Statistics and Application 统计学与应用, 2018, 7(2), 234-240
Published Online April 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.72027

Hans汉斯

N-W Kernel Regression Estimation for Correlation Function of Bivariate Extremes Copula Function

Xiaoyi Jiang, Haomin Zhang, Lifang Liang

College of Science, Guilin University of Technology, Guilin Guangxi
Email: 1532693042@qq.com

Received: Apr. 3rd, 2018; accepted: Apr. 21st, 2018; published: Apr. 28th, 2018

Abstract

This paper gives an estimate of correlation function for bivariate extremes Copula model using kernel regression method. A N-W kernel regression estimator is constructed and we prove that the estimator is asymptotically unbiased. Based on selection of the optimal bandwidth, we compare the N-W kernel regression estimation and OLS estimation by numerical simulation. The result shows that the N-W kernel regression estimator is more stable than the OLS estimator. So, the N-W kernel regression estimation is a relatively favourable non-parametric method.

Keywords

Extremes Copula Function, Correlation Function, N-W Kernel Regression Estimator

二元极值Copula函数的相关函数的N-W核回归估计

蒋晓艺，张浩敏，梁丽芳

桂林理工大学理学院，广西 桂林
Email: 1532693042@qq.com

收稿日期：2018年4月3日；录用日期：2018年4月21日；发布日期：2018年4月28日

摘　要

本文利用核回归估计方法对二元极值Copula函数的相关函数进行估计。构建了相关函数的N-W核回归估

文章引用: 蒋晓艺, 张浩敏, 梁丽芳. 二元极值 Copula 函数的相关函数的 N-W 核回归估计[J]. 统计学与应用, 2018, 7(2): 234-240. DOI: 10.12677/sa.2018.72027

计。在选择最优带宽的前提下，通过数值模拟对比了N-W核回归估计与OLS估计。数值模拟的结果显示N-W核回归估计在一定情况下较之于OLS估计更具有稳定性，是一种相对较优的相关函数非参数估计方法。

关键词

极值Copula函数，相关函数，N-W核回归估计

1. 引言

1959 年 Sklar 首次提出 Copula 的概念并证明了任何一个 n 维联合分布函数都可以用某个 Copula 函数“连接”n 单变量的边际分布函数表达，其中边际分布描述了单个变量的特征，而 Copula 函数用以刻画边际分布之间的“结构”[1]。Copula 函数本质是把多元随机变量的联合分布函数用各一维边际分布函数连接起来的函数。更进一步，如果给出了一组边际分布和某个 Copula 函数，那么就可以用该 Copula 构造一个联合分布。Copula 的这种灵活的特征使得它在金融、环境资源管理等领域得到广泛的应用。

Copula 函数的优良性质和特殊结构使得其在极端事件统计规律的研究中具有重要作用，极值 Copula 函数的相关函数作为极值 Copula 函数的重要推导函数，国外从上个世纪的八十年代开始对相关函数进行估计[2]。Csorgo 与 Revesz 率先提出了经典相关函数的非参数 Pickands 估计[3]。Muller 和 Roeder 通过变量替换和顶点限制的方式获得了二元极值 Copula 相关函数的 CFG-估计，并证明了二元 CFG-估计为非参数无偏估计[4]。2008 年，Zhang，Wells 和 Peng 将相关函数的二元 CFG-估计推广到了多元，并推导出了多元 CFG-估计仍为非参数无偏估计[5]。Peter 和 Nader [6]通过交叉验证的方法获得了 HT-估计。Gordon 和 Johan [7]在 HT-估计的基础上采用最小二乘法获得了相关函数的 OLS-估计。

综合现有的文献可以发现，国内外很多学者关注于极值 Copula 函数的相关函数的研究。受此启发，本文在 OLS-估计和 N-W 核回归估计模型的基础上构建了二元相关函数的 N-W 核回归估计并通过数值模拟验证了 N-W 核回归估计在一定程度上优于 OLS-估计[6]。

2. 极值 Copula 函数的相关函数

假设 (X,Y) 为二元随机变量，令 (X,Y) 的联合分布函数为 H，边缘分布分别为 F，G 且均为连续函数，则 $F(X)$，$G(Y)$ 服从 $[0,1]$ 的均匀分布，C 为二元极值 Copula 函数，令 $U=-\log F(X)$，$V=-\log G(Y)$，任何 C 都可以有如下表达式[2]：

$$\begin{aligned} &P(U\geq u,V\geq v)=P\left(-\log F(X)\geq u,-\log G(Y)\geq v\right)\\ &=C\left(F(X)\leq \mathrm{e}^{-u},G(Y)\leq \mathrm{e}^{-v}\right)=\exp\left\{-(u+v)A\left(\frac{u}{u+v}\right)\right\}\end{aligned} \tag{2.1}$$

上式中 $A(\omega),\omega\in[0,1]$ 为 C 的相关函数。函数 $A(\omega)$ 具有如下性质：

1) $A(\omega)$ 为凸函数；
2) $A(0)=A(1)=1$；
3) $\max\{\omega,1-\omega\}\leq A(\omega)\leq 1$。

$(X_1,Y_1),\cdots,(X_n,Y_n)$为$(X,Y)$的$n$个独立样本取$\xi_i(\omega)=\min\left(\frac{U_i}{\omega},\frac{V_i}{1-\omega}\right)$，$\omega\in[0,1]$，$i=1,2,\cdots,n$，由于$U_i$，$V_i$分别为$U$，$V$的第$i$个分量。

当$x>0$时$\xi_i(\omega)\geq x$的概率表达式为：

$$P(\xi_i(\omega)\geq x)=P\{U_i\geq\omega x,V_i\geq(1-\omega)x\}=\exp\{-xA(\omega)\}.$$

由此可知$\xi_1(\omega),\xi_2(\omega),\cdots,\xi_n(\omega)$服从均值为$1/A(\omega)$的指数分布。因此，$-\log\xi_i(\omega)$服从位置参数为$\log A(\omega)$的Gumbel分布[7]，故有

$$E[-\log\xi_i(\omega)]=\log A(\omega)+\gamma \tag{2.2}$$

其中γ为Euler常数，$\gamma\approx0.5772$。

由(2.2)式得

$$\log\hat{A}_n(\omega)=-\frac{1}{n}\sum_{i=1}^{n}\log\xi_i(\omega)-\gamma,\quad\omega\in[0,1],$$

有

$$E(\log\hat{A}_n(\omega))=E\left(-\frac{1}{n}\sum_{i=0}^{n}\log\xi_i(\omega)-\gamma\right)=E[-\log\xi_i(\omega)]-\gamma=\log A(\omega)$$

成立，所以$\log\hat{A}_n(\omega)$为$\log A_n(\omega)$的渐近无偏估计。

Kendal's τ系数是一个最具有代表性的相关系数，Kendal's τ系数的定义如下[8] [9]：

$$\tau=P((X_1-X_2)(Y_1-Y_2)>0)-P((X-X_2)(Y_1-Y_2)<0)$$

Kendal's τ系数与相关函数的表达式[4]：

$$\tau=\int_0^1\frac{\omega(1-\omega)}{A(\omega)}\mathrm{d}A'(\omega)$$

3. N-W 核估计

设$(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$是来自$(X,Y)$的一个样本，$E(Y)<\infty$，令$m(x)=E(Y\mid X=x)$且$\varepsilon\sim N(0,\sigma^2)$，则$X$与$Y$之间的回归模型为：

$$Y=m(X)+\varepsilon,$$

$m(x)$为未知函数，可以通过权函数方法来拟合，对于样本$(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$，权函数估计就是对$m(x)$进行估计[10]。$m(x)$的非参数回归估计量$\hat{m}_K(x)$可以表示为：

$$\hat{m}_K(x)=\sum_{i=1}^{n}W_{ni}(x;X_1,\cdots,X_n)Y_i\triangleq\sum_{i=1}^{n}W_{ni}(x)Y_i$$

核估计是权函数估计的一种方法，最常见的核估计是Nadaraya和Waston于1964年提出的N-W核权函数回归估计即N-W核估计，N-W核估计得到函数$m(x)$的核光滑方法即[11] [12] [13]

$$\hat{m}_K(x)=\sum_{i=1}^{n}\frac{K_h(x-X_i)}{\sum_{j=1}^{n}K_h(x-X_j)}Y_i, \tag{3.1}$$

其中$K_h(\cdot)=K(\cdot/h)/h$，$K(\cdot)$为核函数，h为带宽或窗宽。核函数$K(x)$具有以下性质：

1) $K(x)\geq 0$；
2) $\int K(x)\mathrm{d}x=1$；
3) $\int xK(x)\mathrm{d}x=0$。

常见的核函数如表 1。

依据 N-W 核回归的定义构建相关函数的核回归的模型：

$$\log \hat{A}_n(\omega)=m(\omega)+\varepsilon,\ \ \varepsilon\sim N(0,\sigma^2)$$

由(3.1)可得到相关函数的 N-W 核估计公式

$$\log \hat{A}_h^{NW}(\omega)=\sum_{i=1}^{n}\frac{K_h(\omega-W_i)}{\sum_{j=1}^{n}K_h(\omega-W_j)}\log \hat{A}_n(W_i),\ \ \omega\in[0,1] \tag{3.2}$$

其中 $W_i(i=1,2,\cdots,n)$ 为 $[0,1]$ 上的随机变量。

核估计的结果与带宽 h 的选择有关所以相关函数的 N-W 核估计结果同样与带宽 h 有关。选择的带宽 h 值越小，核估计的偏差值就会越小，核估计的方差反而越大；反之，选择的带宽 h 值越大，核估计的偏差值就会越大，核估计的方差反而越小。所以要在核估计的偏差与方差之间做一个权衡，使核估计的均方误差最小。选择带宽主要有直接插入法、经验法则、最小平方交叉验证法和惩罚函数法。本文使用广义交叉验证法的最优带宽公式[14]：

$$h^*=\left(\frac{\sigma^2 J_K\int_0^1\left[f(x)\right]^{-1}\mathrm{d}x}{\sigma_K^4\int_0^1\left(m''(x)\right)\mathrm{d}x}\right)^{1/5}n^{-1/5}$$

其中 $J_K=\int_{-1}^{1}K^2(x)\mathrm{d}x$，$\sigma_K^2=\int_{-1}^{1}x^2K(x)\mathrm{d}x$，$f(x)$ 为 $\log \hat{A}_n(\omega)$ 的经验密度函数。

4. 随机模拟

二元极值 Copula 函数的相关函数的模型[15]：

$$A_{\alpha,\beta,r}(\omega)=1-\beta+(\beta-\alpha)\omega+\left[\alpha^r\omega^r+\beta^r(1-\omega)^r\right]^{1/r} \tag{4.1}$$

其中 $0\leq\alpha,\beta\leq 1$，$r\geq 1$。当 $\alpha=\beta=1$ 时(4.1)为：

$$A_r(\omega)=\left[\omega^r+(1-\omega)^r\right]^{1/r} \tag{4.2}$$

Table 1. Common kernel functions
表 1. 常见的核函数

核种类	核函数
Boxcar 核	$K(x)=1/2I(x)$
Gaussian 核	$K(x)=1/\sqrt{2\pi}\mathrm{e}^{-x^2/2}$
Epanechnikov 核	$K(x)=3/4(1-x^2)I(x)$
Tricube 核	$K(x)=70/81(1-\lvert x\rvert^3)^3I(x)$

注：$I(x)$ 为示性函数。

模型(4.1)在除去 $\alpha=\beta=1$ 的情况后为非对称模型，本文选择(4.2)模型与估计值进行比较。在(4.2)模型中的 r 与 Kendal's τ系数关系为 $\tau=1-\frac{1}{r}$。

在本文的模拟过程中相关函数的 N-W 核回归估计均在最优宽带的前提下选择 Gaussian 核函数，其中 Gaussian 核函数的 $J_k=0.2820948$，$\sigma_K=1$ [14]。表 2 为样本量分别为 50，100 和 500 情况下，随机生成 $(X,Y)\sim N(0,0,1,2,1)$ 的二元随机变量，N-W 核估计和 OLS 估计分别与 $A_r(\omega)$ 模型的均方误差。表 3 为样本量分别为 250，500 和 1000 情况下，随机生成 50%的 $(X,Y)\sim N(0,0,1,2,1)$ 和 30%的 $(X,Y)\sim N(0,0,1,3,1)$ 以及 20%的 $(X,Y)\sim N(0,0,2,4,1)$，的混合分布二元随机变量，N-W 核估计和 OLS 估计分别与 $A_r(\omega)$ 模型的均方误差。

如表 2 与表 3 所示的 τ的值均从 0 取到 0.95，间隔为 0.05，相当于 r 的值从 1 取到 20，但间隔不等。如表 2 与表 3 所示在 τ相同样本量不同的情况下，随着样本量的增加相关函数的 N-W 核估计和 OLS 估计与 $A_r(\omega)$ 模型的均方误差几乎均在减小。在如表 2 所示在 τ相同样本量相同的情况下，相关函数的 N-W 核估计与 $A_r(\omega)$ 模型的均方误差均略大于相关函数 OLS-估计与 $A_r(\omega)$ 模型的均方误差。但是在如表 3 所示却相反；如表 2 与表 3 所示均在样本量相同的情况下，随着 r 的增加相关函数的 N-W 核估计与相关函数的 OLS-估计的均方误差都逐渐减小再增加；如表 2 所示相关函数的 N-W 核估计分别在样本量为 50，

Table 2. Mean square error in pure data
表 2. 数据纯净情况下的均方误差

τ	r	N-W			OLS		
		n = 50	n = 100	n = 500	n = 50	n = 100	n = 500
0.00	1.0000	0.146942	0.082552	0.056272	0.051214	0.050122	0.044455
0.05	1.0526	0.123233	0.067013	0.043829	0.039383	0.038107	0.033216
0.10	1.1111	0.102391	0.053682	0.033394	0.029523	0.028136	0.024020
0.15	1.1765	0.084173	0.042356	0.024765	0.021435	0.020007	0.016653
0.20	1.2500	0.068358	0.032843	0.017756	0.014933	0.013532	0.010916
0.25	1.3334	0.054735	0.024965	0.012192	0.009841	0.008534	0.006620
0.30	1.4286	0.043111	0.018555	0.007911	0.005994	0.004846	0.003585
0.35	1.5385	0.033302	0.013456	0.004757	0.003241	0.002313	0.001642
0.40	1.6667	0.025138	0.009520	0.002585	0.001433	0.000787	0.000628
0.45	1.8182	0.018459	0.006606	0.001259	0.000436	0.000125	0.000388
0.50	2.0000	0.013113	0.004579	0.000646	0.000118	0.000191	0.000773
0.55	2.2222	0.008952	0.003309	0.000624	0.000358	0.000851	0.001642
0.60	2.5000	0.005837	0.002671	0.001074	0.001039	0.001972	0.002860
0.65	2.8571	0.003629	0.002542	0.001880	0.002049	0.003422	0.004303
0.70	3.3333	0.002188	0.002803	0.002931	0.003279	0.005070	0.005861
0.75	4.0000	0.001372	0.003345	0.004116	0.004622	0.006791	0.007439
0.80	5.0000	0.001028	0.004068	0.005324	0.005965	0.008479	0.008972
0.85	6.6667	0.000996	0.004898	0.006436	0.007190	0.010066	0.010433
0.90	10.00	0.001103	0.005789	0.007334	0.008171	0.011553	0.011834
0.95	20.00	0.001178	0.006723	0.007933	0.008816	0.012995	0.013142

Table 3. Mean square error in mixed data
表 3. 数据混杂情况下的均方误差

τ	r	N-W			OLS		
		n = 250	n = 500	n = 1000	n = 250	n = 500	n = 1000
0.00	1.0000	0.046526	0.038409	0.038532	0.078072	0.058253	0.052327
0.05	1.0526	0.035004	0.028263	0.028102	0.063104	0.045229	0.040220
0.10	1.1111	0.025533	0.020016	0.019674	0.05032	0.034327	0.030164
0.15	1.1765	0.017901	0.013471	0.013037	0.039519	0.025333	0.021956
0.20	1.2500	0.011904	0.008438	0.007989	0.030507	0.018046	0.015404
0.25	1.3334	0.007356	0.004746	0.004342	0.023105	0.012279	0.010331
0.30	1.4286	0.004075	0.002228	0.001915	0.017148	0.007855	0.006571
0.35	1.5385	0.001893	0.000731	0.000538	0.012477	0.004603	0.003964
0.40	1.6667	0.000646	0.000108	0.000046	0.008942	0.002366	0.002361
0.45	1.8182	0.000181	0.000219	0.000284	0.006402	0.000991	0.001619
0.50	2.0000	0.000346	0.000934	0.001102	0.004721	0.000334	0.001598
0.55	2.2222	0.001003	0.002127	0.002356	0.003766	0.000260	0.000216
0.60	2.5000	0.002017	0.003678	0.003912	0.003528	0.000640	0.003181
0.65	2.8571	0.003267	0.005471	0.005647	0.003410	0.001358	0.004520
0.70	3.3333	0.004643	0.007391	0.007450	0.003999	0.002306	0.006049
0.75	4.0000	0.006055	0.009327	0.009235	0.004711	0.003396	0.007647
0.80	5.0000	0.007442	0.011160	0.001094	0.005565	0.004558	0.009211
0.85	6.6667	0.008776	0.012765	0.012556	0.006490	0.005746	0.010681
0.90	10.00	0.010067	0.014011	0.014093	0.007448	0.006942	0.012058
0.95	20.00	0.011284	0.014809	0.015522	0.008435	0.008087	0.01339

100 和 500 情况下在 $r=6.667$， $r=2.8571$ 和 $r=2.222$ 附近处取得最小。相关函数的 OLS 估计分别在样本量为 50，100 和 500 情况下在均在 $r=2$ 附近处取得最小；如表 3 所示相关函数的 N-W 核估计分别在样本量为 250，500 和 1000 情况下均在 $r=6.667$ 附近处取得最小。相关函数的 OLS 估计分别在样本量为 250，500 和 1000 情况下均在 $r=2.8571$ 和 $r=2.222$ 附近处取得最小；在样本量服从 $X\sim N(0,1)$，$Y\sim N(0,2)$，$\rho=1$ 的情况下，可知在 $r=1.648$ 附近处均方误差为最小。有模拟结果可知在分布已知数据纯净的情况下，相关函数的 OLS 估计效果好，分布未知混杂数据的情况下，相关函数的 N-W 核估计效果好。

5. 总结

本文在二元极值 Copula 函数的相关函数 OLS 估计的基础上，结合具有模型简单，参数少且稳定性高的非参数估计方法 N-W 核估计，提出了相关函数的 N-W 核估计，并证明了该估计的无偏性。通过生成服从不相关的二元正态分布的随机变量数值生成 N-W 核估计与 OLS 估计模拟相关函数。分别与选定的相关函数的模型进行比较，可以得出 N-W 核估计的稳定性在分布未知数据混杂的情况下要高于 OLS 估计。

本文数值分析选择了相关系数单一且样本量较小的情况分析，相关系数的选择和样本量的个数可能会对相关函数的估计会造成影响，在以后的研究中还需进一步的验证方法的适用性。

基金项目

国家自然科学基金项目(71762008)。

参考文献

[1] Sklar, A. (1959) Fonctions de Reparition an Dimensions et Leurs Marges. *Publications de l'Intitut de Statistique de I'Universit de Paris*, **8**, 229-231.

[2] 吴娟. Copula 理论与相关性分析[D]: [博士学位论文]. 武汉: 华中科技大学, 2009.

[3] Congo, M. and Revesz, P. (1981) Strong Approximation in Probability and Statistics. Academic Press, New York, 7-108.

[4] Muller, P. and Roeder, K. (1997) A Nonparametric Estimation Procedure for Bivariate Extreme Value Copulas. *Biometrika*, **84**, 567-577.

[5] Zhang, D., Wells, M.T. and Peng, L. (2008) Nonparametric Estimation of the Dependence Function for a Multivariate Extreme Value Distribution. *Journal of Multivariate Analysis*, **99**, 577-588.

[6] Hall, P. and Tajvidi, N. (2000) Distribution and Dependence-Function Estimation for Bivariate Extreme-Value Distributions. *Bernoulli*, **6**, 835-844.

[7] Gudendorf, G. and Segers, J. (2011) Nonparametric Estimation of an Extreme-Value Copula in Arbitrary Dimensions. *Journal of Multivariate Analysis*, **102**, 37-47.

[8] Fredricks, G.A. and Nelsen, R.B. (2007) On the Relationship between Spearman's Rho and Kendall's Tau for Pairs of Continuous Random Variables. *Journal of Statistical Planning & Inference*, **137**, 2143-2150.

[9] Niewiadomska-Bugaj, M. and Kowalczyk, T. (2005) On Grade Transformation and Its Implications for Copulas. *Brazilian Journal of Probability & Statistics*, **19**, 125-137.

[10] 吴喜之, 王兆军. 非参数统计方法[M]. 北京: 高等教育出版社, 1996: 274-277.

[11] Hardle, W. (1990) Applied Nonparametric Regression: References. Cambridge University Press, 225-226.

[12] Hart, J.D. (1997) Nonparametric Smoothing and Lack-of-Fit Tests. Springer, 6-176.

[13] 陈希孺, 方兆本, 李国英. 非参数统计[M]. 上海: 上海科学技术出版社, 1989: 361-367.

[14] 李艳娟. 核估计量与窗宽选择[J]. 辽宁工程技术大学学报, 2006, 25(3): 478-480.

[15] Tawn, J. (1988) Bivariate Extreme Value Theory: Models and Estimation. *Biometrika*, **75**, 397-451.