第55卷 第5期
1997年10月

气 象 学 报
ACTA METEOROLOGICA SINICA

Vol. 55, No. 5
Oct., 1997

非线性多元样条回归预报模型

严华生 曹 杰 谢应齐

(云南大学地球科学系, 昆明, 650091)

尤 卫红

(云南省气象台, 昆明, 650034)

摘 要

针对异常复杂的非线性预报系统, 提出非线性多元样条回归预报模型。该模型既保留了线性统计预报中多元分析, 逐步筛选预报因子及显著性检验的理论和方法, 又吸取了样条函数分段拟合, 按需要裁剪以适应任意曲线连续变化的优点, 具有处理复杂非线性预报系统的功能。试验结果表明, 该模型具有良好的模拟和预报能力, 值得进一步深入探讨和应用。

关键词: 非线性, 统计预报, 样条函数。

1 引 言

在统计预报中, 过去多采用线性模型。但由于客观世界中普遍存在的非线性和复杂性, 往往难以用线性模型来研究, 于是很有必要发展非线性预报方法。非线性与线性相比, 显然要复杂得多, 有时甚至复杂到难以用初等函数来表示。无论用动力学还是统计学方法要想建立一个较好描述非线性复杂变化特征的预报模型, 都显得困难重重。文中尝试性地应用多项式样条函数来拟合非线性预报系统, 为在预报因子和预报对象间的非线性预报关系复杂到难以表示出来的情况下, 探索性地提供一条解决问题的新途径。

2 样条函数回归模型

设有预报对象 Y 和预报因子 X, 存在非线性预报函数关系, 记为 $Y = f(X)$, 当 $f(X)$ 为未知函数时, 不宜用 X 的高次多项式来逼近未知函数。这是因为当高次多项式描述的曲线在一个小区间上被迫弯曲时, 它在别处就可能剧烈震荡, 即产生龙格(Runge)现象[1, 2]。这就是说, 高次多项式虽然可以使某个区间范围内的拟合残差减小, 但在区间以外, 则往往产生不合理的波动和扭曲, 所以不宜用高次多项式来作为预报拟合模型。但如果不是用一个任意光滑的高次多项式去拟合逼近, 而是用一组分段光滑的低次多项式去拟合逼近它, 就可避免这一问题, 且拟合效果好, 同时也适合作外推预报。文中提出的多元样条逐步回归模型, 其实质就是用这种称之为多项式样条的光滑对接分段多项式来作为

初稿时间: 1996年5月30日; 修改稿时间: 1996年9月2日。

预报拟合逼近函数。

设给定函数点 (x_i, y_i), $i = 1, \cdots, n$。首先在 X 的取值区间$[x_{\min}, x_{\max}]$ 中插入 L 个分点, 称为结点, 把该值区间分为 $L + 1$ 个子区间, 记为: $[x_{min}, r_1]$, $[r_1, r_2]$, $\cdots$, $[r_L, r_{\max}]$, 在每个子区间内, 分别用一个 X 的低次多项式来拟合逼近它, 于是在整个区间上, 拟合函数是一组分段光滑对接的低次多项式, 它的形式如下:

$$y = f(x) = \begin{cases} f_1(x) = a_{1,0} + a_{1,1}x + \cdots + a_{1,p}x^p & x \in [x_{\min}, r_1] \\ \cdots\cdots \\ f_{L+1}(x) = a_{L+1,0} + a_{L+1,1}x + \cdots + a_{L+1,p}x^p & x \in [r_L, x_{\max}] \end{cases} \tag{1}$$

为了使式(1) 简化, 引入半截多项式:

$$u_+^m = \begin{cases} u^m & u > 0 \\ 0 & u \le 0 \end{cases} \tag{2}$$

对 X 轴上给定的一组分点: $r_1, r_2 \cdots, r_L$, 从半截多项式出发, 对它进行移位, 得到一组以 r_j 为跳跃点的半截多项式$(x - r_j)_+^p$, 把它加到一个 X 的 p 次多项式 $\sum_{i=0}^{p} a_i x^j$ 中去, 得到函数

$$S(x) = \sum_{i=0}^{p} a_i x^i + \sum_{j=1}^{L} c_j (x - r_j)_+^p \tag{3}$$

式(3) 称为多项式样条函数, 它具有如下性质:

(1) 在每个子区间内, 它是一个 p 次多项式:

$$S(x) = \begin{cases} S_1(x) = \sum_{i=0}^{p} a_i x^i & x \le r_1 \\ S_2(x) = S_1(x) + c_1(x - r_1)_+^p & x \le r_2 \\ S_{L+1}(x) = S_1(x) + \sum_{j=1}^{L} c_j (x - r_j)_+^p & x \le x_{\max} \end{cases} \tag{4}$$

式(4) 与式(1) 是等价的, 可证明如下:

首先, 在第一区间内等价, 其次在第 j 区间内对式(4) 中 $S(X)$ 的半截多项式展开, 合并 同类项, 与式(1) 比较系数, 即可证明其等价性。不失一般性, 以 $j = 2$ 为例, 对式(4) 中半截多项式展开得:

$$c_1(x - r_1)_+^p = c_1 \sum_{k=0}^{p} d_k x^k r_1^{p-k}$$

于是有: $S_2(x) = \sum_{k=0}^{p} (c_1 d_k r_1^{p-k} + a_k) x^k$, 与式(1) 比较系数得: $a_{2,k} = c_1 d_k r_1^{p-k} + a_k$。

(2) 在分点处 p 阶导数存在并连续。

由以上两点可知, 多项式样条函数由于引进了半截多项式, 成为光滑对接的分段多项式, 用这样的样条函数去拟合逼近一个非线性任意函数, 自然具有更大的转折自如的灵活性, 因此, 适合在非线性复杂预报系统中选作预报拟合逼近函数。

当预报因子有多个时, 设多元样条空间为线性空间的线性组合, 则多元样条函数可写为:

$$S(x_1,\cdots,x_m)=\sum_{k=1}^{m}S(x_k)=\sum_{i=1}^{m}\sum_{j=0}^{p}a_{i,j}x_i^j+\sum_{k=1}^{m}\sum_{k=1}^{L}c_k^{(i)}(x_i-r_k^{(i)})_+^p \tag{5}$$

更一般的多元多项式样条函数可定义为张量积空间的张量积形式[3]

$$S(x_1,\cdots,x_m)=S(x_1)\quad S(x_2)\quad\cdots\quad S(x_m) \tag{6}$$

文中仅考虑在线性空间的多元多项式样条回归模型。在非线性系统中, 由于有误差存在, 因此只需要样条函数具有一定阶数的可微性就可以了, 一般p 取1—3次即可。其中$p=1$ 可称为非线性预报系统的分段线性逼近, $p>1$ 称为分段多项式逼近。

3 建模方法

当给定预报对象 Y 和 m 个预报因子 X 的n 个样本后, 根据样本资料选择性地确定式(5) 中的待估参数: $a_{i,j}, c_k^{(i)}, r_k^{(i)}$, 称为多元样条逐步回归模型建模。

当样本数 n 不足够大时, 尽量采用较低阶数的多项式和较少的分点, 可较好避免龙格现象和减少自由度, 从而增加预报模型的预报能力。

具体建模步骤为:

(1) 确定分点

对给定的预报因子x_i的 n 个样本数据, 对 x_i 按从小到大的顺序排列, 得到新序列x_i', 于是可得到样本区间所有可能分点 $L, L=n-1$。

(2) 生成逐步样条回归备选预报因子集。

当分点确定后, 按式(7) 作变量替换, 得到样条逐步回归备选因子集, 若令:

$$\begin{aligned}
z_1^{(i)}&=x_i\\
z_2^{(i)}&=x_i^2\\
&\cdots\cdots\\
z_p^{(i)}&=x_i^p\\
z_{p+1}^{(i)}&=(x_i-r_1^i)_+^p\\
&\cdots\cdots\\
z_{p+L}^{(i)}&=(x_i-r_L^{(i)})_+^p
\end{aligned} \tag{7}$$

则得因子集, $Z=[z_j^{(i)}]_{j=1,\cdots p+L}^{i=1,\cdots m}$, 总因子数为$(p+L)m$ 个。样条全回归形式为:

$$y=AZ^T \tag{8}$$

(3) 引入逐步回归技术[4]

对因子集 $[z_j^{(i)}]$ 进行筛选, 则可从中自动挑选出最优分点, 合并某些不必要的分段。例如: 若$(x_i-r_4^{(i)})_+^p$ 这一因子在逐步回归中没有入选, 则可认为分点$r_4^{(i)}$ 可以去掉, 即把第 i 个因子的第 4 个子区间$[r_3^{(i)}, r_4^{(i)}]$ 与第 5 个子区间$[r_4^{(i)}, r_5^{(i)}]$ 合并。这就是说, 在多元样条回归中, 应用逐步回归筛选技术, 具有自动挑选最优预报因子、最优分点和最优表达式的能力。最终建立多元样条逐步回归模型为:

$$y'=A'Z'^T \tag{9}$$

(4) 模型检验

建立预报模型后, 可计算复相关系数或方差分析对所建的预报方程进行显著性检验, 以判断模型的性能[4]。

当给出新的观测资料后, 按逐步回归所确定的预报因子及结点分段, 生成新的预报因子集, 即可作出外推预报。

4　应用实例

文中分别用蒙自、昆明、大理3个站1951—1990年5月降水资料作为预报对象, 用1951—1990年1—3月的西北太平洋副高强度指数、极涡中心强度指数、亚洲经向和纬向环流指数、太阳黑子数、乌拉尔山地区平均高度、印缅地区5点高度和等共21个变量作为预报因子。取 $p = 2$, 按前述建模步骤建立蒙自、昆明、大理3个站5月降水的多元样条回归预报模型如下:

$$y'_{985} = -0.05156x_1^2 - 0.59446(x_4 - (-5))_+^2 + 0.00015(x_{14} - 21)_+^2 - 8.9098 \quad (10)$$

$$y'_{778} = -2.92522(x_3 - 32)_+^2 + 0.00172(x_{16} - 452)_+^2 - 0.38579(x_{20} - 13)_+^2 + 0.47714(x_{21} - 25)_+^2 + 26.07025 \quad (11)$$

$$y'_{751} = 0.13160(x_7 - 44)_+^2 + 0.07443(x_9 - 55)_+^2 + 0.00365(x_{11} - 110)_+^2 + 0.0043(x_{11} - 189)_+^2 - 117.12 \quad (12)$$

其复相关系数分别为: $R_{985} = 0.72$, $R_{778} = 0.72$, $R_{751} = 0.79$。均通过 $= 0.001$ 水平的显著性检验。

作为对比, 建立3个站的线性逐步回归方程, 为了使其具有可比性, 采用控制入选剔除 F 值的方法, 使最终入选得到的最优线性回归方程包含与样条回归同样多的因子项数(虽然二者因子形式不同)。得到的线性回归模型如下:

$$y'_{985} = -2.0505x_8 - 1.9282x_9 + 0.38771x_{18} - 95.9 \quad (13)$$

$$y'_{778} = -1.3958x_3 + 4.6608x_5 - 4.8357x_6 + 0.3662x_{18} - 108.36 \quad (14)$$

$$y'_{751} = -1.2134x_3 + 3.3176x_5 - 2.3135x_6 + 2.3928x_9 - 123.67 \quad (15)$$

其复相关系数分别为: $R_{985} = 0.46$, $R_{778} = 0.56$, $R_{751} = 0.43$。通过 $= 0.05$ 水平的显著性检验。

对线性回归和样条回归分别进行历史拟合检验, 样条回归式(10)—(12)的趋势准确率分别为: 35/40, 30/40, 34/40; 线性回归式(13)—(15)的趋势准确率分别为: 29/40, 26/40, 25/40。

对线性回归和样条回归分别作1991—1994年独立样本外推预报检验, 结果样条回归3站的趋势准确率均为3/4, 而线性回归3站的趋势准确率均为2/4。由以上比较可看出, 样条回归比线性回归有较高的历史拟合和外推预报能力。

5　应用问题讨论

(1) 线性与非线性

由半截多项式 $c(x - a)_+^2$ 的导数不变号可以证明它是单调递增或递减函数, 在 $x \geq a$ 的区间对 $c(x - a)_+^2$ 进行展开得 $cx^2 - 2cax + ca^2$。由此可见, 半截多项式既包含非线性函数关系, 又包含了线性函数关系; 再根据建模步骤可知, 参数 a 是首先确定的结点而不是统计估计出来的, 所以实际上半截多项式只包含一个待估参数 c。于是可以认为, 非线性

多元样条回归模型包含了线性回归模型的模拟和预报能力，因而非线性多元样条回归模型与入选因子数相同线性回归模型的相比，更具优越性。下面通过实例计算来进一步验证这一结论。

取文中蒙自应用实例所得样条回归模型与实测值的相关系数0.72；线性回归与实测值的相关系数0.46；样条回归与线性回归拟合值的相关系数0.42，计算线性回归拟合值在考虑了样条回归拟合值后与实测值的偏相关系数为0.25；还达不到 = 0.2的显著性检验水平；再计算样条回归拟合值在考虑了线性回归拟合值后与实测值偏相关系数为0.66，远超过 = 0.001的显著性水平。仿此分析昆明、大理两站得到同样的结论。

这一结果从实例上说明，样条回归已包含了线性回归的模拟预报能力，而线性回归不包含样条回归的模拟预报能力。即样条回归入选因子及表达式已最大限度地把线性回归因子集的综合预报能力整体归纳进去了。

(2) 因子筛选

在以往的线性统计预报中，人们已习惯用线性相关来初选预报因子，把线性相关好的预报因子入选建立线性回归预报模型[4]。而非线性统计预报与线性统计预报不完全相同，不需要预报线性相关好的因子作为备选因子，文中的实例也是如此。非线性统计预报有多种方法，不同的非线性统计预报方法采用不同的因子筛选方法，且与线性相关筛选也不完全相同[5,6]，文中的方法属于一种新的非线性因子筛选方法。

6 总结和展望

把样条函数引入到统计预报中，特别是存在复杂现象的非线性预报系统中加以应用，结果表明具有较好的效果。通过研究得出：

(1) 多元样条函数逐步回归模型是一种可以按需要裁剪，适应任意非线性连续变化形式的模拟方法，其应用范围主要是预报因子和预报对象间存在显著的，难以用标准初等函数表示的复杂非线性预报关系的情况。

(2) 以往样条函数多用于数值逼近中，其主要考虑如何精确地逼近数据点，为了达到这一目的，样条函数的阶数选取较高，分点也较多，它不考虑外推预报的情况。而在统计预报中，主要根据预报因子和预报对象的概率统计规律得出统计预报模型，在这个过程中，追求过多的分段和过高的阶数容易造成模型拟合好而外推预报差的现象。因此，在统计预报中不宜追求过多的分点和过高的阶数。文中的应用例子说明，当样本数在40—50个左右时，取3阶以下，4—5个分点即可达到较好的效果，阶数和分点的选取可通过逐步回归过程中，因子引入剔除显著性检验水平来控制。

(3) 由于非线性和复杂现象的普遍存在，这正是线性统计预报理论难以描述的，也是人们目前认为统计预报水平不高的一个重要原因，因此很有必要发展非线性统计预报理论和方法，本研究就是这方面的探索之一。

(4) 多元样条逐步回归模型保留了线性统计预报模型中多元分析，逐步筛选因子，模型显著性检验等重要理论和方法，吸取了数值逼近中样条函数处理复杂非线性函数的许多优点，应用了对复杂非线性函数采用分段逼近拟合的思想，因此，具有处理多因子，复杂非线性，逐步筛选建立最优预报模型的功能。

(5) 为了使样条回归函数更光滑, 还可引进磨光函数对样条函数进行磨光。

(6) 在样条回归模型中, 某些预报因子的结点表示了该因子在结点以上和以下的变化对预报对象的影响作用有显著不同, 对这些预报因子的结点进行天气气候学和动力学研究, 或许能发现一些重要的物理意义。

(7) 现代样条函数的研究, 已推广到重结点, 指数样条, 三角样条及更广义的抽象样条等, 为非线性复杂预报系统提供了更为丰富的数学模拟的理论和方法。

参考文献

[1] 冯康等. 数值计算方法. 北京: 国防工业出版社, 1978. 1—40.

[2] 李岳生. 数值逼近. 北京: 人民教育出版社, 1978. 65—68, 76—132.

[3] 程正兴. 数据拟合. 西安: 西安交通大学出版社, 1986. 47—169.

[4] 严华生, 王学仁. 多因变量及要素场统计预报. 北京: 气象出版社, 1990. 56—71.

[5] 严华生, 曹杰. 多元门限回归的一种建模方案. 大气科学, 1994, 18(2): 194—199.

[6] 叶笃正等. 当代气候研究. 北京: 气象出版社, 1991. 167—168.

NONLINEAR MU LTIPLE SPLINE REGRESSION MODEL

Yan Huasheng Cao Jie Xie Yingqi

(*Earth Science Department, Yunnan University, Kunming*, 650091)

You Weihong

(*Meteorological Observatory of Yunnan Province, Kunming*, 650034)

Abstract

The nonlinear multiple spline regression model was advanced. The model which retains multivariate analysis, screening factories and significance test of linear statistical prediction, and absorbs the advantage of splines that can simulate data stagewisely and can autoadapte any change of curve can deal with the situation which is abnormally complicated. The results indicate that the model can efficiently simulate and predict nonlinear system, and it is worthy to be studied and applied.

Key words: Nonlinear, Statistical prediction, Splines.