doi:10. 16055/j. issn. 1672-058X. 2022. 0002. 012

基于变系数乘积模型的股指跟踪研究

万 学
(重庆师范大学 数学科学学院, 重庆 401331)

摘 要:随着经济的发展,股票投资进入大众视野,如何选择成分股对股票指数进行跟踪,越来越受到人们的关注,基于此,针对股票指数跟踪问题,提出了利用变系数乘积模型进行变量选择的一种方法。该方法基于 B 样条函数逼近技术,将 LPRE 准则和组 SCAD 惩罚函数结合起来,应用于变系数乘积模型,利用牛顿迭代法和局部二次近似给出了求解估计的实施步骤;为了验证所提方法的有效性,通过数值模拟,将变系数乘积模型 SCAD 惩罚方法(LPRE-S)与变系数模型最小二乘 SCAD 惩罚方法(LS-S)的结果进行了对比,为了验证所提方法的实用性,将 LPRE-S 估计方法与 LS-S 估计方法应用于深证红利指数,对其股指跟踪预测效果进行了比较;结果表明:LPRE-S 估计方法选出真实模型的比率几乎接近 1,能更好地达到变量选择的目的,且在股指跟踪中具有较好的预测效果。

关键词:变系数乘积模型;变量选择;LPRE 估计;SCAD
中图分类号:O212. 7 **文献标志码**:A **文章编号**:1672-058X(2022)02-0083-07

0 引 言

随着我国经济的不断发展,股票投资进入大众视野,成为最热门的投资方式之一。如何选择成分股对股票指数进行跟踪,越来越受到人们的关注。追踪股票指数指以某一股票指数为目标,以该指数的成分股为投资对象,通过购买该成分股所构建的投资组合,用于追踪目标指数的表现。人们感兴趣的是如何用更少的投资来获得更大的回报,这启发了学者们探索如何选择较少的成分股达到跟踪股票指数的目的。

在统计学中,选择较少的成分股追踪股票指数,称为变量选择问题。对于变量选择的方法,国内外已有许多学者对此进行了全面而深入的研究,其中 Tibshirani[1] 在 1996 年提出了一种基于压缩系数的 Lasso(Least Absolute Shrinkage and Selection Operator)方法,克服了传统的逐步回归法、最优子集选择法等方法的不足,为变量选择领域的发展做出了十分重要的贡献。但是,Lasso 方法在很大程度上压缩了变量的系数,致使模型偏差较大,且不具有 Oracle 性质。为了改善这些不足,Fan 等[2] 提出了能同时选出显著变量和得出相应参数估计的 SCAD(Smoothly Clipped Absolute Devation)方法,并在线性模型中证明了该方法的 Oracle 性质;Zou[3] 对不同的系数施加不同的权重进行压缩,提出了 Adaptive Lasso 方法,在一定程度上克服了 Lasso 方法的不足。但是,Adaptive Lasso 方法对于处理具有组效应的数据仍然不理想。为了处理具有组效应的数据,Zou 和 Hastie[4] 提出了 Elastic net 方法,但是该方法不具有 Oracle 性质;为此,Zou 和 Zhang[5] 受 Adaptive Lasso 方法的启发提出了另一种具有 Oracle 性质的方法,即 Adaptive Elastic net 方

收稿日期:2021-05-18;**修回日期**:2021-06-28.
基金项目:国家社会科学基金(17CTJ015);重庆市基础科学与前沿研究技术专项项目(CSTC2018JCYJAX0659).
作者简介:万学(1996—),女,重庆綦江人,硕士研究生,从事统计应用研究.

法。这些选择重要变量的方法已经被研究得相对成熟了,并且被学者们应用于各个领域。

在统计分析中,经常会遇到一些非负数据,例如股票价格、患者的寿命、生存时间等。处理这类数据,通常会考虑如下乘积模型:

$$\boldsymbol{Y}_i = \exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})\varepsilon_i, i = 1,2,\cdots,n \tag{1}$$

其中,$\boldsymbol{X}_i$ 是 p 维协变量,$\boldsymbol{Y}_i$ 是响应变量,$\boldsymbol{\beta}$ 是未知参数向量,ε_i 是严格非负的随机误差。

对于模型式(1)的估计方法,Chen 等[6]基于相对误差思想,提出了最小绝对相对误差(Least Absolute Relative Errors, LARE)准则:

$$LARE(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\{\left|\frac{\boldsymbol{Y}_i - \exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})}{\boldsymbol{Y}_i}\right| + \left|\frac{\boldsymbol{Y}_i - \exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})}{\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})}\right|\right\} \tag{2}$$

通过最小化目标函数式(2)可获得模型式(1)的参数估计。张丹[7]将 LARE 准则和文献[1-3]中提到的变量选择方法结合起来,讨论了模型式(1)的变量选择问题,并对相应的 Oracle 性质进行了证明。虽然 Chen 等[6]提出的 LARE 准则在一定条件下能得到具有相合性和渐近正态性的参数估计,但是 LARE 准则的目标函数式(2)并不光滑,且计算十分复杂,为了克服这些不足,Chen 等[8]考虑将目标函数式(2)中两种相对误差相乘提出了最小乘积相对误差(Least Product Relative Error, LPRE)准则,即最小化以下目标函数:

$$LPRE(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\{\left|\frac{\boldsymbol{Y}_i - \exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})}{\boldsymbol{Y}_i}\right| \times \left|\frac{\boldsymbol{Y}_i - \exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})}{\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})}\right|\right\} = \sum_{i=1}^{n}\{\boldsymbol{Y}_i\exp(-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{Y}_i^{-1}\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}) - 2\} \tag{3}$$

从目标函数式(3)可以看出,相比于 LARE 准则,LPRE 准则的目标函数具有无限可微且严格凸的优点,这使得该目标函数具有唯一的最小值点。Chen 等[8]也通过数值模拟和实例应用证明了在一定条件下 LPRE 估计方法比 LARE 估计方法更有效;李翠平[9]基于 LPRE 准则,通过 Adaptive LASSO, Adaptive Elastic Net,以及 SCAD 方法研究了模型式(1)的变量选择问题,并对相应的 Oracle 性质进行了证明;陈银钧等[10]将 LPRE 准则和 LASSO 方法结合起来研究了模型式(1)的变量选择问题。基于 LARE 和 LPRE 准则,已有许多学者研究了线性乘积模型。但是,仅使用这个模型不能完全反应实际应用中变量之间复杂的潜在关系。胡大海[11]在 LPRE 准则的基础上,研究了变系数乘积模型的非参函数估计问题。

近年来,乘积模型变量选择问题得到了广泛关注,但是对于变系数乘积模型的变量选择问题的研究还鲜少出现。因此,本文将在已有文献的基础上,将 LPRE 和 SCAD 方法应用于变系数乘积模型,研究该模型的变量选择问题,并通过模拟仿真证明所提方法的有效性;最后,利用模拟中的方法追踪深证红利指数,证明所提方法的实用性。

1 模型及方法介绍

1.1 变系数乘积模型简介

当假定参数模型成立时,模型式(1)具有较高的推断精度,且具有容易解释的优点,但是在实际应用中,学者们并不能确定数据服从怎样的模型,如果假定的参数模型与实际情况不相符,对于给定参数模型的估计和统计推断就几乎没有意义。此外,模型式(1)通常是假定 $\log \boldsymbol{Y}$ 与 $\boldsymbol{X}$ 之间呈线性关系,但是有时候这个假定是不成立的。为此,本文考虑适应性更强的变系数乘积模型:

$$\boldsymbol{Y}_i = \exp[\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}(U_i)]\varepsilon_i \tag{4}$$

其中,$\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot),\cdots,\beta_p(\cdot))^{\mathrm{T}}$ 是 $p \times 1$ 维未知函数系数向量,指标变量 $U_i \in [0,1]$,$\boldsymbol{X}_i$ 是协变量,$\boldsymbol{Y}_i$ 是响应变量,ε_i 是严格非负随机误差。

对模型式(4)作对数变换,可将其转换为一般的变系数模型:

$$\boldsymbol{Y}_i^* = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}(U_i) + \varepsilon_i^* \tag{5}$$

其中,$\boldsymbol{Y}_i^* = \log \boldsymbol{Y}_i$,$\varepsilon_i^* = \log \varepsilon_i$。模型式(5)是由 Hastie 和 Tibshirani[12]提出的一类半参数回归模型,它既保持了响应变量 $\boldsymbol{Y}^*$ 与协变量 $\boldsymbol{X}$ 之间的线性关系,又增加了模型的灵活性。

对模型式(4)进行估计,最直接的方法就是将其转换为模型式(5),再利用最小二乘法对其进行

估计,但是最小二乘法具有不稳健的缺点。同样地,对模型式(4)中的响应变量 $\boldsymbol{Y}$ 进行预测时,可以先对模型式(5)中的 $\boldsymbol{Y}^*$ 进行预测,再通过指数变换得到 $\boldsymbol{Y}$ 的预测值,但是在这个估计和预测的过程中始终考虑的是绝对误差,而在实际应用中,对于正响应变量,更多的是关注相对误差而不是绝对误差。因此,本文基于相对误差思想,将 Chen 等[8] 提出的 LPRE 准则应用于变系数乘积模型式(4)。

1.2　估计方法

鉴于 B 样条基函数具有良好的理论性质,类似吕晶[13],本文利用 B 样条基函数去逼近模型式(4)中的未知函数系数 $\boldsymbol{\beta}(\cdot)$。

令 $\boldsymbol{B}(\boldsymbol{u})=(B_1(u),\cdots,B_{K_n}(u))^{\mathrm{T}}$ 为 B 样条基函数,则函数系数 $\beta_j(\cdot)$ 可逼近为如下形式:

$$\beta_j(u)\approx\sum_{l=1}^{K_n}\gamma_{jl}B_{jl}(u),j=1,2,\cdots,p \tag{6}$$

其中,$\boldsymbol{\gamma}_j=(\gamma_{j1},\cdots,\gamma_{jK_n})^{\mathrm{T}}$ 为 B 样条系数向量,$K_n=J+m+1$ 为基函数的个数,J 为内节点的个数,m 为样条的阶,$\{B_k(\cdot),k=1,2,\cdots,K_n\}$ 是线性空间 G_j 的一组基,其中 G_j 由 $[0,1]$ 区间上$(m+1)$阶的 B 样条函数构成。基于函数系数 $\beta_j(\cdot)$ 的近似形式式(6),模型式(4)可表示为如下形式:

$$\boldsymbol{Y}_i\approx\exp\Big(\sum_{j=1}^{p}\sum_{l=1}^{K_n}X_i^{(j)}\gamma_{jl}B_{jl}(U_i)\Big)\varepsilon_i \tag{7}$$

令 $\boldsymbol{\Pi}_i=(X_i^{(1)}B_{11}(U_i),\cdots,X_i^{(1)}B_{1K_n}(U_i),\cdots,X_i^{(p)}B_{p1}(U_i),\cdots,X_i^{(p)}B_{pK_n}(U_i))^{\mathrm{T}}$,则 $\boldsymbol{\Pi}_i^{\mathrm{T}}\boldsymbol{\gamma}=\sum_{j=1}^{p}\sum_{l=1}^{K_n}X_i^{(j)}\gamma_{jl}B_{jl}(U_i)$,其中,$\boldsymbol{\gamma}=(\boldsymbol{\gamma}_1^{\mathrm{T}},\cdots,\boldsymbol{\gamma}_p^{\mathrm{T}})^{\mathrm{T}}$ 是未知样条系数向量。于是模型式(7)可以改写为如下形式:

$$\boldsymbol{Y}_i\approx\exp(\boldsymbol{\Pi}_i^{\mathrm{T}}\boldsymbol{\gamma})\varepsilon_i \tag{8}$$

基于 LPRE 准则,可以通过最小化以下目标函数获得模型式(8)中 $\boldsymbol{\gamma}$ 的估计 $\tilde{\boldsymbol{\gamma}}$,即 $\tilde{\boldsymbol{\gamma}}=\mathrm{argmin}(l(\boldsymbol{\gamma}))$。

$$l(\boldsymbol{\gamma})=\sum_{i=1}^{n}\{\boldsymbol{Y}_i\exp(-\boldsymbol{\Pi}_i^{\mathrm{T}}\boldsymbol{\gamma})+\boldsymbol{Y}_i^{-1}\exp(\boldsymbol{\Pi}_i^{\mathrm{T}}\boldsymbol{\gamma})-2\} \tag{9}$$

由此,求解模型式(4)中未知函数系数的估计就转化为求解模型式(8)中参数向量 $\boldsymbol{\gamma}$ 的估计。

为了选出模型式(4)中的重要变量,需要将不重要变量的系数压缩为 0。由于 Fan 等[2] 提出的 SCAD 惩罚函数具有将较小系数压缩为 0,对较大系数不进行压缩,能使模型偏差更小的优点,且该惩罚函数是一个凸函数,能够得到全局最优解,在优化时不会陷入局部最优解,因此,本文将 SCAD 惩罚函数应用于变系数乘积模型。

令 $p_{\lambda_n}(\cdot)$ 为 SCAD 惩罚函数,其一阶导数定义为如下形式:

$$p'_{\lambda_n}(\theta)=\lambda_n\left\{I(\theta\leqslant\lambda_n)+\frac{(a\lambda_n-\theta)_+}{(a-1)\lambda_n}I(\theta>\lambda_n)\right\}$$

其中,$a>2$,$\theta>0$,λ_n 为调整参数。为此,本文考虑以下惩罚目标函数:

$$Q(\boldsymbol{\gamma})=l(\boldsymbol{\gamma})+n\sum_{j=1}^{p}p_{\lambda_n}(\|\boldsymbol{\gamma}_j\|_2) \tag{10}$$

通过最小化目标函数式(10)即可获得模型(8)的惩罚估计 $\hat{\boldsymbol{\gamma}}$。

2　计算算法和调整参数的选取

2.1　LPRE 参数估计的求解算法

由目标函数式(9),容易看出该目标函数是可微的,所以最小化该目标函数就等价于求解该目标函数的一阶偏导数等于 0 的根,即

$$l'(\boldsymbol{\gamma})=\frac{\partial l(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}=\sum_{i=1}^{n}\boldsymbol{\Pi}_i\{\boldsymbol{Y}_i^{-1}\exp(\boldsymbol{\Pi}_i^{\mathrm{T}}\boldsymbol{\gamma})-\boldsymbol{Y}_i\exp(-\boldsymbol{\Pi}_i^{\mathrm{T}}\boldsymbol{\gamma})\}=0 \tag{11}$$

由于 $l''(\boldsymbol{\gamma})=\frac{\partial l^2(\boldsymbol{\gamma})}{\partial^2\boldsymbol{\gamma}}=\sum_{i=1}^{n}\boldsymbol{\Pi}_i\boldsymbol{\Pi}_i^{\mathrm{T}}\{\boldsymbol{Y}_i^{-1}\exp(\boldsymbol{\Pi}_i^{\mathrm{T}}\boldsymbol{\gamma})+\boldsymbol{Y}_i\exp(-\boldsymbol{\Pi}_i^{\mathrm{T}}\boldsymbol{\gamma})\}$,可知 $l''(\boldsymbol{\gamma})$ 可逆,故式(11)的根可以用牛顿迭代法进行求解。给定初值 $\boldsymbol{\gamma}_0$,则 $l'(\boldsymbol{\gamma})=0$ 的根可由如下逐步迭代解得:

$$\boldsymbol{\gamma}_{n+1}=\boldsymbol{\gamma}_n-\left(\frac{\partial l^2(\boldsymbol{\gamma})}{\partial^2\boldsymbol{\gamma}}\Big|_{\gamma=\gamma_n}\right)^{-1}\left(\frac{\partial l(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\Big|_{\gamma=\gamma_n}\right)$$

当 $\|\boldsymbol{\gamma}_{n+1}-\boldsymbol{\gamma}_n\|<\delta$ 时,例如 $\delta=10^{-8}$,称迭代收敛,其中 $\|\cdot\|$ 表示向量的 Euclidean 范数。

2.2　SCAD 惩罚估计的求解算法

由于 SCAD 惩罚函数非凸且不可导的缺点,导致直接最小化目标函数式(10)十分困难,所以类似

Fan 等[2]利用局部二次近似的方法逼近惩罚函数。给定初值 $\boldsymbol{\gamma}^{(0)}=\tilde{\boldsymbol{\gamma}}$,则惩罚函数 $p_{\lambda_n}(\|\boldsymbol{\gamma}_j\|_2)$ 可以用二次函数局部近似为如下形式:

$$p_{\lambda_n}(\|\boldsymbol{\gamma}_j\|_2)\approx p_{\lambda_n}(\|\boldsymbol{\gamma}_j^{(0)}\|_2)+\frac{1}{2}\{p'_{\lambda_n}(\|\boldsymbol{\gamma}_j^{(0)}\|_2)/\|\boldsymbol{\gamma}_j^{(0)}\|_2\}\{\|\boldsymbol{\gamma}_j\|_2^2-\|\boldsymbol{\gamma}_j^{(0)}\|_2^2\} \tag{12}$$

进一步,去除一些常数部分,则目标函数式(10)可以被近似为以下形式:

$$Q(\boldsymbol{\gamma})\approx l(\boldsymbol{\gamma})+\frac{1}{2}n\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Omega}(\boldsymbol{\gamma}^{(0)})\boldsymbol{\gamma} \tag{13}$$

其中,$\boldsymbol{\Omega}(\boldsymbol{\gamma}^{(0)})=\mathrm{diag}\{p'_{\lambda_n}(\|\boldsymbol{\gamma}_1^{(0)}\|_2)/\|\boldsymbol{\gamma}_1^{(0)}\|_2,\cdots,p'_{\lambda_n}(\|\boldsymbol{\gamma}_p^{(0)}\|_2)/\|\boldsymbol{\gamma}_p^{(0)}\|_2\}$,则 $\boldsymbol{\gamma}$ 的 SACD 惩罚估计 $\hat{\boldsymbol{\gamma}}$ 可以通过最小化目标函数式(13)获得,其可由如下迭代解得:

$$\boldsymbol{\gamma}^{(n+1)}=\boldsymbol{\gamma}^{(n)}-\left\{\frac{\partial l^2(\boldsymbol{\gamma})}{\partial^2\boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(n)}}+n\boldsymbol{\Omega}(\boldsymbol{\gamma}^{(n)})\right\}^{-1}\times\left\{\frac{\partial l(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(n)}}+n\boldsymbol{\Omega}(\boldsymbol{\gamma}^{(n)})\boldsymbol{\gamma}^{(n)}\right\}$$

当 $\|\boldsymbol{\gamma}^{(n+1)}-\boldsymbol{\gamma}^{(n)}\|<\delta$ 时,例如 $\delta=10^{-8}$,称迭代收敛。

注意用二次函数局部近似 $p_{\lambda_n}(\|\boldsymbol{\gamma}_j\|_2)$ 的过程中,为了防止出现分母为 0 的情况,类似式(12)中 $p'_{\lambda_n}(\|\boldsymbol{\gamma}_{(0)}\|_2)/\|\boldsymbol{\gamma}_{(0)}\|_2$ 这样有分母的式子,都在分母上加一个数,即用 $p'_{\lambda_n}(\|\boldsymbol{\gamma}_{(n)}\|_2)/(\|\boldsymbol{\gamma}_{(n)}\|_2+\vartheta)$ 代替 $p'_{\lambda_n}(\|\boldsymbol{\gamma}_{(n)}\|_2)/\|\boldsymbol{\gamma}_{(n)}\|_2$,其中 ϑ 的取值可以参考 Hunter 等[14]的建议。

2.3 调整参数的选取

实际应用中,调整参数的选取会直接影响估计的结果,因此,选择合适的调整参数对于接下来的模拟仿真和实证研究是十分重要的。

首先,本文采用三次 B 样条(即 $m=3$),为了计算更简便,采用等距节点,并且类似明浩等[15]取内节点的个数 $J=[n^{1/(2m+1)}]$,其中 $[c]$ 表示不超过 c 的最大整数;其次,基于 Fan 等[2]的建议,取 $a=3.7$;最后,鉴于贝叶斯信息准则(Bayesian Information Criterion,即 BIC)的良好理论性质,利用 BIC 准则选取最优的 λ_n,即通过最小化以下目标函数来选取 λ_n:

$$BIC(\lambda_n)=\log\left\{\sum_{i=1}^{n}\left\{\left|\frac{\boldsymbol{Y}_i-\exp(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}(U_i))}{\boldsymbol{Y}_i}\right|\times\left|\frac{\boldsymbol{Y}_i-\exp(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}(U_i))}{\exp(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}(U_i))}\right|\right\}\right\}+\left(\frac{\log n}{n}\right)df$$

其中,df 为系数 $\hat{\boldsymbol{\beta}}(\boldsymbol{U})$ 中非零分量的个数。

3 模拟仿真分析

考虑如下变系数乘积模型:

$$\boldsymbol{Y}_i=\exp\{\exp(2U_i-1)X_i^{(1)}+8U_i(1-U_i)X_i^{(2)}+2\cos^2(2\pi U_i)X_i^{(3)}\}\varepsilon_i$$

其中,$(X_i^{(1)},\cdots,X_i^{(p)})$ 服从 P 维多元正态分布,其均值为零,且协方差矩阵中第 (k,l) 个元素为 $0.5^{|k-l|}$,指标变量 U_i 是区间[0,1]上的均匀分布。本文考虑随机误差 ε 服从以下两种分布:$\log\varepsilon\sim N(0,1)$ 和 $\log\varepsilon\sim uniform(-2,2)$。在随机模拟中,取样本容量和维数分别为 $n=300$,$p=20$ 和 $n=500$,$p=50$,对每种设置重复模拟 200 次。为了进行比较,本文采用了两种方法进行模拟,它们分别为变系数乘积模型式(4),基于 LPRE 准则,经过 SCAD 惩罚函数进行压缩估计的方法,即本文所提方法(记为 LPRE-S);变系数模型式(5),基于最小二乘法,经过 SACD 惩罚函数进行压缩估计的方法(记为 LS-S)。为了更好地说明本文所提 LPRE-S 方法的有效性,采用以下 4 个标准来评价拟合效果:均方误差平方根(Square Root of Average Square Errors,RASE):$RASE(\hat{\beta}_j)=\{n^{-1}\sum_{i=1}^{n}(\hat{\beta}_j(U_i)-\beta_j(U_i))^2\}^{\frac{1}{2}}$,其中,在真实模型下所得估计的 RASE 值用 N_{oracle} 表示,惩罚估计的 RASE 值用 $N_{\mathrm{penalized}}$ 表示;零系数被正确估计为零的平均数,其值用"N_{C}"表示;非零系数被错误估计为零的平均数,其值用"N_{IC}"表示;正确拟合模型的比例,其值用"N_{CF}"表示。对于"N_C",越接近 p,说明变量选择效果越好;对于"N_{IC}",越接近 0 越好;对于"N_{CF}",值越接近 1 越好;对于 N_{Oracle} 与 $N_{\mathrm{Penalized}}$,越小越好,且 N_{Oracle} 与 $N_{\mathrm{Penalized}}$ 越接近,说明模型拟合效果越好。具体模拟结果见表 1。

表 1　模拟结果

Table 1　Simulation results

n,p	随机误差	方　法	N_C	N_{IC}	N_{CF}	N_{oracle}	$N_{penalized}$
300,20	$\log\varepsilon \sim N(0,1)$	LS-S	3.41	0.00	0.01	1.092 0	2.720 6
		LPRE-S	16.94	0.00	0.96	0.457 1	0.462 8
	$\log\varepsilon \sim Uniform(-2,2)$	LS-S	3.67	0.00	0.00	1.086 9	2.563 3
		LPRE-S	16.98	0.00	0.98	0.452 2	0.454 6
500,40	$\log\varepsilon \sim N(0,1)$	LS-S	11.74	0.00	0.00	0.925 8	2.976 8
		LPRE-S	37.00	0.00	1.00	0.412 2	0.412 5
	$\log\varepsilon \sim Uniform(-2,2)$	LS-S	11.53	0.00	0.00	0.891 8	2.843 3
		LPRE-S	37.00	0.00	1.00	0.4036	0.4036

从表 1 的模拟结果可以看出:对于给定的模型,两种方法的结果受不同的误差分布影响。首先,当误差的对数服从正态分布时,关于 N_{CF} 与 RASE 值方面,LPRE-S 方法比 LS-S 方法表现得更好,这说明了 LPRE-S 方法比 LS-S 方法更有效,且 LPRE-S 变量选择的结果几乎一致最好。其次,当误差的对数服从 $(-2,2)$ 上的均匀分布时,仍然是 LPRE-S 方法的结果更好,且相比于 $\log\varepsilon$ 服从标准正态分布时,LPRE-S 方法和 LS-S 方法的结果都稍好一点。最后,当样本量增大时,LPRE-S 估计方法选出真实模型的比率随之提高,几乎接近 1,且 $N_{Penalized}$ 与 N_{Oracle} 更加接近,这充分说明了本文所提方法的有效性。

4　实证研究

为了进一步说明所提方法的实用性,将所提 LPRE-S 方法应用于股票指数的跟踪,选取深证红利指数及其成分股作为实证研究对象。深证红利指数是指 40 只能够为深圳股市投资者提供长期稳定回报的股票,是深圳巨潮红利指数的缩影。本文数据来源于西南证券金点子财富管理终端,采用 2019-01-02—2021-02-26 期间,深证红利指数及其 40 只成分股的 522 个日线收盘价数据进行研究。

将深证红利指数作为响应变量 $\boldsymbol{Y}$,40 只成分股作为协变量 $\boldsymbol{X}$,成分股中的鞍钢股份作为指标变量 $\boldsymbol{U}$,考虑随机模拟中的 LPRE-S 和 LS-S 两种方法,同时对所有协变量进行标准化。由于影响股票指数的因素较多,且作用机制较复杂,这使得预测股票指数的长期走势非常困难,但是在短期预测中往往能够取得较好的效果。因此,为了检验模型的预测能力,令 $T=0,1,\cdots,121$,取第 1 天到第 $(400+T)$ 天的数据作为训练集,利用训练集获得参数和非参函数的估计,然后通过训练集上获得的预测模型来预测第 $(400+T+1)$ 天的深证红利指数,从而得到第 401 天到第 522 天的 122 个预测值,其预测效果如图 1、图 2 所示。

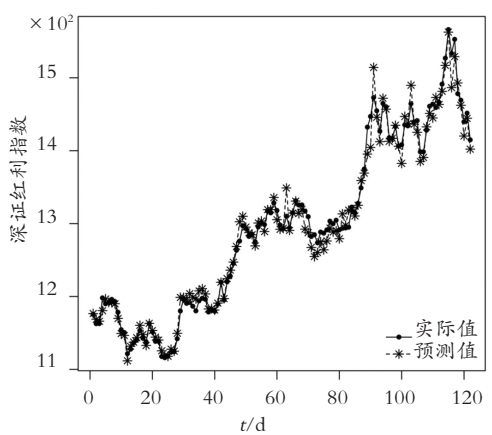

图 1　LPRE-S 方法实际值与预测值走势图

Fig. 1　Chart of actual and predicted values of the LPRE-S method

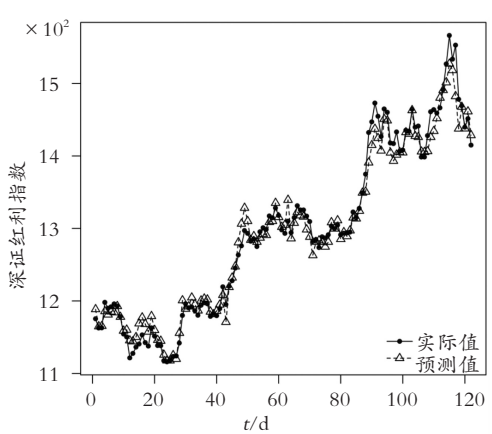

图 2　LS-S 方法实际值与预测值走势图

Fig. 2　Chart of actual and predicted values of the LS-S method

通过观察图 1、图 2,发现在前 60 天的预测中,LPRE-S 方法的实际走势与预测走势几乎一致,易见其预测效果优于 LS-S 方法,而后 62 天,LPRE-S 方法预测效果没有前 60 天预测效果好,且后 60 天两种方法的预测效果差别不是很明显。但是,通过计算得到,LPRE-S 方法在第 401 天到第 522 天所得残差平方和为 2 219 764,LS-S 方法在第 401 天到第 522 天所得残差平方和为 3 322 961,即 LPRE-S 方法的残差平方和要小于 LS-S 方法的残差平方和。

为了进一步对以上两种估计方法的预测效果进行比较,类似 Chen 等[8]建立以下 4 种中位数指标评价 LPRE-S 估计方法和 LS-S 估计方法:

绝对预测误差中位数 R_{MPE}:$\{|Y_i-\hat{Y}_i|\}$;乘积相对预测误差中位数 R_{MPPE}:$\{|\boldsymbol{Y}_i-\hat{\boldsymbol{Y}}_i|^2/(|\boldsymbol{Y}_i\hat{\boldsymbol{Y}}_i|)\}$;可加相对预测误差中位数 R_{MAPE}:$\{|\boldsymbol{Y}_i-\hat{\boldsymbol{Y}}_i|/\boldsymbol{Y}_i+|\boldsymbol{Y}_i-\hat{\boldsymbol{Y}}_i|/\hat{\boldsymbol{Y}}_i\}$;平方预测误差中位数 R_{MSPE}:$\{(\boldsymbol{Y}_i-\hat{\boldsymbol{Y}}_i)^2\}$。具体比较结果见表 2。

表 2 LPRE-S 和 LS-S 两种方法预测误差的中位数比较结果

Table 2 Comparison results of median prediction errors between LPRE-S and LS-S methods

	R_{MPE}	R_{MPPE}	R_{MAPE}	R_{MSPE}
LPRE-S	64.309 8	$2.368\,1\times10^{-5}$	0.009 7	4 148.586 0
LS-S	81.890 5	4.1448×10^{-5}	0.012 9	6 706.068 0

对于表 2 中的 4 种中位数指标,值越小的方法,其预测效果越有效。从表 2 的结果可以看出:LPRE-S 方法在每种中位数指标下的值都比 LS-S 方法的值小,即 LPRE-S 方法的结果优于 LS-S 方法。由此,进一步说明了本文所提方法能更加有效追踪股票指数。

5 结 论

本文基于 B 样条函数逼近技术,将 LPRE 准则和组 SCAD 惩罚函数结合起来,应用于变系数乘积模型,利用牛顿迭代法和局部二次近似给出了所提方法的计算算法,并阐释了如何选取调整参数。通过数值模拟对 LPRE-S 估计方法和 LS-S 估计方法进行了比较,发现 LPRE-S 估计方法选出真实模型的比率几乎接近 1,且 $N_{\mathrm{Penalized}}$ 与 N_{Oracle} 十分接近,这说明了 LPRE-S 估计方法能更好地达到变量选择的目的,证明了所提方法的有效性。为了进一步说明所提方法的实用性,用 LPRE-S 估计方法实现了对深证红利指数的跟踪预测,并与 LS-S 估计方法的预测效果进行了对比。通过比较 122 个预测值的残差平方和与 4 种不同的预测误差中位数指标,发现 LPRE-S 估计方法效果优于 LS-S 估计方法,说明了本文所提方法在股指跟踪中具有较好的预测效果。

参考文献(References):

[1] TIBSHIRANI R. Regression shrinkage and selection via the Lasso[J]. Journal of the Royal Statistical Society: Series B(Statistical Methodology), 1996, 58(1): 276—288.

[2] FAN J Q, LI R Z. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1348—1360.

[3] ZOU H. The adaptive Lasso and its oracle properties[J]. Journal of the American Statistical Association, 2006, 101(476): 1418—1429.

[4] ZOU H, HASTIE T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B(Statistical Methodology), 2005, 67(2): 301—320.

[5] ZOU H, ZHANG H H. On the adaptive elastic net with a diverging number of parameters[J]. The Annals of Statistics, 2009, 37(4): 1733—1751.

[6] CHEN K N, GUO S J, LIN Y Y, et al. Least absolute relative error estimation[J]. Journal of the American Statistical Association, 2010, 105(491): 1104—1112.

[7] 张丹. 乘积模型的变量选择[D]. 开封:河南大学, 2014.
ZHANG Dan. Variable selection of product model[D]. Kaifeng: Henan University, 2014.

[8] CHEN K N, LIN Y Y, WANG Z F, et al. Least product relative error estimation[J]. Journal of Multivariate Analysis, 2016, 144(2): 91—98.

[9] 李翠平. 乘积回归模型的变量选择研究[D]. 开封:河南大学, 2016.
LI Cui-ping. Study on variable selection of product regression model[D]. Kaifeng: Henan University, 2016.

[10] 陈银钧, 刘惠篮. 基于 LPRE 和 LASSO 方法的股指追踪研究[J]. 经济数学, 2020, 37(1): 92—96.
CHEN Yin-jun, LIU Hui-lan. Study on stock index

tracking based on LPRE and Lasso methods[J]. Economic Mathematics,2020,37(1):92—96.
[11] 胡大海. 基于乘积相对误差准则的模型研究[D]. 合肥:中国科学技术大学,2017.
HU Da-hai. Study on the model based on the product relative error criterion[D]. Hefei: University of Science and Technology of China,2017.
[12] HASTIE T,TIBSHIRANI R. Varying-coefficient models[J]. Journal of the Royal Statistical Society: Series B (Methodological Edition), 1993, 55(4): 757—796.
[13] 吕晶. 几类半参数回归模型的稳健估计与变量选择[D]. 重庆:重庆大学,2015.
LYU Jing. Robust estimation and variable selection of several semi-parametric regression models[D]. Chong qing:Chongqing University,2015.
[14] HUNTER D R, LI R. Variable selection using MM algorithms[J]. The Annals of Statistics,2005, 33(4): 1617—1642.
[15] 明浩,刘惠篮. 可加乘积模型的研究[J]. 系统科学与数学,2020,40(3):547—564.
MING Hao, LIU Hui-lan. Research on additive product model [J]. Systems Science and Mathematics, 2020, 40(3):547—564.

Study on Stock Index Tracking Based on Variable Coefficient Product Model

WAN Xue

(School of Mathematical Science, Chongqing Normal University,Chongqing 401331,China)

Abstract: With the development of economy, stock investment has come into public view. How to choose constituent stocks to track the stock index has been paid more and more attention by the people. Based on this, aiming at the problem of stock index tracking, a method of variable coefficient product model for variable selection is proposed. Based on B-spline function approximation technique, this method combines LPRE (Least Product Relative Error) criterion and group SCAD (Class Clipped Absolute Devation) penalty function to apply to the variable coefficient product model. The implementation steps of solving the estimation are given by Newton iterative algorithm and local quadratic approximation. In order to verify the effectiveness of the proposed method, the results of SCAD penalty method with variable coefficient product model (LPRE-S) and the least square SCAD penalty method with variable coefficient model (LS-S) were compared by numerical simulation. In order to verify the practicability of the proposed method,LPRE-S estimation method and LS-S estimation method are compared for the tracking and forecasting effect of dividend index in Shenzhen Stock Exchange. The results show that the ratio of the LPRE-S estimation on the method to select the real model is almost close to 1, which can better achieve the purpose of variable selection and has a good prediction effect in the stock index tracking.

Key words: variable cofficient product model; variable selection; LPRE estimation; SCAD

责任编辑:李翠薇

引用本文/Cite this paper:
万学. 基于变系数乘积模型的股指跟踪研究[J]. 重庆工商大学学报(自然科学版),2022,39(2):83—89.
WAN Xue. Study on stock index tracking based on variable coefficient product model[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(2):83—89.