Part2

刘鋆

一模型构建的基本思想

1 基函数的确定（Bernstein 基函数）

设时间序列资料 X_i, i = 0, 1, ……, n. 构造拟合曲线为

$$\hat{X}(t) = \sum_{j=0}^{m} b_j \varphi_j(t), 0 \le t \le 1, m < n$$

其中 $\varphi_j(t), j = 0, 1, \ldots, m$ 为一组基函数；$b_j, j = 0, 1, \ldots, m$ 为待定系数向量，这里所给定的时间序列 $X_i, i = 0, 1, \ldots, n$ 已经参数化，决定的参数分割为 $\Delta_t : t_0 < t_1 < \cdots < t_n$.

对于基函数的选择，人们首先主意到在各类函数中，多项式函数能较好地满足要求，它能把复杂的现象简单地表达出来，通过改变作为多项式的次数，而具有丰富的表达力，又无穷次可微，对构造的曲线具有足够的光顺性，且容易计算函数值与各阶导数值，及实现可视化。

m 次多项式全体构成 m 次多项式空间。m 次多项式空间中任一组 m+1 个线性无关的多项式都可以作为一组基，因此就有无穷多组基。不同组基之间仅仅差一个线性变换。同一条曲线可以采用不同的多项式基函数表示，由此决定了它们具有不同的性质，而且又不同的优点。

用伯恩斯坦（Bernstein）基函数拟合的曲线方程为

$$\hat{X}(t) = \sum_{j=0}^{m} b_j B_{j,m}(t), 0 \le t \le 1, \tag{1}$$

这里的 $b_j, j = 0, 1, \ldots, m$ 为系数向量。在次称为拟合曲线的控制点。基函数

$$B_{j,m}(t) = C_m^j t^j (1-t)^{m-j}, j = 0, 1, \ldots, m \tag{2}$$

称为 Bernstein 基函数。

它的最大优点是对计算机输入与交互修改拟合曲线带来很大的方便，体现出来数据挖掘的特点，这是由 Bernstein 基函数的性质所决定。其性质与计算公式有：

i）规范性：$B_{j,m}(t) \ge 0, \sum_{j=0}^{m} B_{j,m}(t) \equiv 1$

ii) 对称性：$B_{j,m}(t) = B_{m-j,m}(1-t)$

iii) 函数的递推性：$B_{j,m}(t) = (1-t)B_{j,m-1}(t) + tB_{j-1,m-1}(t)$

iv) 分割性：$B_{j,m}(ct) = \sum_{i=j}^{m} B_{j,i}(c)B_{i,m}(t)$

2 基函数建模

设时间序列数据为 $X_i, i = 0, 1, \ldots, n$ 以 m 次 Bernstein 多项式

$$B_{j,m}(t) = C_m^j t^j (1-t)^{m-j}, j = 0, 1, \ldots, m$$

为基函数，构造曲线为

$$\hat{X}(t) = \sum_{j=0}^{m} b_j B_{j,m}(t), 0 \leq t \leq 1, m < n. \tag{3}$$

拟合这一时间序列数据点，以 Bernstein 基函数建立模型

$$X(t) = \sum_{j=0}^{m} b_j B_{j,m}(t) + \varepsilon(t), \tag{4}$$

其中 $b_j, j = 0, 1, \ldots, m$ 为待定的控制点；$B_{j,m}(t)$ 是 Bernstein 基函数，然后利用所构造曲线的有关性质，对未来的社会现象的发展进行预测，这里需要说明的是：

i)$\hat{X}(t)$ 是拟合数据点 $X_i, i = 0, 1, \ldots, n,$ 在曲线（3）上的值；（4）式中的 $X(t)$ 是经过干扰修正参数以后得到的实际值。建立模型的基本要求是，想用所拟合的曲线来描述参数化以后时间序列数据点的变化情况；

ii）$\varepsilon(t)$ 是误差项，也称为干扰项，它是一个随机变量，干扰项 $\varepsilon(t)$ 包括有被忽略的影响因素、数据的测量误差、随机误差以及模型的关系误差。我们考虑误差项，把它带到所研究的数学模型中，目的在于通过对它的研究，更加确切地说明客观存在的社会现象。

在此，采用最小二乘法估计出控制点 $b_j, j = 0, 1, \ldots, m$ 假设 $\varepsilon(t) \sim N(0, \sigma^2)$; 而且对 $t_1 \neq t_2$ 时，$cov[\varepsilon(t_1), \varepsilon(t_2)] = 0$. 下面具体介绍建立模型问题。

首先对时间序列数据 $X_i, i = 0, 1, \ldots, n$ 进行参数化，由于我们所研究的是间隔相等的时间序列资料，对数据参数化时，不能破坏这一性质，因此，采用等距参数化（均匀参数化）即要求

$$\Delta_i = u_{i+1} - u_i = C.$$

为处理方便取整数序列

$$u_i = i, i = 0, 1, \ldots, n.$$

将上式的参数化结果进行规范化，即得到规范参数化的结果

$$t_i = \frac{u_i}{n}, i = 0, 1, \ldots, n.$$

以下采用规范参数化进行讨论。

我们采用最小二乘法来确定拟合的曲线（3），并建立模型。设所需拟合的曲线为：

$$\hat{X}_i = \sum_{j=0}^{m} b_j B_{j,m}(t_i), i = 0, 1, \ldots, n. \tag{5}$$

模型为

$$X_i = \sum_{j=0}^{m} b_j B_{j,m}(t_i) + \varepsilon(t_i), i = 0, 1, \ldots, n. \tag{6}$$

我们求控制点 $b_j, j = 0, 1, \ldots, m$, 使得

$$J = \sum_{i=0}^{n} (X_i - \hat{X}_i)^2$$

达到最小。

为了明显地表示出 J 为 B_j 的函数，即有

$$J(b_0, b_1, \ldots, b_m) = \sum_{i=0}^{n} (X_i - \sum_{j=0}^{m} b_j B_{j,m}(t_i))^2.$$

根据要求即得到

$$\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{bmatrix} = (\Phi^T \Phi)^{-1} \Phi^T \begin{bmatrix} X_0{}' \\ X_1{}' \\ \vdots \\ X_n{}' \end{bmatrix},$$

其中，

$$\Phi = \begin{bmatrix} B_{0,m}(t_0) & B_{1,m}(t_0) & \ldots & B_{m,m}(t_0) \\ B_{0,m}(t_1) & B_{1,m}(t_1) & \ldots & B_{m,m}(t_1) \\ \vdots & \vdots & \ddots & \vdots \\ B_{0,m}(t_n) & B_{1,m}(t_n) & \ldots & B_{m,m}(t_n) \end{bmatrix},$$

Φ^T 为 Φ 的转置.

这样便估计出了关于模型（4）的 m+1 个控制点 $b_0, b_1, \ldots, b_m$. 从而得到拟合曲线为

$$\hat{X} = \sum_{j=0}^{m} b_j B_{j,m}(t). \tag{7}$$