DOI: 10.13413/j.cnki.jdxblxb.2003.02.001 吉林大学学报(理学版) No. 2
Vol. 41
2003年4月 JOURNAL OF JILIN UNIVERSITY (SCIENCE EDITION) 127~133

缺失数据下 AR(p)模型的估计方法

田 萍[1], 董险峰[2], 王德辉[2], 黄晓薇[2]
(1. 吉林大学商学院, 长春 130012; 2. 吉林大学数学学院, 长春 130012)

提要: 研究在缺失数据条件下, 零均值 AR(p)模型的估计方法. 应用 EM算法, 给出了一个数据和连续两个数据缺失时的具体计算步骤.

关键词: 似然函数; EM算法; 条件概率密度
中图分类号: O212.1 文献标识码: A 文章编号: 1671-5489(2003)02-0127-07

1 引 言

在股票交易活动中, 各种指数均以时间序列的形式表现出来, 而且各种指数间有很强的关联关系. 由于各国的法定假日不同以及突发事件的影响等, 在比较两种指数间的关系(如计算它们的对数差序列)时, 这个序列就出现了缺失数据. 考察上述序列在实际应用中有重要的意义. 因此, 对含缺失数据的时间序列进行建模十分必要.

时间序列模型之一零均值 AR(p)模型[1]:

$$z_t = \varphi_1 z_{t-1} + \cdots + \varphi_p z_{t-p} + X_t, \quad X_t \overset{\text{i.i.d.}}{\sim} N(0,1), \quad \text{且 } X_t \text{与 } \{z_s,\ s < t\} \text{独立}, \tag{1.1}$$

它在时间序列分析和许多问题的研究与处理中都有重要的意义, 其中$\alpha' = (\varphi_1, \cdots, \varphi_p)$是待估参数, 它反映了序列中各变量前后依赖的关系. 因此确定$\alpha' = (\varphi_1, \cdots, \varphi_p)$的估计值对模型至关重要.

关于 AR(p)模型有多种参数估计方法[2~4]. 然而, 当数据量不够大且没有其它信息可以弥补从中间处缺失的数据时, 对参数的估计目前还没有十分有效的方法. 本文讨论当数据从中间处缺失一个或连续两个却又不能舍弃和弥补时的情形, 并分别对其给出实例估计方法. 利用产生的随机数对给出的方法做了优劣验证. 本文对缺失数据的处理, 主要应用 EM算法.

EM算法是一种迭代方法, 主要用来求后验分布的众数(或极大似然估计), 它的每一次迭代由两步组成: E步(求期望)和 M步(极大化).

E步: 将添加数据 y后得到的关于 α 的后验分布密度函数 $L(\alpha \mid z, y)$或对数密度函数 $\ln L(\alpha \mid z, y)$关于给定 $\alpha^{(k)}$和观测数据 z条件下潜在数据 y的条件分布密度函数 $L(y \mid \alpha^{(k)}, z)$求期望, 从而把 y积掉, 得到 $g(\alpha \mid \alpha^{(k)}, z)$. M步: 将 $g(\alpha \mid \alpha^{(k)}, z)$关于 α 极大化得到 $\alpha^{(k+1)}$.

如此形成了一次迭代 $\alpha^{(k)} \to \alpha^{(k+1)}$, 将上述 E步和 M步进行迭代, 直至 $\| \alpha^{(k+1)} - \alpha^{(k)} \|$ 或 $\| g(\alpha^{(k+1)} \mid \alpha^{(k)}, z) - g(\alpha^{(k)} \mid \alpha^{(k)}, y) \|$ 充分小时停止.

关于 EM算法得到的估计序列是否收敛, 收敛结果是否是 $P(\alpha \mid z)$的最大值或局部最大值, 可参见文献[5, 6].

2 一个数据缺失时参数的估计方法

考虑一个数据缺失的情形, 设此时的观测值为 $(z_1, \cdots, z_{i-1}, z_{i+1}, \cdots, z_n)$, 其中 z_i 为缺失值. 为方便, 记

收稿日期: 2002-09-04.
作者简介: 田 萍(1976~), 女, 博士研究生, 从事数量经济研究, E-mail: tianping76@yahoo.com.cn.
基金项目: 吉林大学青年基金.

$$a^2 = 1 + \gamma_1^2 + \gamma_2^2 + \cdots + \gamma_p^2 = \sum_{k=0}^{p} \gamma_k^2, \qquad b = -A_0 + \sum_{k=1}^{p} \gamma_k A_k = \sum_{k=0}^{p} \gamma_k A_k, \qquad c = \sum_{k=0}^{p} A_k^2,$$

其中

$$\gamma_0 = -1, \qquad A_0 = (\gamma_1, \cdots, \gamma_p)(z_{i-1}, \cdots, z_{i-p})',$$

$$A_k = (\gamma_0, \cdots, \gamma_{k-1}, \gamma_{k+1}, \cdots, \gamma_p)(z_{i+k}, \cdots, z_{i+1}, z_{i-1}, \cdots, z_{i+k-p})', \quad k = 1, \cdots, p.$$

定理 2.1 在模型(1.1)下,观测似然函数

$$\tilde{L}(\alpha) = \int_{-\infty}^{\infty} L(\alpha)\,\mathrm{d}z_i = \left|\frac{1}{\sqrt{2\pi}}\right|^n \frac{1}{a}\exp\left|-\frac{c - b^2/a^2}{2}\right| D_{i-1;0} D_{n;i+p+1}, \tag{2.1}$$

其中

$$D_{m;n} = \prod_{j=n}^{m} \exp\left|-\frac{1}{2}(z_j - \gamma_1 z_{j-1} - \cdots - \gamma_p z_{j-p})^2\right|,$$

$$L(\alpha) = \prod_{j=0}^{n} f(z_j - \gamma_1 z_{j-1} - \cdots - \gamma_p z_{j-p}),$$

这里 $f(\cdot)$是标准正态密度函数,$L(\alpha)$为$(z_1, \cdots, z_{i-1}, z_i, z_{i+1}, \cdots, z_n)$对应的似然函数.

证明: 由于

$$\prod_{j=i}^{i+p} f(z_j - \gamma_1 z_{j-1} - \cdots - \gamma_p z_{j-p}) = \exp\left|-\frac{1}{2}(a^2 z_i^2 + 2bz_i + c)\right| \left|\frac{1}{\sqrt{2\pi}}\right|^{p+1}, \tag{2.2}$$

所以

$$\tilde{L}(\alpha) = \int_{-\infty}^{\infty} L(\alpha)\,\mathrm{d}z_i = \left|\frac{1}{\sqrt{2\pi}}\right|^n \frac{1}{a}\exp\left|-\frac{c - b^2/a^2}{2}\right| D_{i-1;0} D_{n;i+p+1}. \tag{2.3}$$

为了应用 EM算法,注意到 z_i 关于 $\alpha^{(k)}$和 $z_i^- = \{z_1, \cdots, z_n\} \backslash \{z_i\}$的条件分布密度为

$$f(z_i | \alpha^{(k)}, z_i^-) = L(\alpha^{(k)}) / \tilde{L}(\alpha^{(k)}) =$$

$$a^{(k)} \prod_{j=i}^{i+p} f(z_j - \gamma_1^{(k)} z_{j-1} - \cdots - \gamma_p^{(k)} z_{j-p}) \exp\left|\frac{c^{(k)} - (b^{(k)}/a^{(k)})^2}{2}\right| \left|\frac{1}{\sqrt{2\pi}}\right|^{-p}, \tag{2.4}$$

其中 $a^{(k)}, b^{(k)}, c^{(k)}$为由第 k步得出 $\alpha^{(k)}$后算出的值,其函数形式同 a, b, c, α 前边所定义的形式.

综上讨论,可以对 EM算法中 E步进行计算.

定理 2.2 在 EM算法中,

$$E(\ln(L(\alpha)) | z_i^-, \alpha^{(k)}) = -(n+1)\ln\sqrt{2\pi} + \ln D_{i-1;0} + \ln D_{n;i+p+1} - \frac{a^2}{2a^{(k)2}}\left|1 + \frac{b^{(k)2}}{a^{(k)2}}\right| - \frac{bb^{(k)}}{a^{(k)2}} - \frac{1}{2}c.$$

证明: 由上述结果可对 EM算法中 E步进行计算:

$$E(\ln(L(\alpha)) | z_i^-, \alpha^{(k)}) =$$

$$\int_{-\infty}^{\infty} \left| -(n+1)\ln(\sqrt{2\pi}) + \sum_{j=0}^{n} \left| -\frac{(z_j - \gamma_1 z_{j-1} - \cdots - \gamma_p z_{j-p})^2}{2}\right| \right| a^{(k)} \left|\frac{1}{\sqrt{2\pi}}\right|^{-p} \cdot$$

$$\exp\left|\frac{c^{(k)} - (b^{(k)}/a^{(k)})^2}{2}\right| \cdot \prod_{j=i}^{i+p} f(z_j - \gamma_1^{(k)} z_{j-1} - \cdots - \gamma_p^{(k)} z_{j-p})\,\mathrm{d}z_i \triangleq$$

$$I_1 + I_2 + I_3 + I_4, \tag{2.5}$$

其中

$$I_1 = -\int_{-\infty}^{\infty} [(n+1)\ln(\sqrt{2\pi}) F(z_i^-, \alpha^{(k)})]\,\mathrm{d}z_i,$$

$$I_2 = \sum_{j=0}^{i-1} \int_{-\infty}^{\infty} \left| -\frac{(z_j - \gamma_1 z_{j-1} - \cdots - \gamma_p z_{j-p})^2}{2}\right| F(z_i^-, \alpha^{(k)})\,\mathrm{d}z_i,$$

$$I_3 = \sum_{j=i+p+1}^{n} \int_{-\infty}^{\infty} \left| -\frac{(z_j - \gamma_1 z_{j-1} - \cdots - \gamma_p z_{j-p})^2}{2}\right| F(z_i^-, \alpha^{(k)})\,\mathrm{d}z_i,$$

$$I_4=\sum_{j=i}^{i+p}\int_{-\infty}^{\infty}\left|-\frac{(z_j-\theta_1z_{j-1}-\cdots-\theta_pz_{j-p})^2}{2}\right|F(z_i^-,\alpha^{(k)})\,\mathrm{d}z_i,$$

这里

$$F(z_i^-,\alpha^{(k)})=a^{(k)}\exp\left|\frac{c^{(k)}-(b^{(k)}/a^{(k)})^2}{2}\right|\left|\frac{1}{\sqrt{2\pi}}\right|^{-p}\prod_{j=i}^{i+p}f(z_j-\theta_1^{(k)}z_{j-1}-\cdots-\theta_p^{(k)}z_{j-p}).$$

下面分别计算 I_1,I_2,I_3,I_4. 由前面的计算易得

$$I_1=-(n+1)\ln\sqrt{2\pi},$$

$$I_2=\sum_{j=0}^{i-1}\left|-\frac{(z_j-\theta_1z_{j-1}-\cdots-\theta_pz_{j-p})^2}{2}\right|=\ln D_{i-1;0},$$

$$I_3=\sum_{j=i+p+1}^{n}\left|-\frac{(z_j-\theta_1z_{j-1}-\cdots-\theta_pz_{j-p})^2}{2}\right|=\ln D_{n;i+p+1},$$

$$I_4=\int_{-\infty}^{\infty}\left|-\frac{1}{2}(z_i-\theta_1z_{i-1}-\cdots-\theta_pz_{i-p})^2\right|\exp\left|-\frac{c^{(k)}-(b^{(k)}/a^{(k)})^2}{2}\right|\cdot$$

$$a^{(k)}\left|\frac{1}{\sqrt{2\pi}}\right|\exp\left|-\frac{(a^{(k)}z_i+b^{(k)}/a^{(k)})^2+c^{(k)}-(b^{(k)}/a^{(k)})^2}{2}\right|\mathrm{d}z_i+$$

$$\sum_{k=1}^{p}\int_{-\infty}^{\infty}\left|-\frac{1}{2}(z_{i+k}-\theta_1z_{i+k-1}-\cdots-\theta_pz_{i+k-p})^2\right|\cdot$$

$$a^{(k)}\left|\frac{1}{\sqrt{2\pi}}\right|\exp\left|-\frac{(a^{(k)}z_i+b^{(k)}/a^{(k)})^2}{2}\right|\mathrm{d}z_i\triangleq I_4^0+\sum_{k=1}^{p}I_4^k.$$

对于 $1\leqslant k\leqslant p$,

$$I_4^k=\int_{-\infty}^{\infty}\left|-\frac{\theta_k^2}{2}z_i^2+\theta_kz_i(z_{i+k}-\theta_1z_{i+k-1}-\cdots-\theta_{k-1}z_{i+1}-\theta_{k+1}z_{i-1}-\cdots-\theta_pz_{i+k-p})-\right.$$

$$\left.\frac{1}{2}(z_{i+k}-\theta_1z_{i+k-1}-\cdots-\theta_{k-1}z_{i+1}-\theta_{k+1}z_{i-1}-\cdots-\theta_pz_{i+k-p})^2\right|\cdot$$

$$a^{(k)}\exp\left|-\frac{1}{2}\left|a^{(k)}z_i+\frac{b^{(k)}}{a^{(k)}}\right|^2\right|\left|\frac{1}{\sqrt{2\pi}}\right|\mathrm{d}z_i\triangleq l_1+l_2+l_3.$$

又由于

$$l_1=-\frac{1}{2}\int_{-\infty}^{\infty}\theta_k^2z_i^2a^{(k)}\exp\left|-\frac{1}{2}\left|a^{(k)}z_i+\frac{b^{(k)}}{a^{(k)}}\right|^2\right|\left|\frac{1}{\sqrt{2\pi}}\right|\mathrm{d}z_i=-\frac{\theta_k^2}{2a^{(k)^2}}-\frac{b^{(k)^2}\theta_k^2}{2a^{(k)4}},$$

$$l_2=\int_{-\infty}^{\infty}\theta_kz_i(z_{i+k}-\theta_1z_{i+k-1}-\cdots-\theta_{k-1}z_{i+1}-\theta_{k+1}z_{i-1}-\cdots-$$

$$\theta_pz_{i+k-p})a^{(k)}\exp\left|-\frac{1}{2}\left|a^{(k)}z_i+\frac{b^{(k)}}{a^{(k)}}\right|^2\right|\left|\frac{1}{\sqrt{2\pi}}\right|\mathrm{d}z_i=$$

$$-\frac{b^{(k)}\theta_k}{a^{(k)^2}}(z_{i+k}-\theta_1z_{i+k-1}-\cdots-\theta_{k-1}z_{i+1}-\theta_{k+1}z_{i-1}-\cdots-\theta_pz_{i+k-p}),$$

$$l_3=-\frac{1}{2}(z_{i+k}-\theta_1z_{i+k-1}-\cdots-\theta_{k-1}z_{i+1}-\theta_{k+1}z_{i-1}-\cdots-\theta_pz_{i+k-p})^2,$$

I_4^0求法同 I_4^k, 结合以上各式, 可知

$$I_4=\sum_{k=0}^{p}I_4^k=-\frac{a^2}{2a^{(k)^2}}\left|1+\frac{b^{(k)^2}}{a^{(k)^2}}\right|-\frac{bb^{(k)}}{a^{(k)^2}}-\frac{1}{2}c,$$

故定理 2.2成立.

当 $p=2$时, EM算法的 E步是二元参数 $\alpha'=(\theta_1,\theta_2)$的函数, 设此函数为 $g(\theta_1,\theta_2)$, 整理可得

$$g(\theta_1,\theta_2)=E(\ln(L(\alpha))|z_i^-,\alpha^{(k)})=$$

$$-(n+1)\ln\sqrt{2\pi}-\frac{1}{2}(m_1\theta_1^2+m_2\theta_2^2-2n_1\theta_1-2n_2\theta_2+2h\theta_1\theta_2+R),\qquad(2.6)$$

其中

$$K = (1/a^{(k)^2})(1 + b^{(k)^2}/a^{(k)^2}), \quad L = b^{(k)}/a^{(k)^2},$$
$$m_1 = z_{-1}^2 + z_0^2 + \cdots + z_{i-1}^2 + z_{i+1}^2 + \cdots + z_{n-1}^2 + K,$$
$$m_2 = z_{-2}^2 + z_{-1}^2 + \cdots + z_{i-1}^2 + z_{i+1}^2 + \cdots + z_{n-2}^2 + K,$$
$$n_1 = z_{-1}z_0 + \cdots + z_{i-2}z_{i-1} + z_{i+1}z_{i+2} + \cdots + z_{n-1}z_n - Lz_{i-1}/2 - Lz_{i+1}/2, \tag{2.7}$$
$$n_2 = z_{-2}z_0 + \cdots + z_{i-3}z_{i-1} + z_{i-1}z_{i+1} + \cdots + z_{n-2}z_n - Lz_{i-2}/2 - Lz_{i+2}/2,$$
$$h = z_{-1}z_{-2} + z_0z_{-1} + \cdots + z_{i-1}z_{i-2} + z_{i+1}z_{i+2} + \cdots + z_{n-1}z_{n-2} - Lz_{i-1}/2 - Lz_{i+1}/2,$$
$$R = z_0^2 + z_1^2 + \cdots + z_{i-1}^2 + z_{i+1}^2 + \cdots + z_n^2 + K.$$

(1) 为了求 $g(T_1, T_2)$的最大值点, 首先应求 $g(T_1, T_2)$的一阶偏导并令其为 0, 然后计算 Hessian阵行列式, 证得 $h^2 - m_1m_2 < 0$, 进而可求得 $g(T_1, T_2)$的极大值点.

事实上, 设

$$s^- = (z_{-1}, z_0, \cdots, z_{i-1}, -L/2, z_{i+1}, \cdots, z_{n-1})',$$
$$s^+ = (z_{-2}, z_{-1}, \cdots, z_{i-2}, z_{i-1}, -L/2, z_{i+1}, \cdots, z_{n-2})',$$

则 $h = s^{-\prime} \circ s^+$,

$$h^2 = (s^- \cdot s^+)^2 \leqslant \| s^- \|^2 \| s^+ \|^2 \leqslant m_1 \cdot m_2,$$

故只需求 T_1, T_2, 满足 $\partial g/\partial T_1 = 0$, $\partial g/\partial T_2 = 0$, 解得

$$T_1^{(k+1)} = \frac{m_2n_1 - hn_2}{m_1m_2 - h^2}, \quad T_2^{(k+1)} = \frac{m_1n_2 - n_1h}{m_1m_2 - h^2}. \tag{2.8}$$

此时给出了当给定 $\alpha^{(k)}$, 可求出新的 $\alpha^{(k+1)}$, 使 $g(T_1, T_2)$达到最大.

算法具体步骤如下: 1) 给出初值 $\alpha^{(0)} = (T_1^{(0)}, T_2^{(0)})$; 2) 利用样本与 $\alpha^{(0)}$计算 (2.7)式中各式值; 3) 计算 (2.8)式求得 $\alpha^{(1)}$; 4) 以 $\alpha^{(1)}$为初值, 从 2)开始循环, 直至达到预先给定的精度为止.

(2) 当有多个参数待估, 且其目标函数满足一定条件, 此时为使函数值达到最大, 可先固定其余 $n-1$个参数 $(T_2, \cdots, T_n)$, 只求 T_1^* 使函数 $f(T_1, \cdots, T_n)$达到最大. 然后再固定 $T_1^*, T_3, \cdots, T_n$, 求 T_2^* 使 $f(T_1^*, T_2, \cdots, T_n)$达到最大. 依此类推, 可求出 $(T_1^*, \cdots, T_n^*)$, 然后再重复上述方法, 直到稳定为止.

由于 $g(T_1, T_2)$是 (T_1, T_2)的凸函数, 于是可采用上述方法求 $g(T_1, T_2)$的最大值点. 固定 T_2时, 由于

$$g(T_1, T_2) = -(n+1)\ln\sqrt{2\pi} - \frac{1}{2}\left| \left| \sqrt{m_1}T_1 + \frac{hT_2 - n_1}{\sqrt{m_1}} \right|^2 - \frac{(hT_2 - n_1)^2}{m_1} + m_2T_2^2 - 2n_2T_2 + c \right|,$$

令 $\partial g(T_1, T_2)/\partial T_1 = 0$, 解得

$$T_1^* = -(hT_2 - n_1)/m_1, \tag{2.9}$$

于是, $g(T_1^*, T_2) = \max\limits_{T_1} g(T_1, T_2)$. 再固定 T_1^*, 同样可以得到 $g(T_1^*, T_2^*) = \max\limits_{T_2} g(T_1, T_2)$, 其中

$$T_2^* = -(hT_1^* - n_2)/m_2. \tag{2.10}$$

算法具体步骤如下: 1) 给出初值 $(T_1^{(0)}, T_2^{(0)})$; 2) 计算 (2.7)式中各式值; 3) 计算 (2.9), (2.10)式求得 $\alpha^* = (T_1^*, T_2^*)$; 4) 以 $\alpha^* = (T_1^*, T_2^*)$代替 $\alpha^{(0)}$重复步骤 2), 3), 直至达到预先给定的精度为止.

3 连续两个数据缺失时的估计方法

当连续两个数据 (z_i, z_{i-1})缺失, 仍用前述方法求 α 估计值时, 由于 Hessian阵行列式值符号不定, 这时, M步得到的解未必是全局极大值点, 但是方法 (2)可以保证每步的极大值是递增的.

在连续两个数据缺失时, 为方便, 记

$$W = -T_1 + T_0T_1 + \cdots + T_{p-1}T_p, \quad e = b - Wz_{i-1},$$
$$A^2 = a^2 - \frac{1}{a^2}W, \quad B = \sum_{k=0}^{p} T_kp_k - \frac{W}{a^2}e, \quad C = \sum_{k=0}^{p} p_k^2 - \frac{e^2}{a^2} + A_p^2,$$
$$p_0 = (T_1, T_2, \cdots, T_p)(z_{i-2}, \cdots, z_{i-p-1})',$$
$$p_k = (T_0, T_1, \cdots, T_{k-2}, T_{k+1}, \cdots, T_p)(z_{i+k-1}, \cdots, z_{i+1}, z_{i-2}, \cdots, z_{i+k-p-1})', \quad k = 1, \cdots, p.$$

定理 **3.1** 在 z_i, z_{i-1}缺失的情况下, 有

$$\tilde{L}(\alpha)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}L(\alpha)\,\mathrm{d}z_{i-1}\mathrm{d}z_i=$$
$$\left|\frac{1}{\sqrt{2\pi}}\right|^{n-1}\frac{1}{aA}\exp\left|-\frac{1}{2}\left|C-\frac{B^2}{A^2}\right|\right|\mathrm{d}z_{i-1}D_{i-2;0}D_{n;i+p+1},\tag{3.1}$$

其中

$$D_{i-2;0}=\prod_{j=0}^{i-2}\exp\left|-\frac{1}{2}(z_j-\varphi_1z_{j-1}-\cdots-\varphi_pz_{j-p})^2\right|.$$

证明与一个数据缺失的方法类似.

设 $(z_{i-1},z_i)^-=\{z_0,\cdots,z_n\}\setminus\{z_{i-1},z_i\}$, $e^{(k)}$, $\omega^{(k)}$函数形式同前面 e,ω为第 k步计算得到的值, $\varphi_1=0$. 由上面的结论, 可得到(z_{i-1},z_i)关于$(z_{i-1},z_i)^-$的条件概率密度为

$$f(z_{i-1},z_i|\,(z_{i-1},z_i)^-,\alpha^{(k)})=\left|\frac{1}{\sqrt{2\pi}}\right|^{-p+1}F(z)\cdot a^{(k)}A^{(k)}\exp\left|\frac{C^{(k)}-B^{(k)2}/A^{(k)2}}{2}\right|\cdot$$
$$\prod_{j=i-1}^{i+p}f(z_j-\varphi_1^{(k)}z_{j-1}-\cdots-\varphi_p^{(k)}z_{j-p}).$$

定理 **3.2**

$$E(\ln(L(\alpha))|\,(z_i,z_{i-1})^-,\alpha^{(k)})=-(n+1)\ln\sqrt{2\pi}+\ln D_{i-2;0}+\ln D_{n;i+p+1}+D,$$

其中

$$D=-\frac{1}{2}U_1a^2-\frac{1}{2}\sum_{k=0}^{p}p_k^2-\frac{1}{2}A_p^2-\frac{B^{(k)}}{A^{(k)2}}\sum\varphi_kp_k-U_2\sum\varphi_{k-1}p_k-U_3\sum\varphi_k\varphi_{k-1}-U_4\varphi_pA_p,$$
$$U_1=\frac{1}{A^{(k)2}}\left|1+\frac{B^{(k)2}}{A^{(k)2}}\right|+\frac{1}{a^{(k)4}}\left|1+\frac{B^{(k)2}}{A^{(k)2}}\right|\frac{1}{A^{(k)2}}\omega^{(k)2}+\frac{1}{a^{(k)2}}+\frac{e^{(k)2}}{a^{(k)4}}-2\frac{e^{(k)}}{a^{(k)4}}\frac{B^{(k)}}{A^{(k)2}}\omega^{(k)},$$
$$U_2=\frac{e^{(k)}}{a^{(k)2}}+\frac{\omega^{(k)}}{a^{(k)2}}\frac{B^{(k)}}{A^{(k)2}},$$
$$U_3=\frac{1}{a^{(k)2}}\frac{1}{A^{(k)2}}\left|-e^{(k)}B^{(k)}+\left|1+\frac{B^{(k)2}}{A^{(k)2}}\right|\omega^{(k)}\right|,$$
$$U_4=\frac{e^{(k)}}{a^{(k)2}}-\frac{\omega^{(k)}}{a^{(k)2}}\frac{B^{(k)}}{A^{(k)2}}.$$

证明与定理 2.2类似.

特别地, 当 $p=2$时, EM算法的 E步是关于二元参数 $\alpha=(\varphi_1,\varphi_2)$的函数, 设此函数为 $\tilde{g}$, 整理得

$$\tilde{g}(\varphi_1,\varphi_2)=E[\ln(L(\alpha))|\,(z_i,z_{i-1})^-,\alpha^{(k)}]=-(n+1)\ln\sqrt{2\pi}-\frac{1}{2}\tilde{m}_1\varphi_1^2-\frac{1}{2}\tilde{m}_2\varphi_2^2-$$
$$\tilde{h}\varphi_1\varphi_2+\tilde{n}_1\varphi_1+\tilde{n}_2\varphi_2-\frac{1}{2}\tilde{R},\tag{3.2}$$

其中

$$\begin{aligned}
\tilde{m}_1&=z_{-1}^2+\cdots+z_{i-2}^2+z_{i+1}^2+\cdots+z_{n-1}^2+U_1,\\
\tilde{m}_2&=z_{-2}^2+\cdots+z_{i-2}^2+z_{i+1}^2+\cdots+z_{n-2}^2+U_1,\\
\tilde{h}&=z_{-2}z_{-1}+\cdots+z_{i-4}z_{i-3}+z_{i-3}z_{i-2}+z_{i+2}z_{i+1}+\\
&\quad\cdots+z_{n-2}z_{n-1}+(B^{(k)}/A^{(k)2})z_{i-2}+U_4z_{i+1}+U_3,\\
\tilde{n}_1&=z_{-1}z_0+z_0z_1+\cdots+z_{i-3}z_{i-2}+z_{i+1}z_{i+2}+\\
&\quad\cdots+z_{n-1}z_n+(B^{(k)}/A^{(k)2})z_{i-2}+U_2z_{i+1}+U_3,\\
\tilde{n}_2&=z_{-2}z_0+\cdots+z_{i-4}z_{i-2}+z_{i+1}z_{i+3}+\cdots+z_{n-2}z_n+\\
&\quad(B^{(k)}/A^{(k)2})z_{i-3}+(B^{(k)}/A^{(k)2})z_{i+1}+U_2z_{i-2}+U_4z_{i+2},\\
\tilde{R}&=z_0^2+\cdots+z_{i-2}^2+z_{i+1}^2+\cdots+z_n^2+U_1^2,
\end{aligned}\tag{3.3}$$

由于不能确定是否有 $\tilde{h}^2<\tilde{m}_1\tilde{m}_2$, 故可采用方法(2)求 $\tilde{g}(\varphi_1,\varphi_2)$的极大值点 (φ_1,φ_2). 为此固定 φ_2, 对

$\tilde{g}(\theta_1,\theta_2)$关于 θ_1 求偏导并令其为零，即 $\tilde{g}'(\theta_1,\theta_2)=-\tilde{m}_1\theta_1-\gamma\theta_2+\tilde{n}_1=0$，解得

$$\theta_1=(-\gamma\theta_2+\tilde{n}_1)/\tilde{m}_1. \tag{3.4}$$

同理，固定 θ_1 时，对 $\tilde{g}(\theta_1,\theta_2)$关于 θ_2 求偏导并令其为零，有 $\tilde{g}'(\theta_1,\theta_2)=-\tilde{m}_2\theta_2-\gamma\theta_1+\tilde{n}_2=0$，解得

$$\theta_2=(-\gamma\theta_1+\tilde{n}_2)/\tilde{m}_2. \tag{3.5}$$

具体算法如下：1) 给出初值 $(\theta_1^{(0)},\theta_2^{(0)})$；2) 利用样本计算(3.3)式中各式值；3) 计算(3.4)，(3.5)式，求得 $\alpha^*=(\theta_1,\theta_2)$；4) 以 3)中结果 $\alpha^*=(\theta_1,\theta_2)$代替 1)中 $\alpha^{(0)}$，重复步骤 2)，3)，直至达到预先给定的精度为止.

4 模拟结果与实例应用

以服从 AR(2)模型的随机数作为样本，当设其中一个数据缺失时，分别利用上述两种方法对AR(2)模型的系数进行估计. 对这两种方法，其真值与得到的估计值分别列于表 1和表 2.

Table 1 List of different true values and their respective estimators based on the method (1)

True value		Estimator	
θ_1	θ_2	$\hat{\theta}_1$	$\hat{\theta}_2$
0.60	0.20	0.563 321	0.238 665
0.50	−0.1	0.537 486	−0.183 769
0.50	0.10	0.481 288	0.012 827
−0.1	0.40	−0.124 946	0.318 359

通过实例可验证方法(1)收敛结果与初值无关.

Table 2 List of different true values, initial values and their respective estimators based on the method (2)

True value		Primitive value		Estimator	
θ_1	θ_2	$\tilde{\theta}_1$	$\tilde{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
0.60	0.20	1.2	0.50	0.537 398	−0.135 910
0.60	0.20	0.2	0.40	0.525 589	−0.109 288
0.60	0.20	0.4	−0.40	0.535 887	−0.150 369
0.50	−0.1	1.2	0.50	0.488 410	−0.230 279
0.50	−0.1	0.2	0.40	0.488 367	−0.226 650
0.50	−0.1	0.4	−0.40	0.489 162	−0.229 817

通过比较数据可知，利用方法(1)得到的估计值相对真值的误差与利用方法(2)得到的估计值相对真值的误差相差不大. 其中方法(1)中 $\hat{\theta}_1$ 的方差为 0.001 7，$\hat{\theta}_2$ 的方差为 0.005 7；方法(2)中 $\hat{\theta}_1$ 的方差为 0.002 3，$\hat{\theta}_2$ 的方差为 0.063 5，而且两种方法均无很强的关于初值的依赖性. 但相对而言，方法(1)的稳定性较强.

在股票交易中，各地指数之间有很强的关联关系，本文以 DOW指与恒生指数为例，利用二者从 2001.9.27~2001.12.21的数据建立 AR(2)模型，并就在一个样本缺失时，对模型系数进行估计.

设 Dow指和恒生指数的时间序列数据分别为 $y_t,x_t(t=-p,\cdots,0,1,\cdots,n)$. 通过观察发现，二者取对数差后趋于稳定.

设

$$O_t=\ln x_t-\ln y_t,\qquad t=-p,\cdots,0,1,\cdots,n.$$

对 O_t 中心化，令 $z_t=O_t-\bar{O}$，其中 $\bar{O}=\frac{1}{n+p+1}\sum_{t=1}^{n}O_t$，$t=-p,\cdots,0,\cdots,n$. 我们关心 z_t 随时间变化规律.

当 z_t 服从 AR(p)模型，即

$$z_t=\theta_1 z_{t-1}+\cdots+\theta_p z_{t-p}+X_t,\qquad X_t\overset{\text{i.i.d}}{\sim}N(0,1). \tag{4.1}$$

需估计出以上模型系数.

利用方法(1)和方法(2)分别建立模型($p=2$). 利用方法(1)得，$\hat{\theta}_1=0.041\ 27$，$\hat{\theta}_2=-0.004\ 2$，此

收敛值与初值无关; 对于方法(2), 当初值不同时, $\hat{\alpha}$ 不尽相同. 结合以上模拟与实例应用的情况, 可以看出方法(2)对初值依赖较强.

衷心感谢宋立新教授的指导.

参 考 文 献

[1] Li Zi-nai(李子奈). Econometrics(计量经济学)[M]. Beijing(北京): Higher Education Press(高等教育出版社), 2000. 21~ 119.

[2] An Hong-zhi(安鸿志). Time Series Analysis(时间序列分析)[M]. Shanghai(上海): Huadong Normal University Press(华东师范大学出版社), 1992. 18~ 31.

[3] Guo Chun-guang(国春光), Lin Zheng-hua(林正华), Lü Xian-rui(吕显瑞). Least Squares Estimation for the Parameter of Autoregressive Models of the Perturbation Term(扰动项序列自回归模型参数的最小二乘估计)[J]. *Acta Scientiarum Naturalium Universitatis Jilinensis*(吉林大学自然科学学报), 1998, (1): 1~ 4.

[4] Lin Zheng-hua(林正华), Feng Ren-zhong(冯仁忠). A Least Square Estimation of Autoregressive Processes(自回归模型参数的最小二乘估计)[J]. *Acta Scientiarum Naturalium Universitatis Jilinensis*(吉林大学自然科学学报), 2001, (2): 1~ 4.

[5] Mao Shi-song(茆诗松). Advanced Statistics(高等数理统计)[M]. Beijing(北京): Higher Education Press(高等教育出版社), 1987. 428~ 443.

[6] Wu C F J. On the Convergence Properties of the EM Algorithm[J]. *The Annals of Statistics*, 1983, **11**: 95~ 103.

Estimation Method of AR(p) Model with One Date or Two Continuous Data Missed

TIAN Ping[1], DONG Xian-feng[2], WANG De-hui[2], HUANG Xiao-wei[2]

(1. *College of Business, Jilin University, Changchun* 130012;
2. *College of Mathematics, Jilin University, Changchun* 130012)

Abstract: The present paper deals with the estimation problem for zero mean value AR(p) model with one date or two continuous data missed. The paper is based on the EM algorithm. By using the iterative method, we have obtained the estimated values of parameters.

Keywords: likihood function; EM algorithm; conditional probable density

(责任编辑: 赵立芹)