DOI: 10.7641/CTA.2013.21262

基于Copula熵的互信息估计方法

韩　敏†, 刘晓欣

(大连理工大学 电子信息与电气工程学部, 辽宁 大连 116023)

摘要: 互信息是一种常用的衡量变量相关性的方法, 但在互信息估计过程中, 联合概率密度的估计往往十分困难. 为了避免联合概率密度的估计, 同时有效提高互信息估计的准确度与效率, 本文提出一种基于Copula熵的互信息估计方法. 利用Copula熵与互信息之间的关系, 将互信息的估计转化为对Copula熵值的估计. 采用基于Kendall秩相关系数的参数估计方法对Copula函数的参数进行估计. 所提算法分别与直方图法、核方法、k近邻法和极大似然法进行比较. 二维高斯数据上的仿真结果表明, 所提方法能够快速准确地对互信息值进行估计.

关键词: 互信息; Copula; 熵; 概率密度函数; 参数估计

中图分类号: TP181　　**文献标识码**: A

Mutual information estimation based on Copula entropy

HAN Min†, LIU Xiao-xin

(Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian Liaoning 116023, China)

Abstract: Mutual information is commonly used in the measure of dependency between variables. However, in the estimation of mutual information, the estimation of joint probability density function is always a hard problem. To avoid the estimation of joint probability density function and to improve both the effectiveness and efficiency of the estimation of mutual information, we propose a novel mutual information estimation method based on the entropy of the Copula density function. The estimation of mutual information is transformed into the estimation of Copula entropy by taking advantages of their relationships. Parameter estimation method based on Kendall's rank correlation coefficient is used for the estimation of the parameters in Copula function. The proposed method is compared with the following four methods: histogram method, kernel method, k nearest neighbor method and maximum likelihood method. Simulation results on the two dimensional Gaussian distribution data substantiate the effectiveness and efficiency of the proposed method.

Key words: mutual information; Copula; entropy; probability density function; parameter estimation

1　引言(Introduction)

互信息是信息理论中一个非常重要的概念, 因互信息能够衡量非线性的相关关系, 目前已广泛应用于变量相关性的评价与变量选择之中[1−2]. 但是由于互信息估计受到联合概率密度估计的限制, 其准确度往往会影响变量相关性的度量[3]. 目前常见互信息估计方法通常从两个角度提高互信息估计的准确度, 一是从概率密度估计的角度尽量提高联合概率密度函数估计的准确度, 如直方图法[4]和核方法[5]; 二是从互信息估计的角度尽量避免对联合概率密度函数的估计, 如k近邻法(k nearest neighbor, KNN)[6].

直方图法[4]是最简单的非参数方法, 计算简单, 且易于理解. 通过将空间分成相等的几个部分, 计算每个部分里包含的元素个数, 从而估计概率密度函数. 该方法计算复杂度低, 精度也低. 高维数据空间中经常出现的稀疏数据分布会大大降低直方图法的可靠性, 因此该方法只适用于低维数据. 直方图法的误差主要来源于空间分割大小的选择, 因此研究者们提出了不同的空间分割形式. 根据其分割形式直方图方法通常可分为等距法、等概率法与自适应法, 3种方法均能得到理论互信息值的合理估计, 其中自适应法随样本规模的变化收敛更快, 等概率法估计较等距法的精度更高一些[7].

核方法[5]采用核函数来进行概率密度的估计, 进而用于计算互信息, 即在特征的每个点上叠置一个基函数, 如高斯核函数或Parzen窗函数. 核方法精度高, 但同时计算复杂度也很高, 适用于数据点较少的情况, 当数据点较多时, 维数较高, 核参数的估计将会变得十分困难.

KNN法[6]通过统计样本点的最近邻信息进行互信息估计, 避免了直接进行概率密度估计, 而且比较容易进行高维互信息的计算. 与核方法相比, k近邻法在

收稿日期: 2012−12−14; 收修改稿日期: 2013−01−29.

†通信作者. E-mail: minhan@dlut.edu.cn; Tel.: +86 0411−84707847.

基金项目: 国家自然科学基金资助项目(61074096).

时间复杂度与精度上均存在优势, 这是因为在搜索邻居样本的过程中数据的有效结构得到了充分利用. k近邻法更稳定, 更不易受噪声影响. 但是参数k的选择对于互信息的估计具有较大的影响, 并且目前尚无合理选取k值的系统方法.

除了以上3种方法之外, 常用的互信息估计方法还有B样条法[8]、小波方法、Edgeworth法、最小二乘法、最大似然法[9]和贝叶斯方法[10]等等. B样条法避免了核密度估计中时间消耗较大的数值积分过程, 因此与核方法相比计算效率更高, 通过初始点的轮换缓解了直方图中存在的初始点选择问题的影响, 而与KNN方法相比, B样条法的计算时间更少. 小波方法能够很好捕捉数据在时域和频域上的局部特性.当数据接近正态分布时, Edgeworth法的估计较为准确, 但当数据的分布类型与整体分布相差较大时, Edgeworth法的估计误差会变得很大. 与之相比, 最大似然法在两种情况下都能得到较为合理的估计结果, 其不足之处在于当与模型的自由度相比数据集规模不足够大时, 最大似然法很容易出现过拟合的现象. 贝叶斯方法很好地克服了这一问题, 它能够确定模型参数的最佳个数. 因此, 贝叶斯法对于无法获得大量数据的数据集很有效. 如前所述, 由于密度估计本身存在的困难导致核方法在很多实际应用中并不可行, 而KNN方法又依赖于k值的选择, 最小二乘法能够克服核方法与KNN法的局限, 并且最小二乘法不受数据分布类型的限制.

综上所述, 每一种互信息估计方法均有其适用的数据类型和优势, 但从互信息估计的角度出发避免估计联合概率密度的方法更为有效. 因此本文通过分析互信息与Copula函数之间的关系, 提出一种基于Copula熵的互信息估计方法, 通过Copula函数对互信息估计的表达式进行化简, 从而避免联合概率密度函数的估计过程. 并与应用较多的直方图法和核方法进行比较, 仿真实验表明所提方法能够实现快速有效的互信息估计.

2 互信息与Copula熵(Mutual information and Copula entropy)

变量X,Y之间的互信息定义[11]如下:

$$I(X,Y)=\iint p_{\rm XY}(x,y)\log\frac{p_{\rm XY}(x,y)}{p_{\rm X}(x)p_{\rm Y}(y)}{\rm d}x{\rm d}y, \tag{1}$$

其中: $p_{\rm XY}(x,y)$表示变量X,Y的联合概率密度, $p_{\rm X}(x)$和$p_{\rm Y}(y)$分别为变量X和Y的边缘概率密度函数. 互信息越大, 则变量X包含了关于Y越多的信息, 即两变量相关性越大.

由二元Copula的Sklar定理[12]可知

$$C(u,v)=P_{\rm XY}(x,y)=P_{\rm XY}(P_{\rm X}^{-1}(u),P_{\rm Y}^{-1}(v)), \tag{2}$$

其中: $u=P_{\rm X}(x)$和$v=P_{\rm Y}(y)$分别为随机变量X和Y的边缘累积分布函数, $P_{\rm XY}(x,y)$为两变量的联合累积分布函数. 由此可推导出Copula密度函数:

$$\begin{aligned}&c(u,v)=\\&\frac{\partial^2C(u,v)}{\partial u\partial v}=\frac{\partial^2C(P_{\rm X}(x),P_{\rm Y}(y))}{\partial P_{\rm X}(x)\partial P_{\rm Y}(y)}=\\&\frac{\partial^2P_{\rm XY}(x,y)}{p_{\rm X}(x)p_{\rm Y}(y)\partial x\partial y}=\frac{p_{\rm XY}(x,y)}{p_{\rm X}(x)p_{\rm Y}(y)}.\end{aligned} \tag{3}$$

根据上式可将互信息的表达式化简如下:

$$\begin{aligned}&I(X,Y)=\\&\iint p_{\rm XY}(x,y)\log\frac{p_{\rm XY}(x,y)}{p_{\rm X}(x)p_{\rm Y}(y)}{\rm d}x{\rm d}y=\\&\iint c(u,v)p_{\rm X}(x)p_{\rm Y}(y)\times\\&\log\frac{c(u,v)p_{\rm X}(x)p_{\rm Y}(y)}{p_{\rm X}(x)p_{\rm Y}(y)}{\rm d}x{\rm d}y=\\&\iint c(u,v)p_{\rm X}(x)p_{\rm Y}(y)\log c(u,v){\rm d}x{\rm d}y=\\&\iint c(u,v)\log c(u,v){\rm d}u{\rm d}v=\\&-H_{\rm c}(u,v),\end{aligned} \tag{4}$$

其中$H_{\rm c}(u,v)$为u和v的Copula熵. 根据文献 [13], Copula熵的定义如下:

$$\begin{aligned}&H_{\rm c}(u_1,u_2,\cdots,u_n)=\\&-\int\cdots\int c(u_1,u_2,\cdots,u_n)\times\\&\log c(u_1,u_2,\cdots,u_n){\rm d}u_1{\rm d}u_2\cdots{\rm d}u_n,\end{aligned} \tag{5}$$

且具有以下性质:

$$\begin{aligned}&H(x_1,x_2,\cdots,x_n)=\\&\sum_{i=1}^{n}H(x_i)+H_{\rm c}(u_1,u_2,\cdots,u_n).\end{aligned} \tag{6}$$

因此, 可利用Copula函数来估计互信息, 从而避免了联合概率密度函数的估计. 同时可将其扩展到高维, 得到下式:

$$\begin{aligned}&I(X_1,X_2,\cdots,X_n)=\\&\int\cdots\int c(u_1,\cdots,u_n)p(x_1),\cdots,p(x_n)\times\\&\log c(u_1,\cdots,u_n){\rm d}x_1\cdots{\rm d}x_n=\\&\int\cdots\int c(u_1,\cdots,u_n)\log c(u_1,\cdots,u_n){\rm d}u_1\cdots{\rm d}u_n,\end{aligned} \tag{7}$$

则互信息与Copula熵存在以下关系:

$$I(X_1,X_2,\cdots,X_n)=-H_{\rm c}(u_1,u_2,\cdots,u_n). \tag{8}$$

3 基于Copula熵的互信息估计方法的实现(Implementation of mutual information estimation based on Copula entropy)

Shannon熵的定义如下:

$$H(X)=-\int p_{\rm X}(x)\log p_{\rm X}(x){\rm d}x. \tag{9}$$

在实际应用中, 目标是根据X的一组样本$(x_1, x_2, \cdots, x_N)$来估计Shannon熵, 因此常用下式对熵进行估计[6]:

$$\hat{H}(X) = -\frac{1}{N}\sum_{i=1}^{N}\log\tilde{p}(x_i). \tag{10}$$

因此, 只要根据样本$(x_1, x_2, \cdots, x_N)$估计出X的概率密度函数$\tilde{p}(x)$, 就能够估计出其熵值.

要估计Copula熵另一个关键在于Copula函数的参数估计, 常用的有效的参数估计方法是基于Kendall τ相关系数的参数估计方法. Copula参数与Kendall τ相关系数的关系如下式[14]:

$$\tau = 4\iint_{[0,1]^2} C(u,v)\mathrm{d}C(u,v) - 1. \tag{11}$$

将不同类型的Copula函数表达式代入式(11), 则可以反推出Copula参数的τ表达式. 例如: Gaussian Copula的分布函数表达式如下:

$$C(u,v) = \int_{-\infty}^{\phi^{-1}(u)}\int_{-\infty}^{\phi^{-1}(v)}\frac{1}{2\pi\sqrt{1-\rho^2}}\times \mathrm{e}^{-x^2-2\rho xy+y^2/2(1-\rho^2)}\mathrm{d}x\mathrm{d}y, \tag{12}$$

则其参数ρ可由Kendallτ相关系数估计:

$$\rho = \sin\frac{\pi}{2\tau}. \tag{13}$$

Clayton Copula的分布函数表达式如下:

$$C(u,v) = \max([u^{-\theta} + v^{-\theta} - 1]^{-1/\theta}, 0), \tag{14}$$

则其参数θ可由Kendallτ相关系数估计:

$$\theta = \frac{2\tau}{1-\tau}. \tag{15}$$

基于Copula熵的互信息估计方法的具体步骤如下所示:

1) 将数据(X, Y)归一化至[0,1], 归一化公式为

$$\bar{X} = \frac{X - X_{\min}}{X_{\max} - X_{\min}},\ \bar{Y} = \frac{Y - Y_{\min}}{Y_{\max} - Y_{\min}}, \tag{16}$$

其中: $X_{\min}$为X的最小值, $X_{\max}$为X的最大值, $Y_{\min}$为Y的最小值, $Y_{\max}$为Y的最大值.

2) 计算$(\bar{X}, \bar{Y})$的Kendall τ相关系数. 将$\bar{X}$中的相同元素组合成s个小集合, U_i表示第i个小集合所包含的元素个数, 将$\bar{Y}$中的相同元素组合成t个小集合, V_i表示第i个小集合所包含的元素个数, 则

$$\begin{aligned} N_1 &= \sum_{i=1}^{s}\frac{1}{2}U_i(U_i - 1),\\ N_2 &= \sum_{i=1}^{t}\frac{1}{2}V_i(V_i - 1). \end{aligned} \tag{17}$$

由此可得$(\bar{X}, \bar{Y})$的Kendallτ相关系数计算公式:

$$\tau = \frac{C - D}{\sqrt{(N_3 - N_1)(N_3 - N_2)}}, \tag{18}$$

其中: C表示$(\bar{X}, \bar{Y})$中拥有一致性的元素对数, D表示$(\bar{X}, \bar{Y})$中拥有不一致性的元素对数,

$$N_3 = \frac{1}{2N(N-1)}.$$

因此该步的计算时间复杂度为$\mathrm{O}(N^2)$.

3) 根据Kendallτ相关系数估计出相应类型Copula函数的参数.

4) 根据Copula函数的参数得到Copula密度函数$c(u, v)$.

5) 根据下式计算得到Copula熵:

$$H_{\mathrm{c}}(u,v) = -\frac{1}{N}\sum_{i=1}^{N}\log c(u_i, v_i), \tag{19}$$

其中$(u_i, v_i),\ i = 1, \cdots, N$是由Copula密度函数$c(u, v)$生成的$N$个样本点.

6) 则互信息为

$$I(X;Y) = -H_{\mathrm{c}}(u,v) = \frac{1}{N}\sum_{i=1}^{N}\log c(u_i, v_i). \tag{20}$$

文献 [9]中最大似然法所得到的互信息估计形式:

$$I(X;Y) = \frac{1}{N}\sum_{i=1}^{N}\log\omega(x_i, y_i), \tag{21}$$

其中

$$\omega(x,y) = \frac{p_{\mathrm{XY}}(x,y)}{p_{\mathrm{X}}(x)p_{\mathrm{Y}}(y)}$$

为密度比函数. 由式(20)与式(21)对比可知, 本文方法与最大似然法所得到的互信息最终表达形式是一致的, Copula密度函数也可以在某种程度上视为一种密度比函数. 在以上各步中, 步骤 1) 与步骤 5) 的计算时间复杂度为$\mathrm{O}(N)$, 步骤 2) 的计算时间复杂度为$\mathrm{O}(N^2)$, 其余各步骤的计算时间复杂度均为$\mathrm{O}(1)$. 因此, 所提方法的整体计算时间复杂度为$\mathrm{O}(N^2)$.

4 仿真结果与分析(Simulation results and analysis)

为了验证基于Copula熵的互信息估计方法的有效性, 分别选取Gaussian Copula、Clayton Copula、Frank Copula和Gumbel Copula四种常见类型的Copula函数进行仿真实验. 随机生成10000×2的高斯分布数据, 期望值分别为$[20, 50]$, 方差为$[1, 1]$, Pearson相关系数r的变化区间为$[0, 1)$. 其二维高斯分布的互信息可由下式给出[7]:

$$I_{\mathrm{Gauss}}(X,Y) = -\frac{1}{2}\log(1 - r^2). \tag{22}$$

基于4种不同类型Copula函数的互信息估计绝对误差如表1所示. 由表1可以看出, Gaussian Copula和Clayton Copula的互信息估计准确度较高, 因此选取这两类Copula的仿真结果分别与直方图法、核方法、KNN法和极大似然法进行比较, 如表2所示, 仿真曲线如图1所示.

表 1 基于4种不同类型Copula函数的互信息估计绝对误差

Table 1 Comparison of absolute errors for mutual information estimation based on four different Copula functions

Pearson相关系数	Gaussian Copula	Clayton Copula	Frank Copula	Gumbel Copula
0.0000	1.3096e-005	1.9074e-004	0.0002	0.0017
0.1000	0.0004	0.0073	0.0027	0.0140
0.2000	0.0001	0.0107	0.0012	0.0208
0.3000	0.0032	0.0081	0.0116	0.0292
0.4000	0.0025	0.0186	0.0204	0.0305
0.5000	0.0021	0.0226	0.0318	0.0389
0.6000	0.0054	0.0398	0.0545	0.0442
0.7000	0.0058	0.0468	0.0711	0.0521
0.8000	0.0127	0.0329	0.1144	0.0500
0.9000	0.0245	0.0989	0.0947	0.0626

表 2 互信息估计仿真结果比较

Table 2 Comparison of estimated results for different mutual information estimation methods

Pearson 相关系数	二维高斯分布	直方图	核方法	KNN	极大似然法	Gaussian Copula	Clayton Copula
0.0000	0.0000	0.0029	0.0096	0.0161	0.0000	0.0000	0.0002
0.1000	0.0050	0.0120	0.0153	0.0156	0.0000	0.0054	0.0123
0.2000	0.0204	0.0285	0.0281	0.0264	0.0172	0.0205	0.0311
0.3000	0.0472	0.0608	0.0594	0.0378	0.0398	0.0440	0.0553
0.4000	0.0872	0.1091	0.1005	0.0962	0.0841	0.0847	0.1058
0.5000	0.1438	0.1865	0.1586	0.1422	0.1345	0.1459	0.1664
0.6000	0.2231	0.2677	0.2381	0.2240	0.2091	0.2177	0.2629
0.7000	0.3367	0.3954	0.3544	0.3472	0.3064	0.3425	0.3835
0.8000	0.5108	0.5686	0.5200	0.5218	0.4642	0.5235	0.5437
0.9000	0.8304	0.8515	0.8313	0.8314	0.8048	0.8059	0.9293

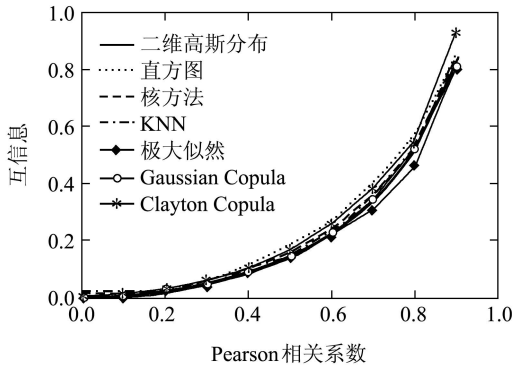

图 1 互信息估计曲线比较

Fig. 1 Comparison of estimated curves for different mutual information estimation methods

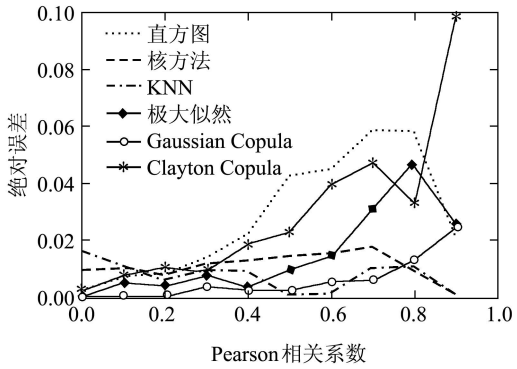

图 2 互信息估计绝对误差曲线比较

Fig. 2 Comparison of the absolute error curves for different mutual information estimation methods

由表2和图1可以看出, 6种方法估计得到的互信息值随着相关系数的变化趋势与二维高斯分布均保持一致. 为了更加直观地分析与比较6种互信息估计方法的仿真结果, 图2给出了6种方法与二维高斯分布的绝对误差曲线.

由图2可以看出, 基于Gaussian Copula的互信息估计方法与核方法得到的互信息值均非常接近二维高斯分布的互信息值, 而直方图法与Clayton Copula估计的结果误差较大. 这说明Copula函数的选择对于互信息估计非常关键. 由于仿真数据为二维高斯

分布, 所以Gaussian Copula的估计准确度更高.

随着Pearson相关系数的增大, 直方图法、极大似然法、Gaussian Copula和Clayton Copula这4种方法估计互信息的绝对误差均有增大的趋势, 其中基于Gaussian Copula的互信息估计方法的绝对误差最小, 且当Pearson相关系数越小, 其优势越明显. 核方法与KNN法表现较稳定, 当Pearson相关系数变化时, 两种方法的估计结果趋于稳定.

5　结论(Conclusions)

针对互信息估计中联合概率密度估计困难且准确度低的问题, 本文提出了一种基于Copula熵的互信息估计方法. 根据 Sklar 定理可以推导出互信息与Copula熵之间的关系式, 从而在互信息估计过程中有效避免了对联合概率密度函数的估计. 然后利用Kendallτ相关系数与Copula累积分布函数之间的关系式, 估计出相应类型的Copula参数. 最后利用Copula密度函数估计出离散化的Copula熵, 从而得到互信息的估计值. 仿真结果表明所提方法不仅计算简便, 计算复杂度低, 而且能够获得较高的估计精度.

参考文献(References):

[1] PENG HC, LONG F, DING C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(8): 1226 – 1238.

[2] 韩敏, 刘晓欣. 基于互信息的分步式输入变量选择多元序列预测研究 [J]. 自动化学报, 2012, 38(6): 999 – 1006.
(HAN Min, LIU Xiaoxin. Stepwise input variable selection based on mutual information for multivariate forecasting [J]. *Acta Automatica Sinica*, 2012, 38(6): 999 – 1006.)

[3] HACINE-GHARBI A, RAVIER P, HARBS R, et al. Low bias histogram-based estimation of mutual information for feature selection [J]. *Pattern Recognition Letters*, 2012, 33(16): 1302 – 1308.

[4] 韩敏, 梁志平. 改进型平均移位柱状图估算概率密度并对互信息作相关分析 [J]. 控制理论与应用, 2011, 28(6): 845 – 850.
(HAN Min, LIANG Zhiping. Correlation analysis of mutual information by probability density estimated from improved averaged-shifted-histogram [J]. *Control Theory & Applications*, 2011, 28(6): 845 – 850.)

[5] MOON Y, RAJAGOPALAN B, LALL U. Estimation of mutual information using kernel density estimators [J]. *Physical Review E*, 1995, 52(3): 2318 – 2321.

[6] KRASKOV A, STOGBAUER H, GRASSBERGER P. Estimating mutual information [J]. *Physical Review E*, 2004, 69(6): 066138.

[7] DARBELLAY G A, VAJDA I. Estimation of the information by an adaptive partitioning of the observation space [J]. *IEEE Transactions on Information Theory*, 1999, 45(4): 1315 – 1321.

[8] DAUB C O, STEUER R, SELBIG J, et al. Estimating mutual information using B-spline functions — an improved similarity measure for analysing gene expression data [J]. *BMC Bioinformatics*, 2004, 5(1): 118.

[9] SUZUKI T, SUGIYAMA M, SESE J, et al. Approximating mutual information by maximum likelihood density ratio estimation [C] //*New Challenges for Feature Selection in Data Mining and Knowledge Discovery, JMLR Workshop and Conference Proceedings*. Brookline, USA: Microtome Publishing, 2008, 4: 5 – 20.

[10] ENDRES D, FOLDIAK P. Bayesian bin distribution inference and mutual information [J]. *IEEE Transactions on Information Theory*, 2005, 51(11): 3766 – 3779.

[11] SHANNON C E. A mathematical theory of communication [J]. *Acm Sigmobile Mobile Computing and Communications Review*, 2001, 5(1): 3 – 55.

[12] ZENG X, DURRANI T S. Estimation of mutual information using copula density function [J]. *Electronics Letters*, 2011, 47(8): 493 – 494.

[13] NELSEN R B. *An Introduction to Copulas* [M]. 2nd Edition. New York: Springer, 2006.

[14] CHERUBINI U, LUCIANO E, VECCHIATO W. *Copula Methods in Finance* [M]. Chichester, England: John Wiley & Sons, 2004.

作者简介:

韩　敏　(1959–), 女, 教授, 博士生导师, 目前研究方向为神经网络、3S系统及混沌序列分析, E-mail: minhan@dlut.edu.cn;

刘晓欣　(1987–), 女, 硕士研究生, 目前研究方向为多变量时间序列相关性分析及变量选择, E-mail: xiaoxinliu@mail.dlut.edu.cn.