DOI: 10.13718/j.cnki.xdzk.2017.11.014

函数型非参数部分自回归模型及其在金融中的应用①

王咪咪,　丁　辉

滁州学院 数学与金融学院,安徽 滁州 239000

摘要: 结合金融市场中的滞后现象以及函数型协变量和响应变量之间的非线性关系提出了函数型非参数部分自回归模型,接着使用 profile 最小二乘方法和非参数核估计方法给出了该模型的估计,并通过统计模拟验证了该方法的有效性,最后通过上证指数的实例验证了模型的预测能力.

关　键　词: 函数型数据;高频数据;非参数部分自回归模型;核估计

中图分类号: O212.7　　**文献标志码**: A　　**文章编号**: 1673-9868(2017)11-0096-06

函数型数据分析是处理和分析高频数据的一个很重要的工具. 国内外已经有一些文献借助于函数型数据分析方法来研究金融市场中的内在规律[1-5]. 但是已有成果都只考虑了函数型协变量对响应变量的影响而没有考虑响应变量的历史时刻对相应变量当前时刻的影响. 本文借用自回归模型的思想,在函数型非参数模型的基础上添加自回归项来体现响应变量的历史时刻对响应变量当前时刻的影响,即函数型非参数部分自回归模型,该模型具有广泛的适用性.

1　函数型非参数部分自回归模型

函数型非参数部分自回归模型的形式如下:

$$Y_i = g(X_i(t)) + \sum_{k=1}^{p} \varphi_k Y_{i-k} + \varepsilon_i \qquad i = p+1, p+2, \cdots, n \tag{1}$$

其中: $X(t)$ 为定义在区间 I 上平方可积且完全观测的函数型协变量, $g(\cdot)$ 为平方可积空间 L^2 到实数域 $\mathbb{R}$ 的实值函数,即 $g: L^2 \longrightarrow \mathbb{R}$, Y 为标量, p 为该模型的滞后阶数, ε 是误差项,满足 $E\varepsilon = 0$, $Var\,\varepsilon = \sigma^2$. 该模型克服了函数型协变量和响应变量的线性约束,且函数型线性模型是该模型的一个特例(即模型没有滞后项 $\varphi_1 = \cdots = \varphi_p = 0$ 且 $g(X(t)) = \int_I X(t)\beta(t)\mathrm{d}t$),自回归模型也是该模型的一个特例(即 $g(X(t)) \equiv 0$). 因此,该模型具有很强的灵活性,在实际研究中有着重要意义.

我们使用 profile 最小二乘估计与非参数核估计相结合的方法对模型进行估计,具体步骤如下:

① 收稿日期: 2016-07-24
基金项目: 全国统计科学研究计划项目(2012LY153);滁州学院科研启动基金资助项目(2014qd012);安徽省自然科学基金研究项目(1508085QA14);安徽省高等学校省级自然科学研究项目(KJ2014A180).
作者简介: 王咪咪(1981-),女,山东博兴人,讲师,主要从事数理统计与决策预测研究.

令

$$\widetilde{Y}_i = Y_i - \sum_{k=1}^{p} \varphi_k Y_{i-k}$$

则模型(1)转化为

$$\widetilde{Y}_i = g(X_i(t)) + \varepsilon_i \tag{2}$$

利用 Nadaraya-Watson 核估计[6]，可得

$$\widetilde{g}(u) = \sum_{j=p+1}^{n} \omega_j(u)\widetilde{Y}_j \tag{3}$$

其中

$$\omega_j(u) = \frac{K_h(d(u, X_j(t)))}{\sum_{s=p+1}^{n} K_h(d(u, X_s(t)))} \qquad K_h(\cdot) = \frac{K\left(\frac{\cdot}{h}\right)}{h}$$

$K(\cdot)$ 是一元核函数，h 是窗宽，$d(\cdot, \cdot)$ 是两个函数型变量的一个距离度量(例如：$d(X(t), Z(t)) = \left\{\int_I (X(t) - Z(t))^2 \mathrm{d}t\right\}^{\frac{1}{2}}$)．于是，由 profile 最小二乘原理，我们可以通过最小化

$$\sum_{i=p+1}^{n} \{Y_i - \sum_{k=1}^{p} \varphi_k Y_{i-k} - \sum_{j=p+1}^{n} [\omega_j(X_i(t))(Y_j - \sum_{k=1}^{p} \varphi_k Y_{j-k})]\}^2 \tag{4}$$

得到参数 $\varphi_1, \cdots, \varphi_p$ 的估计．

令

$$\boldsymbol{Y} = (Y_{p+1}, \cdots, Y_n)^{\mathrm{T}} \qquad \boldsymbol{\Phi} = (\varphi_1, \cdots, \varphi_p)^{\mathrm{T}} \qquad \boldsymbol{W} = (w_{ij})_{(n-p)\times(n-p)}$$

其中

$$w_{ij} = \omega_{j+p}(X_{i+p}(t)) \qquad i, j = 1, \cdots, n-p$$

$$\boldsymbol{Z}_1 = (Y_p, \cdots, Y_{n-1})^{\mathrm{T}}$$

$$\boldsymbol{Z}_2 = (Y_{p-1}, \cdots, Y_{n-2})^{\mathrm{T}}, \cdots$$

$$\boldsymbol{Z}_p = (Y_1, \cdots, Y_{n-p})^{\mathrm{T}}$$

$$\boldsymbol{Z} = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_p)^{\mathrm{T}}$$

$$\boldsymbol{Y}^* = (\boldsymbol{I} - \boldsymbol{W})\boldsymbol{Y} \qquad \boldsymbol{Z}^* = \boldsymbol{Z}(\boldsymbol{I} - \boldsymbol{W})$$

$\boldsymbol{I}$ 为 $n-p$ 阶单位矩阵．则(4)式可以简化为$(\boldsymbol{Y}^* - \boldsymbol{Z}^{*\mathrm{T}}\boldsymbol{\Phi})^{\mathrm{T}}(\boldsymbol{Y}^* - \boldsymbol{Z}^{*\mathrm{T}}\boldsymbol{\Phi})$，于是

$$\hat{\boldsymbol{\Phi}} = (\boldsymbol{Z}^*\boldsymbol{Z}^{*\mathrm{T}})^{-1}\boldsymbol{Z}^*\boldsymbol{Y}^* \tag{5}$$

将(5)式代回(3)式可得

$$\hat{g}(u) = \sum_{j=p+1}^{n} \omega_j(u)(Y_j - \sum_{k=1}^{p} \hat{\varphi}_k Y_{i-k}) \tag{6}$$

于是我们得到非参数函数 $g(\cdot)$ 的估计．在求解最小二乘估计的时候采用 YUAN 和 WEI 的改进的 BFGS 算法[7]．当然，在估计时涉及到参数 h 和 p 的选取，关于 h 和 p 的选择有很多准则(例如 AIC 准则、BIC 准则、GCV 准则和交叉核实准则)，我们使用交叉核实准则来选取这两个参数．函数型非参数部分自回归模型由于克服了函数型协变量和响应变量的线性约束并且又体现了标量型响应变量的历史时刻影响的自回归形式，因此模型本身具有较大的适用性和灵活性．

2 大样本性质

为了给出估计的大样本性质，我们引入记号令 φ_k^0 表示 φ_k 的真值，

$$\boldsymbol{\Phi}^0=(\varphi_1^0,\cdots,\varphi_p^0)^{\mathrm{T}}$$

假设下列条件成立：

1）有界函数 $g(\cdot)\in C^0$，其中 $C^0=\{g:L^2\to\mathbb{R},\lim\limits_{d(x(t),x'(t)\to 0)}g(x'(t))=g(x(t))\}$；

2）$\forall\varepsilon>0$，$P(X(t)\in B(x(t),\varepsilon))=\varphi_{x(t)}(\varepsilon)>0$；

3）窗宽参数 h 满足 $\lim\limits_{n\to\infty}h=0$ 且 $\lim\limits_{n\to\infty}\dfrac{\log(n)}{n\varphi_{x(t)}(\varepsilon)}=0$；

4）核函数 $K(\cdot)$ 在 $[0,1]$ 上有紧支撑，其导数 $K'(\cdot)$ 存在，且存在常数 $c_1<c_2<0$，使得 $c_1\leqslant K'\leqslant c_2$；

5）$\forall m\geqslant 2$，$E(|Y_i-\sum\limits_{k=1}^{p}\varphi_kY_{i-k}|^m\mid X(t)=x(t))<\sigma_m(x(t))<\infty$，其中 $\sigma_m(x(t))$ 关于 $x(t)$ 连续；

6）存在常数 $\lambda>0$，使得 $E(|\varepsilon_i|^{2+\lambda})<\infty$；

7）$\lim\limits_{n\to\infty}\boldsymbol{D}_n=\boldsymbol{D}$，其中 $\boldsymbol{D}_n=\dfrac{1}{n-p}\sum\limits_{i=p+1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}$，$\boldsymbol{\xi}_i=(Y_{i-1},\cdots,Y_{i-p})^{\mathrm{T}}$，$i=p+1,\cdots,n$，$\boldsymbol{D}$ 是一个非负定阵.

则有 $\overset{\wedge}{\varphi}_k\xrightarrow{P}\varphi_k^0$.

证　由假设条件 1）－5）采用类似 FERRATY 和 VIEU 的定理 6.1[8] 证明方法可得

$$\lim_{n\to\infty}\widetilde{g}(u)=g(u),\ a.co. \tag{7}$$

其中 $a.co.$ 代表几乎完全收敛. 则

$$L_n(\boldsymbol{\Phi})=\frac{1}{n-p}\sum_{i=p+1}^{n}\{Y_i-\sum_{k=1}^{p}\varphi_kY_{i-k}-\sum_{j=p+1}^{n}[\omega_j(X_i(t))(Y_j-\sum_{k=1}^{p}\varphi_kY_{j-k})]\}^2=$$

$$\frac{1}{n-p}\sum_{i=p+1}^{n}[Y_i-\sum_{k=1}^{p}\varphi_kY_{i-k}-g(X_i(t))+g(X_i(t))-\widetilde{g}(X_i(t))]^2=$$

$$\frac{1}{n-p}\sum_{i=p+1}^{n}\delta_i^2+\frac{1}{n-p}\sum_{i=p+1}^{n}[g(X_i(t))-\widetilde{g}(X_i(t))]^2+\frac{2}{n-p}\sum_{i=p+1}^{n}\delta_i[g(X_i(t))-\widetilde{g}(X_i(t))]$$

其中

$$\delta_i=Y_i-\sum_{k=1}^{p}\varphi_kY_{i-k}-g(X_i(t))$$

由(7)式及其假设条件 5）和 6），我们有

$$\frac{1}{n-p}\sum_{i=p+1}^{n}[g(X_i(t))-\widetilde{g}(X_i(t))]^2\xrightarrow{P}0$$

$$\frac{2}{n-p}\sum_{i=p+1}^{n}\delta_i[g(X_i(t))-\widetilde{g}(X_i(t))]\xrightarrow{P}0 \tag{8}$$

令

$$L'_n(\boldsymbol{\Phi})=\sum_{i=p+1}^{n}\delta_i^2=\frac{1}{n-p}\sum_{i=p+1}^{n}[Y_i-\sum_{k=1}^{p}\varphi_kY_{i-k}-g(X_i(t))]^2=$$

$$(\boldsymbol{\Phi}^0-\boldsymbol{\Phi})^{\mathrm{T}}\boldsymbol{D}_n(\boldsymbol{\Phi}^0-\boldsymbol{\Phi})+\frac{1}{n-p}\sum_{i=p+1}^{n}\varepsilon_i^2+\frac{2}{n-p}\sum_{i=p+1}^{n}(\boldsymbol{\Phi}^0-\boldsymbol{\Phi})^{\mathrm{T}}\xi_i\varepsilon_i$$

易得

$$L'_n(\boldsymbol{\Phi})\xrightarrow{P}(\boldsymbol{\Phi}^0-\boldsymbol{\Phi})^{\mathrm{T}}\boldsymbol{D}_n(\boldsymbol{\Phi}^0-\boldsymbol{\Phi})+\sigma^2$$

又 $L'_n(\boldsymbol{\Phi})$ 为凸函数，故

$$\sup_{\boldsymbol{\Phi}\in\Omega}\left|L'_n(\boldsymbol{\Phi})-(\boldsymbol{\Phi}^0-\boldsymbol{\Phi})^{\mathrm{T}}\boldsymbol{D}_n(\boldsymbol{\Phi}^0-\boldsymbol{\Phi})-\sigma^2\right|\xrightarrow{P}0$$

对于任意的紧集 $\boldsymbol{\Omega}$ 成立，且 $\hat{\boldsymbol{\Phi}}=O_P(1)$，因此 $\hat{\boldsymbol{\Phi}}\xrightarrow{P}\boldsymbol{\Phi}^0$，进而 $\hat{\varphi}_k\xrightarrow{P}\varphi_k^0$.

3 统计模拟

为了说明估计的有效性，我们下面进行统计模拟. 按照下面的模型来生成数据：

$$Y_i=g(X_i(t))+0.4Y_{i-1}-0.3Y_{i-2}+0.3Y_{i-3}+\varepsilon_i\qquad i=4,5,\cdots,n+3$$

其中

$$g(X(t))=\sin\left(\int_0^1X(t)\beta(t)\mathrm{d}t\right)$$

对于 $\int_0^1X(t)\beta(t)\mathrm{d}t$ 部分，我们按照下面的方式随机生成函数型数据和斜率函数，即

$$\beta(t)=\sum_{k=1}^{50}b_k\varphi_k(t)$$

$$X_i(t)=\sum_{k=1}^{50}\xi_k\theta_{ik}\varphi_k(t)$$

$$b_1=0.3\qquad b_k=4(-1)^{k+1}k^{-2}\qquad k\geqslant 2$$

$$\varphi_k(t)=1\qquad \varphi_k(t)=\sqrt{2}\cos[(k-1)\pi t]\qquad k\geqslant 2$$

$$\xi_k=(-1)^{k+1}k^{-1}$$

θ_{ik} 独立且服从区间$(-\sqrt{3},\sqrt{3})$上的均匀分布；$Y_1=0.6$，$Y_2=0.3$，$Y_3=0.4$，误差项 ε_i 独立且均服从正态分布 $N(0,0.1^2)$.

我们考虑样本量 $n=100$，$n=200$ 和 $n=300$ 这3种情况，并且假设 $X_i(t)$ 在$[0,1]$区间上的100个等间隔的时间点被完全观测到. 通过使用交叉核实准则，我们得到滞后阶数 $\hat{p}=3$，窗宽参数 $\hat{h}=0.526$，参数 $\varphi_1,\cdots,\varphi_3$ 的估计值的好坏我们通过估计均值和标准差来体现，非参数 $g(X(t))$ 估计的好坏我们通过均方误差 $RASE$ 来体现，其中

$$RASE(\hat{g}(X(t)))=\left\{\frac{1}{n}\sum_{i=4}^{n+p}[\hat{g}(X_i(t))-g(X_i(t))]^2\right\}^{\frac{1}{2}}$$

由表1可以看出，参数 $\varphi_1,\cdots,\varphi_3$ 的估计的偏差与标准差及其非参数 $g(X(t))$ 估计均方误差 $RASE$ 随着样本量的增加而减少. 模拟说明我们的估计方法是有效的.

表1 统计结果表

n	均值			均方误差
	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$	$\hat{g}(u)$
100	0.381(0.104)	−0.319(0.105)	0.260(0.116)	0.513(0.024)
200	0.395(0.073)	−0.307(0.063)	0.284(0.082)	0.511(0.019)
300	0.401(0.059)	−0.306(0.060)	0.294(0.069)	0.510(0.014)

注：括号内为标准差.

4 实证分析

我们对上证指数数据进行函数型非参数部分自回归模型建模来说明我们模型的优良性. 选取自2015年6月1日至2016年3月1日共183个交易日的上证指数数据. 此数据含有上证指数开盘价 Y，每个交易日每

5 分钟的上证指数价格 $X(t)$，因交易日内的交易时间均为 9：30－11：30 及 13：00－15：00，故每个交易日每 5 min 的上证指数数据共 48 个. 首先对数据进行预处理：将 Y 和 $X(t)$ 取对数. 该实证分析的目的是找出相对合适的模型对开盘价进行预测. 由于上证指数第 i 天开盘价 Y_i 受到历史时期开盘价格 $Y_{i-1},\cdots,Y_{i-p}$ 的影响，且 Y_i 与第 i 天每 5 min 上证价格 $X_i(t)$ 之间不一定能够满足线性关系，故利用函数型非参数部分自回归模型来刻画 Y_i 与 $X_i(t)$ 和 $Y_{i-1},\cdots,Y_{i-p}$ 的关系，即

$$Y_i = g(X_i(t)) + \sum_{k=1}^{p} \varphi_k Y_{i-k} + \varepsilon_i \qquad i = p+1, p+2, \cdots, n$$

为了比较不同的模型对开盘价格的预测能力，考虑 4 个模型：自回归模型、函数型线性模型、函数型部分线性自回归模型、函数型非参数部分自回归模型. 我们使用滚动预测法来体现模型的预测效果：从前 173 个数据用作训练集去预测第 174 个数据，然后再把前 174 个数据用作训练集，预测第 175 个数据，以此类推，直至前 182 个数据用作训练集，预测第 183 个数据结束. 我们使用平均预测误差

$$ARPE = \frac{1}{10}\sum_{i=174}^{183} (Y_i - \hat{Y}_i)^2$$

作为比较不同模型的预测好坏的准则. 4 个模型及其对应的平均预测误差结果见表 2.

由表 2 可见函数型非参数部分自回归模型的平均预测误差最小，比函数型部分线性自回归模型提高了 12.6%，比线性模型提高了 23.7%，比自回归模型提高了 95%. 故这 4 个模型中函数型非参数部分线性模型的预测效果最好，因此该模型可以为处理和分析金融高频数据提供新的分析方法和思路.

表 2　不同模型及其相对应的平均预测误差

模　型	平均预测误差 / $\times 10^{-5}$	模　型	平均预测误差 / $\times 10^{-5}$
$Y_i = \alpha + \int_0^1 X_i(t)\beta(t)\mathrm{d}t + \varepsilon_i$	5.58	$Y_i = \alpha + \int_0^1 X_i(t)\beta(t)\mathrm{d}t + \sum_{k=1}^{p} \varphi_k Y_{i-k} + \varepsilon_i$	4.91
$Y_i = \sum_{k=1}^{p} \varphi_k Y_{i-k} + \varepsilon_i$	93.9	$Y_i = g(X_i(t)) + \sum_{k=1}^{p} \varphi_k Y_{i-k} + \varepsilon_i$	4.29

5　总　结

函数型非参数部分自回归模型克服了函数型协变量与响应变量的线性约束，同时引入自回归效应，用来刻画响应变量的历史时刻对响应变量的当前时刻的影响. 该模型是函数型线性模型和函数型部分线性自回归模型及其自回归模型的一种推广，具有很大的灵活性和适用性. 因此函数型非参数部分自回归模型可以为今后研究金融市场中高频数据提供一种新的思路和模式.

参考文献：

[1] 程丽娟. 上证指数的函数型主成分分析预测 [J]. 岭南师范学院学报，2016，37(3)：39－43.

[2] 蔺顺锋，易丹辉，肖宏伟. 基于函数型数据分析视角的我国副省级城市年平均工资差异研究 [J]. 现代管理科学，2015(3)：27－29.

[3] 龙　文，李　楠，王惠文，等. 金融危机过程中不同类型国家经济发展的差异性比较——基于函数数据分析方法 [J]. 管理评论，2014，26(3)：3－10.

[4] 许　梁，孙　涛，徐　箭，等. 基于函数型非参数回归模型的中长期日负荷曲线预测 [J]. 电力自动化设备，2015，35(7)：89－94.

[5] 马晓波，冯凌秉，李　玮. 高频数据日内波动特征的函数型分析 [J]. 企业导报，2011(22)：76－77.

[6] WASSERMAN L. All of Nonparametric Statistics [M]. New York: Springer, 2006.

[7] YUAN G, WEI Z. Non Monotone Backtracking Inexact BFGS Method for Regression Analysis [J]. Communications in Statistics-Theory and Methods, 2013, 42(2): 214－238.

[8] FERRATY F, VIEU P. Nonparametric Functional Data Analysis: Theory and Practice [M]. New York: Springer, 2006.

[9] RAMSAY J O, SILVERMAN B W. Functional Data Analysis [M]. New York: Springer, 1997.

Functional Nonparametric Partial Auto-regression Model and Its Applications in Finance

WANG Mi-mi, DING Hui

School of Mathematics and Finance, Chuzhou University, Chuzhou Anhui 239000, China

Abstract: Functional data analysis is an important method of analyzing high-frequency data of the financial market. Combining the lag phenomenon on the financial market and the nonlinear relationship between the functional covariate and the response variable, this paper proposes a functional nonparametric partial auto-regression model. Then, the profile least square method and the nonparametric kernel estimation are used to obtain the estimators of the model. Statistical simulation verified its validity. A real example about Shanghai Stock Index data is used to demonstrate the good prediction ability of the model.

Key words: functional data; high frequency data; partial nonparametric auto-regression model; kernel estimation

责任编辑 张 枸