# 第6章　样条函数回归

东北财经大学统计学院

2014年10月14日

## 1　引言

给定一组观测数据$(x_1, y_1), ..., (x_n, y_n)$, 我们想研究**响应变量** $Y$和**协变量** $X$之间的关系，这个关系用一个函数来刻画，于是我们考虑如下模型：

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2), \ i = 1, ..., n, \tag{1}$$

若令$\mathbf{y} = (y_1, ..., y_n)^T$, $f(\mathbf{x}) = (f(x_1), ..., f(x_n))^T$, $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$. 则模型(1)可写成：

$$\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\epsilon}, \ \boldsymbol{\epsilon} \sim (\mathbf{0}_n, \sigma^2 \mathbf{I}_n). \tag{2}$$

本章我们将采用样条函数逼近的方法去估计回归函数$f(\cdot)$.

样条(函数)是一种分段多项式，各相邻段上的多项式之间又具有某种连接性质，因而它既保持了多项式的简单性和逼近的可行性，又在各段之间保持了相对独立的局部性质。理论和实践证明，样条是一类特别有效的逼近工具。

本章的大部分内容来自Wu and Zhang (2006) 和Ruppert, Wand, and Carroll (2003).

# 2 回归样条(regression spline)

## 2.1 介绍

设$a = \min(x_i), b = \max(x_i)$, 考虑区间$[a, b]$上的一个分割：$a = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = b$. 我们通常称这些分割点为节点(knots),并称$\tau_r, r = 1, 2, ..., K$为内部节点(interior knots). 于是这些节点将$[a, b]$分割成$K$个子区间(local neighborhoods):

$$[\tau_r, \tau_{r+1}), r = 0, 1, ..., K,$$

那么我们可以在任意子区间$[\tau_r, \tau_{r+1})$内做Taylor展开。

回归样条是一个分段多项式：(1) 在节点$\tau_r$和$\tau_{r+1}$之间是一个多项式；(2) 在节点处恰当地连接在一起，但允许在节点处存在不连续的导数。

那么如何构造一个回归样条函数呢？我们首先定义一组基函数(basis functions)，然后考虑由这组基函数生成的线性空间。

## 2.2 Truncated Power Basis

给定一组节点$\tau_1, \tau_2, ..., \tau_K$, 定义如下基函数：

$$1, x, \cdots, x^k, (x - \tau_1)_+^k, \cdots, (x - \tau_K)_+^k, \tag{3}$$

其中

$$w_+^k = (w_+)^k.$$

这里$w_+ = \max(0, w)$. 我们称式(3)为次数(degree)为$k$的truncated power basis. 显然，式(3)中前$(k + 1)$个基函数为多项式函数(次数 $\leq k$), 而

其余的均为$k$次Truncated Power Basis函数。通常，我们称次数为$k = 0, 1, 2, 3$的truncated power basis分别为“常数”、“线性”、“二次”和“三次”truncated power basis. 一般$k = 3$足够了,即cubic splines。

于是一个回归样条函数(regression spline)可表示为：

$$f(x) = \sum_{s=0}^{k} \beta_s x^s + \sum_{r=1}^{K} \beta_{k+r}(x - \tau_r)_+^k, \tag{4}$$

其中$\beta_0, \beta_1, ..., \beta_{k+K}$为一组系数参数。为方便，我们常称其为次数为$k$，节点为$\tau_1, ..., \tau_K$的回归样条函数, 而当$k = 1, 2, 3$时，分别称为“线性”、“二次”和“三次”回归样条函数。

显然，在任意子区间(local neighborhood)$[\tau_r, \tau_{r+1})$内，都有：

$$f(x) = \sum_{s=}^{k} \beta_s x^s + \sum_{l=1}^{r} \beta_{k+l}(x - \tau_l)^k,$$

其恰为一$k$次多项式。然而对于$r = 1, 2, ..., K$, 由于：

$$\begin{aligned} f^{(k)}(\tau_r-) &= k!(\beta_k + \sum_{l=1}^{r-1} \beta_{k+l}), \\ f^{(k)}(\tau_r+) &= k!(\beta_k + \sum_{l=1}^{r} \beta_{k+l}), \end{aligned}$$

因而

$$f^{(k)}(\tau_r+) - f^{(k)}(\tau_r-) = k!\beta_{k+r}.$$

即：$f^{(k)}(x)$在节点$\tau_r, r = 1, 2, ..., K$处有一个跳跃(值为$k!\beta_{k+r}$). 换句话说：次数为$k$的回归样条函数的$k$阶导函数是不连续的，而低于$k$阶的导函数都是连续的；此外，系数$\beta_{k+r}$则度量了跳跃值的大小。

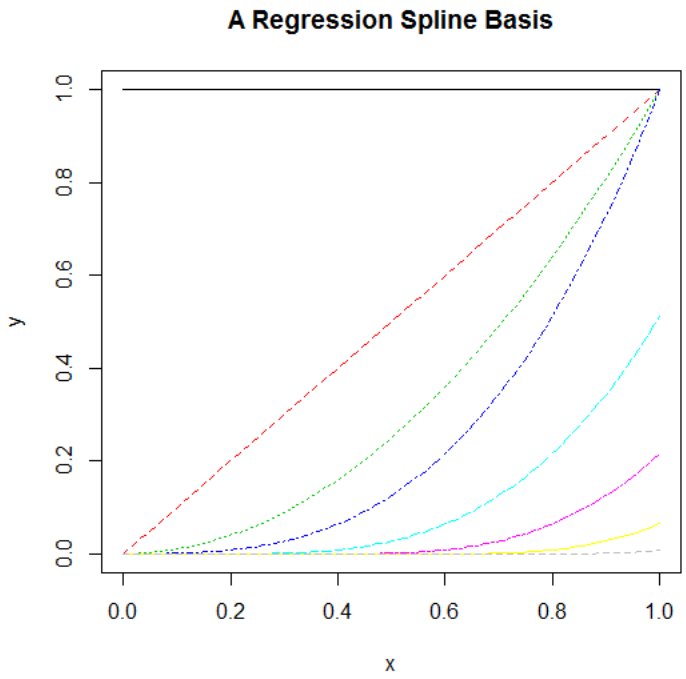



Figure 1: 三次truncated power basis基函数.

图1是一个节点为$\{0.2, 0.4, 0.6, 0.8\}$的三次truncated power basis函数的示例。

## 2.3 Regression Spline Smoother

为方便，记

$$\mathbf{\Phi}_p(x) = [1, x, \cdots, x^k, (x-\tau_1)_+^k, \cdots, (x-\tau_K)_+^k]^T, \tag{5}$$

其中$p = K + k + 1$表示基函数的个数. 类似地，记

$$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{t+K})^T.$$

则回归样条函数(4)可写成

$$f(x) = \mathbf{\Phi}_p(x)^T \boldsymbol{\beta}.$$

于是模型

$$\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\epsilon},$$

可写为：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

其中，

$$\begin{aligned} \mathbf{y} &= (y_1, ..., y_n)^T, \\ \mathbf{X} &= (\mathbf{\Phi}_p(x_1), ..., \mathbf{\Phi}_p(x_n))^T, \\ \boldsymbol{\epsilon} &= (\epsilon_1, ..., \epsilon_n)^T. \end{aligned}$$

注意到$\mathbf{\Phi}_p(x)$是基底，因而当$n \geq p$时，$\mathbf{X}$是满秩矩阵，从而，$(\mathbf{X}^T\mathbf{X})^{-1}$存在，于是可得最小二乘估计(OLSE):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

从而，$f(t)$的回归样条估计为：

$$\hat{f}_p(x) = \mathbf{\Phi}_p(x)^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{6}$$

我们常称式(6)为regression spline smoother. 对于那些观测点$x_i, i = 1, ..., n$, 我们有拟合值：

$$\hat{\mathbf{y}}_p = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \equiv \mathbf{A}_p\mathbf{y},$$

其中$\hat{\mathbf{y}}_p = (\hat{y}_1, ..., \hat{y}_n)^T$, $\hat{y}_i = \hat{f}_p(x_i)$, $\mathbf{A}_p = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. 易证矩阵$\mathbf{A}_p$为对称幂等阵，即$\mathbf{A}_p^T = \mathbf{A}_p, \mathbf{A}_p^2 = \mathbf{A}_p$且有trace$(\mathbf{A}_p) = p$. 我们通常称trace$(\mathbf{A}_p)$为regression spline smoother的自由度(degree-of-freedom).

## 2.4 节点的位置和个数的选择

Regression spline smoother的优劣跟节点的位置和个数有着很重要的关系，而与regression spline basis函数的次数$k$关系不大，为计算方便，我们通常取$k = 1, 2, 3.$ 对于选取节点的位置，通常有三种方法：

- 等距离选取:

  在区间$[a, b]$内等距离地选取$K$个点作为节点，即：

  $$\tau_r = a + (b - a)r/(K + 1),\ r = 1, 2, ..., K.$$

  这种方法通常用于观测值$x_i, i = 1, 2, ..., n$分布比较均匀时。

- 样本分位数的等距离选取:

  此方法是选取等距离的样本分位数作为节点。设顺序统计量为$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, 则选取的$K$个节点为：

  $$\tau_r = x_{(1+[rn/(K+1)])}, r = 1, 2, ..., K,$$

  这里$[a]$表示$a$的整数部分。这种方法通常用于$X_i, i = 1, 2, ..., n$分布不均匀的情形，观测值比较多的区域，节点放的多一些。

- 基于模型选择的方法:

  这种方法是将所有的观测点$X_i, i = 1, 2, ..., n$都做为候选节点，由于删除一个节点等价于删除了一个相应的基函数，于是我们可以利用模

型选择的方法(例如forward selection, backward elimination, or stepwise regression)来删除节点。

**需要指出的是**：这里我们将节点的个数$K$看作smoothing parameter, 从而引出了smoothing parameter selector的问题，我们会在后面详细讨论这个问题。

## 2.5 实例

1. 继续研究摩托车碰撞试验的数据，该数据集来源于R程序包**MASS**. 此数据集由一系列在模拟摩托车事故中头部加速度的观测值构成,目的主要是用来检测头盔的性能。

2. 继续研究发动机排放气体中的nitric oxides的浓度和equivalence ratio(一种空气和ethanol混合程度的度量)的关系。响应变量是$NOx$，协变量是$E$ (发动机的equivalence ratio)和$C$ (压缩比例). 我们去拟合$NOx$和$E$之间的非线性关系。

## 2.6 General Basis-Based Smoother

前面我们选取的$\mathbf{\Phi}_p(x)$为Truncated power basis 函数，事实上，我们还可以选择其他的基函数，例如B-spline basis, wavelet bases, Fourier Series, polynomial bases 等等，如果选择polynomial bases $\{1, x, x^2, ..., x^p\}$,就是我们经常用的多项式回归。

## 2.7 Natural cubic splines

尽管前面讲的cubic splines 被广泛地运用，然而它有一定的局限性: 它只拟合了节点之间的数据，而当数据落在两个边界节点之外时我们用3次多项式来拟合，由于在两个边界节点之外我们没有观测数据，因而这种推

断会导致很严重的错误，特别是在$X$的两个极限点附近。而Natural cubic splines则很好地解决了这个问题。Natural cubic splines 基函数可以看作是cubic splines 基函数的一个变形，另外加上了一个限制：在边界的节点(boundary knots)外为线性函数。实践证明：natural cubic splines 往往优于cubic splines.

- Cubic splines:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \tau_1)_+^3 + \beta_5 (x - \tau_2)_+^3.$$

  注意：在边界的节点外，函数依然是三次多项式。如果有$K$个节点，则我们需要$K + 4$个参数（包括截距）。

- Natural cubic splines.

$$f(x) = \beta_0 + \beta_1 x + \beta_4 (x - \tau_1)_+^3 + \beta_5 (x - \tau_2)_+^3.$$

  注意：在边界的节点外，函数为一次多项式。如果有$K$个节点，则我们需要$K + 2$个参数（包括截距）。

也就是说：natural cubic splines则通过在自变量$x$的定义域的两个边界处各增加一个节点，然后在这两个新增加的边界节点与其相邻节点之间用一个线性一次函数来拟合数据，换句话说，这种方法迫使样条函数在边界处为线性一次函数，从而避免了三次样条函数在边界处拟合很差的情形，进而在一定程度上改进了样条逼近。

## 2.8 B-spline basis

由于(natural) cubic splines基函数是高度相关(highly correlation)的，很容

易导致矩阵$\mathbf{X}$各列之间出现大量的共线性(collinearity), 这种共线性会导致$\mathbf{X}^T\mathbf{X}$为近似奇异阵(singular matrix), 从而估计的不准确(Ruppert, Wand, and Carroll 2003)。于是人们提出了更稳定的B-样条(B-splines)基函数。

给定一组节点$a = \tau_0 < \tau_1 < ... < \tau_K < \tau_{K+1} = b$, $q$次(即degree=$q$)的B-样条基函数是由$q+1$个$q$次分段多项式在$q$个内部节点$\tau_{j_1} < ... < \tau_{j_q}$处连接起来，且具有如下的特点(Eilers and Marx, 1996):

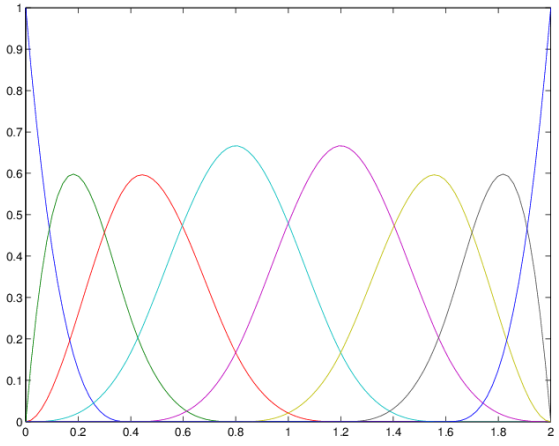
- it consists of $q+1$ polynomial pieces, each of degree $q$;
- the polynomial pieces join at $q$ inner knots;
- at the joining points, derivatives up to order $q-1$ are continuous;
- the B-spline is positive on a domain spanned by $q+2$ knots; everywhere else it is zero;
- except at the boundaries, it overlaps with $2q$ polynomial pieces of its neighbors;
- at a given $x$, $q+1$ B-splines are nonzero.

因为$k$次多项式函数为$k+1$阶多项式,因而我们有时说B-样条基函数的阶数(order)为多少，例如order=4, 此即为3次B-样条(degree=3). 在区间$[a, b]$内

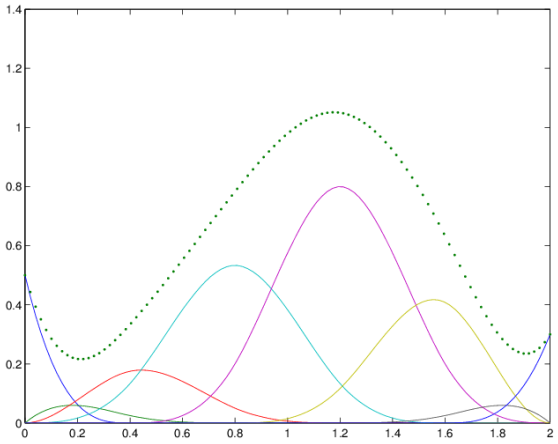
若插入$K-1$个内部节点，从而将$[a,b]$分成$K$个子区间，则在每个子区间上将有$q+1$个非零的次数为$q$的B-样条函数，而B-样条总数为$K+q$.

De Boor (1978) 给出了一种循环算法去计算任意degree的B-样条基函数。图(2) 展示了三种B-样条基函数:(a) degree=1 or order=2; (b) degree=2 or order=3; (c) degree=3 or order=4。图(2.8)展示的是一个B-样条函数(即B-样条基函数的线性组合)。

(a) cubic B-spline basis



(b) weighted cubic B-splines



## 2.9 Radial Basis Functions

这类基函数的定义如下：

$$1, x, ..., x^p, |x-\tau_1|^p, ..., |x-\tau_K|^p.$$

这里$p$为奇数。图(3)展示了$p=1$时的一组基函数。

由于

$$|x-\tau_k|^p = r(|x-\tau_k|),$$

其中$r(u)=u^p$，因而基函数$|x-\tau_k|^p (1 \le k \le K)$只依赖于距离$|x-\tau_k|$和一个

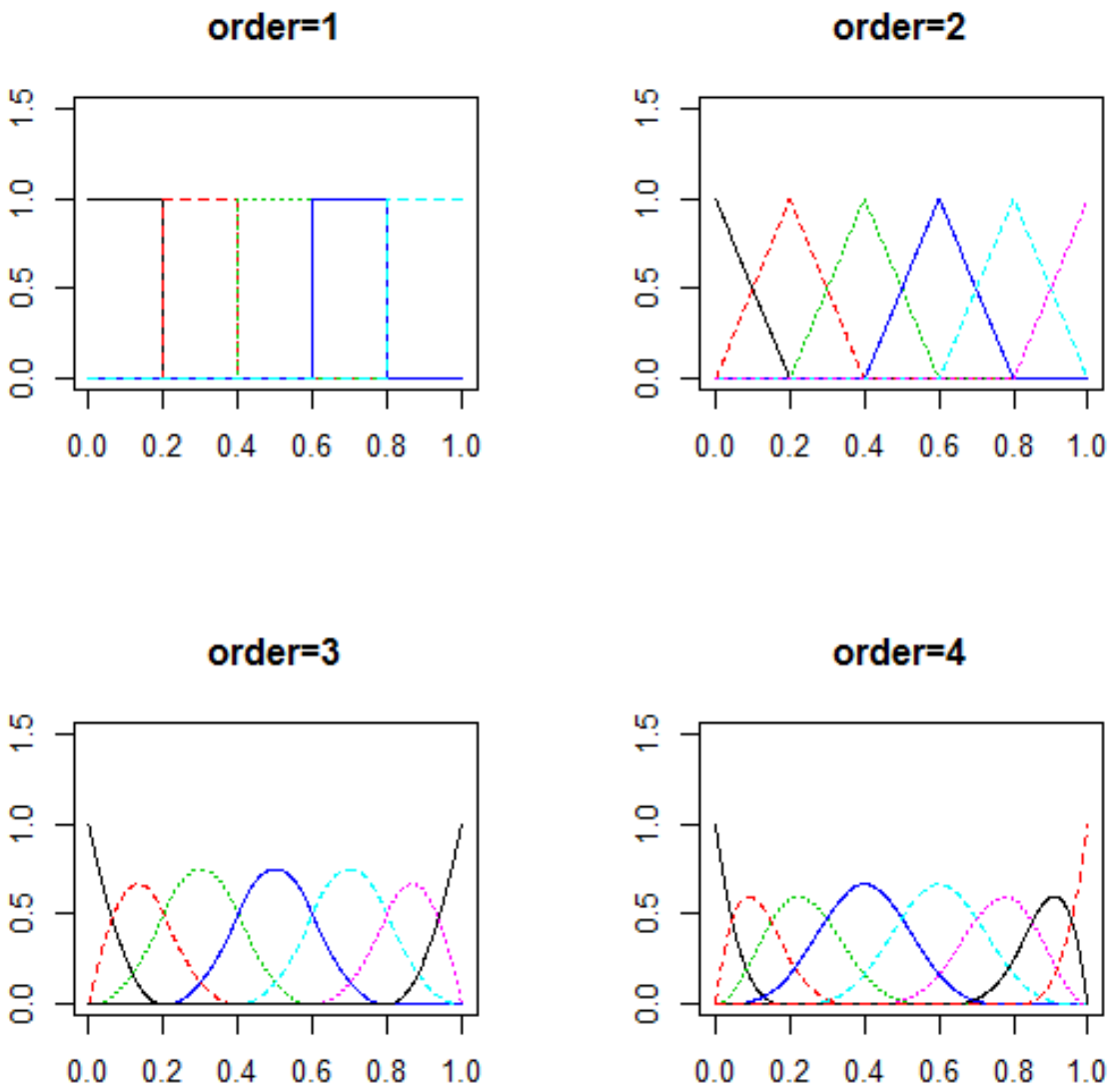



Figure 2: B-spline basis functions

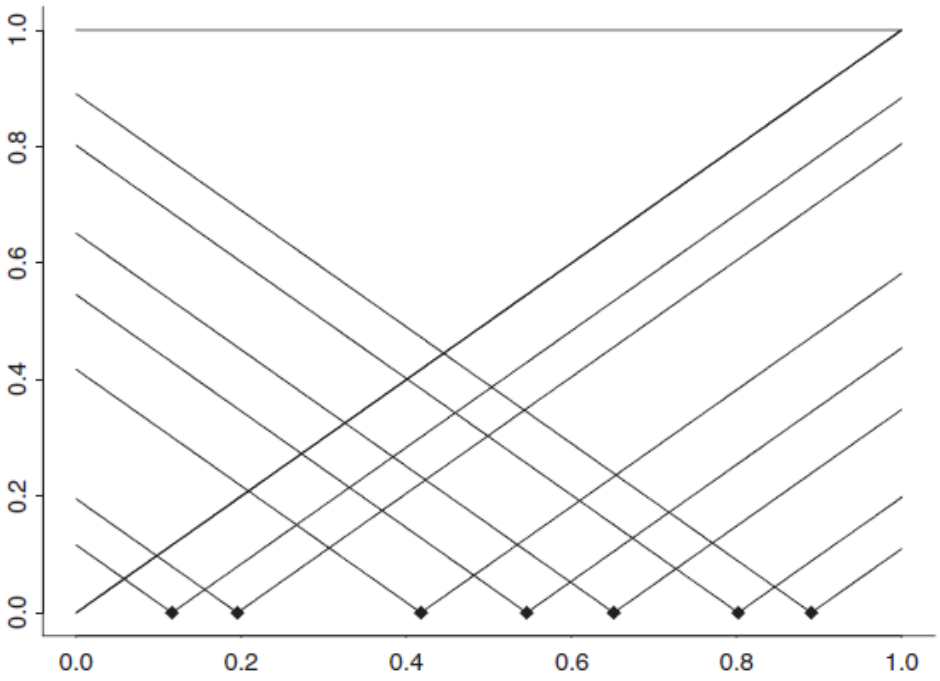



Figure 3: linear radial basis functions

一元函数$r(\cdot)$. 此外，我们很容易将其推广到多维的情况。假定$\mathbf{x} \in R^d$,节点为$\boldsymbol{\tau}_1, ..., \tau_K \in R^d$, 则基函数为：

$$r(\|\mathbf{x} - \boldsymbol{\tau}_k\|);$$

其中，$\|\mathbf{v}\| = \sqrt{\mathbf{v}^T\mathbf{v}}$表示向量$\mathbf{v}$的长度, $r(\mathbf{u})$为一多元函数。更多内容参见Ruppert, Wand, and Carroll (2003)。

## 2.10 Fourier basis

傅立叶展开也是我们常用的一种手法：

$$f(t) = \beta_0 + \beta_1 \sin(\omega t) + \beta_2 \cos(\omega t) + \beta_3 \sin(2\omega t) + \beta_4 \cos(2\omega t) + ...$$

因而我们定义基函数：

$$\phi_0(t) = 1, \phi_{2r-1}(t) = \sin(r\omega t), \phi_{2r}(t) = \cos(r\omega t), r = 1, 2, ...$$

这种基函数是周期函数，周期为$2\pi/\omega$. 所以对于那些周期性的观测数据，我们可以考虑傅立叶基函数。此外，傅立叶基函数除第一个为常数1外，其余的基函数都是成对出现的，即我们一般取$2k+1$个基函数。

# 3 SMOOTHING SPLINES

在做regression spline时，大部分工作是放在节点个数$K$的选取上，我们通常可以用AIC、BIC等准则来选取，而且$K$一般是小于样本容量$n$的，但由于$K$必须为整数，可供我们选择的空间毕竟有限，无法做更精细的调整。

事实上，我们还可以考虑另外一种方法：将所有观测点都做为节点，然而如果对样条函数不加任何限制的话，相当于我们只做了插值。为了消除由此带来的undersmoothed问题，我们可以考虑引入一个惩罚项，从而提高拟合曲线的光滑程度。

我们考虑如下的惩罚函数：

$$\int_a^b \left\{f^{k)}(u)\right\}^2 du, \tag{7}$$

其中$k \geq 1$. 则模型(1)的smoothing spline smoother为下列惩罚最小二乘(PLS)准则的最小值解$\hat{f}_\lambda(x)$:

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_a^b \left\{f^{(k)}(x)\right\}^2 dx, \tag{8}$$

其中$f(\cdot) \in \mathcal{W}_2^k[a,b]$, $\mathcal{W}_2^k[a,b]$ 为$k$阶Sobolev空间：

$$\left\{ f : f^{(r)} \text{ 绝对连续 }, 0 \le r \le k-1, \int_a^b \{f^{(k)}(x)\}^2 dx < \infty \right\}.$$

$\lambda > 0$是用来控制$\hat{f}_\lambda(x)$光滑程度的smoothing parameter。显然，当$k = 2$时，$\lambda = 0$相当于我们只做了插值工作；而如果$\lambda = +\infty$, 则$\hat{f}_\lambda(x) = a + bx$为线性函数. 所以(8)中的第一项相当于是拟合优度问题或者说是衡量模型的偏差；而第二项是控制函数的光滑程度以防止出现undersmoothed或oversmoothed的情况发生。当$k = 2$时，所得的的$\hat{f}_\lambda(x)$称为natural cubic smoothing spline (NCSS). 关于smoothing splines的更多内容, 请参考Eubank (1988,1999), Wahba (1990), Green and Silverman (1994) 和Gu (2002)等.

## 3.1 Cubic Smoothing Splines

下面我们重点研究$k = 2$的情形。设$\tau_1 < ... < \tau_K$为全体观测点(即所有节点)，并令

$$h_r = \tau_{r+1} - \tau_r, r = 1, 2, ..., K-1.$$

我们定义几个矩阵：$\mathbf{A} = (a_{rs})_{K\times(K-2)}$, $\mathbf{B} = (b_{rs})_{(K-2)\times(K-2)}$, $\mathbf{G} = \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^T$,

其中：

$$\begin{cases} a_{rr} = h_r^{-1}, \\ a_{r+1,r} = -(h_r^{-1} + h_{r+1}^{-1}), \\ a_{r+2,r} = -h_{r+1}^{-1}, \\ \text{其它的 } a_{rs} \text{ 均为0}, \\ r = 1, 2, ..., K-2. \end{cases} \qquad \begin{cases} b_{11} = (h_1 + h_2)/3 \\ b_{21} = h_2/6, \\ b_{r,r+1} = h_{(r+1)}/6, \\ b_{r+1,r+1} = (h_{(r+1)} + h_{(r+2)})/3, \\ b_{r+2,r+1} = h_{(r+2)}/6, \\ b_{K-3,K-2} = h_{(K-2)}/6, \\ b_{K-2,K-2} = (h_{(K-2)} + h_{(K-1)})/3, \\ \text{其它的 } b_{rs} \text{ 均为0}, \\ r = 1, 2, ..., K-4. \end{cases}$$

令$\mathbf{f} = (f_1, ..., f_K)^T$, 其中$f_r = f(\tau_r), r = 1, 2, ..., K$. 则当$k = 2$时(7)可表示为

$$\int_a^b [f''(x)]^2 dx = \mathbf{f}^T \mathbf{G} \mathbf{f}. \tag{9}$$

于是问题(8)转化为：

$$(\mathbf{y} - \mathbf{W}\mathbf{f})^T (\mathbf{y} - \mathbf{W}\mathbf{f}) + \lambda \mathbf{f}^T \mathbf{G} \mathbf{f},$$

其中$\mathbf{W} = (w_{ir})_{n \times K}$, $w_{ir} = 1$若$x_i = \tau_r$，否则为0。则最小二乘估计$\hat{\mathbf{f}}_\lambda$为：

$$\hat{\mathbf{f}}_\lambda = (\mathbf{W}^T \mathbf{W} + \lambda \mathbf{G})^{-1} \mathbf{W}^T \mathbf{y}.$$

及

$$\hat{\mathbf{y}}_\lambda = \mathbf{W}(\mathbf{W}^T \mathbf{W} + \lambda \mathbf{G})^{-1} \mathbf{W}^T \mathbf{y}.$$

需要指出的是：我们这里只是估计了$f(x)$在各观测点$\{x_1, ..., x_n\}$的函数值，

对于其它的点$x$所对应的函数值$f(x)$，可以利用Green and Silverman(1994)里的方法或者利用插值的方法（例如三次样条插值)去得到。

# 4 Penalized splines

Smoothing splines将所有的不同的观测点都作为节点，然后利用惩罚函数来控制拟合的函数的光滑程度。然而，当观测点非常多的时候，所要估计的参数就会非常多，从而导致计算量非常大，计算效率低。此外，在计算惩罚函数时也很复杂。而penalized spline (P-spline) 方法则很好地解决了这个问题。P-splines利用一些预先定义的节点来定义一组基函数，同时增加了一个惩罚函数，而这个惩罚函数跟基函数有关。

## 4.1 Penalized Spline Smoother

考虑优化问题(8), 如果我们能用一组基函数的线性组合来逼近$f(x)$，则惩罚函数就可以具体写出表达式了。

我们先以truncated power basis为例，假设$\mathbf{\Phi}_p(x)$为一组truncated power basis函数，次数为$k$，内部节点为$\tau_1, ..., \tau_K$. 则$f(x)$可近似写为

$$f(x) \approx \mathbf{\Phi}_p(x)^T \boldsymbol{\beta},$$

其中$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{k+K})^T$. 令矩阵$\mathbf{G}$为

$$\mathbf{G} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K \end{pmatrix}$$

则优化问题(8)可写为：

$$\sum_{i=1}^{n}\left\{y_i-\mathbf{\Phi}_p(x_i)^T\boldsymbol{\beta}\right\}^2+\lambda\boldsymbol{\beta}^T\mathbf{G}\boldsymbol{\beta}. \tag{10}$$

于是有：

$$\hat{f_\lambda}(x)=\mathbf{\Phi}_p(x)^T(\mathbf{W}^T\mathbf{W}+\lambda\mathbf{G})^{-1}\mathbf{W}^T\mathbf{y}.$$

其中$\mathbf{W}=(\mathbf{\Phi}_p(x_1),...,\mathbf{\Phi}_p(x_n))^T$. 从而

$$\hat{\mathbf{y}}_\lambda=\mathbf{W}(\mathbf{W}^T\mathbf{W}+\lambda\mathbf{G})^{-1}\mathbf{W}^T\mathbf{y}.$$

下面我们再以B-样条为例讲解P-splines smoother。设我们有$J$个次数为$q$的B-样条基函数：

$$\mathbf{\Phi}(x)=(B_1(x),...,B_J(x))^T.$$

并设$f(x)$可表示成$\mathbf{\Phi}(x)$的一个线性组合：

$$f(x)=\sum_{j=1}^{J}\beta_jB_j(x)=\mathbf{\Phi}(x)^T\boldsymbol{\beta}.$$

则在平方损失函数下，欲优化的目标函数为：

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^{n}\left\{y_i-\mathbf{\Phi}(x_i)^T\boldsymbol{\beta}\right\}^2.$$

一般我们取的节点个数会很多，从而导致估计量的方差比较大(undersmoothed). 为了提高拟合优度，O'Sullivan(1986,1988)提出增加一项如下的惩罚函数：

$$\mathrm{PEN}(\boldsymbol{\beta}) = \lambda \int_a^b \left\{ \sum_{j=1}^{J} \beta_j B_j''(x) \right\}^2 dx. \tag{11}$$

这里$\lambda > 0$为光滑参数(smoothing parameter). Eilers and Marx (1996)则提出了一种基于高阶(higher-order)的相邻B-样条的系数的有限差分公式的惩罚函数，即：

$$\int_a^b \left\{ \sum_{j=1}^{J} \beta_j B_j^{(k)}(x) \right\}^2 dx = \sum_{j=k+1}^{J} (\Delta^k \beta_j)^2,$$

其中$\Delta a_j = a_j - a_{j-1}$, $\Delta^k a_j = \Delta^{(k-1)} a_j - \Delta^{(k-1)} a_{j-1}$. 可以证明：存在矩阵$\mathbf{D}_k$, 使得：

$$\int_a^b \left\{ \sum_{j=1}^{J} \beta_j B_j^{(k)}(x) \right\}^2 dx = \boldsymbol{\beta}^T \mathbf{D}_k^T \mathbf{D}_k \boldsymbol{\beta}, \tag{12}$$

从而$\boldsymbol{\beta}$的最小二乘估计为：

$$\ell^*(\boldsymbol{\beta}|\lambda) = \sum_{i=1}^{n} \{ y_i - \boldsymbol{\Phi}(x_i)^T \boldsymbol{\beta} \}^2 + \lambda \boldsymbol{\beta}^T \mathbf{D}_k^T \mathbf{D}_k \boldsymbol{\beta}.$$

等价于

$$\ell^*(\boldsymbol{\beta}|\lambda) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \mathbf{D}_k^T \mathbf{D}_k \boldsymbol{\beta}. \tag{13}$$

其中$\mathbf{X}^T = \boldsymbol{\Phi}(\mathbf{x})$. 上式对$\boldsymbol{\beta}$求导并令之为0，解得：

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{D}_k^T \mathbf{D}_k)^{-1} \mathbf{X}^T \mathbf{y}. \tag{14}$$

而$f(x)$的最小二乘估计为：

$$\hat{f}(x) = \boldsymbol{\Phi}(x)^T \hat{\boldsymbol{\beta}} = \boldsymbol{\Phi}(x)^T (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{D}_k^T \mathbf{D}_k)^{-1} \mathbf{X}^T \mathbf{y}.$$

y的拟合值为：

$$\hat{\mathbf{y}} = \mathbf{\Phi}(x)^T\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D}_k^T\mathbf{D}_k)^{-1}\mathbf{X}^T\mathbf{y}.$$

一般我们取$k = 2$即可，即考虑惩罚函数：

$$\lambda\int_a^b \left\{\sum_{j=1}^{J}\beta_j B_j''(x)\right\}^2 dx$$

## 4.2 Connection between a Penalized Spline and a LME Model

我们可以将penalize spline问题视为一个linear mixed-effects(LME)模型. 我们以truncate power basis 为例。

记$\mathbf{\Phi}(x) = (1, x, x^2, ..., x^k)^T$, $\mathbf{\Psi}(x) = ((x-\tau_1)_+^k, ..., (x-\tau_K)_+^k)^T$. 令$\mathbf{A}_i = \mathbf{\Phi}(x_i), \mathbf{Z}_i = \mathbf{\Psi}(x_i)$. 则有

$$(\mathbf{y} - \mathbf{A}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T(\mathbf{y} - \mathbf{A}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) + \lambda\mathbf{b}^T\mathbf{b},$$

其中$\mathbf{A} = (\mathbf{A}_1, ..., \mathbf{A}_n)^T$, $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)^T$, $\mathbf{b} = (\mathbf{b}_1, ..., \mathbf{b}_n)^T$. 因而我们可以考虑LME模型：

$$y_i = \mathbf{A}_i^T\boldsymbol{\beta} + \mathbf{Z}_i^T\mathbf{b}_i + \epsilon_i, i = 1, ..., n,$$

其中$\epsilon_i \sim N(0, \sigma^2)$, $\mathbf{b}_i \sim N(\mathbf{0}, \frac{\sigma^2}{\lambda}\mathbf{I}_K)$.

于是我们可以借用各种统计软件里的LME模型的各种方法来估计参数$\boldsymbol{\beta}$和$\mathbf{b}_i$.

# 5 SMOOTHING PARAMETER SELECTION

在样条回归里，smoothing parameter 的选择至关重要，它既要兼顾拟合优度问题(goodness-of-fit)，还要兼顾模型的效率(model complexity)。

## 5.1 LINEAR SMOOTHER

首先我们说，前面介绍的这些方法所得到的$f(x)$的估计(smoother)$\hat{f}(x)$都是$\mathbf{y} = (y_1, ..., y_n)^T$的线性函数(linear smoother). 换句话说，$\mathbf{y}$的拟合值$\hat{\mathbf{y}}_\rho$满足：

$$\hat{\mathbf{y}}_\rho = \mathbf{H}_\rho \mathbf{y},$$

这里$\rho > 0$为smoothing parameter, 而$\mathbf{H}_\rho$的下角标为$\rho$表示该矩阵依赖于$\rho$.

## 5.2 拟合优度(Goodness-of-fit)

显然，评价一个估计量$\hat{f}(x)$的优劣程度的一个准则是考虑残差平方和(Sum of Squared Errors, 简记为SSE):

$$\mathrm{SSE}_\rho = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathbf{y}^T (\mathbf{I}_n - \mathbf{H}_\rho)^T (\mathbf{I}_n - \mathbf{H}_\rho) \mathbf{y},$$

其中$\hat{y}_i = \hat{f}_\rho(x_i)$. $\mathrm{SSE}_\rho$ 刻画了模型对观测数据的拟合程度。显然当$\rho = 0$时，$\mathrm{SSE}_\rho = 0$取最小值。

另外一个常用的评价标准是log-likelihood. 若假设随机误差$\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, 则可以证明：

$$\mathrm{Loglik} = -\frac{n}{2} \log \left( \frac{2\pi e^1}{n} \mathrm{SSE}_\rho \right).$$

## 5.3 Model Complexity

我们称smoother matrix $\mathbf{A}_\rho$的迹(trace)为linear smoother的自由度(degrees of freedom,简记为df),即：

$$\text{df} = \text{trace}(\mathbf{H}_\rho) = \sum_{i=1}^{n} a_{ii},$$

其中$a_{ii}$为$\mathbf{H}_\rho$的对角元。这个df 常被用作近似刻画多少个有效的参数用到了模型里。例如对于regression splines, 所用的基函数的个数$p$即为我们的smoothing parameter $\rho$, 易证：

$$\text{df} = \text{trace}(\mathbf{H}_\rho) = \text{trace}\{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\} = p,$$

即我们的回归模型中恰用了$p$个参数(基函数的系数). 显然，$p$越大，所使用的参数越多，模型也就越复杂。

需要指出的是：一般地，df 不一定是整数；此外，当smoothing parameter $\rho$变小时，df 的值会变大(表示模型的复杂程度增加了)；而此时$SSE_\rho$是变小了，表明模型更接近于观测数据了。

## 5.4 AIC and BIC

Akaike 信息准则(AIC) (Akaike 1973) 是一个模型选择的准则，被广泛用于线性模型、时间序列等问题中，其表现形式为：

$$\text{AIC}(\rho) = -2\text{Loglik} + 2\text{df}.$$

选取使得AIC达到最小的$\rho$为最佳的smoothing parameter。

Bayesian 信息准则(BIC)(Schwarz 1978)为AIC的一种变体，其表达形式

为:

$$\mathrm{BIC}(\rho) = -2\mathrm{Loglik} + \log(n)\mathrm{df}.$$

一般而言，AIC方法比BIC方法更保守些，即选择的模型会含有较多的参数。

## 5.5 Cross-Validation

考虑如下的cross-validation (CV)准则：

$$\mathrm{CV}(\rho) = n^{-1}\sum_{i=1}^{n}\left\{y_i - \hat{f}_\rho^{(-i)}(x_i)\right\}^2, \tag{15}$$

其中$\hat{f}_\rho^{(-i)}(\cdot)$是基于观测数据$\{(x_1,y_1),...,(x_{i-1},y_{i-1}),(x_{i+1},y_{i+1}),...,(x_n,y_n)\}$的$f(\cdot)$的估计。使得CV函数(15)达到最小的$\rho$即为最佳的smoothing parameter. 这种方法被称为"leave-one-out"策略。

然而，由于当样本容量$n$比较大时，式(15)的计算量非常大，故我们通常使用它的一个变体或近似表达式。可以证明，式(15) 可简化为：

$$\mathrm{CV}(\rho) = n^{-1}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_\rho(x_i)}{1-h_{ii}}\right\}^2 = n^{-1}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{y}_i}{1-h_{ii}}\right\}^2, \tag{16}$$

其中$h_{11},...,h_{nn}$为smoother matrix $\mathbf{H}_\rho$的对角元，且

$$\hat{y}_i = \hat{f}_\rho(x_i), \qquad \begin{pmatrix} \hat{f}_\rho(x_1) \\ \hat{f}_\rho(x_2) \\ \vdots \\ \hat{f}_\rho(x_n) \end{pmatrix} = \mathbf{H}_\rho\mathbf{y}.$$

这样的话，对于每一个$\rho$值，我们只需要做一次局部回归就可以了。

### 5.6 Generalized Cross-Validation

作为cross-validation 准则的一个近似，Generalized cross-validation (GCV) 由Wahba (1977) 和Craven and Wahba (1979) 首先提出的。其主要思想是用平均值$\frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\text{trace}(\mathbf{H}_\rho)$来代替$h_{ii}$. 于是有：

$$\text{GCV}(\rho) = \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{y}_i}{1 - n^{-1}\text{trace}(\mathbf{H}_\rho)} \right\}^2, \tag{17}$$

显然分子刻画的是模型的拟合优度，而分母刻画的是模型的复杂程度。使得GCV函数(17)达到最小的$\rho$为最佳的smoothing parameter.

## 6 Statistical inference

在做非参数分析时，我们不仅要估计出回归函数$f(x)$, 我们还有得到$\hat{f}(x)$的标准差和置信区间。

### 6.1 Confidence intervals

### 6.2 Hypothesis test

### 6.3 Derivative estimates

## 7 How to choose a smoother?

下面这段话来自Ruppert, Wand, and Carroll (2003，page 87)

(1) **Convenience**. Is it available on the analyst's favorite computer package?

(2) **Implementability**. If not immediately available, how easy is it to

implement in the analyst's favorite programming language?

(3) **Flexibility**. Is the smoother able to handle a wide range of types of relationships that may exist among the variables of interest?

(4) **Simplicity**. Is it easy to understand how the technique processes the data to obtain answers?

(5) **Tractability**. Is it easy to analyze the mathematical properties of the technique?

(6) **Reliability**. Can the answers be trusted?

(7) **Efficiency**. Does the technique use the data in the most efficient way?

(8) **Extendibility**. Is the technique easily extended to more complicated settings?

# 8 Bayesian P-splines

此方法是由Lang and Brezger (2004) 提出的。

前面我们以truncated power basis函数为例讲到：penalized splines相当于一个linear mixed-effects 模型。下面我们研究一下基于B-样条基函数的penalized splines.

我们采用Eilers and Marx (1996)的差分方法，有：

$$\lambda \int_a^b \left\{ \sum_{j=1}^{J} \beta_j B_j'(x) \right\}^2 dx = \lambda \sum_{j=2}^{J} (\Delta \beta_j)^2 = \boldsymbol{\beta}^T \mathbf{D}_1^T \mathbf{D}_1 \boldsymbol{\beta}, \tag{18}$$

和

$$\lambda \int_a^b \left\{ \sum_{j=1}^{J} \beta_j B_j''(x) \right\}^2 dx = \lambda \sum_{j=3}^{J} (\Delta^2 \beta_j)^2 = \boldsymbol{\beta}^T \mathbf{D}_2^T \mathbf{D}_2 \boldsymbol{\beta}, \tag{19}$$

其中

$$\mathbf{D}_1 = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{(K-1)\times K} , \mathbf{D}_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ 0 & 0 & 1 & -2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{(K-2)\times K}$$

Lang and Brezger (2004) 提出：Eilers and Marx(1996)的一阶和二阶差分等价于如下的一阶和二阶随机游动：

$$\beta_j = \beta_{j-1} + u_j, \quad \beta_j = 2\beta_{j-1} - \beta_{j-2} + u_j,$$

其中$u_j \sim N(0, \lambda^{-1})$. 对于$\beta_1$ 和$\beta_2$，我们可以赋予无信息先验：

$$\pi(b_1) \propto const, \quad \pi(b_2) \propto const.$$

于是对于给定的$\lambda > 0$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)^T$的先验分布为：

$$\pi(\boldsymbol{\beta}|\lambda^{-2}) \propto \exp\left(-\frac{\lambda}{2}\boldsymbol{\beta}^2\mathbf{D}_k^T\mathbf{D}_k\boldsymbol{\beta}\right), \quad k = 1, 2,$$

此即相当于式(18)和(19).

这种方法的一个优势在于：我们将smoothing parameter $\lambda$ 视为了超参数(hyper-parameter), 从而可以赋之先验分布，例如伽玛分布，从而在sampling过程中将之做为一个参数去估计。

# 9 其它应用

## 9.1 极大似然估计

首先我们回忆一下参数模型的极大似然方法。

假定$\{(X_i, Y_i), i = 1, ..., n\}$是来自总体$(X, Y)$的一个样本，我们想估计总体函数$\theta(\cdot)$。令$\ell\{f(X_i), Y_i\}$为第$i$个观测数据$(X_i, Y_i)$的似然函数的对数(log-likelihood),其中$g$为待估函数，则有：

$$\ell(f; \mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \ell\{f(x_i), y_i\}.$$

若函数$f$可参数化为$f(x) = f_\theta(x)$, 其中$\theta$为未知参数，则$\theta$的极大似然估计(MLE)为

$$\hat{\theta}_{\text{MLE}} = \text{argmin}_\theta \sum_{i=1}^{n} \ell\{f_\theta(X_i), Y_i\}.$$

下面我们考虑$f(\cdot)$的非参数估计。以B-样条的penalzied splines 为例，假设：

$$f(x) \approx \mathbf{\Phi}(x)^T \boldsymbol{\beta},$$

其中$\mathbf{\Phi}(x)$为一组B-spline basis functions, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T$。于是我们得到如下的penalized log-likelihood:

$$\ell_p(\boldsymbol{\beta}; \lambda) = \sum_{i=1}^{n} \ell(\mathbf{x}_i^T \boldsymbol{\beta}, y_i) - \frac{\lambda}{2} \boldsymbol{\beta}^T \mathbf{D}_k \mathbf{D}_k \boldsymbol{\beta}. \tag{20}$$

将上式关于$\boldsymbol{\beta}$求最大值，即得$\hat{\boldsymbol{\beta}}$, 从而$\hat{f}(x) = \mathbf{\Phi}(x)^T \hat{\boldsymbol{\beta}}$。

下面我们研究几个具体例子。

（1）考虑如下模型：

$$Y = f(X) + \epsilon,$$

其中$\epsilon \sim N(0, \sigma^2)$, 且$X$与$\epsilon$相互独立。则log-likelihood为

$$-n\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i - f(x_i)\}^2.$$

若采用B-样条的penalized splines，则penalized log-likelihood为

$$-n\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right\}^2 - \frac{\lambda}{2}\boldsymbol{\beta}^T\mathbf{D}_k^T\mathbf{D}_k\boldsymbol{\beta}.$$

然后解其关于$\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T$的最大值问题。

（2）假设我们想研究吸烟跟肺癌的关系，$X$表示烟龄，$Y$表示是否有肺癌，我们想估计吸烟得肺癌的概率，这个概率用$\theta \in (0, 1)$表示，则我们的模型为：

$$\log(\theta/(1-\theta)) = f(x).$$

此时，我们通常假定$Y$服从二项分布$b(1, \theta)$，于是我们可以写出log-likelihood：

（3）普阿松分布：

设$Y_i|X_i \sim Poisson(\lambda_i), i = 1, ..., n,$ 其中$\lambda_i$满足$\log(\lambda_i) = f(x_i)$. 则log-likelihood函数为：

# 10 多元样条回归

对于多维数据，同样存在样条回归问题，只是我们需要考虑多元样条函数。以二元函数为例，假定协变量为$\mathbf{X} = (X_1, X_2)$, 我们可以定义二元B-样条基

函数$B(x_1, x_2)$, 但为了方便，我们通常采用：

$$B(x_1, x_2) = B_1(x_1) * B_2(x_2),$$

即由两个一元的B-样条基函数的乘积构成二元B-样条基函数。

### 10.1 维数诅咒(curse-of-dimensionality)

## 11 应用

例9.1 考虑如下模型

$$Y_i = g(X_i) + \sigma\epsilon_i, i = 1, ..., n,$$

其中$g(x) = x + 2e^{-16x^2}$, $\sigma = 0.4$, $X_i \sim U(-2, 2)$, $\epsilon_i \sim N(0, 1)$. 随机产生数据$\{(X_i, Y_i), i = 1, ..., n\}$, 试估计$g(x)$.

例9.2 考虑如下模型

$$Y_i = g(X_i) + \sigma\epsilon_i, i = 1, ..., n,$$

其中$g(x) = \sin(2x) + 2e^{-16x^2}$, $\sigma = 0.3$, $X_i \sim U(-2, 2)$, $\epsilon_i \sim N(0, 1)$. 随机产生数据$\{(X_i, Y_i), i = 1, ..., n\}$, 试估计$g(x)$.

例9.3 考虑如下模型

$$Y_i = g(X_i) + \sigma\epsilon_i, i = 1, ..., n,$$

其中$g(x) = 0.3\exp\{-4(x+1)^2\} + 0.7\exp\{-16(x-1)^2\}$, $\sigma = 0.1$, $X_i \sim$

$U(-2,2)$, $\epsilon_i \sim N(0,1)$. 随机产生数据$\{(X_i,Y_i), i=1,...,n\}$, 试估计$g(x)$.

例9.4 考虑如下模型

$$Y_i = g(X_i) + \sigma\epsilon_i, i = 1,...,n,$$

其中$g(x) = 0.4x + 1$, $\sigma = 0.15$, $X_i \sim N(0,1)$, $\epsilon_i \sim N(0,1)$. 随机产生数据$\{(X_i,Y_i), i=1,...,n\}$, 试估计$g(x)$.

例9.5 研究加拿大工人收入(age.income数据集)的年龄(age)和收入(income)的关系，该数据来源于R程序包"SemiPar"。总共调查了$n = 205$个加拿大工人的年龄(age)和收入(income), 所有工人的都是高中毕业, 取log.income = log(income，试拟合age与log.income之间的函数关系。

例9.6 我们来研究摩托车碰撞试验的数据集mcycle，该数据集来源于R程序包**MASS**. 试拟合time与acceleration之间的函数关系。

例9.7 我们来研究ethanol数据集，该数据集来源于R程序包**locfit**. 试拟合E与NOx之间的函数关系。

# 参考文献


Fan, J. Q. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC, Boca Raton, USA.

Green, P. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London, United Kingdom.

Hardle, W., Muller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models* . Springer, New York, USA.

Li, Q. and Racine, J. S. (2006). *Nonparametric Econometrics: Theory and Practice* . Princeton University Press, USA.

Ritz, C. and Streibig, J. C. (2008). *Nonlinear Regression with R.* Springer, New York, USA.

Ruppert, D., Wand, M. P. , and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press, United Kingdom.

Simonoff, J. S. (1998). *Smoothing Methods in Statistics* . Springer, New York, USA.

Wang, M. P. and Jones, M. C. (1994). *Kernel Smoothing.* Chapman & Hall/CRC, Boca Raton, USA.

Wu, H. L. and Zhang, J. T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches.* Wiley-Interscience, USA.