第 42 卷　第 2 期　　计　算　机　工　程　　2016 年 2 月
Vol. 42　No. 2　　Computer Engineering　　February 2016

·人工智能及识别技术·　　文章编号：1000-3428(2016)02-0175-09　　文献标识码：A　　中图分类号：TP301.6

基于形态特征与因果岭回归的股市态势预测算法

姚宏亮，马晓琴，王　浩，李俊照

（合肥工业大学计算机与信息学院，合肥 230009）

摘　要：基于股票波动典型的 M 形态，提出一种基于因果关系的岭回归股市态势预测算法。根据 M 形态的波动特征，引入能量思想，以 M 形态的边、波峰和波谷为结点，构建 M 形态的贝叶斯网络结构模型。利用马尔科夫毯算法和非对称信息熵，得到 M 形态的局部因果结构。采用因果强度的度量标准，将 M 形态因果关系引入到岭回归模型中，对股市态势进行预测。该模型通过将股票形成和能量波动的因果关系相结合，可以有效地发现股市的突变点。真实数据集上的实验结果表明，相比标准的岭回归算法和基于径向基的神经网络算法，该算法具有更好的预测效果。

关键词：贝叶斯网络；能量模型；因果分析；岭回归模型；预测算法

中文引用格式：姚宏亮，马晓琴，王　浩，等. 基于形态特征与因果岭回归的股市态势预测算法[J]. 计算机工程，2016，42(2)：175-183.

英文引用格式：Yao Hongliang，Ma Xiaoqin，Wang Hao，et al. Stock Market Trend Prediction Algorithm Based on Morphological Characteristics and Causal Ridge Regression[J]. Computer Engineering，2016，42(2)：175-183.

Stock Market Trend Prediction Algorithm Based on Morphological Characteristics and Causal Ridge Regression

YAO Hongliang，MA Xiaoqin，WANG Hao，LI Junzhao

（School of Computer and Information，Hefei University of Technology，Hefei 230009，China）

【Abstract】Based on the typical M form of the block volatility，this paper puts forward a ridge regression stock market trend prediction algorithm based on causality. Stock form reflects the stock fluctuations of energy change. According to the characteristics of the fluctuations in the form of M introducing energy ideas，based on edge，peaks and troughs in the form of M nodes，it builds a Bayesian network structure model in the form of M. By using Markov blanket algorithm and asymmetric information entropy，it gets a local causal structure in the form of M. The introduction of the strength of causal metrics is introduced to the M shape causality in ridge regression model for its stock market trend prediction of the model through stock form and causation of energy fluctuations，which can effectively find the abrupt change point of the stock market. Results on real data sets show that，compared with ridge regression algorithm and radial basis neural network algorithm，the proposed algorithm has better prediction effect.

【Key words】Bayesian network；energy model；causal analysis；ridge regression model；prediction algorithm

DOI：10.3969/j.issn.1000-3428.2016.02.032

1　概述

股票价格的走势虽然受到多种因素影响[1]，股票价格态势变化仍然具有内在规律和特点，股票价格的形态是各因素综合结果在态势上的表现，如常见的 M 形态。

现有的关于股票价格预测的方法主要采用的是一些传统的时间序列模型[2]，如常用的 AR，ARMA，ARIMA 模型等，可是这些模型的前提是股票时序数据呈现正太分布且趋于平稳的情况[3]，因而不能有效地对股票数据进行预测[4]。基于传统时间序列模型的不足，文献[4-5]对其进行改进，分别提出了 ARCH 模型和 GARCH 模型，都可以应用于股票时序数据处于非正态分布情况下的预测，然而，ARCH 和 GARCH 模型简单、单一，致使模型预测精度较低[6-7]。

股票的价格能体现出很多有价值的信息，若能

基金项目：国家自然科学基金资助项目(61175051，61070131，61175033)。

作者简介：姚宏亮(1972－)，男，副教授、博士、CCF 会员，主研方向为人工智能、知识工程；马晓琴，硕士；王　浩，教授、CCF 会员；李俊照，讲师、博士研究生。

收稿日期：2015-01-12　　**修回日期：**2015-03-13　　**E-mail：**dmicyhl@163.com

有效地分解、聚合和利用这些信息,则能够提高股票的总体预测精度[8]。文献[9]通过成交量构建股票市场的价格能量体系来预测股市后期走势,仅从价格出发,没有结合影响价格的因素以及价格包含的其他信息,预测模型具有局部单一性。股市的走势易受到短期的政策影响,尤其是股市的波动,但随着政策事件对股市冲击力的逐步减弱,股票市场逐步趋于成熟[10-11]。同时,由于股票市场具有自身的变化趋势,因此可以从股票走势具有的指数形态特征角度出发,进而更深入地研究股票市场内部的波动现象,并且充分利用股市价格具有的特征信息以及影响因素及形态特征具有的因果关系。

岭回归法是一种能统一诊断和处理多重共线性问题的特殊方法。将岭回归分析用于股票价格线性回归分析中,建立股票开盘价、最高价、最低价之间的岭回归方程,对长期股价预测有重要意义[12]。文献[13]提出一种基于径向基神经网络的非线性岭回归建模方法,该方法较标准的岭回归算法具有较好的稳定性和预测精度,但模型结构化比较薄弱。股市态势预测是从形态的局部结构出发,结合影响因素与形态特征进行预测,从而提出一种基于因果关系的岭回归算法。

本文提出的基于因果关系的岭回归算法从指数的 M 形态出发,对股市的形态走势进行预测。该算法从因果分析的角度,结合贝叶斯网络和岭回归模型对股市 M 形态的二峰的出现进行预测。首先根据顶点和与边特征,建立能量模型,并构建 M 形态的贝叶斯网络,得到目标结点的马尔科夫毯。根据结点变量的非对称信息熵以及边的熵因果关系,得到 M 形态边和结点的局部因果结构,进而通过因果强度将因果关系加入岭回归,构建因果关系的岭回归预测模型,从而对在指数的 M 形态的第一峰和谷底出现之后,通过对 M 形态的第二峰的能量值预测第二峰的出现。

2 能量模型

2.1 股票指数 M 形态描述

股市指数形态常出现如图 1 所示的特征形态,由于该特征的形成过程中有两峰且在两峰之间存在谷底,形如英文字母 M,故称为指数的 M 形态。

图 1 指数 M 形态

指数 M 形态的每条边上升或者下降的幅度,是由该边以及顶点的作用力决定的,如 M 形态的第一条边的支撑力度大,其上升的幅度就大,相应的峰顶就高;同样地,下降边的作用力越大,其相应的下跌幅度就大。这里将边和顶点的作用力称为边和顶点的能量,并定义出相应的能量计算模型来衡量边和顶点的作用力。

2.2 M 形态能量计算

根据指数 M 形态各顶点与边表现出来的特征,对案例进行分析并定义边和顶点的能量模型。

2.2.1 谷底顶点的能量定义

谷底顶点的能量定义如下:

lr 表示前面一条边的下跌幅度。

rd_{-30} 表示顶点实体最低点与 30 日线的相对距离。

rd_{-60} 表示顶点实体最低点与 60 日线的相对距离。

va_{ij} 表示第 i 个 j 连阴 K 线形态的平均成交量,如 va_{i3} 是第 i 个三连阴的平均成交量。

lr_a_{ij} 表示第 i 个 j 连阴线的平均涨跌幅,如 lr_a_{i3} 为第 i 三连阴的平均涨跌幅。

v_j 表示第 j 个巨量长阴实体的成交量。

lr_j 表示第 j 个巨量长阴实体的涨跌幅。

$$E = \partial_1 \cdot lr + \partial_2 \cdot rd_{-30} + \partial_3 \cdot rd_{-60} + \partial_4 \cdot \left(\sum_{i=1}^{n} va_i \cdot |lr\_a_{ij}| \right) + \partial_5 \cdot \left(\sum_{j=1}^{m} v_j \cdot |lr_j| \right) \tag{1}$$

其中,$lr = (C_h - C_l)/C_l \cdot 100\%$;$C_h$ 为区间最高收盘价;C_l 为区间最低收盘价。

$$rd_{-30} = \begin{cases} (c_{30} - o_i)/c_5 & \text{if } (o_i < c_i) \\ (c_{30} - c_i)/c_5 & \text{if } (o_i > c_i) \end{cases}$$

其中,c_{30} 为 30 日线的收盘价;o_i 为第 i 天的开盘价;c_i 为第 i 天的收盘价。

$$rd_{-60} = \begin{cases} (c_{60} - o_i)/c_5 & \text{if } (o_i < c_i) \\ (c_{60} - c_i)/c_5 & \text{if } (o_i > c_i) \end{cases}$$

其中,c_{60} 为 30 日线的收盘价;o_i 为第 i 天的开盘价;c_i 为第 i 天的收盘价。有:

$$va_{ij} = (v_1 + v_2 + \cdots + v_j)/j$$

$$lr\_a_{ij} = (lr_1 + lr_2 + \cdots + lr_j)/j$$

v_1, v_2, v_3 为三连阴三天的成量;lr_1, lr_2, lr_3 为三连阴三天的涨跌幅。有:

$$lr_j = (c_j - o_j)/o_j \cdot 100\%$$

2.2.2 峰顶顶点的能量定义

峰顶顶点的能量定义要素有:顶点实体的最高点与 30 日线的相对距离,顶点实体的最高点与60 日线的相对距离,放量滞涨情况。同时考虑到顶点的

形态各异,有尖和平缓之分,为了能量定义的一般化,将顶点的平均成交量作为顶点的能量定义要素。

峰顶顶点的初步能量定义形式化如下:

va 表示每个顶点的平均成交量。

rd_{-30}表示顶点实体最低点与 30 日线的相对距离。

rd_{-60}表示顶点实体最低点与 60 日线的相对距离。

lr_i 表示放量滞涨实体的涨幅。

$lr_{长阳}$表示前期对比的巨量长阳实体涨幅。

v 表示每天的成交量。

ps_d_i 表示放量滞涨 i 的程度。

$$E = \alpha_1 \cdot rd_{-30} + \alpha_2 \cdot rd_{-60} + \alpha_3 \cdot va + \alpha_4 \cdot \sum_{i=1}^{n}(ps\_d_i \cdot v_i) \tag{2}$$

其中,$va = (v_1 + v_2 + \cdots + v_t)/t$;$t$ 为顶点持续的交易日;$v_1, v_2, \cdots, v_t$ 为顶点持续交易日的成交量。

$$rd_{-30} = \begin{cases} (c_{30} - o_i)/c_5 & \text{if } (o_i < c_i) \\ (c_{30} - c_i)/c_{5i} & \text{if } (o_i > c_i) \end{cases}$$

其中,c_{30}为 30 日线的收盘价;o_i 为第 i 天的开盘价;c_i 为第 i 天的收盘价。

$$rd_{-60} = \begin{cases} (c_{60} - o_i)/c_5 & \text{if } (o_i < c_i) \\ (c_{60} - c_i)/c_5 & \text{if } (o_i > c_i) \end{cases}$$

其中,c_{60}为 30 日线的收盘价;o_i 为第 i 天的开盘价;c_i 为第 i 天的收盘价。

$$ps\_d_i = \frac{(lr_i/v_i)}{(lr_{长阳}/v_{长阳})}$$

2.2.3　上升边的能量定义

上升边的能量定义要素如下:

hr 表示边的上涨幅度。

va_{ij}表示第 i 个 j 连阳实体的平均成交量,如 va_{i3} 为第 i 个三连阳的平均成交量。

hr_a_{ij}表示第 i 个 j 连阳实体的平均涨跌幅,如 hr_a_{i3}为第 i 个三连阳的平均涨跌幅。

v_j 表示第 j 个巨量长阳实体的成交量。

hr_j 表示第 j 个巨量长阳实体的涨跌幅。

$$E = \beta_1 \cdot hr + \beta_2 \cdot \left(\sum_{i=1}^{n} va_i \cdot |\, hr\_a_{ij} \,|\right) + \beta_3 \cdot \left(\sum_{j=1}^{m} v_j \cdot |\, hr_j \,|\right) \tag{3}$$

其中,$hr = (C_h - C_l)/C_l \cdot 100\%$;$C_h$ 为区间的最高收盘价;C_l 为区间最低收盘价。

$va_{ij} = (v_1 + v_2 + \cdots + v_j)/j$($v_1, v_2, v_3$ 是三连阳 i 的三天成交量)。

$hr\_a_{ij} = (hr_1 + hr_2 + \cdots + hr_j)/j$($hr_1, hr_2, hr_3$ 分别是三连阳 i 的三天涨跌幅)。

$$hr_j = \begin{cases} (c_j - o_j) & \text{if } (c_j > o_j) \\ (o_j - c_j) & \text{if } (c_j < o_j) \end{cases}$$

其中,c_j 为长阳实体 j 的收盘价;o_j 为长阳实体的开盘价。

2.2.4　下降边的能量定义

下降边的能量定义要素有:边的下跌幅度,下跌过程中的缩量比,三连阴个数,长阴实体个数。

初步的能量形式化定义如下:

lr 表示边的下跌幅度。

fd 表示下跌过程的缩量比。

va_{ij}表示第 i 个 j 连阴的平均成交量。

hr_a_{ij}表示第 i 个 j 连阴的平均涨跌幅。

v_j 表示第 j 个巨量长阴实体的成交量。

hr_j 表示第 j 个巨量长阴实体的涨跌幅。

$$E = \varepsilon_1 \cdot lr + \varepsilon_2 \cdot fd + \varepsilon_3 \cdot \sum_{i=1}^{n} (va_i \cdot |lr\_a_{ij}|) + \varepsilon_4 \cdot \sum_{j=1}^{m} (v_i \cdot |lr_i|) \tag{4}$$

其中,$lr = (C_h - C_l)/C_l \cdot 100\%$;$C_h$ 为区间最高收盘价;C_l 为区间最低收盘价。

$va_{ij} = (v_1 + v_2 + \cdots + v_j)/j$

$lr\_a_{ij} = (lr_1 + lr_2 + \cdots + lr_j)/j$

其中,v_1, v_2, v_3 为三连阴三天的成交量;lr_1, lr_2, lr_3 为三连阴三天的涨跌幅。

$lr_j = (c_j - o_j)/o_j \cdot 100\%$

其中,c_j 为长阴 j 的收盘价;o_j 为长阴 j 的开盘价。

$$fd = \frac{(v_{min1} + v_{min2})}{(v_{max1} + v_{max2})} \times 100\% \tag{5}$$

其中,v_{min1}和 v_{min2}为下降区间中连续两天最小的成交量;v_{max1} 和 v_{max2} 为下降区间中连续两天最大的成交量)。

3　局部因果结构

3.1　马尔科夫毯

贝叶斯网络(Bayesian Network,BN)是一个有向无环图,可以用一个二元组表示 $B = (G, \theta)$,贝叶斯网络主要是用于体现结点变量之间的概率依赖关系。假设 $G = \langle V, E \rangle$ 表示一个有向无环图,其中 $V = \{X_1, X_2, \cdots, X_n\}$ 是随机变量集,而 E 表示结点集合 V 中任意结点之间的对应关系,即 E 是有向无环图中的有向边集。条件概率分布集为 $\theta = \{\theta_1, \theta_2, \cdots, \theta_n\}$,$\theta_i = P(X_i \mid Pa(X_i))$ 为结点 X_i 的概率分布,$Pa(X_i)$为 X_i 的父结点集合[14-16]。若有贝叶斯网络 $G = \langle V, E \rangle$ 与联合概率分布 $P(V)$,以及给定网络中任意一个结点的父结点,且满足该结点与它的非子孙结点独立,则可称 $\langle G, P \rangle$ 是满足因果马尔科夫条件[14-15]。

定义 1　假设给定贝叶斯网络 $G = \langle V, E \rangle$ 以及联合概率分布 $P(V)$,如果 G 中所体现的结点变量

条件独立性与 P 中表示的马尔科夫条件是一一对应的话,则可称 G 和 P 是 faithful。

定义 2 若对于给定的一个变量 $T \in V$ 以及变量子集 $S \subseteq U, T \notin S$,在变量 T 的马尔科夫毯(记作 $MB(T)$)给定的情况下,有 $I(S, T \mid MB(T))$ 存在,则称 G 是最小特征子集。

定理 1 在给定贝叶斯网络 $G = \langle V, E \rangle$ 与联合概率分布 $P(V)$ 中,如果图 G 和联合概率 $P(V)$ 是符合忠实性假设的,则对于图 $MB(T)$ 中的任意一个结点 T 的马尔科夫毯 $MB\{T\}$ 都存在且是唯一的。

定理 2 在满足忠实性假设条件的贝叶斯网络中,任意结点变量 T 的马尔科夫毯 $MB(T)$ 集中包含的是变量 T 的父结点、子结点以及子结点的父结点。

3.2 非对称信息熵确定的因果关系

在贝叶斯网络结构中,对于其中任意 2 个结点 A 和 B,若 $A \rightarrow B$(表示方向边从结点 A 到结点 B);$B \rightarrow A$(表示方向边是结点 B 到 A);$A \perp B$(表示结点 A 和结点 B 之间没有边)。则在可用数据集 D 以及领域知识 K 上,对于边的概率可以作如下定义:

$$\Pr(A \rightarrow B \mid D, K) = \sum_{A \rightarrow B \in E(G)} \Pr(G \mid D, K) \tag{6}$$

且有 $\Pr(G \mid D, K)$,则为在给定数据集 D 以及领域知识 K 下,网络结构 G 中出现的概率,结点 B 则为在网络 G 中不同于结点 A 的其他结点。

不对称信息熵[17-18]定义如下:

$$U_{NS}(A) = \sum_{B} ((-\Pr(A \rightarrow B) \cdot \log(\Pr(A \rightarrow B)) - (1 - \Pr(A \rightarrow B)) \cdot \log(1 - \Pr(A \rightarrow B))) \tag{7}$$

其中,A 为网络 G 中的任意一个结点,而结点 B 则是为了学习到网络结构不同于结点 A 的所有结点集。不对称信息熵 U_{NS} 考虑到结点变量 A 与其他结点变量间的 2 个状态,即从 A 到其他变量边的概率是否存在。而 U_s 则考虑的是结点变量 A 到其他结点变量的 3 种状态:$A \rightarrow B$,$A \leftarrow B$ 和 $A \perp B$。利用扰动[19]和非对称信息熵确定局部结构的边的存在性以及方向,即结点之间的因果关系。

3.3 局部网络结构的构建

算法描述如下:

输入 多个 M 形态各前 3 个顶点和前 3 条边对应的成交量,开盘价,收盘价,最高价,最低价数据集

输出 具有因果关系的局部结构

Step1 根据顶点与边的能量定义计算出 M 形态案例的边与顶点的能量值。

Step2 由 Step1 中的能量值,运用 K2 算法构建 M 结点的贝叶斯网络,并得到目标节点的马尔科夫毯局部结构。

Step3 根据非对称信息熵计算公式(式(7))计算得到局部结构中结点变量的非对称信息熵,判断边的存在性以及方向性。

Step4 根据边的熵计算公式(式(6))计算出局部结构中存在的边的熵。

Step5 由 Step3 与 Step4 判断因果关系的确定性,得到具有因果关系的局部结构。

图 2 是 M 形态结点的贝叶斯网络。图 3 是目标结点的马尔科夫毯。目标节点是 7(M 形态的第4 个顶点),节点 7 的马尔科夫毯是 1(M 形态的第1 个顶点)、2(M 形态的第 1 条边),3(M 形态的第 2 个顶点),4(M 形态的第 2 条边),5(M 形态的第 3 个顶点),6(M 形态的第 3 条边)。并在算法的 Step3 ~ Step5 得到的结点变量的非对称信息熵以及边的熵判定图 3 所示的目标节点的马尔科夫毯是具有因果关系的局部结构。实验结果证明,由结点变量的非对称信息熵以及边的熵值判定,目标结点的马尔科夫毯中的结点变量之间存在因果关系。

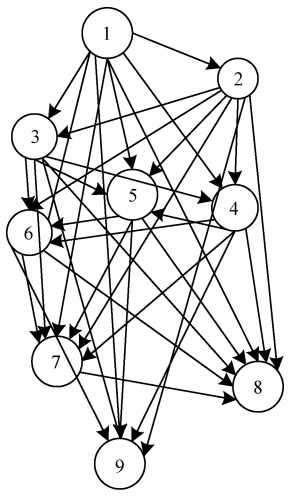

图 2 M 形态顶点与边之间的贝叶斯网络

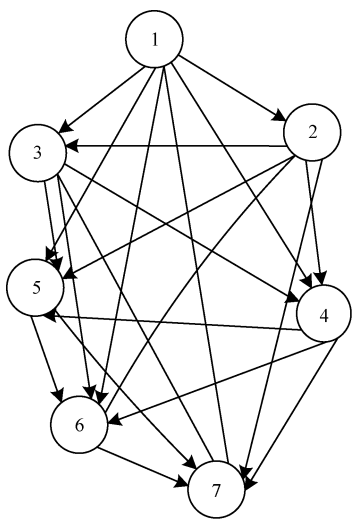

图 3 目标结点 7 的马尔科夫毯

4　基于结点能量的因果强度

可以用因果强度度量标准作为评价因果属性的规范标准[20-21]。

4.1　因果强度理论

因果关系的概率理论开始于概率对比:$\Delta P_c = P(e|c) - P(e|\neg c)$,这种情况下 c 只是一个表面原因。因果度理论主要起源于严格的因果结构需求,并规定效果 e 与候选原因 c 的协方差必须是独立于 e 与其他原因(令其为 a)的协方差。同时,c 的发生也必须是独立于 a 的,亦即 a 和 c 是独立发生的。假设 p_c 和 p_a 分别代表 c 和 a 对于 e 产生的因果强度,假设 e 是由 c 和 a 导致的结果,由此导出公式:

$$P(e) = P(c)P_c + P(a)P_a - P(c)P(a)P_cP_a \tag{8}$$

因果强度公式为:

$$P_c = \frac{-\Delta P_c}{P(e|\neg c)} \tag{9}$$

4.2　因果强度计算

指数 M 形态的顶点与边的能量大小主要体现在边的涨跌幅度以及顶点的位置,即 M 形态边的涨跌幅以及顶点的高低受到前面顶点与边的作用力的影响,即为顶点与边的能量。

从能量角度,考虑结点 i 对于结点 j 的影响程度作为结点 i 对结点 j 的因果作用强度。结合因果强度理论,从 M 形态结点的能量特征出发,定义指数 M 形态结点特有的具有能量思想的因果强度计算模型。设 $F(i\rightarrow j)$ 记为 P_{E_i},为结点 i 对结点 j 的因果强度,从结点变量的能量特征出发,E_j 为其他结点对第 j 个结点的效果,E_i 为候选原因,E_k 为与 E_i 独立的其他原因结点,且对结点 j 有因果关系。

设 P_{E_i} 和 P_{E_k} 分别为 E_i 和 E_k 产生 E_j 的因果强度,则有公式:

$$P(E_j) = P(E_i)\cdot P_{E_i} + P(E_k)\cdot P_{E_k} - P(E_i)\cdot P(E_k)\cdot P_{E_i}\cdot P_{E_k} \tag{10}$$

因果强度公式为:

$$F(i\rightarrow j) = P_{E_i} = \frac{-\Delta P_{E_i}}{P(E_j|\neg E_i)} \tag{11}$$

且 $F(i\rightarrow j)$ 是单向的,即 F 是非对称的。

选取上证指数和深证成指的 M 形态案例各 5 个,表 1 和表 2 时所选取的 M 形态案例的结点能量值与变化幅度。

表 1　上证指数 M 形态案例能量值与变化幅度

案例时间	第 1 个顶点	第 1 条边	第 2 个顶点	第 2 条边	第 3 个顶点	第 3 条边	第 4 个顶点	幅度
2011-08-09—2011-09-05	2 534.02	2 575.34	2 617.47	2 562.19	2 534.75	2 580.64	2 585.06	83.45
2009-09-30—2010-02-02	2 754.54	3 009.31	3 281.10	3 225.32	3 062.15	3 186.98	3 264.33	526.57
2008-11-03—2008-12-26	1 737.39	1 895.68	1 983.54	1 924.03	1 885.14	1 979.77	2 038.70	301.31
2007-07-06—2008-04-22	3 698.61	4 897.45	5 995.93	5 473.66	4 881.64	5 145.93	5 475.46	2 297.32
2005-06-06—2005-07-06	1 021.36	1 077.99	1 131.05	1 095.22	1 079.43	1 101.37	1 116.68	109.69

表 2　深证成指 M 形态案例能量值与变化幅度

案例时间	第 1 个顶点	第 1 条边	第 2 个顶点	第 2 条边	第 3 个顶点	第 3 条边	第 4 个顶点	幅度
2012-12-04—2013-06-25	7 806.22	9 072.62	9 892.51	9 147.70	8 704.79	9 114.28	9 370.81	2 086.29
2012-01-06—2012-07-31	8 617.33	10 427.24	10 116.39	9 410.26	9 410.26	10 612.89	10 524.68	1 795.56
2010-02-05—2010-04-20	11 944.98	12 271.10	12 491.50	12 303.96	11 926.05	12 367.19	12 735.35	809.30
2009-11-02—2009-11-23	12 715.65	13 930.28	13 980.28	12 647.51	12 813.29	13 699.97	13 533.54	1 332.77
2003-07-01—2003-08-25	3 232.69	3 336.74	3 443.70	3 313.25	3 276.50	3 326.94	3 352.81	120.12

图 4 和图 5 分别是选取上证指数和深证成指具有典型的 M 形态的各结点能量变化图,主要是为了说明前面结点的能量值对后面结点的影响以及整个过程中的能量变化幅度。从表 1 和表 2 的案例结点能量值分析得到:前面结点的能量值影响后面结点的能量值的大小,即前面结点的能量大,后面结点的能量值也大。从图 4 和图 5 中可以看出,案例的第一个结点的能量值大小,对该 M 形态形成过程中结点的能量值是有影响的,而且结点能量值大,后面结点的能量值变化大,并且整个 M 形态的形成过程中能量值变化幅度也很大。

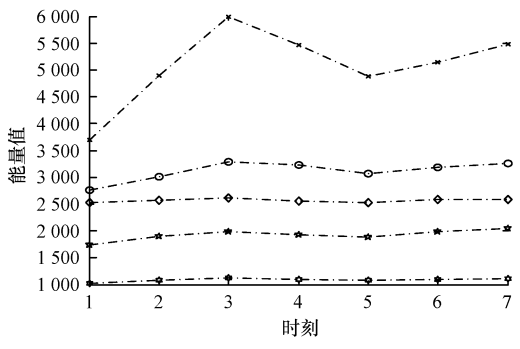

图 4　上证指数 M 形态结点能量变化

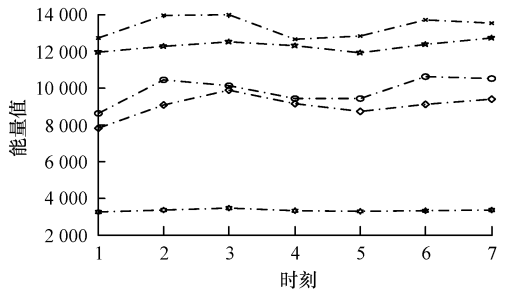

图 5 深证成指 M 形态结点能量变化

5 基于因果关系的岭回归算法

5.1 岭回归算法

岭回归算法[22-23]解决的是一组复共线性的变量之间的回归分析,体现的是变量与因变量之间的函数关系,是根据变量和因变量的案例数据学习得到变量的函数模型。通常的岭回归模型为:

$$Y_j=\beta_0+\beta_1X_{1j}+\beta_2X_{2j}+\cdots+\beta_mX_{mj}+\varepsilon_j \tag{12}$$

具有:

$$X=\begin{bmatrix}1 & X_{11} & X_{21} & \cdots & X_{m1}\\ 1 & X_{12} & X_{22} & \cdots & X_{m2}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & X_{1n} & X_{2n} & \cdots & X_{mn}\end{bmatrix}_{n\times(m+1)}$$

$$\beta=\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_m\end{bmatrix}_{(m+1)\times1}$$

$$Y=\begin{bmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{bmatrix}_{n\times1}$$

5.2 基于因果关系的岭回归算法计算步骤

基于因果关系的岭回归算法首先生成目标结点变量的马尔科夫毯,并组成具有较强依赖关系的局部结构。然后利用结点变量的非对称信息熵进行因果关系的发现,从而得到一个关于目标结点的局部因果结构。再利用结点之间因果强度计算局部因果结构变量之间的因果强度,并作为岭回归模型的重要参数,构建具有因果关系的岭回归能量模型,用于对目标节点的能量值预测,作为判别 M 形态第二峰出现的依据。

设变量 $\boldsymbol{X}=(x_1,x_2,x_3,x_4,x_5,x_6)$,其中存在 x_1,x_2,x_3,x_4,x_5,x_6 在图 4 中的结构关系。矩阵 $\boldsymbol{X}$ 为 M 形态已出现的 3 个顶点和 3 条边的能量值,x_1 为第 1 个顶点的能量值,x_2 为第 1 条边的能量值,x_3 为第 2 个顶点的能量值,x_4 为第 2 条边的能量值,x_5 为第 3 个顶的能量值,x_6 为第 3 条边的能量值。矩阵 $\boldsymbol{Y}$ 为已有 M 形态案例的第 4 个顶点的能量值向量,亦即要预测能量值对应的顶点即为图 4 中的目标结点 7。

$\boldsymbol{Y}$ 与 $\boldsymbol{X}$ 之间的原始岭回归回归模型为 $\boldsymbol{Y}=\boldsymbol{X}\times\boldsymbol{B}$,$\boldsymbol{B}$ 为回归模型的系数矩阵,但是由于变量 $\boldsymbol{X}$ 的各向量之间存在明显的局部因果结构关系。为了能使顶点与边之间具有因果关系的能量作为一个整体融入到受前面边和顶点影响的第 4 个顶点(即为 M 形态的第二峰)的能量预测模型中,选用结点变量之间的因果关系强度,如:$P(x_i\rightarrow x_j)$ 表示变量 x_i 对 x_j 的因果关系强度。

将因果关系强度融入到岭回归算法模型中,得到具有因果关系强度的岭回归算法模型为:

$$\begin{aligned}\boldsymbol{Y}=&\beta_0+\beta_1\times x_1+[\beta_2+P(x_1\rightarrow x_2)]\times x_2\\&+[\beta_3+P(x_1\rightarrow x_2)+P(x_2\rightarrow x_3)]\times x_3\\&+\cdots+[\beta_m+\sum_{i=1}^{m-1}P(x_i\rightarrow x_m)]\times x_m\end{aligned} \tag{13}$$

其中:

$$\boldsymbol{X}=\begin{bmatrix}1 & X_{11} & X_{21} & \cdots & X_{m1}\\ 1 & X_{12} & X_{22} & \cdots & X_{m2}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & X_{1n} & X_{2n} & \cdots & X_{mn}\end{bmatrix}_{n\times(m+1)}$$

$$\boldsymbol{\beta}=\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\times[1+P(x_1\rightarrow x_2)]\\ \vdots\\ \beta_m\times[1+\sum_{i=1}^{m-1}P(x_i\rightarrow x_m)]\end{bmatrix}_{(m+1)\times1}$$

$$\boldsymbol{Y}=\begin{bmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{bmatrix}_{n\times1}$$

该模型中回归系数 $\boldsymbol{\beta}$ 的最小平方估计为:

$$b=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}=(b_0,b_1,\cdots,b_m)' \tag{14}$$

岭回归分析首先要对 $\boldsymbol{X}$ 变数作中心化与标量化处理,为使得处于同样数量级上的不同自变数便于比较,这就是引入新变数 Z,令:

$$Z_{ij}=(X_{ij}-\bar{x}_i)/\sqrt{\sum x_i^2} \tag{15}$$

其中,$i=1,2,\cdots,m$;$j=1,2,\cdots,n$。

于是式(12)变为:

$$\begin{aligned}Y_j-\bar{y}=&\beta_1^ZZ_{1j}+\{\beta_2\times[1+P(x_1\rightarrow x_2)]Z\}Z_{2j}\\&+\cdots+\{\beta_m\times[1+\sum_{i=1}^{m-1}P(x_i\rightarrow x_m)]\}^ZZ_{mj}\\&+\varepsilon_j\end{aligned} \tag{16}$$

进一步有:

$$\boldsymbol{Z}=\begin{bmatrix} Z_{11} & Z_{21} & \cdots & Z_{m1} \\ Z_{12} & Z_{22} & \cdots & Z_{m2} \\ \vdots & \vdots & & \vdots \\ Z_{1n} & Z_{2n} & \cdots & Z_{mn} \end{bmatrix}_{n\times m}$$

$$\boldsymbol{\beta}^{Z}=\begin{bmatrix} \beta_1^{Z} \\ \{\beta_2\times[1+P(x_1\rightarrow x_2)]\}^{Z} \\ \vdots \\ \{\beta_m\times[1+\sum_{i=1}^{m-1}P(x_i\rightarrow x_m)]\}^{Z} \end{bmatrix}_{m\times 1}$$

$$(\boldsymbol{Y}-\bar{y}\,\boldsymbol{I}_n)=\begin{bmatrix} Y_1-\bar{y} \\ Y_2-\bar{y} \\ \vdots \\ Y_n-\bar{y} \end{bmatrix}_{n\times 1}$$

其中,$\boldsymbol{\beta}^{Z}$表示回归系数,$\boldsymbol{\beta}$ 是 $\boldsymbol{Z}$ 变数估计,统计上又称它们为标准化回归系数。由于 $\boldsymbol{Z}'\boldsymbol{I}_n=0$,$\boldsymbol{\beta}^{Z}$的最小平方估计为:

$$\begin{aligned} b^{Z}&=(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{Y}-\bar{y}\boldsymbol{I}_n) \\ &=(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\ &=(b_1{}^{Z},b_2{}^{Z},\cdots,b_m{}^{Z})' \end{aligned} \tag{17}$$

因而在实际分析中,因变数仍可用观察值向量 $\boldsymbol{Y}$ 而不用中心化向量($\boldsymbol{Y}-\bar{y}\boldsymbol{I}$),只要最后在回归方程中记:

$$\bar{y}=b_0{}^{Z} \tag{18}$$

且 b_i 和 $\beta_i{}^{Z}$具有关系:

$$b_i=b_i{}^{Z}/\sqrt{\sum x_i^2},b_0=\bar{y}-\sum_1^m b_i\bar{x}_i \tag{19}$$

5.3　基于因果关系的岭回归算法描述

算法描述如下:

输入　M 形态对应的数据集 Dataset

输出　M 形态第二峰值的能量值

Step1　对 M 形态的边和点的特征进行分析,得出顶点与边各特征因素之间的权重关系。

Step2　根据特征分析的结果,定义边和顶点的能量 E。

Step3　根据式(1)~式(4),计算得到案例的边和顶点的能量值,离散化之后,构建相应的 BN。

Step4　从 Step3 中的 BN 中,得到目标结点 i 的 $MB(i)$,通过计算结点的非对称信息熵 U_{NS}以及扰动学习[15],得到局部因果结构 G。

Step5　根据 Step4 中的 G,Step2 中得到的能量的定义,对 Dataset 案例数据进行处理,运用式(10)得到结点的因果强度 P_{E_i}。

Step6　将结点的因果关系强度 P_{E_i}以及顶点与边的能量数据作为式(13)的输入,得到具有因果关系的岭回归顶点与边的能量模型,对第二峰的值进行预测。

Step7　预测结果评价,分别在 Data1 和 Data2 上进行预测及评价。

6　实验数据处理

6.1　实验数据与环境

本文的实验环境是 32 位 Windows 7 Ultimate, Pentium(R) Dual-Core PCU E6700 @ 3.20 GHz, RAM 2.00 GB, Matlab7.11.0。实验中所用的工具包为贝叶斯 FullBNT-1.0.4。实验数据来自国元领航软件下载的上证指数与深证成指的 M 形态案例边和顶点对应的相关数据 Data1 与 Data2。

6.2　预测模型的建立

用数据 Data1 建立基于因果关系的岭回归模型为:

$$\begin{aligned} y=&9.989\,0+0.098\,3x_1-0.119\,9x_2 \\ &+0.039\,7x_3+0.448\,8x_4 \\ &-1.393\,8x_5+1.904\,6x_6 \end{aligned} \tag{20}$$

用数据 Data2 建立的基于因果关系的岭回归模型为:

$$\begin{aligned} y=&-11.689\,1-0.020\,8x_1+0.020\,7x_2 \\ &-0.005\,8x_3+0.879\,5x_4 \\ &-0.757\,6x_5+0.888\,2x_6 \end{aligned} \tag{21}$$

6.3　对比实验算法

关于对比实验算法部分使用 2 个对比算法,一个使用原始的岭回归算法,即没有加入因果关系强度的岭回归算法,采用相同的数据集作为算法的输入向量;另一个使用基于径向基的神经网络岭回归算法模型。

6.4　评价标准

平均相对误差:

$$r_{\mathrm{RAE}}=\frac{\sum_{i=1}^{n}\left|\frac{y-y'}{y}\right|}{n} \tag{22}$$

其中,n 表示预测集中的整体样本总个数;y 是实际的真实值;y'为实验预测值;r_{RAE}可以用来表示预测值偏离实际值的大小,它的值越小表明偏离度越小,说明预测结果的精确度越高。

7　基于因果关系的岭回归模型预测结果分析

7.1　实验数据选择

大盘 M 形态数据 Data1,建立模型数据 2002-11-27—2010-11-27,预测数据 2010-12-01—2014-11-20。

深证成指 M 形态数据 Data2,建立模型数据 2003-07-01—2009-10-31,预测数据 2009-11-02—2014-11-20。

7.2 实验对比

基于因果关系的岭回归(Ridge Regression based on Causality, RRC)算法与标准的岭回归(Ridge Regression,RR)算法比较如图 6、图 7 所示。基于因果关系的岭回归(RRC)算法与基于径向基神经网络的岭回归(Ridge Regression Based on Radial Basis Function neural network,RR-RBF)算法比较如图 8、图 9 所示。

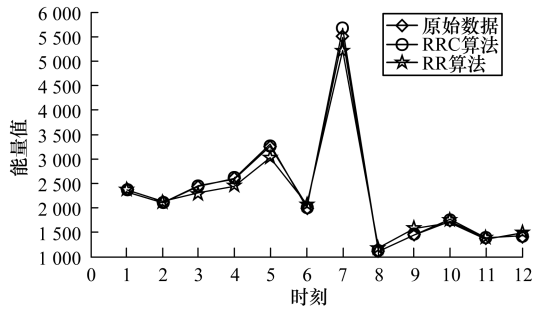

图 6 数据 Data1 上的结果对比 1

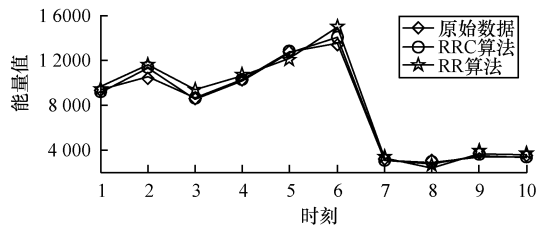

图 7 数据 Data2 上的结果对比 1

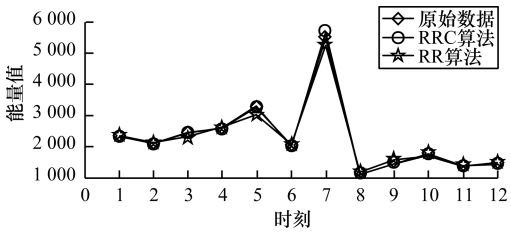

图 8 数据 Data1 上的结果对比 2

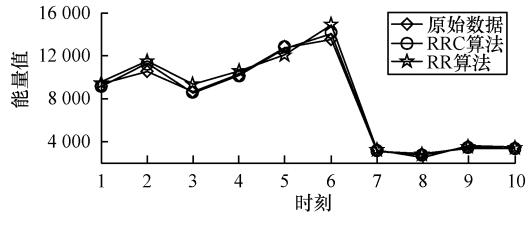

图 9 数据 Data2 上的结果对比 2

选取案例的原始数据的真实值,分别与基于径向基神经网络的岭回归模型、标准岭回归算法和基于因果关系的岭回归算法的预测结果值进行对比分析,结果如表 3 ~ 表 5 所示。

表 3 数据 Data1 的预测结果

序号	原始数据	基于径向基神经网络的岭回归算法	基于因果关系的岭回归算法	标准岭回归算法
1	2 315.92	2 354.2	2 330.1	2 364.90
2	2 105.73	2 127.59	2 098.6	2 130.59
3	2 450.95	2 319.83	2 433.0	2 309.83
4	2 585.06	2 633.70	2 599.0	2 433.70
5	3 264.33	3 026.53	3 305.7	3 029.96
6	2 038.70	2 078.90	2 002.3	2 100.30
7	5 475.46	5 254.38	5 664.3	5 214.38
8	1 116.68	1 186.70	1 106.5	1 179.36
9	1 447.99	1 573.70	1 441.3	1 593.20
10	1 770.28	1 698.80	1 726.9	1 679.31
11	1 386.65	1 365.70	1 384.5	1 355.71
12	1 427.51	1 487.90	1 432.1	1 492.90

表 4 数据 Data2 的预测结果

序号	原始数据	基于径向基神经网络的岭回归算法	基于因果关系的岭回归算法	标准岭回归算法
1	9 370.81	9 587.94	9 135	9 687.94
2	10 524.68	11 537.60	11 329	11 593.60
3	8 667.22	8 357.90	8 524	9 375.90
4	10 284.24	10 576.90	10 178	10 625.90
5	12 735.35	12 073.90	12 663	12 173.90
6	13 533.54	14 876.50	14 053	14 950.50
7	3 066.91	3 205.50	3 071	3 305.50
8	2 892.61	2 501.70	2 776	2 401.90
9	3 366.04	3 621.90	3 503	3 651.00
10	3 352.81	3 487.90	3 321	3 587.90

表 5 算法预测结果的平均相对误差

算法	Data1	Data2
基于径向基神经网络的岭回归算法	0.036 7	0.062 5
基于因果关系的岭回归算法	0.010 9	0.026 4
标准岭回归算法	0.047 9	0.081 0

7.3 结果分析

由实验结果以及相对误差的比较来看,基于因果关系的岭回归算法在数据集 Data1 与 Data2 上的预测效果比基于径向基神经网络的岭回归算法以及标准的岭回归算法要好,尤其是在高点处的效果更明显。在数据集 Data1 的实验对比图上,在第 3,5,7,9,12 时刻处,基于因果关系的岭回归算法的预测效果明显好于基于径向基神经网络的岭回归算法和标准岭回归算法,其他点两种算法的预测结果与实际值相比较都差不多。在数据集 Data2 的实验对比

图上,第 2,3,4,5,6,8 时刻处,基于因果关系的岭回归算法的预测效果明显好于基于径向基神经网络的岭回归算法和标准岭回归算法,另外点处两种算法的预测结果与实际值的偏差都很小。基于因果关系的岭回归算法预测值与实际值的误差远小于标准岭回归算法以及基于径向基神经网络的误差值。总体上来看,基于因果关系的岭回归算法的预测效果好于基于径向基神经网络的岭回归算法和标准岭回归算法的预测效果。

8　结束语

本文提出一种基于因果关系的岭回归算法,并根据股市 K 线走势表现的显著特征 M 形态作为研究对象,分析 M 形态的点与边表现出来的特征因素,然后构建顶点与边之间的贝叶斯网络,从中学习得到 M 形态边和顶点之间的因果关系,根据那些体现因果关系的因素定义出谷底顶点、顶峰顶点、上升边、下降边的能量。然后将 M 形态的前 3 条边和前 4 个顶点的能量作为岭回归模型的输入,构建 M 形态的前面 3 条边和 4 个顶点能量之间的因果岭回归模型,从而对 M 形态的第二峰的能量值进行预测,进而可以根据预测的能量值预测 M 形态的第二峰出现的大概位置。从实验图形中可以看出,该算法对股市中的 M 形态第二峰的能量值预测与实际值比较接近,但由于中国经济政策对股市的走势影响比较大,致使有些突变点难以预测,因此预测的结果和实际还是有偏差的,但是从 K 线的形态出发以及整体的实验结果来看,本文算法的预测效果高于前期提出的股市预测算法的效果。根据实验分析,对比其他算法,RRC 算法有更好的预测效果,后期的工作还将从形态出发,结合弹性系统和 M 形态的能量演化,在预测整个 M 形态的形成过程时更进一步提高预测精度。

参考文献

[1] Hadavandi E, Shavandi H, Ghanbari A. Integration of Genetic Fuzzy Systems and Artificial Neural Networks for Stock Price Forecasting [J]. Knowledge-based Systems, 2010, 23(8): 800-808.

[2] 姚宏亮,杜明超,李俊照,等. 一种基于流特征模式的股市跟踪预测算法[J]. 计算机科学, 2013, 40(12): 45-51.

[3] Kazem A, Sharifi E, Hussain F K, et al. Support Vector Regression with Chaos-based Firefly Algorithm for Stock Market Price Forecasting[J]. Applied Soft Computing, 2013, 13(2): 947-958.

[4] Yi Zuo, Kita E. Stock Price Forecast Using Bayesian Network[J]. Expert Systems with Applications, 2012, 39(8): 6729-6737.

[5] Engle R F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation[J]. Econometrica, 1981, 50(4): 987-1008.

[6] Bollerslev T. Generalized Autoregressive Conditional Heteroskedasticity[J]. Journal of Econometrics, 1986, 31(3): 307-327.

[7] Wang Jujie, Wang Jianzhou, Zhang Zhe, et al. Stock Index Forecasting Based on a Hybrid Model [J]. Omega, 2012, 40(6): 758-766.

[8] Zhang Tao, Sai Ying, Zheng Yuan. Research of Stock Index Futures Prediction Model Based on Rough Set and Support Vector Machine [C]//Proceedings of IEEE International Conference on Granular Computing. Washington D. C., USA: IEEE Press, 2008: 797-800.

[9] 孙　彬,王立民,李铁克. 股票市场能量体系研究[J]. 技术经济与管理研究, 2010, (6): 3-8.

[10] 王　浩,陈　娟,姚宏亮,等. 基于离群特征模式的股市波动预测模型[J]. 计算机工程与应用, 2014, 50(22): 243-249.

[11] 徐君华,李启亚. 宏观政策对我国股市影响的实证研究[J]. 经济研究, 2001, (9): 12-21.

[12] 张谊然. 股票价格线性回归分析——基于 Matlab 岭回归分析[J]. 时代金融, 2013, (3): 198.

[13] 周　峰,孟秀云. 基于岭回归径向基神经网络的 MIMU 误差建模[J]. 系统仿真学报, 2010, 22(9): 2056-2060.

[14] Daly R, Shen Q, Aitken S. Learning Bayesian Networks: Approaches and Issues [J]. Knowledge Engineering Review, 2011, 26(2): 99-157.

[15] Cheng Jie, Bell D, Liu Weiru. Learning Bayesian Networks from Data: An Efficient Approach Based on Information Theory [J]. Artificial Intelligence, 2002, 37(12): 43-90.

[16] 周冬梅,王　浩,姚宏亮,等. 一种基于因果强度的局部因果结构主动学习方法[J]. 计算机科学, 2012, 39(11): 237-242.

[17] Tong S, Koller D. Active Learning for Structure in Bayesian Networks [C]//Proceedings of International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2001: 863-869.

[18] 王　浩,刘向南,姚宏亮,等. 基于扰动和因果强度的因果贝叶斯网络结构的主动学习研究[D]. 合肥: 合肥工业大学, 2012.

[19] 姚宏亮,吴立辉,王　浩,等. 基于局部因果关系分析的隐变量发现算法. 计算机科学与探索[J]. 2014, 8(4): 456-466.

[20] Aliferis C F, Statnikov A, Tsamardinos I, et al. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation [J]. Journal of Machine Learning Research, 2010, 11(1): 171-234.

[21] Li Guoliang, Leong T Y. Active Learning for Causal Bayesian Network Structure with Non-symmetrical Entropy [C]//Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Germany: Springer, 2009: 290-301.

[22] Cai C X, Kyaw K, Zhang Qi. Stock Index Return Forecasting: The Information of the Constituents [J]. Economics Letters, 2012, 116(1): 72-74.

[23] Deng Wen-shuenn, Chu Chih-kang, Cheng Ming-yen. A Study of Local Linear Ridge Regression Estimators[J]. Journal of Statistics Planning and Inference, 2001, 93(1-2): 225-238.

编辑　顾逸斐