，需要设计一个计算简单，利用时空相关性对信息进行压缩的信源编码方案，即实时分布式信源编码方案。

1.1 基于字典的信源编码方法

该无损编码方法最初是针对文字信息的，它利用之前接收到的字符信息生成字典对之后的字符信息进行压缩编码。对于非字符信息，该方法可拓展为利用测量参数的时间相关性，计算当前测量值与前一时刻测量值的差值，对该差值进行压缩编码。对于惰性变化参数，该压缩编码方法能够大量减少信息冗余。该方法没有利用空间相关性。

1.2 分布式信源编码 (distributed source coding (DSC))

其主要思想在于独立编码，联合解码。根据 slepian-wolf 定理，两个相关信源 (x, y) 传输信息到一个 sink 节点，对两个信源进行编码时，可以根据其相关性分别独立压缩编码，不需要两信源之间进行额外通信。若 y 作为 x 的边信息，则 x 需要的比特位为 $H(x)$，而 y 只需要传输 $H(x|y)$ 即可。

DSC 利用信源之间的空间相关性对信息进行压缩编码，但各节点自身时间相关性并没有加以利用。由于其编解码方法的非对称性，相关信源在传输路径选择灵活性上受到了限制。在多节点网络中，可以通过分簇的方式，簇头作为该簇各节点的信息汇聚节点，簇内其他节点直接与簇头进行通信，簇头采集信息作为边信息，簇内其他各节点直接传输与簇头采集信息的相关性熵编码即可。

由于对于一些环境参数的编码，就其码字较短的角度来考虑，SW 编码较为适用，问题在于要合理分配节点通信量，尽量平均节点能耗；相关性模型建立。对于例如温湿度等参数，空间相关性通常认为其符合典型的多元高斯分布，相应的其边缘概率密度函数分布也应该为高斯分布，但实际情况中并非如此。有一种更为合适的相关性模型:*Copula-Function-Based Correlation Model*。另外，目前的 DSC 方法并没有能够真正实现平均分配节点能耗，需要考虑能否尽量平均分配节点能耗。

文章 [1] 中提出了将信源建模呈隐形马尔科夫过程 (Hidden Markov Processes (HMPs))，使用 raptor 编码方法对信源进行编码。

1.3 相关性模型

多元高斯相关性模型 (Multivariate Gaussian Correlation Model).

多元高斯相关性模型中，联合概率密度分布函数：

$$f(x_1, x_2, \ldots, x_N) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}} \times exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{1.3.1}$$

但其对于实际应用参数（温湿度等）并不能真实建模。

基于 CF 相关性模型 (CF-Based (Copula Function) correlation model) 连续边缘概率密度函数可以使用 *kernel density estimation (KDE)* 方法来进行估计：

$$f_n(x_n) = \frac{1}{M \times h_n}\sum_{i=1}^{M} K\left(\frac{x-x_i}{h_n}\right). \tag{1.3.2}$$

其中，M 是样本数量。该方法使用高斯核 $K(\mathfrak{v}) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\mathfrak{v}^2\right)$ 进行曲线估计。对于每一个传感器采集样本的 $f_n(x_n)$ 都会选择一个适当的平滑参数 h_n。

相关性矩阵 Γ 中的非对角线元素等于估算 Spearman 的 rho 值，对角线元素为 1。Spearman 参数由每个簇内传感器节点通过“训练”过程进行估计。

SW 编码利用的节点间采集参数的空间相关性，利用时间相关性的话，也许可以将时间变换为空间。

问题：时间变换为空间的模型建立，如何能够联合利用时空相关性。

1.4 分布式压缩感知 (Distributed Compressive Sensing(DCS))

压缩感知是对变换编码的重大改进，传统变换编码 (transform coding(TC)) 流程：

1. 变换编码需要采集全部 N 个采样；
2. 计算全部集合的变换系数 $\{\mathfrak{v}(n)\}$;
3. 计算完毕后定位 K 个重要系数，并丢弃其他不重要的系数；
4. 将 K 个系数进行编码与定位。

但 TC 编码，相比 DCS，本质上的低能效表现在：

1. 必须对参数进行全部采样；
2. 编码器必须计算全部 N 个采样的变换系数 $\{\mathfrak{v}(n)\}$，即使最后仅保留 K 个系数；
3. 编码器必须对重要系数所在位置进行编码，由于每个信号的重要系数的位置是不同的，这样增加了编码率。

所以，通过以上对比，若要使用基于变换的编码方法时，应选择压缩感知。如果可以发现一个计算简单的测量矩阵的话，对于较短码字也可以使用压缩感知的方法，实现更大的压缩率。

弄清楚流程，搞清楚其如何利用时空相关性。最初的 DCS 并不能平均分配节点能耗，后续文章有作者提出了可以平均能能耗的方法，且须注意计算复杂度。

1.5 问题总结和思考

有一点需要注意，节点之间的额外通信也可以考虑在内。例如，虽然在方案设计中不考虑，但实际应用中，在 Ad-hc 网络中，拓扑发现过程中相邻节点间的通信是不可避免的，这样，是可以利用这些通信过程附加少量参数的，可使相邻接点互相发现空间相关性。

对于 SW 编码，变换编码的区别，SW 编码本质上是熵编码，而变换编码在于将信息量化编码后使用向量/矩阵计算的方法得到码字。之前对于 SW 编码理解有所偏差，有文章提到可以在大规模传感网络中使用 SW 编码，需要了解其思路及应用方法。压缩感知在于利用采样值内部本质特征对其进行压缩，所使用方法是基于变换编码的，即要将原始数据看做向量 x 在 R^n 空间中，使用矩阵 $A_{m\times n}, t = Ax$ 映射到 R^m 空间中，$m << n$。

分簇方法缺点：对于 ad-hoc 网络，路径选择无法最优化，且簇头变化有可能导致两簇簇头之间无法通信。

Chapter 2

信源相关性建模 Correlation Modeling

使用 SW 进行编码，节点的时空相关性究竟如何体现？个人认为，需要先有时间空间“协作”概念，需要考虑的是如何“协作”，个人理解:

> 假设一个簇内所有节点都通过储存前一次或几次采集的信息时间相关信息，簇内其他节点都利用簇头节点信息作为边信息 (空间相关信息)。
>
> 基于简单 SW 编码 [2]，利用相关性将码字分 8 个组，簇内其他节点只需要传输 index 即可，我这里将其定义为一维相关性；若有可能，将时间相关性同时利用的话，利用二维相关性，可否将分组减少，这样传输比特位将会更少。
>
> 关于相关性思考，如果利用 CF 方法，假设我们采集环境参数。应该将环境参数随着距离，时间的变化是呈现怎样的分布？

2.1 Basis

两个随机变量 X,Y，他们的分布函数分别为 $F(x)=P[X\le x]$ 和 $G(y)=P[Y\le y]$，联合分布函数为 $H(x,y)=P[X\le x,Y\le y]$。每一对实测值对应的 $F(x),G(y),H(x,y)$ 取值范围都在 $[0,1]$ 之间。即点 $(F(x),G(y))$ 位于单位平方形 $[0,1]\times[0,1]$ 中。

一个*“2-increasing”*函数：一个变量的非减函数的一个二维类比。一个单位矩形 $\mathbf{I}^2=\mathbf{I}\times\mathbf{I}$，其中 $\mathbf{I}=[0,1]$。在一个矩形平面 $\bar{\mathbf{R}}^2$ 中两相邻点的直积 (笛卡尔积 (Cartesian product))$B=[x_1,x_2]\times[y_1,y_2]$，可得 B 的顶点：$(x_1,y_1),(x_1,y_2),(x_2,y_1),(x_2,y_2)$。一个两点实函数 (*2-place real function*) H 的域为 Dom$H$，是 $\bar{\mathbf{R}}^2$ 的一个子集，它的数值域 Ran$H$ 是 $\mathbf{R}$ 的一个子集。

Definition 2.1.0.1. *Let S_1 and S_2 be noempty subsets of $\bar{\mathbf{R}}$, and let H be a 2-place real function such that Dom$H=S_1\times S_2$. Let $B=[x_1,x_2]\times[y_1,y_2]$ be rectangle all of whose vertices are in Dom$H$. Then the* H-volume of B *is given by*

$$V_H(B)=H(x_2,y_2)-H(x_2,y_1)-H(x_1,y_2)+H(x_1,y_1) \tag{2.1.1}$$

如果定义矩形平面 B 中 H 的一阶差分 (first order differences) 为

$$\triangle_{x_1}^{x_2}H(x,y)=H(x_2,y)-H(x_1,y)\quad \triangle_{y_1}^{y_2}H(x,y)=H(x,y_2)-H(x,y_1),$$

then the H-volume of a rectangle B is the *second order difference* of H on B,

$$V_H(B)=\triangle_{y_1}^{y_2}\triangle_{x_1}^{x_2}H(x,y).$$

Definition 2.1.0.2. *A 2-place real function H is 2-increasing if $V_H(B) \geq 0$ for all rectangles B whose vertices lie in $DomH$.*

判断 H 是否为 2-increasing，根据 eq. (2.1.1)，利用 $[x_1, x_2], [y_1, y_2]$ 产生的四个点来进行计算。

2.2 Copulas Function (CF)

2.2.1 CF 基本概念

[3] 根据上节，再引入一个 "零基面的"(*grounded*) 概念：

设 S_1, S_2 是 $\bar{\mathbf{R}}$ 的非空子集，S_1 中最小的元素为 a_1，S_2 中最小元素为 a_2，S_1, S_2 的一个联合分布函数 H，对于 $S_1 \times S_2$ 的所有 (x, y)，都满足 $H(x, a_2) = 0 = H(a_1, y)$ 的话，则 H 是 "零基面的的"。

将 copulas 看作在单位矩形平面 $\mathbf{I}^2$ 中的 subcopulas

Definition 2.2.1.1. *一个二维 subcopula(2-subcopula 或简言之 subcopula) 是一个具有以下性质的函数:*

1. *$DomC' = S_1 \times S_2$(C' 的操作域)，其中 S_1 和 S_2 是单位向量 $\mathbf{I} = [0, 1]$ 的子集;*
2. *C' 是触底且 2-increasing;*
3. *对于 S_1 中任意元素 u，S_2 中任意元素 v，满足*

$$C'(u, 1) = u \quad 和 \quad C'(1, v) = v \tag{2.2.1}$$

接下来将操作域是 $\mathbf{I}$ 子集的 subcopula C' 拓展至操作域为整个 $\mathbf{I}$ 的 copula C：

Definition 2.2.1.2. *一个取值空间为 $\mathbf{I}^2$ 的 subcopula 成为 copula，$C' = \mathbf{I}^2$。Copula 具有以下性质:*

1. *对于 $\mathbf{I}$ 中的每个 u, v，都有*

$$C(u, 0) = \quad 0 \quad = C(0, v) \tag{2.2.2}$$

$$C(u, 1) = 1 \qquad and \qquad C(1, v) = v; \tag{2.2.3}$$

2. *对于 $\mathbf{I}$ 中的每个 u_1, u_2, v_1, v_2，且 $u_1 \leq u_2, v_1 \leq v_2$，则:*

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \tag{2.2.4}$$

从 eq. (2.1.1) 可得，$C(u, v) = V_C([0, u] \times [0, v]) = C(u, v) - C(u, 0) - C(0, v) + C(0, 0)$，可以将 $C(u, v)$ 看做利用一个在 $\mathbf{I}$ 中的数生成一个 $[0, u] \times [0, v]$ 的矩形平面。definition 2.2.1.2中第二点给出了一个依托 "容斥 (inclusion-exclusion)"[1]原理并使用 C 的一个不等式，将给定的 u_1, u_2, v_1, v_2 分配到 $\mathbf{I}^2$ 中的矩形平面 $[u_1, u_2] \times [v_1, v_2]$。

接下来引入 copula 和 subcopula 的几个性质：

Theorem 2.2.1.1. *设 C' 是一个 subcopula，对于 $DomC'$ 中的每一对 (u, v) 都有:*

$$\max(u + v - 1, 0) \leq C'(u, v) \leq min(u, v) \tag{2.2.5}$$

[1] "容斥" 原理是一种计数方法，使计算结果既无遗漏又无重复 (类似拓扑学中的遍历最短路径？)

Proof. (u,v) 为 DomC' 中任意一点。有

$$\begin{aligned} C'(u,v) &\leq C'(u,1) = u \\ &and \\ C'(u,v) &\leq C'(u,v) = v \\ &\Rightarrow \\ C'(u,v) &\leq \min(u,v) \end{aligned}$$

$V_{C'}([u,1]\times[v,1]) \geq 0$eq. (2.2.4) 可引申出 $C'(u,v) \geq u+v-1$，再结合 $C'(u,v)\geq 0$ 可以得出 $C'(u,v)\geq \max(u+v-1,0)$。 □

对于 copula，使 $M(u,v)=\min(u,v)$，$W(u,v)=\max(u+v-1,0)$，则对于每一个 C 和每一个 $\mathbf{I}^2$ 中的 (u,v)，

$$W(u,v)\leq C(u,v)\leq M(u,v) \tag{2.2.6}$$

Theorem 2.2.1.2. 设 C' 是一个 *subcopula*，对于 *Dom*C' 中的每一对 $(u_1,u_2),(v_1,v_2)$ 都有：

$$|C'(u_2,v_2)-C'(u_1,v_1)|\leq |u_2-u_1|+|v_2-v_1| \tag{2.2.7}$$

因此，C' 在它的域中是均匀连续的。

Definition 2.2.1.3. C 是一个 *copula*，a 是 $\mathbf{I}$ 中的一个任意数值。C 在 a 点的水平截面是一个从 $\mathbf{I}$ 到 $\mathbf{I}$ 的函数：$t\mapsto C(t,a)$；垂直截面：$t\mapsto C(a,t)$；对角截面是从 $\mathbf{I}$ 到 $\mathbf{I}$ 的一个函数 $\delta_C = C(t,t)$。

Corollary 2.2.1.1. *Copula* C 的水平截面，垂直截面以及对角截面在 $\mathbf{I}$ 上都是非减，均匀连续的。

Theorem 2.2.1.3. C 为一个 *copula*，对于 $\mathbf{I}$ 中的任意 v，对于几乎所有的 u，存在偏微分 $\partial C(u,v)/\partial u$，并且对于 v 和 u，有

$$0\leq \frac{\partial}{\partial u}C(u,v)\leq 1 \tag{2.2.8}$$

相似的，对于 $\mathbf{I}$ 中的任意 u，对于几乎所有的 v，存在偏微分 $\partial C(u,v)/\partial v$，且对于 u 和 v，有

$$0\leq \frac{\partial}{\partial v}C(u,v)\leq 1 \tag{2.2.9}$$

函数 $u\mapsto \partial C(u,v)/\partial v$ 和函数 $v\mapsto \partial C(u,v)/\partial u$ 在 $\mathbf{I}$ 几乎任意点都是非减的。

Theorem 2.2.1.4. C 是 *copula*，如果 $\partial C(u,v)\partial v$ 和 $\partial^2 C(u,v)/\partial u\partial v$ 在 $\mathbf{I}^2$ 上是连续的，且当 $v=0$，对于所有 $u\in(0,1)$，都有 $\partial C(u,v)/\partial u$，那么在 $(0,1)^2$ 上存在 $\partial C(u,v)/\partial u$ 和 $\partial^2 C(u,v)/\partial v\partial u$，并且 $\partial^2 C(u,v)/\partial u\partial v = \partial^2 C(u,v)/\partial v\partial u$。

2.2.2 Sklar's Theorem

[3] Sklar 定理阐明了，Copula 函数如何描述多变量联合分布和其各自单独边缘分布之间的关系。Skalr 定理给出了可使用 copula 的单变量以及多变量分布的特征：

Definition 2.2.2.1. 域 $\overline{\mathbf{R}}$ 中的单变量分布函数 F 满足：

1. F 是非减的，
2. $F(-\infty)=0$ 且 $F(\infty)=1$。

Definition 2.2.2.2. 域 $\overline{\mathbf{R}}^2$ 中的联合变量分布 H 满足：

1. H 是 *2-increasing(section 2.1)* 的，

2. $H(x,-\infty)=H(-\infty,y)=0$, 且 $H(\infty,\infty)=1$。

H 是“触底的”, 由于 $DomH=\overline{\mathbf{R}}$, 可由联合分布 H 推出各变量各自的边缘分布$F(x)=H(x,\infty)$,$G(y)=H(\infty,y)$。

Theorem 2.2.2.1 (**Sklar's Theorem**). *Let H be a joint distribution function with margins F and G. Then there exists a copula C such that for all x,y in $\overline{\mathbf{R}}$,*

假设 H 为两个边缘分布分别为 F 和 G 的随机变量的联合分布函数, 那么存在一个 *copula* 函数 C, 使 $\overline{\mathbf{R}}$ 中的所有 x,y 满足,

$$H(x,y)=C(F(x),G(y)). \tag{2.2.10}$$

$$C(x,y)=H(F^{(-1)}(x),G^{(-1)}(y)) \tag{2.2.11}$$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on $RanF\times RanG$. Conversely, if C is a copula and F and G are distribution fuctions, then the function H defined by eq. (2.2.10) is a joint distribution function with margins F and G.

如果 F 和 G 是连续的, 则 C 是唯一的; 否则, C 则仅由 $RanF\times RanG$ 决定。反之来说, 如果 C 是一个 *copula*, F 和 G 是分布函数, 那么, 由 *eq.* (2.2.10) 定义的联合分布 H 的边缘分布为 F 和 G。

举例说明 H,F,G 以及 C 的关系 :

Example 2.2.2.1. 联合分布函数 H:

$$H(x,y)=\begin{cases}\dfrac{(x+1)(e^y-1)}{x+2e^y-1}, & (x,y)\in[-1,1]\times[0,\infty],\\ 1-e^{-y}, & (x,y)\in(1,\infty]\times[0,\infty],\\ 0, & elsewhere\end{cases}$$

其边缘分布 F 和 G 分别为:

$$F(x)=H(x,\infty)=\begin{cases}0, & x<-1,\\ (x+1)/2, & x\in[-1,1],\\ 1, & x>1\end{cases}\ and\ G(y)=H(\infty,y)=\begin{cases}0, & y<0,\\ 1-e^{-y}, & y\geq 0.\end{cases}$$

F 和 G 的准逆函数 *(quasi-inverses)* 分别为 $F^{(-1)}(u)=2u-1$, $G^{(-1)}(v)=-\ln(1-v)$, $u,v\in\mathbf{I}$。则根据 *eq.* (2.2.11):

$$C(u,v)=\frac{uv}{u+v-uv} \tag{2.2.12}$$

对于在 $\overline{\mathbf{R}}^2$ 中的联合分布 H_C, 可小结之 :

$$H_C(x,y)=\begin{cases}0, & x<0\ or\ y<0,\\ C(x,y) & (x,y)\in\mathbf{I}^2,\\ x, & y>1,x\in\mathbf{I},\\ y, & x>1,y\in\mathbf{I},\\ 1, & x>1\ and\ y>1.\end{cases} \tag{2.2.13}$$

Definition 2.2.2.3. 设 F 是一个变量的分布函数, 那么 F 的准逆/拟逆 (quasi-inverse) 函数是定义域为 $\mathbf{I}$, 满足以下情况的任意 $F^{(-1)}$,

1. 如果 t 在 $RanF$ 内, 那么 $F^{(-1)}(t)$ 是任意一个 $x\in\overline{\mathbf{R}}$ 满足 $F(x)=t$, $t\in RanF$

$$F(F^{(-1)}(t))=t;$$

2. 如果 $t \notin RanF$，那么

$$F^{(-1)}(t) = \inf x | F(x) \geq t = \sup x | F(x) \leq t.$$

如果 F 是严格递增的，那么它有且仅有一个准逆函数，即它的逆函数 F^{-1}。即当 F 严增时，F 的准逆函数即是它的逆函数。

2.2.3 Copula 的重要性质

Theorem 2.2.3.1. *Le X and Y be continous random variables. Then X and Y are independent if and only if $C_{XY} = \Pi$*

Theorem 2.2.3.2. *Let X and Y be continuous random variables with copulas C_{XY}. If α and β are strictly increasing on RanX and RanY, respectively, then $C_{\alpha(X)\beta(Y)} = C_{XY}$. Thus C_{xy} is invariant under strictly increasing tranformations of X and Y.*

以上定理阐述了，若两个连续随机变量 X,Y 的在 RanX 和 RanY 中的变换 α, β 是严格递增的，则变换前后的 copula 函数是一样的。

Theorem 2.2.3.3. *Let C be a copula. For any v in $\mathbf{I}$, the partial derivative $\partial C(u,v)/\partial u$ exists for almost all u, and for such va and u,*

$$0 \leq \frac{\partial}{\partial u} C(u,v) \leq 1 \tag{2.2.14}$$

Similarly, for any u in $\mathbf{I}$, the partial derivative $\partial C(u,v)/\partial v$ exists for almost all v, and for such u and v,

$$0 \leq \frac{\partial}{\partial v} C(u,v) \leq 1 \tag{2.2.15}$$

Furthermore, the function $u \mapsto \partial C(u,v)/\partial v$ and $v \mapsto \partial C(u,v)/\partial u$ are defined and nondecreasing almost everywhere on $\mathbf{I}$.

Theorem 2.2.3.4. *Let C be a copula. If $\partial C(u,v)/\partial v$ and $\partial^2 C(u,v)/\partial u \partial v$ are continous on $\mathbf{I}^2$ and $\partial C(u,v)/\partial u$ exists for all $u \in (0,1)$ when $v = 0$, then $\partial C(u,v)/\partial u$ and $\partial^2 C(u,v) \partial v \partial u$ exist in $(0,1)^2$ and $\partial^2 C(u,v)/\partial u \partial v = \partial^2 C(u,v)/\partial v \partial u$.*

Theorem 2.2.3.5. *X 和 Y 是两个连续随机变量，copula 函数为 C_{XY}。α, β 分别在 RanX 和 RanY 严格单调:*

1. 如果 α 严格递增，β 严格递减，则

$$C_{\alpha(X)\beta(Y)}(u,v) = u - C_{XY}(u, 1-v)$$

2. 如果 α 严格递减，β 严格递增，则

$$C_{\alpha(X)\beta(Y)}(u,v) = v - C_{XY}(1-u, v)$$

3. 如果 α, β 都是严格递减的，则

$$C_{\alpha(X)\beta(Y)}(u,v) = u + v - 1 + C_{XY}(1-u, 1-v)$$

Example 2.2.3.1. *The support of the Fréchet-Hoeffding upper bound M is the main diagonal of $\mathbf{I}^2$, i.e., the graph of $v = u$ for u in $\mathbf{I}$, so that M is singular. This follows from the fact that the M-measure of any open rectangle that lies entirely above or below the main diagonal is zero. Also note that $\partial^2 M; /\partial u \partial v = 0$ everywhere in $\mathbf{I}^2$ except on the main diagonal. Similarly, the support of the Fréchet-Hoeffding lower bound W is the secondary diagonal of $\mathbf{I}^2$, i.e., the graph of $v = 1 - u$ for u in $\mathbf{I}$, and thus W is singular as well.*

Fréchet-Hoeffding 上界 M 的支集[2]是 $\mathbf{I}^2$ 主对角线。以上是根据：任意整体位于注对角线上方或下方的开方矩形，其 M-测量值为 0。同样需要注意的是，除了在主对角线上，$\mathbf{I}^2$ 中任一点都有 $\partial^2 M/\partial u\partial v = 0$。相似的，*Fréchet-Hoeffding* 下界 W 的支集是 $\mathbf{I}^2$ 的次对角线，即图形 $v = 1-u, u \in \mathbf{I}$，$W$ 也是非奇异的。

“Copula Models” 的优点有 [4]：

1. 它可以将任意的非同一分布族的单变量边缘分布进行组合；
2. Copula 中特殊的一类模型，称为 “elliptical copula(EC)”，其性质是：相比其他一些多变量概率模型，随着维数的增加 (这里应该理解为变量数)，EC 模型复杂度增加幅度要远小于其他模型。
3. 它的通用性良好，可以应用包含多种多变量模型，并且提供了一个可以生成更多多变量模型的框架。

Cupols 目的在于将多变量的分布函数累积耦合出它们的边缘分布。对于传统多变量分布，一旦参数化分布函数选定，可以通过积分来获得其边缘分布。这样，由原来多变量自身分布可以决定其边缘分布特征。这种方法限制在于这些多变量要属于同一个分布族。而 Copula 建模则可以利用 copula function 将不属于同一分布族的变量 “粘贴 (glued together)” 起来确定其边缘分布。(2.1 of [4])

CF 仅仅确定随机变量之间的相关性，并不影响变量本身的分布特。CF 的一个思想是将原始随机变量 X_j，通过变换的方法得到均匀分布的随机变量 $U_j = F_j(X_j)$(在一个向量空间中分布不均匀的变量映射为另一个向量空间中均匀分布的变量)。前提是，变换后变量间相关性分布于原始变量保持一致。该方法优势在于，变换之后，变量间相关性分布更容易被获取。

2.2.4 Fréchet-Hoeffding Bounds for Joint Distribution Functions

从式 $W(u,v) = max(u+v-1, 0) \le C(u,v) \le min(u,v) = M(u,v)$ 可得，设随机变量 X 和 Y，其联合分布函数为 H，H 的边缘分布分别为 F 和 G，对于所有 $x, y \in \overline{\mathbf{R}}$，有

$$max(F(x)+G(y)-1, 0) \le H(x,y) \le min(F(x), G(y)) \tag{2.2.16}$$

对于 H 的两个边界可称为 Freéchet-Hoeffding 边界。

Definition 2.2.4.1. 一个 $\overline{\mathbf{R}}^2$ 的子集 S，如果任意两对观测值 $(x,y),(u,v) \in S$，若 $x < u$ 则 $y \le v$，那么 S 是非减的；若 $x < u$ 则 $y \ge v$，则 S 是非增的。

Lemma 2.2.4.1. $\overline{\mathbf{R}}^2$ 的一个子集 S，S 是非减的条件是，当且仅当每一对 $(x,y) \in \overline{\mathbf{R}}^2$，都有

1. 对于所有的 $(u,v) \in S$，$u \le x$ 则表示 $v \le y$; 或者
2. 对于所有的 $(u,v) \in S$，$v \le y$ 则表示 $u \le x$。

Lemma 2.2.4.2. 随机变量 X,Y 的联合分布函数 H。H 等于 *Fréchet-Hoeffding* 上界的条件是，当且仅当对于每一个 $(x,y) \in \overline{\mathbf{R}}$，都有 $P[X > x, Y \le y] = 0$ 或者 $P[X \le x, Y > y] = 0$。

Theorem 2.2.4.1. *Let X and Y be random variables with joint distribution fuction H. Then H is identically equal to its Fréchet-Hoeffding upper bound if and only if the support of H is a nondecreasing subset of $\overline{\mathbf{R}}^2$.*

随机变量 X,Y 的联合分布函数 H。当且仅当 H 的定义域为一个非减的 $\overline{\mathbf{R}}^2$ 子集时，H 恒等于 *Fréchet-Hoeffding* 上界，也就是图形 (直线)$v = u, u \in \mathbf{I}$，所以 M 是非奇异的。

[2] 支集就是带入该函数中满足此条件的元素

2.2.5 Survival Copulas

$\overline{F}(x) = P[X > x] = 1 - F(x)$, $\overline{H}(x,y) = P[X > x, Y > y]$, $\overline{H}$ 的边缘分布分别为

$$\bar{F} = \bar{H}(x, -\infty), \tag{2.2.17}$$
$$\bar{G} = \bar{H}(-\infty, y) \tag{2.2.18}$$
$$\tag{2.2.19}$$

且

$$\bar{H}(-\infty, -\infty) = 1 \tag{2.2.20}$$

，根据 theorem 2.2.2.1，可得

$$\begin{aligned}\overline{H}(x,y) &= 1 - F(x) - G(y) + H(x,y) \\ &= \overline{F}(x) + \overline{G}(y) - 1 + C(F(x), G(y)) \\ &= \overline{F}(x) + \overline{G}(y) - 1 + C(1 - \overline{F}(x), 1 - \overline{G}(y)),\end{aligned}$$

根据 theorem 2.2.3.5，下面定义了一个 $\hat{C}$ 并得出 C 和 $\hat{C}$ 的关系：

$$\overline{H}(x,y) = \hat{C}(\overline{F}(x), \overline{G}(y)). \tag{2.2.21}$$

$$\hat{C}(u,v) = u + v - 1 + C(1-u, 1-v), \tag{2.2.22}$$

需要注意的还有，要区分 $\overline{C}$ 和 $\hat{C}$，从 eq. (2.2.22) 推得

$$\begin{aligned}\overline{C}(u,v) &= P[U > u, V > v] \\ &= 1 - u - v + C(u,v) \\ &= \hat{C}(1-u, 1-v)\end{aligned}$$

而从 eq. (2.2.21) 推得

$$\hat{C}(u,v) = \overline{H}(\bar{F}^{-1}(u), \bar{G}^{-1}(v)) \tag{2.2.23}$$

通过一个例子，来充分说明 $\overline{H}, \overline{F}, \overline{G}, \hat{C}$ 之间的关系

Example 2.2.5.1. *A bivariate Pareto distribution. Let X and Y be random variables whose joint survival function is given by*

$$\overline{H}_\theta(x,y) = \begin{cases} (1+x+y)^{-\theta}, & x \geq 0, y \geq 0, \\ (1+x)^{-\theta}, & x \geq 0, y < 0, \\ (1+y)^{-\theta}, & x < 0, y \geq 0. \\ 1, & x < 0, y < 0; \end{cases}$$

where $\theta > 0$.

可以算出其边缘分布 $\overline{F}$ 和 $\overline{G}$ 为

$$\overline{F}(x) = \begin{cases} (1+x)^{-\theta}, & x \geq 0, \\ 1, & x < 0, \end{cases} \quad \text{and} \quad \overline{G}(y) = \begin{cases} (1+y)^{-\theta}, & y \geq 0, \\ 1, & y < 0, \end{cases}$$

根据 eq. (2.2.21)，算出 $\overline{F}^{-1}(u) = u^{-1/\theta} - 1$ 和 $\overline{G}^{-1} = v^{-1/\theta} - 1$，根据 eq. (2.2.23)

$$\begin{aligned}\hat{C}(u,v) &= \overline{H}(\overline{F}^{-1}(u), \overline{G}^{-1}(v)) \\ &= (u^{-1/\theta} + v^{-1/\theta} - 1)^{-\theta}\end{aligned}$$

其中 $u \geq 0, v \geq 0$。进一步可验证 section 2.2.5和 eq. (2.2.20)：

$$\begin{aligned} \overline{H}(u, -\infty) &= \overline{H}(u, v < 0) = (1 + u^{-1/\theta} - 1) = u \\ \overline{H}(-\infty, v) &= \overline{H}(u < 0, v) = v \\ \overline{H}(-\infty, -\infty) &= \overline{H}(u < 0, v < 0) = 1 \end{aligned}$$

The dual of a copula C is the function $\tilde{C}$ defined by $\tilde{C}(u,v) = u + v - C(u,v)$; the co-copula is the function C^* defined by $C^*(u,v) = 1 - C(1-u, 1-v)$。$\tilde{C}$ 和 C^* 都不是一个 copupla，但当 C 是一对随机变量 X, Y 的 copula 时，

$$P[X \leq x, Y \leq y] = C(F(x), G(y)) \text{ and } P[X > x, Y > y] = \hat{C}(\bar{F}(x), \bar{G}(y)),$$

还有

$$P[X \leq x \text{ or } Y \leq y] = \tilde{C}(F(x), G(y)), \tag{2.2.24}$$

和

$$P[X > x \text{ or } Y > y] = C^*(\bar{F}(x), \bar{G}(y)) \tag{2.2.25}$$

2.2.6 Symmetric and Ordering

一个随机变量 X 关于 a 对称，则 $P[X - a] \leq x = P[a - X] \leq x$，可得

$$F(a + x) = \bar{F}(a - x) \tag{2.2.26}$$

Definition 2.2.6.1. *X 和 Y 是两个随机变量，(a,b) 是 $\mathbf{R}^2$ 中的一个点，*

1. *若 X 和 Y 分别关于 a 和 b 是对称的，则 (X,Y) 是关于 (a,b) 边缘对称 (marginally symmetric) 的。*
2. *若 $X - a$ 和 $Y - b$ 的联合分布函数与 $a - X$ 和 $b - Y$ 联合分布函数一样，则 (X,Y) 是关于 (a,b) 径向对称的 (radially symmetric)。*
3. *若四对随机变量:$(X - a, Y - b), (X - a, b - Y), (a - X, Y - b), (a - X, b - Y)$ 共用一个联合分布，则称 (X,Y) 关于 (a,b) 是联合对称的 (jointly symmetric)。*

Theorem 2.2.6.1. *X 和 Y 是连续随机变量，其联合分布函数为 H，H 的边缘分布分别为 F 和 G。(a,b) 是 $\mathbf{R}^2$ 上一点。当且仅当*

$$H(a + x, b + y) = \overline{H}(a - x, b - y) \text{ for all } (x, y) \text{ in } \mathbf{R}^2 \tag{2.2.27}$$

时，(X,Y) 是关于 (a,b) 径向对称的。

2.2.7 随机变量的生成

为了获取分布函数为 F 的随机变量 X 的一个观测值 x，

1. 生成一个变量 u，其为 (0,1) 上的均匀分布；
2. 令 $x = F^{(-1)}(u)$，其中 $F^{(-1)}$ 是任意的 F 的准逆/拟逆 (quasi-inverse) 函数

关于准逆函数，definition 2.2.2.3给出了定义。

得益于 skalr 定理，仅需要在 (0,1) 均匀分布的随机变量 (U,V) 生成一对观测值 (u,v)，(U,V) 的联合分布函数为关于原始随机变量 X,Y 的 copula C。一种生成这样的 (u,v) 的方法，叫做条件分布法 (conditional distribution method)。对于该方法，需要得出在 $U = u$ 条件下的 V，记为 $c_u(v)$，

$$c_u(v) = P[V \leq v | U = u] = \lim_{\triangle u \to 0} \frac{C(u + \triangle u, v) - C(u, v)}{\triangle u} = \frac{\partial C(u,v)}{\partial u} \tag{2.2.28}$$

[从 theorem 2.2.1.3知道 $v \longmapsto \partial C(u,v)/\partial u$，$v$ 存在且 $v \in \mathbf{I}$ 是非减的，这里记为 $c_u(v)$。]

方法流程为：

1. 生成两个 $(0,1)$ 上均匀分布的变量 u 和 t;
2. 令 $v = c_u^{(-1)}(t)$，其中 $c_u^{(-1)}$ 为 c_u 的准逆函数；
3. (u,v) 即是所求。

Example 2.2.7.1. 这里拿 *example 2.2.2.1*继续举例说明，根据 *eq.* (2.2.12) 得

$$C(u,v) = \frac{uv}{u+v-uv},$$

然后计算 c_u 和 $c_u^{(-1)}$:

$$c_u(v) = \frac{\partial}{\partial u}C(u,v) = \left(\frac{v}{u+v-uv}\right)^2 \quad and \quad c_u^{-1}(t) = \frac{u\sqrt{t}}{1-(1-u)\sqrt{t}}$$

接下来的算法用于生成随机变量 (x,y)：

1. 生成两个独立的 (0,1) 均匀分布变量 u 和 t；
2. 令 $v = \frac{u\sqrt{t}}{1-(1-u)\sqrt{t}}$,
3. 令 $x = 2u-1$ 以及 $y = -\ln(1-v)$。(从 example 2.2.2.1中计算的 $F^{(-1)}(u) = 2u-1$ 和 $G^{(-1)}(v) = -\ln(1-v)$)
4. (x,y) 即是所求。

Survival copulas 也可以用在条件分布法中，从一个给定 survival 函数的分布中来生成随机变量。从 theorem 2.2.3.5和 eq. (2.2.22) 中，知随机变量 U,V 的 $C(u,v)$ 相应的 survival copula$\hat{C}(u,v) = u+v-1+C(1-u,1-v)$，是 $(1-U,1-V)$ 的分布函数。如果 U 是 $(0,1)$ 均匀分布，则 $1-U$ 也是。下面的算法用来生成 (U,V),

1. 生成两个独立的 (0,1) 均匀分布变量 u 和 t；
2. 令 $v = \hat{c}_u^{(-1)}(t)$，其中 $\hat{c}_u^{(-1)}$ 表示 $\hat{c}_u(v) = \partial\hat{C}(u,v)/\partial u$ 的准逆函数；
3. (u,v) 即是所求。

2.2.8 多变量 copula

<+To be continued+>

2.2.9 边缘统计 (Marginal Statistics)

对于联合统计 x,y，$F_x(x)$ 为边缘分布，$f_x(x)$ 为 x 的边缘概率密度函数。

$$F_x(x) = F(x,\infty) \qquad F_y(y) = F(\infty,y) \tag{2.2.29}$$

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy \qquad f_y(y) = \int_{-\infty}^{\infty} f(x,y)dx \tag{2.2.30}$$

2.2.10 Copula 模型估计

建立一个 Copula 模型，有两组参数需要估算 [4]:

1. 第一组是每个选择的边缘分布 $\Theta = \{\theta_1,\ldots,\theta_m\}$ 的参数；
2. 第二组是所选 CF 的相关性参数。

常用方法使用了两步法对这些参数进行估算：先分别估算边缘分布的参数；然后基于以上的估算对相关矩阵 Γ 进行估算。前一步骤可以通过多种方式实现，如最大似然 (maximum likelihood(ML))、贝叶斯 (Bayesian)、或者基于时刻的方法 (a method of moments based approach)。但是 Γ 的估算就要复杂的多了，并且取决于随机变量是连续或离散的。

2.3 Dependence(相关性)

2.3.1 基本概念

随机变量 x，方差：

$$\mathrm{var}(x) = \mathbf{D}x = \mathbf{E}(x - \mathbf{E}x)^2 = \mathbf{E}x^2 - (\mathbf{E}x)^2 \tag{2.3.1}$$

两个随机变量 x, y 的方差：

$$\begin{aligned} \mathbf{D}(x+y) &= \mathbf{E}[(x - \mathbf{E}x) + (y - \mathbf{E}y)]^2 \\ &= \mathbf{D}x + \mathbf{D}y + 2\mathbf{E}(x - \mathbf{E}x)(y - \mathbf{E}y) \\ &= \mathbf{D}x + \mathbf{D}y + 2\mathrm{cov}(x, y) \end{aligned} \tag{2.3.2}$$

其中

$$\mathrm{cov}(x, y) = \mathbf{E}(x - \mathbf{E}x)(y - \mathbf{E}y) \tag{2.3.3}$$

是 x, y 的协方差，相关系数：

$$\rho(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathbf{D}x \times \mathbf{D}y}} \tag{2.3.4}$$

若 x, y 独立，则 $\mathrm{cov}(x, y) = 0$。若 $\rho(x, y) = \pm 1$，则 x, y 线性相关：

$$y = ax + b$$

2.3.2 Concordance(调和性)

假设两个随机变量，(x_i, y_i) 和 (x_j, y_j) 表示两个向量 (X, Y) 的两对观测值，如果 $x_i < x_j, y_i < y_j$ 或者 $x_i > x_j, y_i > y_j$，则称 $(x_i, y_i), (x_j, y_j)$ 是“调和”(concordant) 的。若 $x_i < x_j, y_i > y_j$ 或者 $x_i > x_j, y_i < y_j$，则称 $(x_i, y_i), (x_j, y_j)$ 是“不调和”(discordant) 的。以上可总结为: (x_i, y_i) and (x_j, y_j) are concordant if $(x_i - x_j)(y_i - y_j) > 0$ and disconcordant if $(x_i - x_j)(y_i - y_j) < 0$。

Definition 2.3.2.1. *A numeric measure κ of association between two continuous random variables X and Y whose copula is C is a measure of concordance if it satisfies the following properties (again we write $\kappa_{X,Y}$ or κ_C when convenient):*

两个 *copula* 为 C 的随机变量 X 和 Y，一个对它们的相关性的数值测量 κ 是一致性测量的条件是其满足以下性质:

1. *κ is defined for every pair X, Y of continuous random variables;*
2. *$-1 \le \kappa_{X,Y} \le 1$, $\kappa_{X,X} = 1$, and $\kappa_{X,-X} = -1$;*
3. *$\kappa_{X,Y} = \kappa_{Y,X}$;*
4. *if X and Y are independent, then $\kappa_{X,Y} = \kappa_\Pi = 0$;*
5. *$\kappa_{-X,Y} = \kappa_{X,-Y} = -\kappa_{X,Y}$;*
6. *if C_1 and C_2 are copulas such that $C_1 \prec C_2$, then $\kappa_{C_1} \le \kappa_{C_2}$;*
7. *if $\{(X_n, Y_n)\}$ is a sequence of continuous random variables with copulas C_n, and if $\{C_n\}$ converges pointwise to C, then $\lim_{n\to\infty} \kappa_{C_n} = \kappa_C$.*

Theorem 2.3.2.1. *Let κ be a measure of concordance for continuous random variables X and Y:*

1. *if Y is almost surely an increasing function of X, then $\kappa_{X,Y} = \kappa_M = 1$;*

2. *if Y is almost surely a decreasing function of X, then $\kappa_{X,Y} = \kappa_W = -1$;*

3. *if α and β are almost surely strictly monotone functions on RanX and RanY, respectively, then $\kappa_{\alpha(X),\beta(Y)} = \kappa_{X,Y}$.*

根据以上的定义，从下面的定理可以看到 tau 和 rho 都是一致性的测量工具：

Theorem 2.3.2.2. *If X and Y are continuous random variables whose copula is C, then the population versions of Kendall's tau eq. (2.3.7) and Spearman's rho theorem 2.3.2.6 satisfy the properties in definition 2.3.2.1 and theorem 2.3.2.1 for a measure of concordance.*

2.3.2.1 Kendall's (τ)

两个随机变量 X, Y。

Kendall's τ可以对变量之间这种联合的一致性进行采样量化。它描述了两个随机独立均匀分布向量的一致性与非一致性概率的差值 [5]:

$$\tau = \tau_{X,Y} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0] \tag{2.3.5}$$

引入基本概念定理：

Theorem 2.3.2.3. *[3] Let (X_1, Y_1) and (X_2, Y_2) be independent vectors of continuous random variables with joint distribution functions H_1 and H_2, respectively, with common margins F (of X_1 and X_2) and G (of Y_1 and Y_2). Let C_1 and C_2 denote the copulas of (X_1, Y_1) and (X_2, Y_2), respectively, so that $H_1(x, y) = C_1(F(x), G(y))$ and $H_2(x, y) = C_2(F(x), G(y))$. Let $\mathcal{Q}$ denote the difference between the probabilities of concordance and discordance of (X_1, Y_1) and (X_2, Y_2), i.e., let*

(X_1, Y_1) 和 (X_2, Y_2) 分别是两个随机变量的两对向量，其联合分布函数分别为 H_1 和 H_2。C_1 和 C_2 分别表示 (X_1, Y_1) 和 (X_2, Y_2) 的 *copula* 函数，则 $H_1(x, y) = C_1(F(x), G(y))$，$H_2(x, y) = C_2(F(x), G(y))$。用 $\mathcal{Q}$ 表示两对取值调和概率与不调和概率的差值:

$$\mathcal{Q} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0] \tag{2.3.6}$$

Then

$$\mathcal{Q} = \mathcal{Q}(C_1, C_2) = 4 \iint_{\mathbf{I}^2} C_2(u, v) dC_1(u, v) - 1 \tag{2.3.7}$$

where u, v are the transformations of $F(x), G(y)$, $u = F(x)$ and $v = G(y)$.

Corollary 2.3.2.1. *Let C_1, C_2 and $\mathcal{Q}$ be as given in theorem 2.3.2.3. Then*

1. *$\mathcal{Q}$ is symmetric in tis arguments: $\mathcal{Q}(C_1, C_2) = \mathcal{Q}(C_2, C_1)$;*

2. *$\mathcal{Q}$ is nondecreasing in each argument: if $C_1 \prec C_1'$ and $C_2 \prec C_2'$ for all (u, v) in $\mathbf{I}^2$, then $\mathcal{Q}(C_1, C_2) \leq \mathcal{Q}(C_1', C_2')$.*

3. *Copulas can be replaced by survival copulas in $\mathcal{Q}$, i.e., $\mathcal{Q}(C_1, C_2) = \mathcal{Q}(\hat{C}_1, \hat{C}_2)$.*

Theorem 2.3.2.4. *Let X and Y be continuous random variables whose copula is C. Then the population version of Kendall's tau for X and Y (which we will denote either $\tau_{X,Y}$ or τ_C) is given by*

$$\tau_{X,Y} = \tau_C = \mathcal{Q}(C, C) = 4 \iint_{\mathbf{I}^2} C(u, v) dC(u, v) - 1 \tag{2.3.8}$$

Corollary 2.3.2.2. *Let X and Y be random variables with an Archimedean copula C generated by φ in Ω. The population version τ_C of Kendall's tau for X and Y is given by*

$$\tau_C = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt \tag{2.3.9}$$

Theorem 2.3.2.5. *Let C_1 and C_2 be copulas. Then*

$$\iint_{\mathbf{I}^2} C_1(u,v)dC_2(u,v) = \frac{1}{2} - \iint_{\mathbf{I}^2} \frac{\partial}{\partial u}C_1(u,v)\frac{\partial}{\partial v}C_2(u,v)dudv \tag{2.3.10}$$

2.3.2.2 Pearson's ϱ

Person's ϱ:

$$\varrho^p_{X,Y} = \mathrm{cov}(X,Y)/\sqrt{\mathrm{var}(X)\mathrm{var}(Y)} \tag{2.3.11}$$

$\varrho^p_{j_1 j_2}$ 就，应该看做是在变量服从正态分布情况下，自然产生的总体相关系数。需要注意的是，当变量不服从正态分布的情况下，用它来作为变量的相关性系数则不适用 [4]。公式2.3.11可变为：

$$\rho^s_{X,Y} = \mathrm{corr}(F_X(X), F_Y(Y)) \tag{2.3.12}$$

为了对相关性进行建模，该测量方法与 copula 方法联系紧密。因为它简单描述了通过变换后的随机变量 $U_j = F_j(j)$ 的相关性，这些变量本身就可以使用 copula 函数 C 描述他们的分布关系。由于 U_j 服从 $[0,1]$ 均匀分布，其方差为 $1/12$，期望为 $1/2$。因此 eq. (2.3.12) 可简写为

$$\begin{aligned}\rho^s_{X,Y} &= \mathrm{cov}(U_X,U_Y)/(\mathrm{var}(U_X)\mathrm{var}(U_Y))^{1/2} \\ &= 12E\left[\left(U_X - \frac{1}{2}\right)\left(U_Y - \frac{1}{2}\right)\right] \\ &= 12E[U_X U_Y] - 3\end{aligned} \tag{2.3.13}$$

$\mathrm{cov}(U_X,U_Y)$ 不再有原始变量边缘分布来决定。因此相比 Spearman's ϱ，Pearson's ρ 对原始变量边缘分布不敏感，就不再受到原始变量必须服从正态分布的限制。

2.3.2.3 Spearman's rho (ρ)

Spearman's ρ是变换后变量 $F(X)$ 和 $G(Y)$ 的 Pearson 系数 ϱ [5]，它不是根据原始变量来计算，而是通过变量的秩次来计算 ϱ，即得到 ρ。F 和 G 分别是变量 X 和 Y 的分布函数:

$$\rho = \varrho(F(X), G(Y)) \tag{2.3.14}$$

通过关联性测量，当 X,Y 是正（负）关联时，τ,ρ 也是正（负）关联的。当 X,Y 相互独立时，它们为 0。当 X,Y 通过非线性变换后是严格递增的，它们的值保持不变。

三对独立随机向量 $(X_1,Y_1),(X_2,Y_2),(X_3,Y_3)$，它们共用的联合分布函数为 H(边缘分布为 F 和 G) 以及 copula 函数 C。Spearman's rho 的总体版本 $\rho_{X,Y}$ 被定义为与两个向量 $(X_1,Y_1),(X_2,Y_3)$ 一致的概率减去不一致的概率值成比例。

$$\rho_{X,Y} = 3(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]) \tag{2.3.15}$$

这两对向量特点是：两对边缘分布相同的变量，但其中一对的两个变量有联合分布函数 H，但另一对的两个变量是相互独立的，即 (X_1,Y_1) 的联合分布为 $H(x,y)$，但 (X_2,Y_3) 的联合分布为 $F(x)G(y)$(X_2,Y_3 是相互独立的)。因此 X_2,Y_3 的 copula 是 Π，根据 theorem 2.3.2.3和 corollary 2.3.2.1的第一部分，引出定理：

Theorem 2.3.2.6. *Let X and Y be continous random variables whose copula is C. Then the population version of Sperman's rho for X and Y (which we will denote by either $\rho_{X,Y}$ or ρ_C) is given by*

$$\rho_{X,Y} = \rho_C = 3\mathcal{Q}(C,\Pi), \tag{2.3.16}$$

$$= 12\iint_{\mathbf{I}^2} uv\ dC(u,v) - 3, \tag{2.3.17}$$

$$= 12\iint_{\mathbf{I}^2} C(u,v)\ dudv - 3 \tag{2.3.18}$$

τ 和 ρ 是基于秩次的关联性测量值。从文章 [6] 关于相关性结构的统计推断应该总是基于秩次的，因为它们在 X,Y 的单增变换条件下是不变统计量。

在文献 [7] 中，作者表明了相关矩阵和相关系数的关系：$\rho_{lj}^{(p)} = \dfrac{\mathrm{Cov}(X_l, X_j)}{\sqrt{(\mathrm{Var}(X_l)\mathrm{Var}(X_j))}}$，其中 $\mathrm{Var}(X_l)$ 和 $\mathrm{Var}(X_j)$ 分别表示随机变量 X_l 和 X_j 的方差，$l, j \in \{1, 2, \ldots, N\}$；$\mathrm{Cov}(X_l, X_j)$ 是相关矩阵 Σ 的第 (l, j) 个元素。

在文章 [6]，作者给出了 Spearman's ρ 优于 pearson's ϱ 的几个方面：

1. 当且仅当随机变量 X，Y 是函数相关时，$E(\rho) = \pm 1$;
2. 当且仅当 X，Y 是线性相关时，才满足 $E(\varrho) = \pm 1$，相比 ρ 更受限制，且
3. ρ 较为通用，对于不同分布都可以估算出一个意义明确的总体参数，然而当 ϱ 应用在 "重尾分布" 的情形时 (heavy-tailed distributions)，例如柯西分布，是无法得到一个理论上的相关参数的。

2.3.2.4 τ 和 ρ 的关系

Theorem 2.3.2.7. *X 和 Y 是连续随机变量，则 τ 和 ρ:*

$$-2 \le 3\tau - 2\rho \le 1 \tag{2.3.19}$$

Theorem 2.3.2.8. *theorem 2.3.2.2中定义的 X, Y, τ, ρ，那么*

$$\frac{1+\rho}{2} \ge \left(\frac{1+\tau}{2}\right)^2 \tag{2.3.20}$$

and

$$\frac{1-\rho}{2} \ge \left(\frac{1-\tau}{2}\right)^2 \tag{2.3.21}$$

同样是 X, Y, τ, ρ

Corollary 2.3.2.3.

$$\frac{3\tau - 1}{2} \le \rho \le \frac{1 + 2\tau - \tau^2}{2},\ \tau \ge 0, \tag{2.3.22}$$

and

$$\frac{\tau^2 + 2\tau + 1}{2} \le \rho \le \frac{1 + 3\tau}{2},\ \tau \le 0. \tag{2.3.23}$$

2.3.2.5 其他的一致性测量方法

Theorem 2.3.2.9. *Let X and Y be continuous random variables whose copula is C. Then the population version of Gini's measure of association for X and Y (which we will denote by either $\gamma_{X,Y}$ or γ_C) is given by*

$$\gamma_{X,Y} = \gamma_C = \mathcal{Q}(C, M) + \mathcal{Q}(C, W). \tag{2.3.24}$$

关于 ρ 和 γ，可以这样理解：$\rho = 3\mathcal{Q}(C, \Pi)$ 测量的一种一致性关系是；使用 copula C 来表达的 X 和 Y 的分布，与使用 copula Π 来表达 X 和 Y 不相关之间“距离”。而 $\gamma = \mathcal{Q}(C, M) + \mathcal{Q}(C, W)$ 测量的一致性关系是：C 和单调相关 M 和 W 之间的“距离”。

Corollary 2.3.2.4. *Under the hypotheses of theorem 2.3.2.9,*

$$\gamma_C = 4\left[\int_0^1 C(u, 1-u)du - \int_0^1 [u - C(u, u)]du\right]. \tag{2.3.25}$$

还有一种测量参数为 β <+ 未完待续 +>

2.3.3 相关性的性质

之前章节已经涉及了很多关于相关性的性质，如：当 X 和 Y 独立时，根据 theorem 2.2.3.1，它们的 copula 函数为 $C_{XY} = \Pi = uv$；当这两个变量的联合分布函数是 Fréchet-Hoeffding 的一个边界 M 获 W 时，几乎可以确定它们是呈线性相关的。因此一对随机变量的“相关性性质”可以认为是它们所有的联合分布函数组成的集中的一个子集。很多相关性质，根据变量的分布函数，可被直接描述为 copulas，或者 copulas 的简单性质。

但多数情况下变量的分布并不是相互独立的，也很少是呈线性相关的，这才是我们要研究的重点。

为了引出以下内容，先定义两个基本概念，“正相关”与“负相关”。正相关与之前描述的一致性概念相对应，而负相关则与不一致概念相对应。

2.3.3.1 象限相关 (Quadrant Dependence)

Definition 2.3.3.1. *Let X and Y be random variables. X and Y are positively quadrant dependent (PQD) if for all (x, y) in $\mathbf{R}^2$,*

$$P[X \le x, Y \le y] \ge P[X \le x]P[Y \le Y]. \tag{2.3.26}$$

or equivalently

$$P[X > x, Y > y] \ge P[X > x]P[Y > y]. \tag{2.3.27}$$

如果 X,Y 正象限相关，则记为 $PQD(X, Y)$。负象限相关 (negative quadrant dependent (NQD)) 的概念则是与 PQD 相反，记为 $NQD(X, Y)$。

Theorem 2.3.3.1. *连续随机变量 X,Y，联合分布为边缘分布分别为 F,G 的 H，copula C，如果 X 和 Y 是满足 PQD，则*

$$3\tau_{X,Y} \ge \rho_{X,Y} \ge 0, \text{ and } \beta_{X,Y} \ge 0$$

2.3.3.2 (尾单调)Tail Monotonicity

式 eq. (2.3.26) 可以写作：

$$P[Y \leq y|X \leq x] \geq P[Y \leq y]$$

或者

$$P[Y \leq y??|X \leq x] \geq P[Y \leq y|X \leq \infty] = P[y|\infty]$$

Definition 2.3.3.2. *X 和 Y 为随机变量*

1. *Y 在 X 上是左拖尾递减的 (记为 LTD(Y/X))，如果*

$$P[Y \leq y|X \leq x] \text{对于所有 } y \text{ 是 } x \text{ 递减的。} \tag{2.3.28}$$

2. *X 在 Y 中是左拖尾递减的 (LTD(X/Y))，如果*

$$P[X \leq x|Y \leq y] \text{对于所有 } x \text{ 是 } y \text{ 递减的。} \tag{2.3.29}$$

3.

<++>

<++>

2.4 关于变换

Rosenblatt Transformation

Rosenblatt 变换 (RT) 是将环境变量从物理空间映射到一个变量是独立标准正太 (standard normal variables) 分布的变换空间中。然后通过已定义的，在变换空间中给定超概率分布的一个独立标准正太变量集，映射回到物理空间中。 [5]

该变换通过映射:$\mathcal{R} : \mathbb{R}^d \to (0,1)^d$，将物理空间中的一个随机向量 $\mathbf{X} = (x_1, \ldots, x_d)$，映射到随机变量独立均匀分布的空间上。映射之后:$\mathbf{E} = (E_1, \ldots, E_d)$ [5],

$$\begin{aligned} E_1 &= F_1(X_1) \\ E_2 &= F_{2|1}(X_2|X_1) \\ &\vdots \\ E_d &= F_{d|1,2,\ldots,d-1}(X_d|X_1, \ldots, X_{d-1}) \end{aligned} \tag{2.4.1}$$

其中，$F_i(x_i)$ 是 X_i 的边缘累积分布函数 (marginal cumulative distribution function)。$F_{i|1,\ldots,i-1}(x_i|x_1, \ldots, x_{i-1})$ 是给定 $X_1 = x_1, \ldots, X_{i-1} = x_{i-1}$ 条件下 X_i 的条件分布。然后，利用 $\mathbf{E}$ 来得到所谓约化空间 (reduced space)中的一个随机变量集 $\mathbf{Z} = (Z_1, \ldots, Z_d)$：

$$\Phi(Z_j) = E_j, \qquad j = 1, \ldots, d \tag{2.4.2}$$

其中，Φ 是标准正太变量的累积分布函数。

根据 [5] 中使用 copulas 获取环境变量变化轮廓 (environmental contous)，使用Inverting Rosenblatt transformation，通过将向量 $\mathbf{V}$ 映射到物理空间中环境变量 $\mathbf{X}$ 来实现：

$$\begin{aligned} x_1 &= F_1^{-1}(e_1) \\ x_2 &= F_{2|1}^{-1}(e_2|x_1) \\ &\vdots \\ x_d &= F_{d|1,2,\ldots,d-1}^{-1}(e_d|x_1, x_2, \ldots, x_{d-1}) \end{aligned} \tag{2.4.3}$$

Copula $C(\mathbf{u})$ 定义：

$$C(\mathbf{u}) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)) \tag{2.4.4}$$

相应的 copula 密度为：

$$c(\mathbf{u}) = \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d} \tag{2.4.5}$$

通常，$\mathbf{X}$ 的边缘分布可以通过计算：

$$F_{i|1,\ldots,i-1}(x_i|x_1, \ldots, x_{i-1}) = C_{i|1,\ldots,i-1}(u_i|u_1, \ldots, u_{i-1}) \tag{2.4.6}$$

其中,

$$C_{i|1,\ldots,i-1}(u_i|u_1, \ldots, u_{i-1}) = \frac{\partial^{i-1} C(u_1, \ldots, u_i, 1, \ldots, 1)/\partial u_1 \ldots \partial u_{i-1}}{\partial^{i-1} C(u_1, \ldots, u_{i-1}, 1, \ldots, 1)/\partial u_1 \ldots \partial u_{i-1}} \tag{2.4.7}$$

对于二维变量来说, 设 $u_1 = u,\ u_2 = v$

$$F_{x_2|x_1}(x_2|x_1) = C(v|u) = \frac{\partial C(u, u)}{\partial u} \tag{2.4.8}$$

公式2.4.3可以写为

$$\begin{aligned} u_1 &= e_1 \\ u_2 &= C_{2|1}^{-1}(e_2|u_1) \\ &\vdots \\ u_d &= C_{d|1,2,\ldots,d-1}^{-1}(e_d|u_1, u_2, \ldots u_{d-1}) \end{aligned} \tag{2.4.9}$$

接下来，如何确定 Copula

2.5 几种实际建模的流程

2.5.1 Independent Copula

2.5.2 Gaussian Copula

分位函数 (quantile function) 也被成为点函数 (percent point function) 或逆累计分布函数 (inverse cumulative distribution function)。假设 $F_X(x) := \Pr(X \leq x) = p$，其分位函数 $Q(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\}$，即求得使累计函数 F 的值大于 p 的最小 x 值，可记为 $Q = F^{-1}$。

使 $p(u)$ 和 $p(v)$ 定义了联合标准正太分布函数 Φ 的分位函数 (quantile fuction), 即 $\Phi(p(u)) = u$ 和 $\Phi(p(v)) = v$。$p(u)$ 和 $p(v)$ 都是实际无力变量在 *Nataf* 空间上的变换值。高斯 copula 定义为：

$$C(u, v) = B(p(u), p(v); \varrho) \tag{2.5.1}$$

其中 B 是标准二元正太累计分布 (bivariate normal cumulative distribution)，ϱ 是 $p(u)$ 和 $p(v)$ 之间的 Pearson 系数 (Pearson coefficient)。从公式2.4.8可得到：

$$C(v|u) = \frac{\partial C(u, v)}{\partial u} = \Phi\left(\frac{p(v) - \varrho p(u)}{\sqrt{1 - \varrho^2}}\right) \tag{2.5.2}$$

2.5.3 Archimedean Copula

2.5.4 Frank Copula

2.5.5 Gumbel Copula

2.5.6 Farlie-Gumbel-Morgenstern Copula

Chapter 3

Practical Coding Methods

3.1 Enhanced Correlation Estimators for Distributed Source Coding in Large Wireless Sensor Networks

In paper [8], a detailed coding scheme was introduced, including side information generation, encoding method in sensor nodes and so on. The coding scheme in this paper includes two parts:

1. **The training phase** of length N: a sensing node maps its l-bit reading $x_s(n)$ according to the alphabet $\mathcal{A} = \{a_i\}_{i=1,2,\ldots,2^l}$, with a quantization step of $|a_{i+1} - a_i| = \triangle$, and sends an uncompressed version of its data coded in l-bits. After collecting the N snapshots of the training phase, the fusion center estimates the correlation parameters for each sensor.

2. **The coding phase**: A given side-information $y(n)$ is available at the fusion center and the sensing node can encode its reading using only $b(n) \leq l$ bits. Hence, the sensor transmits only the index B of a sub-codebook $\mathcal{A}_B \subseteq \mathcal{A}$ (B is codified in $b(n)$ bits) that contains the mapped reading $x_s(n)$. The fusion center receives the sub-codebook identifier B, and selects the symbol in $\mathcal{A}_B$ closer to the side-information $y(n)$,

$$x_s(n) = \arg \min_{a_i \in \mathcal{A}_B} |y(n) - a_i| \tag{3.1.1}$$

3.1.1 Math basis

3.1.1.1 Stieltjes transform

In mathematics, the the **Stieltjes transformation** $S_\rho(z)$ of a measure of density ρ on a real interval $\mathcal{I}$ is the function of the complex variable z defined outside $\mathcal{I}$ by the formula

$$S_\rho = \int_I \frac{\rho(t)dt}{z-t}, \qquad t \in I \subset \mathbb{R}, z \in \mathbb{C}\backslash\mathbb{R} \tag{3.1.2}$$

Under certain conditions we can reconstitute the density function ρ starting from its Stieltjes transformation thanks to the inverse formula of Stieltjes-Perron. For example, if the density ρ is continuous throughout $\mathcal{I}$, one will have inside this interval

$$\rho(x) = \lim_{\varepsilon \to 0^+} \frac{S_\rho(x - i\varepsilon) - S_\rho(x + i\varepsilon)}{2i\pi} \tag{3.1.3}$$

3.1.2 Compute the Side-information $y(n)$

: Observation vector $\mathbf{x}(n) \in \mathbb{R}^{M \times M}$ as the information available at the fusion center and $\mathbf{r}_x$ is the cross-correlation vector, $\mathbf{r}_x = \mathbb{E}[\mathbf{x}(n)x_s(n)]$. The vector $\mathbf{x}(n)$ collects:

1. the K past readings of the sensors;
2. the readings of the set $\mathcal{S}'$ of already-decoded sensors in time slot n (where $\mathcal{S}' \subset \mathcal{S}$ with cardinality S'), hence $M = K + S'$.

<++>

参考文献

[1] M. Fresia, L. Vandendorpe, and H. V. Poor, "Distributed source coding using raptor codes for hidden markov sources," *Ieee Transactions on Signal Processing*, vol. 57, no. 7, pp. 2868–2875, 2009. [Online]. Available: <GotoISI>://WOS:000267379200041

[2] T. Srisooksai, K. Keamarungsi, P. Lamsrichan, and K. Araki, "Practical data compression in wireless sensor networks: A survey," *Journal of Network and Computer Applications*, vol. 35, no. 1, pp. 37–59, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804511000555

[3] B. Ravens, "An introduction to copulas," *Technometrics*, vol. 42, no. 3, 2000.

[4] M. S. Smith and P. J. Danaher, "Modeling multivariate distributions using copulas: Applications in marketing," *Marketing Science*, vol. 30, no. 1, pp. 4–21, 2011.

[5] R. Montes-Iturrizaga and E. Heredia-Zavoni, "Environmental contours using copulas," *Applied Ocean Research*, vol. 52, pp. 125–139, 2015. [Online]. Available: <GotoISI>://WOS:000360419100012

[6] C. Genest and A.-C. Favre, "Everything you always wanted to know about copula modeling but were afraid to ask," *JOURNAL OF HYDROLOGIC ENGINEERING*, vol. 12, no. 4, pp. 347–368, JUL-AUG 2007, Conference on Copula Modeling in Hydrology, Quebec City, CANADA, MAY, 2004.

[7] N. Deligiannis, E. Zimos, D. M. Ofrim, Y. Andreopoulos, and A. Munteanu, "Distributed joint source-channel coding with copula-function-based correlation modeling for wireless sensors measuring temperature," *Ieee Sensors Journal*, vol. 15, no. 8, pp. 4496–4507, 2015. [Online]. Available: <GotoISI>://WOS:000357802000043

[8] J. Enric Barcelo-Llado, A. Morell Perez, and G. Seco-Granados, "Enhanced Correlation Estimators for Distributed Source Coding in Large Wireless Sensor Networks," *IEEE SENSORS JOURNAL*, vol. 12, no. 9, pp. 2799–2806, SEP 2012.