基函数模型

钱洋

2016年9月10日

目录

样条基函数与基函数加权和

➢样条基函数

✓经典回归模型

$$E(y|X) = X\beta \Rightarrow y = X\beta + \varepsilon$$

✓由于线性函数的局限，有时需要与非线性函数组合使用，回归中的$X_i\beta$用$\mu(X_i)$表示：

$$\mu(x) = \sum_{h=1}^{H} \beta_h b_h(x),$$

$b = \{b_h\}_{h=1}^{H}$ 表示预先设定的一组基函数，其中 b_h (h=1,2, ⋯H)是一组线性无关组函数，一般是x的非线性函数。

$\beta = (\beta_1, \ldots, \beta_h)$是基函数的系数，即回归系数。

样条基函数与基函数加权和

➢样条基函数

- ✓在数学中，基函数是函数空间一组特殊的基的元素。对于函数空间中的连续函数都可以表示成一系列基函数的线性组合，就像是在向量空间中每个向量都可以表示成基向量的线性组合一样。
- ✓泰勒级数展开是一个非常熟悉的例子，其是将一个函数表示成多个函数的无限累加。通过泰勒级数展开，可以逼近目标函数。对于一个模型，恰当的选择有限基函数的集合是非常有必要的。
- ✓经常的选择是高斯径向基函数族(Gaussian radial basis functions)。

$$b_h(x) = \exp\left(-\frac{|x - x_h|^2}{l^2}\right),$$

x_h表示基函数的中心点，l表示基函数的宽度参数(width parameter),基函数的数量和宽度参数控制函数的范围(scale)。

x_h

样条基函数与基函数加权和

$$y(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

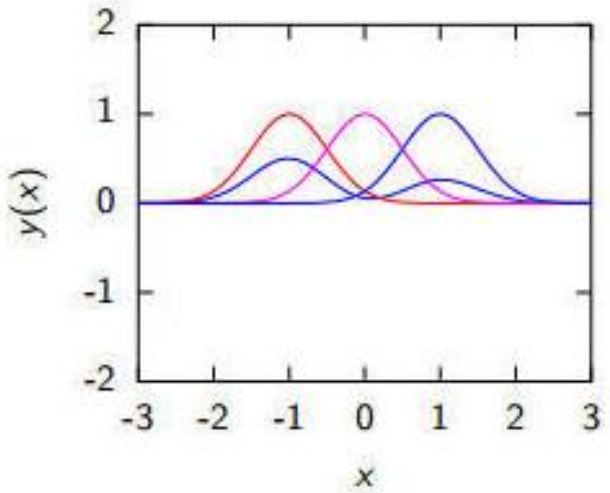


Figure: Function from radial basis with weights $w_1 = 0.50596$, $w_2 = -0.046315$, $w_3 = 0.26813$.

Figure: Radial basis functions.

$$y(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

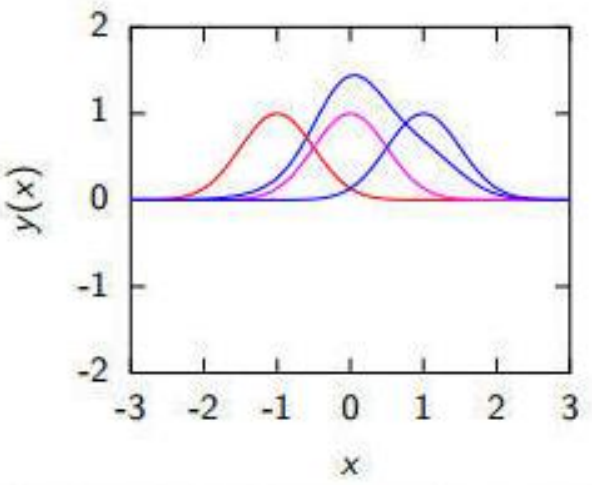


Figure: Function from radial basis with weights $w_1 = 0.07179$, $w_2 = 1.3591$, $w_3 = 0.50604$.

$w_2 = -0.18924$, $w_3 = -1.8183$.

x_h

样条基函数与基函数加权和

➢样条基函数

✓**样条函数：**一类分段（片）光滑、并且在各段交接处也有一定光滑性的函数。

给定一组平面上顶点 (x_i,y_i) $(i=0，1，\ldots，n)$，并设在区间$[a,b]$上的Δ：$a=x_0<x_1<\ldots<x_{n-1}<x_n=b$,那么在〔$a,b$〕上的一个函数 $S(x)$ 称为*K*阶连续样条函数，如果它满足下面两个条件：

(1)在每个小区间〔x_{i-1},x_i〕$(i=1，2，\ldots，n)$内，$S(x)$是具有*K*阶或*K*阶以上连续函数。

(2)在x_i $(i=1，2，\ldots，n-1)$处成立

x_h

样条基函数与基函数加权和

➢样条基函数

✓三次样条函数

假设在区间〔a，b〕上给定一个分割

$$\Delta:\quad a=x_0<x_1<\ldots<x_{n-1}<x_n=b,$$

在〔a,b〕上的一个函数$S(x)$称为插值三次样条函数，
如果满足下列条件:

(1)在每一小区间〔x_{i-1},x_i〕(i=1，2，…，n)内$S(x)$分别是三次多项式函数;

(2)在节点x_i(i=1，2，…，n-1)处成立：

$$S^{(k)}(x_i-0)=S^{(k)}(x_i+0),k=0,1,2,$$

即小区间上的三次多项式函数，在拼接点处x_i具有二阶连续拼接。

(3)满足插值条件$y_i=S(x_i),i$=0，1，…，n.

x_h

样条基函数与基函数加权和

➢样条基函数

✓经常选择的基函数族是B-样条。

✓B-样条是指定义在一些有条件节点上的连续分段函数。当节点等距离时，即 $x_{h+k} = x_h + \delta k$ ，为均匀(uniform)B样条。一个高阶的B样条基函数可以由低阶的B样条基函数迭代得到。

✓三次B-样条基础函数被定义为分段的三次方多项式。

$$b_h(x) = \begin{cases} \frac{1}{6}u^3 & \text{for } x \in (x_h, x_{h+1}), \quad u = (x - x_h)/\delta \\ \frac{1}{6}(1 + 3u + 3u^2 - 3u^3) & \text{for } x \in (x_{h+1}, x_{h+2}), \quad u = (x - x_{h+1})/\delta \\ \frac{1}{6}(4 - 6u^2 + 3u^3) & \text{for } x \in (x_{h+2}, x_{h+3}), \quad u = (x - x_{h+2})/\delta \\ \frac{1}{6}(1 - 3u + 3u^2 - u^3) & \text{for } x \in (x_{h+3}, x_{h+4}), \quad u = (x - x_{h+3})/\delta \\ 0 & \text{otherwise.} \end{cases} \tag{20.2}$$

✓基函数的宽度由节点之间的距离δ控制。模型的最大弹性，由数据范围内的均匀节点的数量控制。同时，节点也可以是非等距的，可以构造非均匀B样条。

x_h

样条基函数与基函数加权和

➢样条基函数

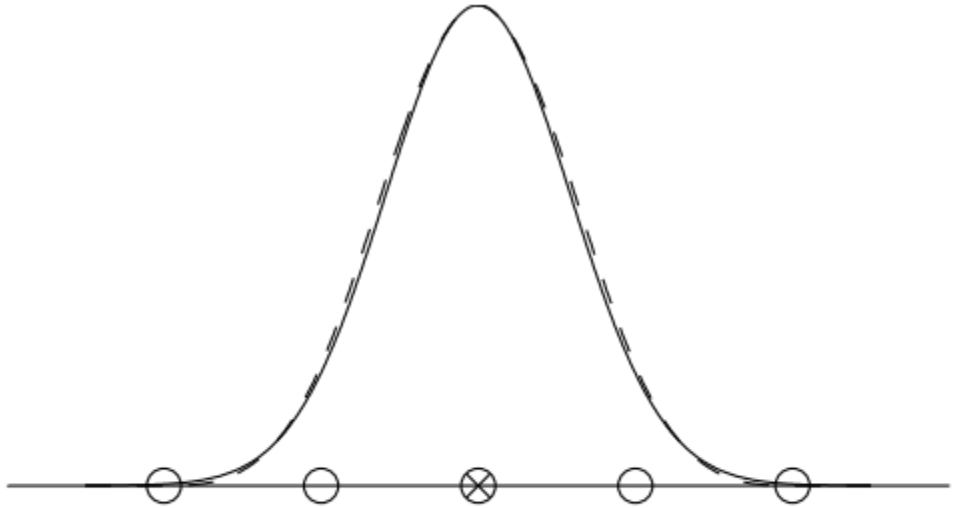


Figure 20.1 *Single Gaussian (solid line) and cubic B-spline (dashed line) basis functions scaled to have the same width. The X marks the center of the Gaussian basis function, and the circles mark the location of knots for the cubic B-spline.*

实线表示高斯函数曲线，虚线表示相同宽度的B样条基函数曲线。这种形状加权和的形式(权重可以是负，可以是正，也可以是0)可以用来模拟平滑函数。

样条基函数与基函数加权和

➢样条基函数

$$\mu(x) = \sum_{h=1}^{H} \beta_h b_h(x),$$

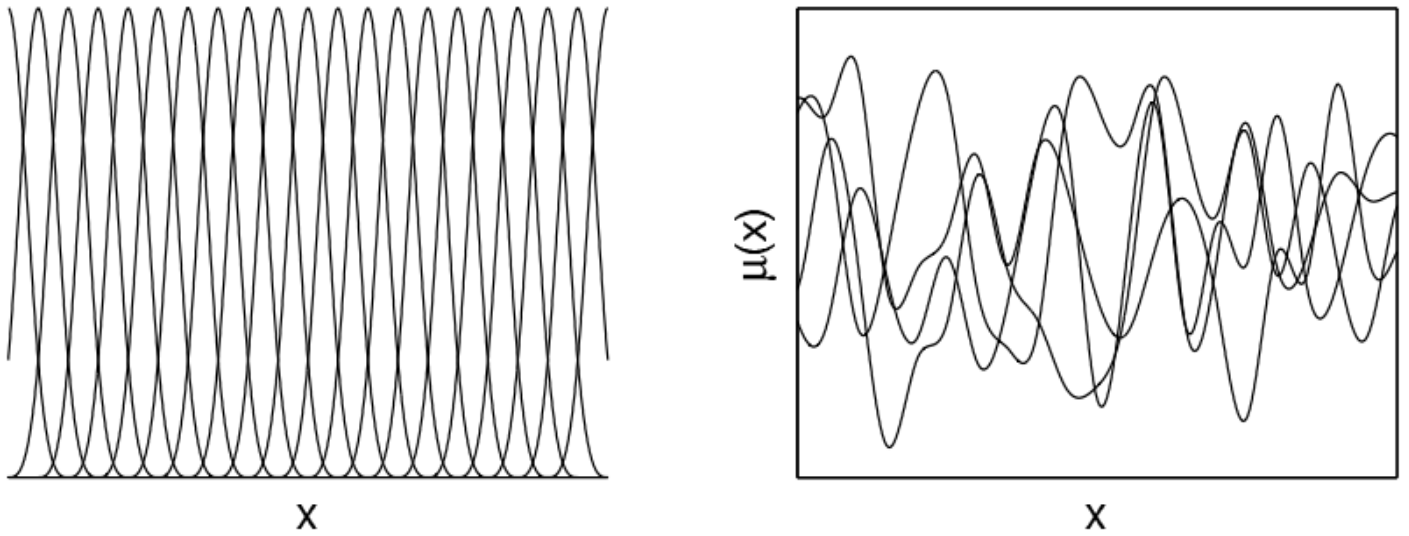


Figure 20.2 *(a) A set of cubic B-splines with equally spaced knots. (b) A set of random draws from the B-spline prior for $\mu(x)$ based on the basis functions in the left graph, assuming independent standard normal priors for the basis coefficients.*

图显示了一组B样条基函数，基系数独立服从标准正太分布，随机抽取基系数β_h作为权重。样条函数的数量H影响结果模型$\mu(x)$的灵活性。

样条基函数与基函数加权和

➢样条基函数

✓ 考虑到选择的基$b = \{b_h\}_{h=1}^{H}$，模型相对参数是线性的，重新将模型表示成：

$$y_i = \mu(x_i) + \varepsilon_i = w_i\beta + \varepsilon_i \qquad w_i = (b_1(x_i), \ldots, b_H(x_i))$$

✓ 我们将模型看成是关于β的线性模型，这样就可以用线性回归的方法来拟合模型。例如，(β, δ^2)先验服从多元正太逆-x^2，在给定数据$(x_i, y_i)_{i=1}^{n}$，(β, δ^2)共轭后验分布仍然服从多元正太逆-x^2。

$$\begin{aligned} \mu|\sigma^2 &\sim \mathrm{N}(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2), \end{aligned}$$

样条基函数与基函数加权和

➢样条基函数

联合先验概率为：

$$p(\mu,\sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2]\right).$$

联合后验概率为：

$$\begin{aligned} p(\mu,\sigma^2|y) &\propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2]\right) \times \\ &\quad \times (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) \\ &= \text{N-Inv-}\chi^2(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2), \end{aligned}$$

$$\begin{aligned} \mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2 \end{aligned}$$

样条基函数与基函数加权和

➢样条基函数

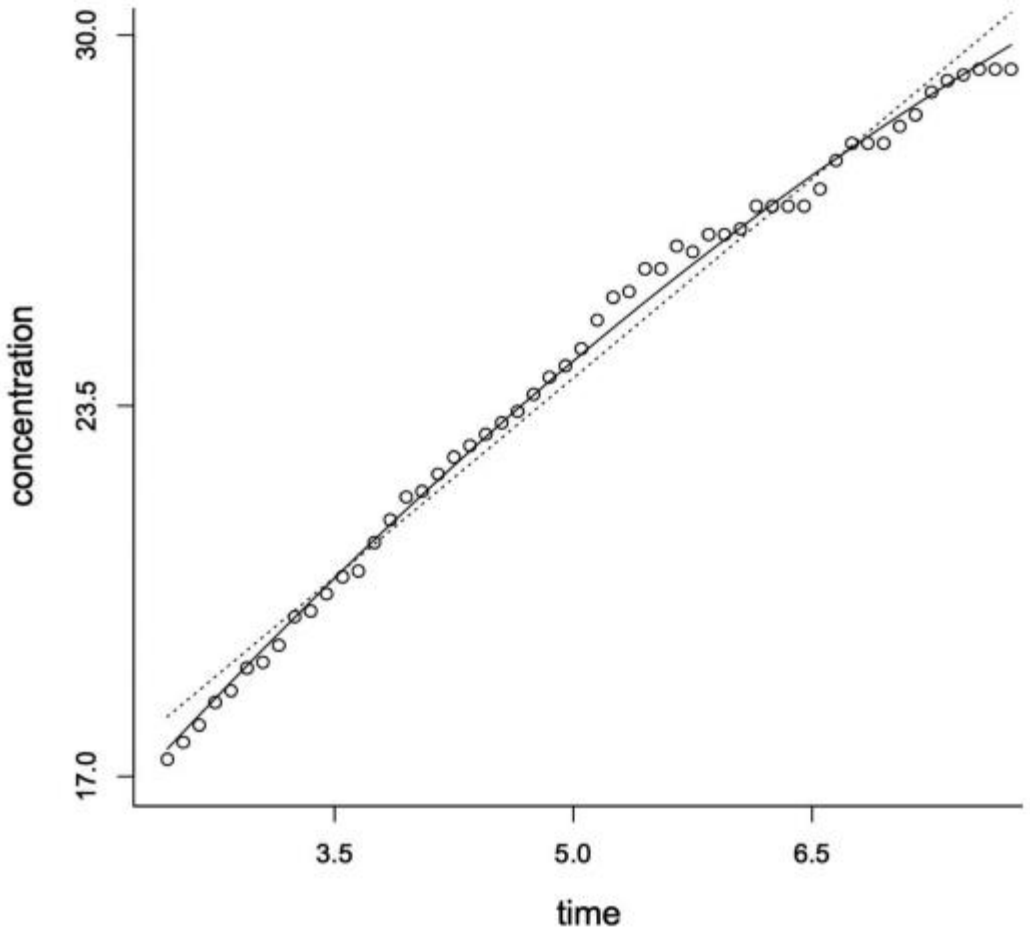
✓ 尽管模型是关于参数β线性的，但是模型可以通过基函数的线性组合精确逼近。通常使用的是基函数模型与线性模型的组合。

$$\mu(x) = \beta_1 + \beta_2 x + \sum_{h=3}^{H+2} \beta_h b_h(x).$$

样条基函数与基函数加权和

➢ 氯离子浓度案例

实验所用数据集来自于一个生物实验，测量的是一小段时间间隔内的氯离子浓度，包括54个测量值。



图中的小圆点表示测量的氯离子浓度。可以看出数据很接近一条直线，但是在局部却存在明显偏差。虚线表示用线性回归拟合的直线，弯曲的实线表示采用B样条得到的后验均值曲线。

$E(\mu(x)|y)$表示后验均值

样条基函数与基函数加权和

- 氯离子浓度案例

✓ 在本案例中，有21个参数需要估计，即$(\beta_1, \ldots, \beta_{21})$，但却只有54个数据点。如果我们不引入先验信息去估计基的系数，这样是有问题的。有不少策略可以调解数据贫乏问题(data sparsity)。一种可行策略是对$\mu(x)$的参数方程引入一个非参数先验。

$$\beta|\sigma \sim N(\beta_0, \sigma^2\lambda^{-1}I_H)$$

$$\sigma^2 \sim Inv-gamma(a_0, b_0)$$

✓ 在x处的预测值，曲线的先验期望可以表示为：

$$\mu_0(x) = E\mu(x) = \sum_{h=1}^{H} \beta_{0h}\, b_h(x)$$

样条基函数与基函数加权和

- 氯离子浓度案例

✓ 假设 $\mu_0(x) = \alpha + \psi x$ 先验的均值是线性的。我们通过最小二乘法估计 β_0 的值，生成 $\mu_0(x)$ 使其尽可能的接近$\alpha + \psi x$，我们发现在这个案例中H=21时，$\mu_0(x)$与 $\alpha + \psi x$ 的值是没有区别的。为了简单起见，我们可以将α与ψ最小二乘估计带入，得到后验的均值是:

$$\hat{\mu}(x) = \mathrm{E}\left(\mu(x)|(x_1, y_1), \ldots, (x_n, y_n)\right) = (W^T W + \lambda I_H)^{-1}(W^T y + \lambda \hat{\mu}_0(x)),$$

✓ $\hat{\mu}_0(x)$表示最小二乘法得到的回归曲线上的值，$W = (w_1, \ldots, w_n)^T$，后验均值$\hat{\mu}_0(x)$收缩于线性回归估计。应对数据贫乏性问题，线性回归拟合中，允许非参偏差。为了更加完整的分析，我们可以对α与ψ设置一个超参数先验或者选择平滑先验。一阶自回归模型中，经常使用的是贝叶斯惩罚样条(Bayesian penalized (P) splines)。

✓ 在应用样条函数时，一个很重要的方面是节点数量以及节点的位置。一般地，我们通过均匀或者回归变量相等概率的样本分位数来确定节点的位置。这样只剩下确定节点的数目。若节点数目过多，则会过分的拟合数据（over-fitting）,若节点数目过少，则会过分光滑数据（under fitting）。

样条基函数与基函数加权和

- 氯离子浓度案例

✓ 对于节点的数目的选择，通常有两种处理办法。其一，是在一定的标准下，最优化节点数目，此种方法一般难于计算，只能在一定范围内达到局部最优解；其二，是惩罚样条（penalized (P) splines）。惩罚样条方法先选择较大的节点数目，通过惩罚项来防止模型过分拟合数据，因此惩罚样条方法一般选择的节点数目比较保守，参数数目较多。

✓ 当假定节点数目是随机的，因此模型的维数是不确定的，我们可以对节点的数目及位置设置一个先验，运用适应维数变化的可逆的跳MCMC方法(reversible jump MCMC)。但在实际中，设计有效的可逆的跳算法是具有挑战性的。

✓ 通过先验的选择，放松变量的选择。不把 β_h 的先验严格设置为0，而是通过收缩系数使其接近于0。重尾(heavy tails)是为了避免过度收缩重要的基函数的系数。

样条基函数与基函数加权和

- 贝叶斯模型平均估计及算法

$$y = f(x) + \varepsilon \quad \varepsilon \sim N(0,\ \sigma^2)$$

- 假设模型的光滑回归函数是三次B样条函数，样条函数的基 $\pi(x)$

$$y_i = \pi^{\tau}(x_i)\beta + \varepsilon_i \qquad \varepsilon_i \sim N(0,\ \sigma^2) \qquad \text{观测数据} D = (y_i, x_i)_{i=1}^{n} \qquad \theta=(\beta, \sigma^2)$$

- 给定结点个数，用回归变量等概率的样本分位数来确定结点的位置，如选择3个结点，则回归变量选择回归变量的25th,50th,75th样本分位数。在给定样条阶数和确定结点的位置，样条函数的基由结点个数唯一确定。

$$M = \{0, \cdots, k, \cdots\} \qquad k\text{表示回归样条结点的个数}$$

- 选择模型k的先验为$p(k)$,给定模型k的情况下，参数的先验为$p(\theta|k)$
- 在贝叶斯回归中，$\theta=(\beta, \sigma^2)$的有效而且易于计算的先验是共轭的正太逆 $gamma$ 分布。

$$p(\beta, \sigma^2|k) = N(0, \sigma^2\lambda I_N) IG(\sigma^2; a_0, b_0) \qquad IG(x; a, b) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} \exp\left(-\frac{b}{x}\right)$$

λ, a_0, b_0为常数，$IG(x; a, b)$参数为a, b的逆$gamma$概率密度函数,I_N表示N*N的单位阵，N为回归样条的维数。

样条基函数与基函数加权和

- 贝叶斯模型平均估计及算法
- 给定模型k，参数向量θ的联合后验分布为：

$$p(\theta|D,k)=p(\beta,\sigma^2|D,k)\propto p(\beta,\sigma^2|X,Y,k)=\frac{p(\beta,\sigma^2,X,Y,k)}{p(X,Y,k)}$$

$$=\frac{p(Y|\beta,\sigma^2,X,k)p(\beta,\sigma^2,X,k)}{p(X,Y,k)}=\frac{p(Y|\beta,\sigma^2,X,k)p(\beta|\sigma^2,X,k)p(\sigma^2,X,k)}{p(X,Y,k)}$$

$$=\frac{p(Y|\beta,\sigma^2,X,k)p(\beta|\sigma^2,X,k)p(\sigma^2|X,k)p(X,k)}{p(Y|X,k)p(X,k)}$$

$$\propto p(Y|\beta,\sigma^2,X,k)p(\beta|\sigma^2,k)p(\sigma^2)$$

$$p(\beta,\sigma^2|D,k)=p(\beta|\sigma^2,D,k)p(\sigma^2|D,k)$$

其中：$p(\beta|\sigma^2,D,k)$ 与 $p(\sigma^2|D,k)$ 为 β 与 σ^2 的后验分布，后验分布仍是正太逆 $gamma$分布。

样条基函数与基函数加权和

- 贝叶斯模型平均估计及算法

$$p(X=(\pi(x_1),\cdots,\pi(x_n))$$
$$Y=(y_1,\cdots,y_n)$$

$$p(\beta,\sigma^2\mid D,k)\propto(V=\left(X^TX+\frac{1}{\lambda}I\right)^{-1}$$
$$=(\sigma^2)^{-\left(\frac{n}{2}+\frac{N}{2}+\alpha_0+1\right)}\mathrm{e}\,H=X^TY$$

$$a=\frac{n}{2}+a_0$$

$$b=\left(\frac{1}{2}Y^TY+\frac{1}{2}H^TVH+\frac{1}{b_0}\right)^{-1}$$

$$\beta)-\frac{1}{2\lambda\sigma^2}\beta^T\beta-\frac{1}{b_0\sigma^2}\Big\}$$
$$VH-\frac{1}{2\sigma^2}Y^TY-\frac{1}{b_0\sigma^2}\Big\}$$

样条基函数与基函数加权和

- 贝叶斯模型平均估计及算法

$$p\left(\beta,\sigma^2 \mid D,k\right)\propto\left(\sigma^2\right)^{-\left(\frac{n}{2}+\frac{N}{2}+\alpha_0+1\right)}\exp\left\{-\frac{1}{2\sigma^2}\left(Y-X\beta\right)^T\left(Y-X\beta\right)-\frac{1}{2\lambda\sigma^2}\beta^T\beta-\frac{1}{b_0\sigma^2}\right\}$$

$$=\left(\sigma^2\right)^{-\left(\frac{n}{2}+\frac{N}{2}+\alpha_0+1\right)}\exp\left\{-\frac{1}{2\sigma^2}\left(Y-VH\right)^T V^{-1}\left(Y-VH\right)+\frac{1}{2\sigma^2}H^TVH-\frac{1}{2\sigma^2}Y^TY-\frac{1}{b_0\sigma^2}\right\}$$

$$p\left(\beta \mid \sigma^2,D,k\right)\propto\exp\left(-\frac{1}{2\sigma^2}\left(\beta-VH\right)^T V^{-1}\left(\beta-VH\right)\right)$$

$$p\left(\sigma^2 \mid \beta,D,k\right)\propto\left(\sigma^2\right)^{-\left(\frac{n}{2}+\alpha_0+1\right)}\exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{2}Y^TY-\frac{1}{2}H^TVH+\frac{1}{b_0}\right)\right\}$$

$$p\left(\beta \mid \sigma^2,k,D\right)=N\left(VH,\sigma^2V\right)$$

$$p\left(\sigma^2 \mid D,\ \ k\right)=IG\left(\sigma^2;a,b\right)$$

样条基函数与基函数加权和

- 贝叶斯模型平均估计及算法

取k为无信息先验，$p(k) = U(0,1,\cdots k_{max})$,$k_{max}$是表示节点个数可能的最大值。$\theta, k$的联合先验为：

$$p(\theta,k) = p\left(\beta \mid \sigma^2,k\right) p\left(\sigma^2\right) p(k)$$

- 若给定模型的k,θ取正太倒$gamma$先验，则θ, k的联合后验为：

$$\begin{aligned} p(\theta,k \mid D) &\propto p(Y \mid \theta,k)\, p(\theta,k) = p(Y \mid \theta,k)\, p\left(\beta \mid \sigma^2,k\right) p\left(\sigma^2\right) p(k) \\ &= N\left(X\beta, I_n\sigma^2\right) N\left(0,\lambda I_N\sigma^2\right) IG\left(\sigma^2;a_0,b_0\right) \end{aligned}$$

样条基函数与基函数加权和

- 贝叶斯模型平均数值例子

$$f_1(x)=2\sin(0.15\pi x),x\in[0,10].$$

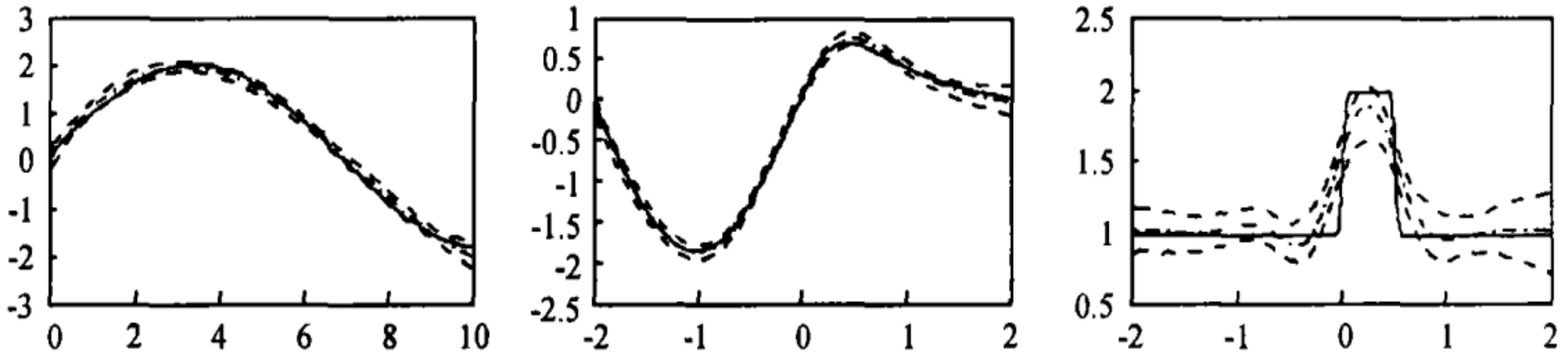
$$f_2(x)=\frac{2\sin\left(\frac{\pi}{2}x\right)}{1+2x^2(\mathrm{sgn}(x))+1},x\in[-2,2].$$

$$H(x)=\frac{1}{1+\exp(-x)},\quad f_3(x)=H(100x)+H(-100(x-0.5)),x\in[-2,2].$$

- 其中，$f_1(x)$中的x来自于均匀分布$U(0,\ 1)$,产生50组样本。$f_2(x),f_3(x)$中的x来自$N(0,\ 1)$正太分布。三个函数的$\varepsilon\sim N(0,\ 0.15^2)$。
- 在贝叶斯模拟运算中，假设$\sigma^2\sim IG(\sigma^2,1,\ 1)$,样条基的参数服从$\beta\sim N(0,\ 1000\sigma^2 I)$，将函数的定义区间100等分，从后验分布$p(\theta,k|D)$抽样2000次，舍弃前500次,计算出B样条在这些等分位点的后验均值和后验分位数，进而得到函数的估计及后验区间估计。

样条基函数与基函数加权和

- 贝叶斯模型平均数值例子



- 其中图中的实线为真实的函数。

变量选择和系数收缩

- 变量选择：$(X_1, X_2 \ldots X_H)$是H个回归自变量，这些回归自变量可能有许多与响应变量y无关。这就需要从这些可能的回归自变量中找出与y有关的回归自变量的子集，寻找回归自变量子集的过程称之为变量选择。

变量选择和系数收缩

- 原文献George E I, McCulloch R E. Variable selection via Gibbs sampling[J]. Journal of the American Statistical Association, 1993, 88(423): 881-889.
- 在回归情境下，包含观测变量Y与X。

$$Y \mid \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I\right)$$

- Y表示$n*1$的向量，$X=[X_1, X_2, \ldots, X_p]$表示$n*p$的矩阵，$\beta=[\beta_1, \beta_2, \ldots, \beta_p]$。其中，$\beta$，$\sigma^2$均是未知的。
- 假设模型的β来自于两个正太混合，引入隐变量$\gamma_i = 0\ or\ 1$，我们可以将这种正太混合表示成：

$$\beta_i \mid \gamma_i \sim (1-\gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i^2\tau_i^2) \qquad (2)$$

$$P(\gamma_i = 1) = 1 - P(\gamma_i = 0) = p_i. \qquad (3)$$

变量选择和系数收缩

- 为了获取(2)式，我们假设$\beta_i|\gamma_i$的先验为多元正太先验。

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_p(0, \mathbf{D}_\gamma \mathbf{R} \mathbf{D}_\gamma), \tag{4}$$

- 其中，$\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$，R是相关矩阵,一般取$I$。

$$\mathbf{D}_\gamma \equiv \operatorname{diag}[a_1\tau_1, \ldots, a_p\tau_p], \tag{5}$$

- 如果$\gamma_i = 0, a_i = 1$;如果$\gamma_i = 1, a_i = c_i$
- 对于残差σ^2

$$\sigma^2|\boldsymbol{\gamma} \sim \mathrm{IG}(\nu_\gamma/2, \nu_\gamma\lambda_\gamma/2), \tag{6}$$

- γ的边际后验分布

$$f(\boldsymbol{\gamma}|\mathbf{Y}) \propto f(\mathbf{Y}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma}),$$

变量选择和系数收缩

- 获取$f(\gamma)$

$$f(\boldsymbol{\gamma}) = \prod p_i^{\gamma_i}(1-p_i)^{(1-\gamma_i)}. \qquad (7)$$

- Gibbs抽样

$$\begin{aligned}\boldsymbol{\beta}^j &\sim f(\boldsymbol{\beta}^j|\mathbf{Y}, \sigma^{j-1}, \boldsymbol{\gamma}^{j-1}) \\ &= N_p(\mathbf{A}_{\boldsymbol{\gamma}^{j-1}}(\sigma^{j-1})^{-2}\mathbf{X}'\mathbf{X}\hat{\beta}_{\mathrm{LS}}, \mathbf{A}_{\boldsymbol{\gamma}^{j-1}}), \qquad (13)\end{aligned}$$

$$\mathbf{A}_{\boldsymbol{\gamma}^{j-1}} = ((\sigma^{j-1})^{-2}\mathbf{X}'\mathbf{X} + \mathbf{D}_{\boldsymbol{\gamma}^{j-1}}^{-1}\mathbf{R}^{-1}\mathbf{D}_{\boldsymbol{\gamma}^{j-1}}^{-1})^{-1}.$$

$$\begin{aligned}\sigma^j &\sim f(\sigma^j|\mathbf{Y}, \boldsymbol{\beta}^j, \boldsymbol{\gamma}^{j-1}) \\ &= \mathrm{IG}\left(\frac{n+\nu_{\boldsymbol{\gamma}^{j-1}}}{2}, \frac{|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^j|^2 + \nu_{\boldsymbol{\gamma}^{j-1}}\lambda_{\boldsymbol{\gamma}^{j-1}}}{2}\right), \qquad (14)\end{aligned}$$

变量选择和系数收缩

- Gibbs抽样

$$\gamma_i^j \sim f(\gamma_i^j \mid \mathbf{Y}, \beta^j, \sigma^j, \gamma_{(i)}^j) = f(\gamma_i^j \mid \beta^j, \sigma^j, \gamma_{(i)}^j), \quad (15)$$

$$P(\gamma_i^j = 1 \mid \beta^j, \sigma^j, \gamma_{(i)}^j) = \frac{a}{a+b}, \quad (16)$$

where

$$a = f(\beta^j \mid \gamma_{(i)}^j, \gamma_i^j = 1) \times f(\sigma^j \mid \gamma_{(i)}^j, \gamma_i^j = 1) f(\gamma_{(i)}^j, \gamma_i^j = 1) \quad (16a)$$

and

$$b = f(\beta^j \mid \gamma_{(i)}^j, \gamma_i^j = 0) \times f(\sigma^j \mid \gamma_{(i)}^j, \gamma_i^j = 0) f(\gamma_{(i)}^j, \gamma_i^j = 0). \quad (16b)$$

- 当v_γ，λ_r为常量时，即$v_\gamma = v$，$\lambda_r = \lambda$

$$a = f(\beta^j \mid \gamma_{(i)}^j, \gamma_i^j = 1) p_i \quad (16c) \qquad b = f(\beta^j \mid \gamma_{(i)}^j, \gamma_i^j = 0)(1 - p_i). \quad (16d)$$

变量选择和系数收缩

- 模拟例子
- 问题一
- $p=5, X=(X_1, X_2, \ldots X_5)$60个观测值，独立服从$N_{60}(0,1)$。

$$Y = X_4 + 1.2X_5 + \varepsilon \quad \varepsilon \sim N_{60}(0, \sigma^2 I) \quad \sigma = 2.5$$

- $\beta = (0,0,0,1,1.2)$。基于最小二乘估计得到的结果为$\hat{\beta} = (0.03, -0.45, 0.25, 0.84, 1.29)$。
- $\hat{\sigma}_\beta = (0.36, 0.40, 0.36, 0.31, 0.33)$； $\hat{\sigma}$ =2.58。
- 问题二
- $X_3^* = \mathrm{X_5} + 0.15\mathrm{Z}$ $\mathrm{Z} \sim \mathrm{N}(0,\ 1)$ X_3是可以表示X_5的。问题二打算解决的是共线性问题。基于最小二乘估计的结果为$\hat{\beta} = (0.01, -0.38, 0.34, 0.83, 0.95)$.
- $\hat{\sigma}_\beta = (0.35, 0.39, 2.33, 0.31, 2.35)$; $\hat{\sigma}$ =2.59。
- 设置$f(\gamma) \equiv \frac{1^5}{2}$； $\tau_1, \ldots, \tau_5 = 0.33$； $c_1, \ldots c_5 = 10$； $R = I$; $v_\gamma \equiv 0$;

变量选择和系数收缩

- 模拟例子
- 采用Gibbs抽样5000次，获取$\gamma^1, \gamma^2 \ldots, \gamma^{5000}$，制作列表，并统计频率，并取top-n个最高频率对应的模型。

Table 1. High Frequency Models, Example 4.1

Problem 1		*Problem 2*	
Model variables	*Proportion*	*Model variables*	*Proportion*
5	.258	3	.146
4 5	.242	5	.123
2 5	.070	3 4	.098
2 4 5	.055	4 5	.086

变量选择和系数收缩

- 关注具有高斯残差的非参数回归模型，设置 $b = \{b_h\}_{h=1}^{H}$ 表示预先设定的潜在基函数集合。有：

$$y_i \sim \mathrm{N}(w_i\beta, \sigma^2), \quad w_i = (b_1(x_i), \ldots, b_H(x_i)).$$

- 在实际应用中，存在着不确定性，所以基函数确实是有必要的。使用贝叶斯做变量选择时，考虑基函数是否应该被排除在模型之外。我们引入模型的索引(index)，$\gamma = (\gamma_1, \ldots, \gamma_H) \in \Gamma$

  $\gamma_h = 1$表示基函数 b_h是被包括的。$\gamma_h = 0$ 表示基函数应该被排除。 Γ 表示模型空间，有 2^H种可能。非零系数 $\beta_\gamma = \{\beta_h : \gamma_h = 1\}$ 需要设置先验。针对整个模型空间，设置如下先验：

$$\beta_h \sim \pi_h\delta_0 + (1-\pi_h)\mathrm{N}(0, \kappa_h^{-1}\sigma^2), \quad \sigma^2 \sim \text{Inv-gamma}(a, b), \qquad (20.3)$$

δ_0表示退化分布(随机变量，以概率1取常数P(X=C)=1,则称这个分布为退化分布或单点分布)。20.3中设置 $\beta_h = 0$ 的概率为 π_h，设置非零系数服从先验分布 $N(0, \kappa_h^{-1}\sigma^2)$。对 $h = 1, \ldots, H$ γ_h独立服从 $\gamma_h \sim Bernoulli(1-\pi_h)$ 并且 $\beta_\gamma \sim N_{p_\lambda}(0, V_\lambda\sigma^2)$，$p_\gamma = \sum_h \gamma_h$ 表示模型中基函数的数量，$V_\gamma = diag(\kappa_h : \gamma_h = 1)$ 。模型20.3成为变量选择混合先验。

变量选择和系数收缩

- 在缺乏先验信息的情况下，某些基系数更有可能被考虑在内。我们设置 $\pi_h = \pi$，选择一个超参数先验，$\pi \sim Beta(a_\pi, b_\pi)$。这种先验也会引起自动贝叶斯多重性的调整，这会导致基系数为0的情况增加，从而避免了很多不必要的基系数。这种调整从π的全条件后验分布看非常清晰。π的全条件后验分布的简单形式如下：

$$\pi \mid - \sim Beta\left(a_\pi + \sum_h (1-\gamma_h),\ b_\pi + \sum_h \gamma_h\right)$$

- 为了产生基函数系数的重尾柯西先验，设置κ_h独立服从 $\kappa_h \sim Gamma(0.5, 0.5)$ 对于非零系数不应该选择不恰当的先验，因为这会导致高后验概率(Section 7.4)。方差的先验的参数可以设置 $a, b \to 0$

为了简单起见，可以设置固定π和$\kappa_h = \kappa$，联合后验分布取共轭分布。

$$\Pr(\gamma \mid y, X) = \frac{p(X,\gamma)\,p(y \mid X,\gamma)}{p(y,X)} = \frac{\pi^{k-p_\gamma}(1-\pi)^{p_\gamma}\,p(y \mid X,\gamma)}{\sum_{\gamma^* \in \Gamma} \pi^{k-p_{\gamma^*}}(1-\pi)^{p_{\gamma^*}}\,p(y \mid X,\gamma^*)}$$

变量选择和系数收缩

- $p(y|X,\gamma)$表示在γ下的似然

$$p(y|X,\gamma)=\int\prod_{i=1}^{n}\mathrm{N}(y_i|w_{i,\gamma}\beta_\gamma,\sigma^2)\mathrm{N}(\beta_\gamma|0,V_\gamma\sigma^2)\text{Inv-gamma}(\sigma^2|a,b)d\beta_\gamma d\sigma^2$$

- 很遗憾，尽管模型γ的后验能够得到，但却很难计算出来。除非基函数H的数量很小。因为有大量的不同的模型（2^H）在分母中相加计算。所以，除非在低维的情形之下，可采用这种计算。否则，近似计算是必须使用的。一种可采用的策略是基于MCMC的随机搜索算法来确定 Γ 中的高后验概率模型，然后对这些模型取平均。

- 另一种，可行的方法是利用Gibbs抽样更新γ_h

$$p\left(\gamma_h=1\,|\,\gamma_{(-h)},\pi\right)=\frac{(1-\pi)\,p\left(y\,|\,X,\gamma_h=1,\gamma_{(-h)}\right)}{(1-\pi)\,p\left(y\,|\,X,\gamma_h=1,\gamma_{(-h)}\right)+\pi\,p\left(y\,|\,X,\gamma_h=0,\gamma_{(-h)}\right)}$$

变量选择和系数收缩

- 从上述抽样中，我们能够引导模型的选择，抛弃不必要的基函数，从而简化模型。给定一个0-1损失函数，如果选择了不型则损失为1，否则为0。由此估计γ的边际后验分布，使得边际后验分布出现频率最大的模型就是我们要选择的模型。不幸的是，除非H很小，否则会出现很多模型，这些模型的后验模型概率与取得最大后验概率模型的很相似。这会误导我们对基函数选择的推断。针对这个问题，γ的后验模型平均更能反映基函数选择的不确定性，如果对选择单个模型感兴趣，中位数概率模型[变量指示器后验概率大于后者等于0.5，即$Pr(\gamma_h = 1|data) > 0.5$]可能比最大后验概率模型有更好的性能。

变量选择和系数收缩

- 氯离子浓度案例(继续)
- 我们对氯离子浓度案例再次进行分析，使用贝叶斯变量选择来解释不确定性，并使用样条基函数来刻画曲线。
- 如果模型21个基函数全部被包含在内，并且基系数的弱信息先验为$N(0,\ I)$或者$N(0,\ 2^2I)$。我们得到的拟合极其不好，后验均值曲线显著的向下背离数据，向x轴靠拢。我们考虑到含有21个基函数的模型是全模型，我们指定每个基函数被包含在内的概率为0.5。基函数的系数独立服从$N(0,\ 2^2)$的先验。残差σ^2独立服从$Inv-gamma(1,1)$先验，我们采用Gibbs抽样更新β_h和σ，在每一次迭代中，β_h的不同的子集都会自动分配值0。运行Gibbs抽样到近似收敛，基函数数量的后验均值为12.0，95%的后验区间[8.0，16.0]，标准差的后验均值为$\hat{\sigma}$=0.27,95%的后验区间为[0.23,0.33]，这表明误差很小。
- 使用贝叶斯变量选择方法来解释潜在的缺点是可能对初始基的选择有些敏感。例如，H =21，我们使用3次-B样条，传达了一些的隐含信息，即曲线会被完全光滑没有很急剧的转折点和峰值点。在许多的应用中，这是合理的，但是当峰值函数需要先验时，可以使用小波或者其他的基函数。

变量选择和系数收缩

- 先验收缩
- 适当的允许基函数的系数为0，进而减少模型的基函数，这种方式是非常不错的，但是却带来了计算上的代价。当模型的数量空间2^H非常大时，使用MCMC算法不能够很好的收敛，在成千上万次迭代过程中，只有一小部分模型被访问到。每次更新γ时，马尔科夫链混合速度较慢。另外，当非共轭先验用于精确计算时，这也会变得很困难。这些问题不会出现在氯离子数据应用中，因其实际计算时间远小于1分钟，1分钟可以跑很多次MCMC迭代，并且我们考察的每一种情况混合的都很好。但是，模型扩展到多元预测因子问题时，问题可能就会出现。
- 针对此问题，一种可行的解决方案是避免设置任何系数严格的等于0，而是使用正则化或者收缩先验(Section 14.6),这种恰当的先验是在0附近有很高的密度，对应的基函数就能够很有效的排除在外，同时对于重尾也避免了过度收缩。一种常用的收缩先验如下高斯混合的形式：

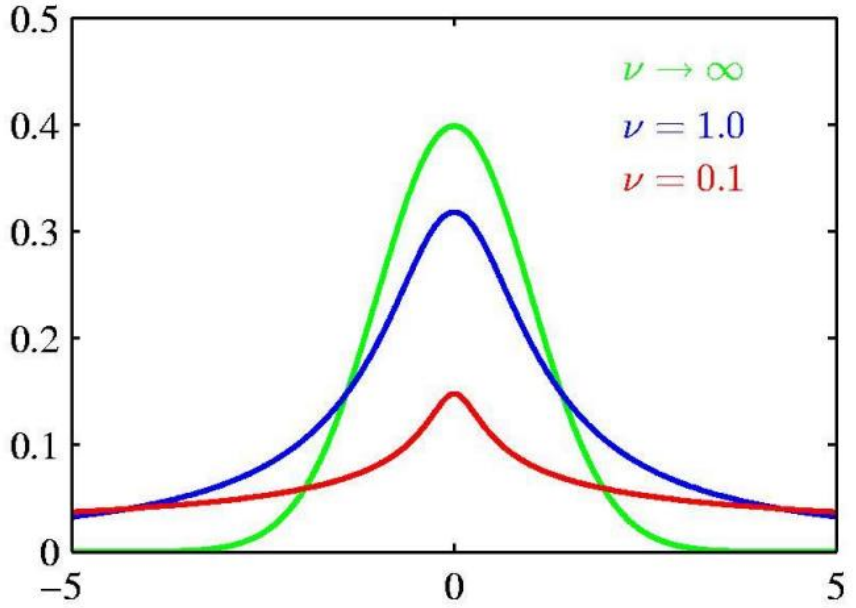
$$\beta_h \sim \mathrm{N}(0, \sigma_h^2), \quad \sigma_h^2 \sim G, \qquad G = \text{Inv-gamma}(\tfrac{\nu}{2}, \tfrac{\nu}{2}).$$

- G表示方差所服从的混合分布。设置$G = Inv - gamma\left(\frac{v}{2}, \frac{v}{2}\right)$,基于高斯混合，能够获得自由度为$v$的$t$分布。在机器学习非参数回归的相关文献中，基系数的常见收缩先验设置为t分布，其自由度接近于0，在极限的情况下，我们能够获得正太Jeffreys先验。但通过这种方式，所获得的后验不太恰当，并且不能很好的适应不确定性。

变量选择和系数收缩

- 先验收缩
- 为了获得恰当的先验和调解不确定性，一种常用的方法是使自由度v取一个很小的非零值，如$v = 10^{-6}$。对于$v > 0$,后验的众数不会恰好等于0，但是对于不必要的基系数，β_h的后验仍然会集中在0附近。在默认情况下，我们可以设置$v = 1$,等价于柯西先验，这种方式能够很好的估计函数μ。
- 基于正太分布的混合，能够产生Laplace(双峰)先验，这种先验可以用于Lasso方法中。若假设回归系数服从Laplace(双峰)先验，则Lasso方法得到的系数估计与其极大后验相一致。 Lasso方法一个很吸引人的特性是能同时实现变量选择和系数收缩。

$$p(\beta) = \prod_{i=1}^{n} \frac{\lambda}{2} e^{-\lambda|\beta_i|}$$



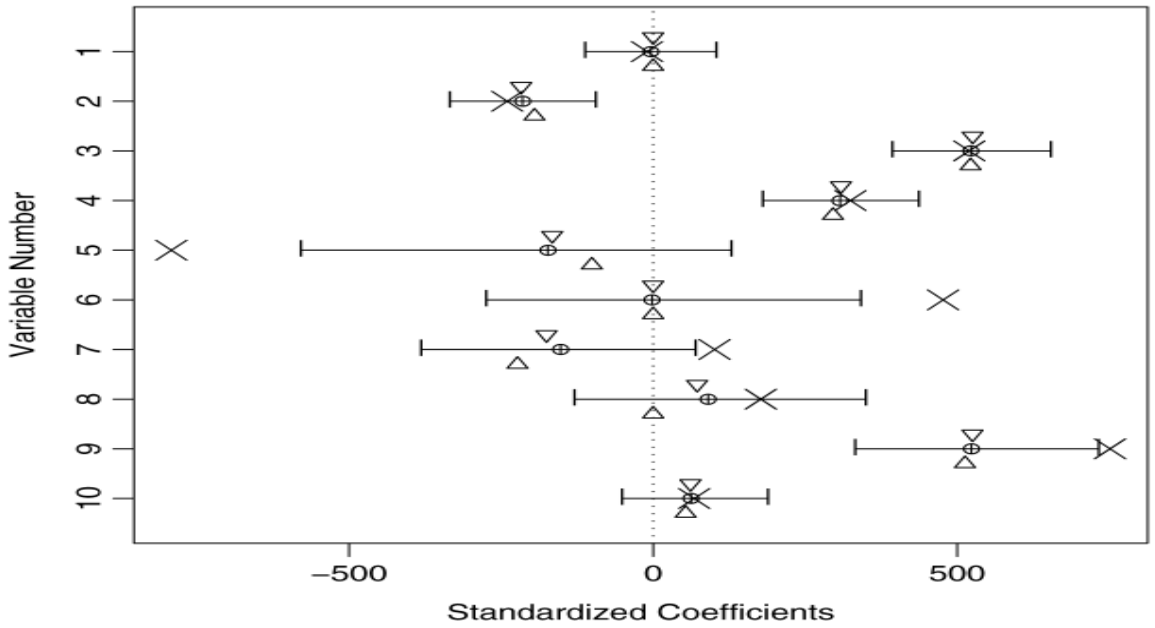
	$\nu = 1$	$\nu \to \infty$
$\mathrm{St}(x\|\mu,\lambda,\nu)$	Cauchy	$\mathcal{N}(x\|\mu,\lambda^{-1})$

变量选择和系数收缩

- 先验收缩

$$\hat{\beta}_{Lasso} = \arg\min_{\beta}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\Sigma_{i=1}^{p}|\beta_i|,$$

- 其中$\lambda \geq 0$为调和参数，随着λ的增大将变量回归系数向0的方向连续压缩且使得回归系数严格等于0，进而达到变量选择的目的，当$\lambda=0$时，Lasso估计量，即为最小二乘估计。通过调节λ的大小，控制选入模型变量的个数。



变量选择和系数收缩

- 先验收缩

$$
\begin{aligned}
p(x|\mu,a,b) &= \int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)\,\mathrm{d}\tau \\
&= \int_0^\infty \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\
&= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2} \\
&= \mathrm{St}(x|\mu,\lambda,\nu)
\end{aligned}
$$

where

$$\lambda = a/b \qquad \eta = \tau b/a \qquad \nu = 2a.$$

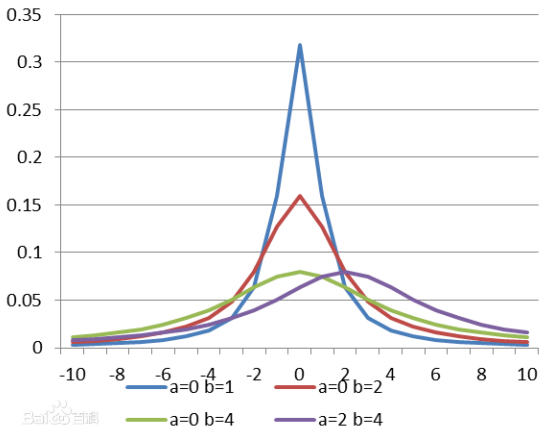
Infinite mixture of Gaussians.

变量选择和系数收缩

- 先验收缩（参考文献:**GENERALIZED DOUBLE PARETO SHRINKAGE**）
- 回归系数服从广义双Pareto先验分布。

$$\mathrm{gdP}(\beta|\xi,\alpha)=\frac{1}{2\xi}\left(1+\frac{|\beta|}{\alpha\xi}\right)^{-(\alpha+1)}$$

- 尺度参数$\xi>0$,形状参数$\alpha>0$。假设$\beta\sim N(0,\ \sigma^2)$,$\sigma^2\sim expon\left(\lambda^2/2\right)$,$\lambda\sim Gamma(\alpha,\eta)$，则$\beta$的边缘密度函数就是$gdP(\beta|\xi,\alpha)$,其中，$\xi=\eta/\alpha$。默认情景下，设置$\alpha=\eta=1$，这就会产生类似柯西分布一样的尾部。在非参回归中，使用双Pareto先验分布作为基系数的收缩先验，可以表示为：



$$p(\beta|\sigma)=\prod_{h=1}^{H}\frac{\alpha}{2\sigma\eta}\left(1+\frac{|\beta_h|}{\sigma\eta}\right)^{-(\alpha+1)}$$

变量选择和系数收缩

- 先验收缩
- 基系数的收缩先验为双Pareto先验分布。

$$p(\beta|\sigma) = \prod_{h=1}^{H} \frac{\alpha}{2\sigma\eta}\left(1+\frac{|\beta_h|}{\sigma\eta}\right)^{-(\alpha+1)}$$

- 这就等价于$\beta_h \sim N(0,\ \sigma^2\tau_h)$,$\tau_h \sim expon\left(\lambda_h^2/2\right)$,$\lambda_h \sim Gamma(\alpha,\eta)$，设置$p(\sigma) \propto 1/\sigma$。我们可以从条件后验概率中进行抽样获取各参数：

$$\begin{aligned}
\beta|- &\sim \mathrm{N}((W^TW+T^{-1})^{-1}W^Ty,\ \sigma^2(W^TW+T^{-1})^{-1}) \\
\sigma^2|- &\sim \text{Inv-gamma}((n+k)/2,\ (y-W\beta)^T(y-X\beta)/2+\beta^TT^{-1}\beta/2) \\
\lambda_h|- &\sim \text{Gamma}(\alpha+1,\ |\beta_h|/\sigma+\eta) \\
\tau_h^{-1}|- &\sim \text{Inv-Gaussian}(\mu=(\lambda_h\sigma/\beta_h,\ \rho=\lambda_h^2)
\end{aligned}$$

$$W=(w_1,\ldots,w_n) \text{ and } T=\text{Diag}(\tau_1,\ldots,\tau_H).$$

变量选择和系数收缩

- 先验收缩
- Gibbs抽样迭代很多次后，能过获得很好的收敛和混合特性。收敛之后，我们从后验分布中抽取非参回归曲线$\mu(x)$,$\mu(x)$是一系列基函数的线性组合，通过广义双Pareto先验分布使得基函数的系数收缩于0。在高维环境下，涉及到大量的潜在的基函数，倾向于设置许多基函数的系数接近于0，然而一点也不收缩重要基函数的系数。