第 10 讲：贝叶斯线性与广义线性模型

张伟平

目录

1.1 贝叶斯线性模型

- 线性模型是至今最流行的统计模型
- 它包括 t test 和 ANOVA 这两个特例
- 多元线性回归模型

$$Y_i \sim N\left(\beta_0 + X_{i1}\beta_1 + \ldots + X_{ip}\beta_p, \sigma^2\right), i = 1, \ldots, n$$

  假设相互独立.

- 在 $n \gg p$ 情形下, 经典回归和贝叶斯回归 (先验取为无信息先验) 非常相似
- 但是对一些问题, 结果可以是不同的, 解释也是不同的

最小二乘回顾

- $\beta = (\beta_0, \ldots, \beta_{p-1})'$ 的最小二乘估计为

$$\hat{\beta}_{ols} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \mu_i)^2$$

  其中 $\mu_i = \beta_0 + X_{i,1}\beta_1 + \ldots + X_{i,p-1}\beta_{p-1}$

- $\hat{\beta}_{ols}$ 是无偏估计, 即使误差不服从正态分布

- 如果误差服从正态分布, 则似然正比于

$$\prod_{i=1}^{n} \exp\left[-\frac{(Y_i - \mu_i)^2}{2\sigma^2}\right] = \exp\left[-\frac{\sum_{i=1}^{n}(Y_i - \mu_i)^2}{2\sigma^2}\right]$$

- 因此, 如果误差服从正态, 则 $\hat{\beta}_{ols}$ 也是 MLE

- 记 $Y = (Y_1, \ldots, Y_n)^T$, X 为 $n \times p$ 设计阵
- 则 $EY = X\beta$, 最小二乘解为

$$\hat{\beta}_{ols} = \underset{\beta}{\operatorname{argmin}}(Y - \mathbf{X}\beta)^T(Y - \mathbf{X}\beta) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y$$

- 如果误差服从正态 $N(0, \sigma^2)$, 则

$$\hat{\beta}_{\text{ols}} \sim N\left[\beta, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]$$

- 方差 σ^2 使用均方残差来估计

$$\hat{\sigma}^2 = \frac{1}{n-p}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

贝叶斯回归

- 假设相互独立服从正态

$$Y_i \sim N\left(X_i'\beta, \sigma^2\right), i = 1, \ldots, n, rank(X) = p$$

似然函数仍为

$$\begin{aligned} l(\beta, \sigma | y, X) &\propto \sigma^{-n} \exp\left\{-(y - X\beta)'(y - X\beta)/2\sigma^2\right\} \\ &\propto \sigma^{-n} \exp\{-[\nu s^2 + (\beta - \hat{\beta}_{ols})' X' X (\beta - \hat{\beta}_{ols})]/2\sigma^2\} \end{aligned}$$

其中 $\nu = n - p$, $\nu s^2 = (y - X\hat{\beta}_{ols})'(y - X\hat{\beta}_{ols})$,

- 贝叶斯分析需要指定 β 和 σ^2 的先验分布
- 对 β 的先验有多种选择: 不正常先验, 高斯先验, 双指数先验, ...
- 对 σ 的先验也有多种选择: 不正常先验, 逆伽马先验, ...

不正常先验

- Jeffreys 先验: $p(\beta) = 1$, 后验分布正常
- 如果 σ 已知, 则

$$\beta|Y \sim N\left[\hat{\beta}_{ols}, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]$$

- 因此结果应该类似于最小二乘
- σ 一般是未知的, 此时多假设共轭先验 $\sigma^2 \sim InvGamma(a, b)$, 其中超参数 a, b 取较小值, 比如 $a = b = 0.01$.

$$p\left(\sigma^2\right) = \frac{b^a}{\Gamma(a)}\left(\frac{1}{\sigma^2}\right)^{a+1}\exp\left(-\frac{b}{\sigma^2}\right), \sigma^2 > 0$$

则后验分布为

$$\beta|\mathbf{Y} \sim t(\hat{\beta}_{OLS}, \frac{b + \|Y - X\hat{\beta}_{OLS}\|^2/2}{a + n/2}(X'X)^{-1}, 2a + n)$$

- 如果 $p(\beta,\sigma)\propto 1/\sigma$, 则联合后验分布为

  $$p(\beta,\sigma|D)\propto\sigma^{-(n+1)}\exp\left\{-\left[\nu s^2+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]/2\sigma^2\right\}$$

  其中 $\nu=n-p-1$, $\nu s^2=(y-X\beta)'(y-X\hat{\beta})$

- 因此记 D 为样本数据, 则

  $$p(\beta|D)\propto\left\{\nu s^2+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right\}^{-(v+p)/2}$$

  以及

  $$p(\sigma|D)\propto\sigma^{-(\nu+1)}\exp\left\{-\nu s^2/2\sigma^2\right\}$$

  即 $\beta|D\sim t_p(\hat{\beta}_{ols},(X'X)^{-1}s^2,\nu)$, 以及 $\sigma|D\sim InvGamma(\nu/2,\nu s^2/2)$

- 因此

  $$t_\nu=\left(\beta_i-\hat{\beta}_i\right)/s_{ii}$$

  服从一元 t 分布, 其中 $s_{ii}=(X'X)^{-1}_{ii}s^2$, $\hat{\beta}_i=\hat{\beta}_{ols,i}$

正态逆伽马 (NIG) 先验

- 正态逆伽马先验 $NIG\left(\mu_\beta, V_\beta, a, b\right)$:

$$p\left(\beta, \sigma^2\right) = p\left(\beta|\sigma^2\right) p\left(\sigma^2\right) = N\left(\mu_\beta, \sigma^2 V_\beta\right) \times IG(a,b) = NIG\left(\mu_\beta, V_\beta, a, b\right)$$

$$= \frac{b^a (\sigma^2)^{-(a+p/2+1)}}{(2\pi)^{p/2} \left|V_\beta\right|^{1/2} \Gamma(a)} \exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}\left(\beta - \mu_\beta\right)^T V_\beta^{-1}\left(\beta - \mu_\beta\right)\right\}\right]$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{a+p/2+1} \times \exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}\left(\beta - \mu_\beta\right)^T V_\beta^{-1}\left(\beta - \mu_\beta\right)\right\}\right]$$

- 积分掉 σ, 则

$$p(\beta) = \frac{\Gamma\left(a + \frac{p}{2}\right)}{\pi^{p/2} \left|(2a)\frac{b}{a} V_\beta\right|^{1/2} \Gamma(a)} \left[1 + \frac{\left(\beta - \mu_\beta\right)^T \left[\frac{b}{a} V_\beta\right]^{-1} \left(\beta - \mu_\beta\right)}{2a}\right]^{-\left(\frac{2a+p}{2}\right)}$$

- 即

$$\beta \sim t(\mu_\beta, \frac{b}{a} V_\beta, 2a)$$

- 因此可以得到后验分布

$$p(\beta, \sigma^2|y) \propto \left(\frac{1}{\sigma^2}\right)^{a+(n+p)/2+1} \times \exp\left\{-\frac{1}{\sigma^2}\left[b^* + (\beta-\mu^*)^T V^{*-1}(\beta-\mu^*)/2\right]\right\}$$

其中

$$\mu^* = \left(V_\beta^{-1} + X^T X\right)^{-1}\left(V_\beta^{-1}\mu_\beta + X^T y\right)$$
$$V^* = \left(V_\beta^{-1} + X^T X\right)^{-1}$$
$$a^* = a + n/2$$
$$b^* = b + \frac{1}{2}(y - X\mu_\beta)^T\left(I + XV_\beta X^T\right)^{-1}(y - X\mu_\beta)$$

- 即

$$\beta|Y \sim t(\mu^*, \frac{b^*}{a^*}V^*, 2a^*)$$

- $\mathbf{Y}$ 的边际分布

$$
\begin{aligned}
p(y) &= \int p\left(y|\beta,\sigma^2\right) p\left(\beta,\sigma^2\right) d\beta d\sigma^2 \\
&= \int N\left(X\beta,\sigma^2 I_n\right) \times NIG\left(\mu_\beta, V_\beta, a, b\right) d\beta d\sigma^2 \\
&= t\left(X\mu, \frac{b}{a}\left(I + XVX^T\right), 2a\right)
\end{aligned}
$$

Zellner's g-prior

- 假设

$$\beta \sim N\left[0, \frac{\sigma^2}{g}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]$$

  其中 X 为列满秩.

- 此时后验均值为

$$\frac{1}{1+g}\hat{\beta}_{OLS}$$

- 这是一种压缩估计, g 控制了压缩程度. 常用 $g = 1/n$, 称为 UIP(unit information prior)

贝叶斯岭估计

- 如果协变量之间存在共线性性, 则常使用独立先验

$$\beta_j \sim N\left(0, \sigma^2/g\right), j = 1, \ldots, p$$

此时后验众数为

$$\arg\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \mu_i\right)^2 + g \sum_{j=1}^{p} \beta_j^2$$

即为岭估计, 其中 $\mu_i = \beta_0 + X_{i1}\beta_1 + \ldots + X_{ip}\beta_p$.

- 此时 β 给定 σ^2 的后验分布由前可知

$$\beta|\mathbf{y}, \sigma^2 \sim N((X'X + gI)^{-1}X'Y, (X'X + gI)^{-1}\sigma^2)$$

BLASSO

- 当回归系数中有很多零时, LASSO 估计为

$$\arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \mu_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- 其可以视为 Bayesian LASSO prior: $\beta_j \sim DE(g)$ 下的后验估计

$$p(\beta_j) = \frac{g}{2} \exp\left(-g|\beta_j|\right)$$

  固定 σ^2 时候, 在 $g = \lambda/(2\sigma^2)$ 时的后验众数即为 LASSO 解.

- 同时进行变量选择和估计, g 控制了压缩量, g 增加偏差减少, 方差减少

- R 包 BLR 可以方便进行贝叶斯岭估计和 LASSO 估计

- 注意到双指数分布密度可以表示为正态-指数混合:

$$\frac{g}{2}\exp(-g|\beta|)=\int_0^\infty \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\beta^2/2\tau^2\right)\frac{g^2}{2}\exp\left(-\frac{g^2\tau^2}{2}\right)\mathrm{d}\tau^2,$$

因此 BLASSO 可以视为是一个层次模型

1. $p\left(\beta_0\right)\propto 1$
2. $p\left(\beta_j\right)\sim N\left(0,\tau_j^2\right)$
3. $p\left(\tau_j^2\right)\sim Exp\left(g^2/2\right)$
4. $Y_i\sim N(\mu_i,\sigma^2)$, i=1,...,n

因此可以方便地在 WINBUGS 中实施.

- Park and Casella (2008) 指出如果 σ^2 指定一个独立先验, 则后验分布是多峰的

在 σ^2 未知时候, 完全的贝叶斯框架:

- 考虑先验: $\beta_j \sim DE(g), j = 1, \ldots, p$ (忽略掉截距项)

$$p(\beta_j|\sigma) = \frac{g}{2\sigma} \exp\left(-g|\beta_j|/\sigma\right)$$

以及 $p(\sigma^2) \propto 1/\sigma^2$. 这样设置很重要, 以保证后验分布 $\beta, \sigma^2|\mathbf{y}$ 是单峰的. 缺乏单峰性会使得 Gibbs 抽样器收敛速度较慢, 以及估计可能没有意义

- 从而对数后验为

$$\ln\left(\pi\left(\sigma^2\right)\right) - \frac{n+p-1}{2} \ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\|\tilde{y} - \mathbf{X}\beta\|_2^2 - \lambda\|\beta\|_1/\sqrt{\sigma^2}$$

- 使用一一变换

$$\phi \leftrightarrow \beta/\sqrt{\sigma^2}, \quad \rho \leftrightarrow 1/\sqrt{\sigma^2}$$

得到

$$\ln\left(\pi\left(1/\rho^2\right)\right)+(n+p-1)\ln(\rho)-\frac{1}{2}\|\rho\tilde{y}-\mathbf{X}\phi\|_2^2-\lambda\|\phi\|_1$$

只要 $\ln\left(\pi\left(1/\rho^2\right)\right)$ 是凹的, 那么上述函数就是一个关于 (ϕ,ρ) 的二次凹函数. 从而后验分布是单峰的. 只要 σ^2 的先验具有刻度不变性, $\ln\left(\pi\left(1/\rho^2\right)\right)$ 就是凹的, 这样的先验包括 $p(\sigma^2)\propto 1/\sigma^2$, $IG(a,b)$ 等.

- 层次模型表示

1. $p\left(\beta_0\right)\propto 1$
2. $p\left(\sigma^2\right)\propto\frac{1}{\sigma^2}$
3. $p\left(\beta_j\right)\sim N\left(0,\tau_j^2\right)$
4. $p\left(\tau_j^2\right)\sim Exp\left(g^2/2\right)$
5. $Y_i\sim N(\mu_i,\sigma^2)$, i=1,...,n

1.1.1 贝叶斯回归诊断

- 视 $u = y - X\beta$ 为一个我们希望推断的未知量. 如果 u 已知, 则可以用来检查模型假设, 独立性等.

- 记 $\hat{u}$ 为 u 的后验均值, 即

$$\hat{u} = E(u|D) = y - XE[\beta|D]$$

其中 $D = (y, X)$. 在无信息先验下, $E(\beta|D) = \hat{\beta} = (X'X)^{-1} X'y$, 以及 $\hat{u} = y - X\hat{\beta}$.

- 对 u 的一个分量, $u_i = y_i - x_i'\beta$, 则在无信息先验 $p(\beta, \sigma) \propto 1/\sigma$ 下

$$t = (u_i - \hat{u}_i) / s_{u_i}$$

服从一元后验自由度为 $\nu = n - p$ 的 t 分布, 其中 $s_{u_i}^2 = x_i' (X'X)^{-1} x_i s^2$

- 从而 u_i 的 $1-\alpha$ 贝叶斯可信区间为

$$\hat{u}_i \pm t_{\alpha/2} s_{u_i}$$

  通过观察各个点是否在此可信区间里来判断模型是否合适

- 影响点分析 /异常点分析: 可以通过对每个异常点引入或剔除时模型系数估计的变化来判断

- 将数据分成不同的子集, 研究每个子集下参数估计的变化情况

1.2 贝叶斯广义线性回归

- 广义线性模型 (GLM) 是线性模型的一种推广. GLM 由三个部分组成
  - Systematic Component: 线性预测量 $\theta = X\beta$
  - Link Function: 联系函数 g 将响应变量期望与线性预测量联系起来

$$g(E[Y|X]) = \theta$$

  - Stochastic Component: 指定响应变量 Y 的条件分布服从指数族分布 (自然形式)

$$f(y|\theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

包括 Normal, Binomial, Poisson, Categorical, Multinomial, Poisson, Beta.

- Residuals: 尽管可以表示成类似线性模型中观测结果减去预测结果值, 但是更有用的是使用 Deviance residuals.

正态分布的指数族形式.

↑Example
↓Example

$$
\begin{aligned}
f\left(y|\mu,\sigma^2\right) &= \left(2\pi\sigma^2\right)^{-1/2}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \\
&= \exp\left[-\frac{1}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}(y-\mu)^2\right] \\
&= \exp[(\underbrace{y\mu}_{y\theta}-\underbrace{\frac{\mu^2}{2}}_{b(\theta)})/\underbrace{\sigma^2}_{a(\phi)}+\underbrace{\frac{-1}{2}\left(\frac{y^2}{\sigma^2}+\log\left(2\pi\sigma^2\right)\right)}_{c(y,\phi)}]
\end{aligned}
$$

Table 1: Natural Link Function Summary for Example Distributions

Distribution		Canonical Link: $\theta = g(\mu)$	Inverse Link: $\mu = g^{-1}(\theta)$
Poisson		$log(\mu)$	$\exp(\theta)$
Binomial	*logit link:*	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{\exp(\theta)}{1+\exp(\theta)}$
	probit link:	$\Phi^{-1}(\mu)$	$\Phi(\theta)$
	cloglog link:	$\log(-\log(1-\mu))$	$1-\exp(-\exp(\theta))$
Normal		μ	θ
Gamma		$-\frac{1}{\mu}$	$-\frac{1}{\theta}$
Negative Binomial		$\log(1-\mu)$	$1-\exp(\theta)$

残差与模型拟合

- 偏差 $D = \sum_{i=1}^{n} d(\theta, y_i)$
- 线性模型残差: $\mathbf{R}_{\text{standard}} = \mathbf{y} - X\beta$
- 响应变量残差: $\mathbf{R}_{\text{Response}} = \mathbf{y} - g^{-1}(X\beta)$
- Pearson 残差: $\mathbf{R}_{Pearson} = \frac{\mathbf{y} - \hat{E}[Y|X]}{\sqrt{\widehat{Var}[Y|X]}}$
- $\mathbf{R}_{\text{Working}} = (\mathbf{y} - \hat{\mu}) \frac{\partial \mu}{\partial \eta}|_{\beta=\hat{\beta}} = (\mathbf{y} - \hat{\mu})/g'(\hat{\mu})$

个体偏差函数

$$R_{\text{Deviance}} = \frac{(y_i - \mu_i)}{|y_i - \mu_i|} \sqrt{|d(\theta, y_i)|}$$

其中 $d(\theta, y_i) = -2\left[\ell\left(\hat{\theta}, \psi|y_i\right) - \ell\left(\tilde{\theta}, \psi|y_i\right)\right]$

不同分布的偏差函数

Distribution	Canonical Parameter	Deviance Function
Poisson(μ)	$\theta = log(\mu)$	$2\sum\left[y_i \log\left(\frac{y_i}{\mu_i}\right) - y_i + \mu_i\right]$
Binomial(m,p)	$\theta = log\left(\frac{\mu}{1-\mu}\right)$	$2\sum\left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \mu_i}\right)\right]$
Normal(μ, σ)	$\theta = \mu$	$\sum\left[y_i - \mu_i\right]^2$
Gamma(μ, δ)	$\theta = -\frac{1}{\mu}$	$2\sum\left[-\log\left(\frac{y_i}{\mu_i}\right)\frac{y_i - \mu_i}{\mu_i}\right]$
Negative Binom(μ, p)	$\theta = \log(1-\mu)$	$2\sum\left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (1+y_i)\log\left(\frac{1+\mu_i}{1+y_i}\right)\right]$

GLMs 的贝叶斯推断

后验分布为

$$\begin{aligned}\pi(\beta,\phi|y) &= \frac{L(y;\mathbf{X},\beta,\phi)\pi(\beta,\phi)}{\int\int L(y;\mathbf{X},\beta,\phi)\pi(\beta,\phi)d\beta d\phi} = \frac{L(y;\mathbf{X},\beta,\phi)\pi(\beta,\phi}{\pi(y;\mathbf{X})} \\ &\propto f(y;\mathbf{X},\beta,\phi)\pi(\beta,\phi) \\ &= \exp\left[\sum_{i=1}^{n}\left\{y_i\theta_i - b\left(\theta_i\right)\right\}/a(\phi) + c\left(y_i,\phi\right)\right]\pi(\beta,\phi)\end{aligned}$$

- 一般来说, 后验分布没有简单形式
- 先验分布 $\pi(\beta,\phi)$ 可以选择有信息或无信息先验

Binomial Logistic Regression

- $Y \sim B(n, \pi)$, π 为成功的概率
- $\theta = g(\pi)$, $\theta = \alpha + \mathbf{x}'\beta$. 常用的联系函数包

$$\begin{aligned} g(\pi) &= logit(\pi) = log(\pi/(1-\pi)) \\ &= \alpha + \mathbf{x}'\beta \end{aligned}$$

  或者

$$g(\pi) = \Phi^{-1}(\pi) = \alpha + \mathbf{x}'\beta$$

- 在一组样本量 N 下, 似然函数分别为

$$\begin{aligned} &\prod_{i=1}^{N} \binom{n}{y_i} \left(\text{logit}^{-1}(\theta)\right)^{y_i} \left(1 - \text{logit}^{-1}(\theta)\right)^{n-y_i} \\ &= \prod_{i=1}^{N} \binom{n}{y_i} \left(\frac{e^{\theta}}{1+e^{\theta}}\right)^{y_i} \left(\frac{1}{1+e^{\theta}}\right)^{n-y_i} \end{aligned}$$

或者

$$\prod_{i=1}^{N} \binom{n}{y_i} \left(\Phi(\theta)\right)^{y_i} \left(1-\Phi(\theta)\right)^{n-y_i}$$

- 我们需要指定 α, β 的先验分布, 根据问题背景, 可以使用各种有信息或无信息先验分布. 在 logit 联系函数下, 研究表明一个厚尾分布比较合适, 包括 t 分布, Cauchy 分布等.

- 在独立先验下, 后验分布为

$$\pi\left(\alpha, \beta | y, \mathbf{X}\right) \propto \pi\left(\alpha\right) \times \prod_{k=1}^{K} \pi\left(\beta_k\right) \times \prod_{i=1}^{N} \left[g^{-1}\left(\theta_i\right)\right]^{y_i} \left(1-g^{-1}\left(\theta_i\right)\right)^{n_i-y_i}.$$

Bayesian Logistic random effects regression models

- 设 Y_{ij} 表示第 i 个个体在第 j 个中心是否具有某个特征, x_{ij} 表示协变量, u_j 表示第 j 个中心的随机效应, 则

$$\ln\left(\frac{P\left(Y_{ij}=1|x_{ij},u_j\right)}{P\left(Y_{ij}=0|x_{ij},u_j\right)}\right)=\alpha_1+\sum_{k=1}^{K}\beta_k x_{kjj}+u_j$$
$$u_j\sim N\left(0,\sigma^2\right)\quad j=1,2,\ldots,J\quad i=1,2,\ldots,n_j$$

- 如果 Y_{ij} 的值是有序的, 则

$$\ln\left(\frac{P\left(Y_{1j}\leqslant m|x_{1j},u_j\right)}{P\left(Y_{ij}>m|x_{ij},u_j\right)}\right)=\alpha_m+\sum_{k=1}^{K}\beta_k x_{klj}+u_j\quad(m=1,2,\ldots,M)$$
$$u_1\sim N\left(0,\sigma^2\right),\quad j=1,2,\ldots,J;\quad i=1,2,\ldots,n_j$$

其他广义线性模型

- Poisson 回归

$$y_i \sim Pois(\lambda)$$
$$log(\lambda) = x_i'\beta$$

- Probit 回归

$$y_i \sim B(1, p_i)$$
$$\Phi^{-1}(p_i) = x_i'\beta$$

- beta 回归

$$y_i \sim Beta(\alpha_i, \beta_i)$$
$$\alpha_i = \mu_i\phi, \beta_i = (1-\mu_i)\phi$$
$$g(\mu_i) = x_i'\beta$$

Poisson 对数回归例子

↑Example

飞机受损数据 (Montgomery et al., 2006). 数据包含了越南战争中 30 次空袭任务中飞机受损的数据, 共 4 个变量:

- damage: 飞机上损坏位置的个数
- type: 0-1 变量表示飞机的类型 (0 for A4; 1 for A6)
- bombload: 载弹情况 (吨)
- airexp: 飞行员飞行经验总月数

↓Example

使用模型

$$\text{damage}_i \sim \text{Poisson}\ (\lambda_i)$$

$$\log \lambda_i = \beta_1 + \beta_2\ \text{type}_i + \beta_3\ \text{bombload}_i + \beta_4\ \text{airexp}_i$$

$$\text{for } i = 1, 2, \ldots, 30$$

估计的模型为

$$\log \lambda_i = -0.77 + 0.58\ \text{type}_i - 0.18\ \text{bombload}_i - 0.011\ \text{airexp}_i$$

后验 95% 区间表明只有 *bombload* 的系数是远离 0 的. 所有参数估计汇总如下

node	mean	sd	MC error	2.5%	median	97.5%	start	sample	harmonic
beta[1]	-0.766	1.089	0.1762	-3.168	-0.835	1.619	1001	1000	
beta[2]	0.580	0.466	0.0513	-0.302	0.584	1.537	1001	1000	
beta[3]	0.177	0.068	0.0099	0.040	0.177	0.308	1001	1000	
beta[4]	-0.011	0.010	0.0015	-0.033	-0.010	0.007	1001	1000	
B[1]	0.862	1.221	0.1829	0.042	0.434	5.050	1001	1000	0.465
B[2]	1.993	0.996	0.1050	0.739	1.793	4.652	1001	1000	1.786
B[3]	1.197	0.081	0.0118	1.041	1.193	1.360	1001	1000	1.194
B[4]	0.989	0.010	0.0015	0.968	0.990	1.007	1001	1000	0.989

B_j 的调和均值使用 β 的后验均值在 WinBUGS 外计算

模型参数解释 可以直接基于 $B_j = exp(\beta_j)$ 的值解释参数

- A6 的期望受损数是 A4 的两倍 (当飞行员飞行经验时长和飞机载弹量相同), 相同飞行员经验时长和载弹量情况下, A6 比 A4 的受损程度从后验上预期高 79%

- 每吨载弹增加 20% 飞机的期望受损位置数
- 飞行经验多一年减少%1 的飞机期望受损位置数

估计特定轮廓 两种飞机受损位置数的最大, 最小, 平均和中位数轮廓容易计算. WINBUGS 程序如下

```
# profiles
# values for bombload
profiles[1,1] <- ranked( bombload[], 1 ) # minimum of bombload
profiles[2,1] <- mean(bombload[])         # mean of bombload
profiles[3,1] <- 0.5*( ranked( bombload[], 15 )+ranked( bombload[], 16 )) #median
profiles[4,1] <- ranked( bombload[], 30 )  #max
# values for airexp
profiles[1,2] <- ranked( airexp[], 30 ) #max experience
profiles[2,2] <- mean(airexp[])          #mean
profiles[3,2] <- 0.5*( ranked( airexp[], 15 )+ranked( airexp[], 16 )) #median
profiles[4,2] <- ranked( airexp[], 1 )  #min experience

for (k in 1:4){
    a4.profile[k] <- exp( beta[1] + beta[3]*profiles[k,1] + beta[4]*profiles[k,2] )
    a6.profile[k] <- a4.profile[k]*B[2]
    # this is equivalent to setting exp( beta[1] + beta[3]*profile[k,1] + beta[4]*profile[
}
```

使用 DIC 选择变量 这里共有 3 个自变量, 因此有 8 种模型. 所有模型可以在 WINBUGS 中方便同时估计. 最后选择 DIC 最小的模

型.

	Dbar	Dhat	pD	DIC	Model
y1	108.6	107.6	1.01	109.6	Constant
y2	94.0	92.0	1.99	96.0	Type
$y3$	84.8	82.9	1.91	86.7	Bombload
$y4$	85.3	82.4	2.95	88.3	Type + Bombload
$y5$	106.2	104.3	1.97	108.2	Airexp
$y6$	88.9	85.9	3.01	92.0	Type + Airexp
$y7$	83.9	81.0	2.92	86.9	Bombload + Airexp
$y8$	83.7	79.7	3.98	87.7	Type + Bombload + Airexp
total	735.6	715.9	19.76	755.4	

Burnin-10,000; iterations kept=10,000

由最小的 DIC(86.7) 可以得出只需要 bombload 一个自变量即可. DIC=86.9 非常靠近 86.7, 表明两个模型预测能力相似, 因此飞行员经验时长可能也是一个重要变量.

考虑英超足球 2006-2007 赛季数据, 使用泊松对数线性模型预测结果. 考虑主 (HT) 客 (AT) 场进球数目.

↑Example

↓Example

$$
\begin{aligned}
Y_{ij} &\sim \text{Poisson}\left(\lambda_{ik}\right) \\
\log\left(\lambda_{i1}\right) &= \mu + \text{ home } + a_{\text{HT}_i} + d_{\text{AT}_i} \\
\log\left(\lambda_{i2}\right) &= \mu \quad\; + a_{\text{AT}_i} + d_{\text{HT}}, \qquad \text{for } i = 1, 2, \ldots, n
\end{aligned}
$$

n 为比赛场数, home 表示主场效应, HT_i 和 AT_i 表示第 i 场比赛的主场和客场对, a_k 和 d_k 为第 $k(k = 1, \ldots, 20)$ 个球队的进攻和防御效应. 满足

$$
\sum_{k=1}^{K} a_k = 0 \text{ 和 } \sum_{k=1}^{K} d_k = 0
$$

(1) 估计参数 (2) 预测将来比赛 (3) 重新生成联赛结果.

1.3 贝叶斯预测

对一组新观测 $\tilde{X}$, 我们希望预测其响应 $\tilde{y}$.

- 如果 β, σ^2 均已知, 则 $\tilde{y} \sim N\left(\tilde{X}\beta, \sigma^2 I_m\right)$
- β, σ^2 服从 NIG 时候

$$
\begin{aligned}
p(\tilde{y}|\mathbf{y}) &= \int p\left(\tilde{\mathbf{y}}|\beta, \sigma^2\right) p\left(\beta, \sigma^2|\mathbf{y}\right) d\beta d\sigma^2 \\
&= \int N\left(\tilde{X}\beta, \sigma^2 I_m\right) \times NIG\left(\mu^*, V^*, a^*, b^*\right) d\beta d\sigma^2 \\
&= t_{2a^*}\left(\tilde{X}\mu^*, \frac{b^*}{a^*}\left(I + \tilde{X}V^*\tilde{X}^T\right)\right)
\end{aligned}
$$

- β, σ^2 服从无信息先验 $p\left(\beta, \sigma^2\right) \propto 1/\sigma^2$ 时候这等价于在上述 NIG 系统中, 令 $V_\beta^{-1} \to 0$, $a \to -p/2$, $b \to 0$, 此时后验分布为

$NIG\left(\mu^*, V^*, a^*, b^*\right)$, 其中

$$\mu^* = \hat{\beta} = \left(X^T X\right)^{-1} X^T \mathbf{y}$$
$$V^* = \left(X^T X\right)^{-1}$$
$$a^* = \frac{n-p}{2}$$
$$b^* = \frac{(n-p)s^2}{2}$$

其中 $s^2 = \frac{1}{n-p}(\mathbf{y} - X\hat{\beta})^T(\mathbf{y} - X\hat{\beta}) = \frac{1}{n-p}\mathbf{y}^T\left(I - P_X\right)\mathbf{y}$, $P_X = X\left(X^T X\right)^{-1} X^T$

- 此时, 边际后验分布

$$\sigma^2 | Y, X \sim IG((n-p)/2, (n-p)s^2/2)$$

- β 的边际后验分布为多元 $t_{n-p}(\hat{\beta}, s^2 X'X)$.

- 后验预测分布为多元 t 分布 $t_{n-p}\left(\tilde{X}\hat{\beta}, s^2\left(I+\tilde{X}\left(X^TX\right)^{-1}\tilde{X}^T\right)\right)$, 注意到

$$\begin{aligned} E\left(\tilde{\mathbf{y}}|\sigma^2, y\right) &= E\left[E\left(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}\right)|\sigma^2, \mathbf{y}\right] \\ &= E\left[\tilde{X}\beta|\sigma^2, \mathbf{y}\right] \\ &= \tilde{X}\hat{\beta} = \tilde{X}\left(X^TX\right)^{-1}X^T\mathbf{y} \end{aligned}$$

以及给定 σ^2 有

$$\begin{aligned} \operatorname{var}\left(\tilde{\mathbf{y}}|\sigma^2, \mathbf{y}\right) &= E\left[\operatorname{var}\left(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}\right)|\sigma^2, \mathbf{y}\right] + \operatorname{var}\left[E\left(\tilde{\mathbf{y}}|\beta, \sigma^2, \mathbf{y}\right)|\sigma^2, \mathbf{y}\right] \\ &= E\left[\sigma^2 I_m\right] + \operatorname{var}\left[\tilde{X}\beta|\sigma^2, \mathbf{y}\right] \\ &= \left(I_m + \tilde{X}\left(X^TX\right)^{-1}\tilde{X}^T\right)\sigma^2 \end{aligned}$$

即给定 σ^2, 后验预测方差由两部分组成: 抽样波动性和关于 β 的不确定性

1.4 贝叶斯非参数回归

- 贝叶斯参数模型: 数据 = 隐含的模式 + 独立噪音
  - 参数个数固定有限个
  - 贝叶斯问题的解 = 关于参数的后验分布
- 非参数贝叶斯模型是指定义在无穷维空间上的贝叶斯模型
- 其参数空间 $\mathcal{T}$ 是**所有可能的模式组成的集合**, 例如

问题	$\mathcal{T}$
密度估计	概率分布
回归	光滑函数
聚类	划分

贝叶斯问题的解 =**关于模式的后验分布**

考虑非参数回归模型

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n$$

其中 $E\epsilon_i = 0$, m 为未知的光滑函数

- 频率方法核估计 m 为

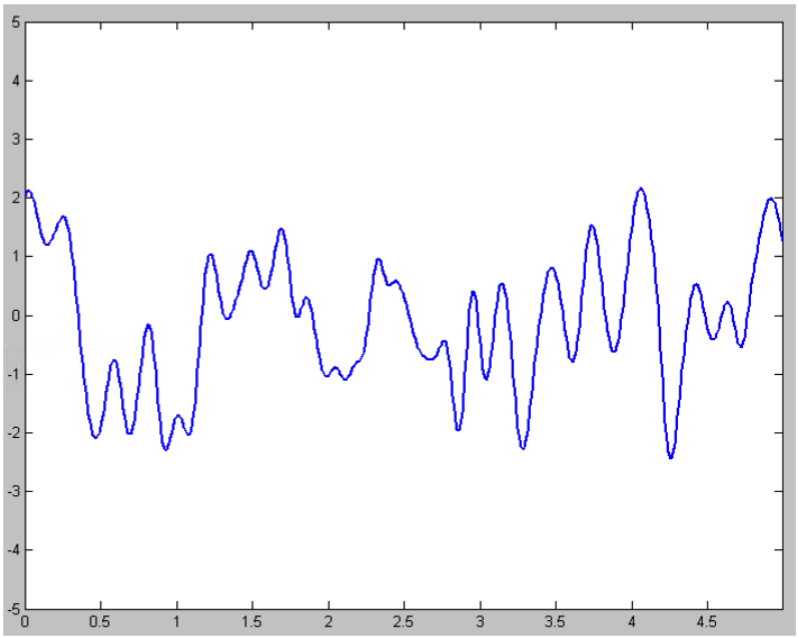
$$\widehat{m}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\|x - X_i\|}{h}\right)}{\sum_{i=1}^n K\left(\frac{\|x - X_i\|}{h}\right)}$$

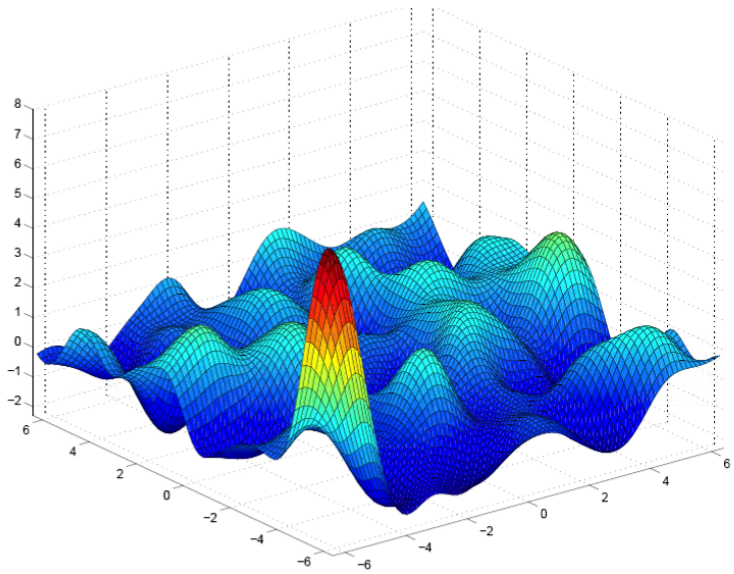
  其中 K 为核函数, h 为带宽.

- 贝叶斯方法则需要在回归函数类 $\mathcal{M}$ 上指定先验分布 π. 常见的选择是高斯过程.

- 在有限维场合, $m(X_i) = \phi(X_i)'w$, ϕ 为基函数. 我们一般假设先验 $w \sim N(0, \Sigma_p)$, 因此 $m|(X_1, \ldots, X_n) \sim N(0, \Phi'\Sigma\Phi)$. 如

果增加 m 向量的长度为无穷, 则协方差就转换为函数 $K(X_i, X_j)$. 无穷维向量 m 变为未知函数 $m(x)$, $m(x)$ 的极限变成高斯过程.





一维和二维高斯过程样本路径 (RBF 核)

一个随机过程 $m(x), x \in \mathcal{X} \subset R^d$ 称为高斯过程，如果对任意 $x_1, \ldots, x_n \in \mathcal{X}$ ，向量 $(m(x_1), m(x_2), \ldots, m(x_n))$ 服从正态分布

$$(m(x_1), m(x_2), \ldots, m(x_n)) \sim N(\mu(x), K(x))$$

Definition

其中 $K_{ij}(x) = K(x_i, x_j)$ 是一个 Mercer 核 (即得到的有限维协方差矩阵是非负定的). 记为

$$m(\cdot) \sim \mathcal{GP}(\mu(\cdot), K(\cdot, \cdot))$$

- 协方差矩阵 $K(x) = (K_{ij}(x))$ 通过核函数 K 来得到, 需要满足非负定性: $v'Kv \geq 0$.
- 常见核函数包括

- RBF 核: $K\left(x, x'\right)=\sigma_0^2 \exp\left[-\frac{1}{2}\left(\frac{x-x'}{\lambda}\right)^2\right]$
- 线性协方差: $K\left(x, x'\right)=\sigma_0^2+xx'$
- 布朗运动 (Wiener 过程): $K\left(x, x'\right)=\min\left(x, x'\right)$
- 周期协方差: $K\left(x, x'\right)=\exp\left(-\frac{2\sin^2\left(\frac{x-x'}{2}\right)}{\lambda^2}\right)$
- 神经网络协方差 $k\left(x, x'\right)=\tanh\left(ax\cdot x'+b\right)$(不是一个有效的核)

$$k_{\mathrm{NN}}\left(\mathrm{x}, \mathrm{x}'\right)=\frac{2}{\pi}\sin^{-1}\left(\frac{2\tilde{\mathrm{x}}'\Sigma\tilde{\mathrm{x}}'}{\sqrt{\left(1+2\tilde{\mathrm{x}}'\Sigma\tilde{\mathrm{x}}\right)\left(1+2\tilde{\mathrm{x}}'^{\top}\Sigma\tilde{\mathrm{x}}'\right)}}\right)$$

- 可以通过求和, 乘积和卷积核函数得到新的协方差.

- 对回归模型 $Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n,$ 贝叶斯模型假设未知函数 m 的先验分布为零均值高斯过程

$$m(\cdot) \sim \mathcal{GP}(0, K(\cdot, \cdot))$$

- 则对任意给定的 $\mathbf{x} = (x_1, x_2, \ldots, x_n), \mathbf{m} = (m(x_1), \ldots, m(x_n))$ 的先验分布为 $\mathbf{m} \sim N(0, K),$ 即有密度

$$\pi(\mathbf{m}) = (2\pi)^{-n/2} |K|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{m}^T K^{-1} \mathbf{m}\right)$$

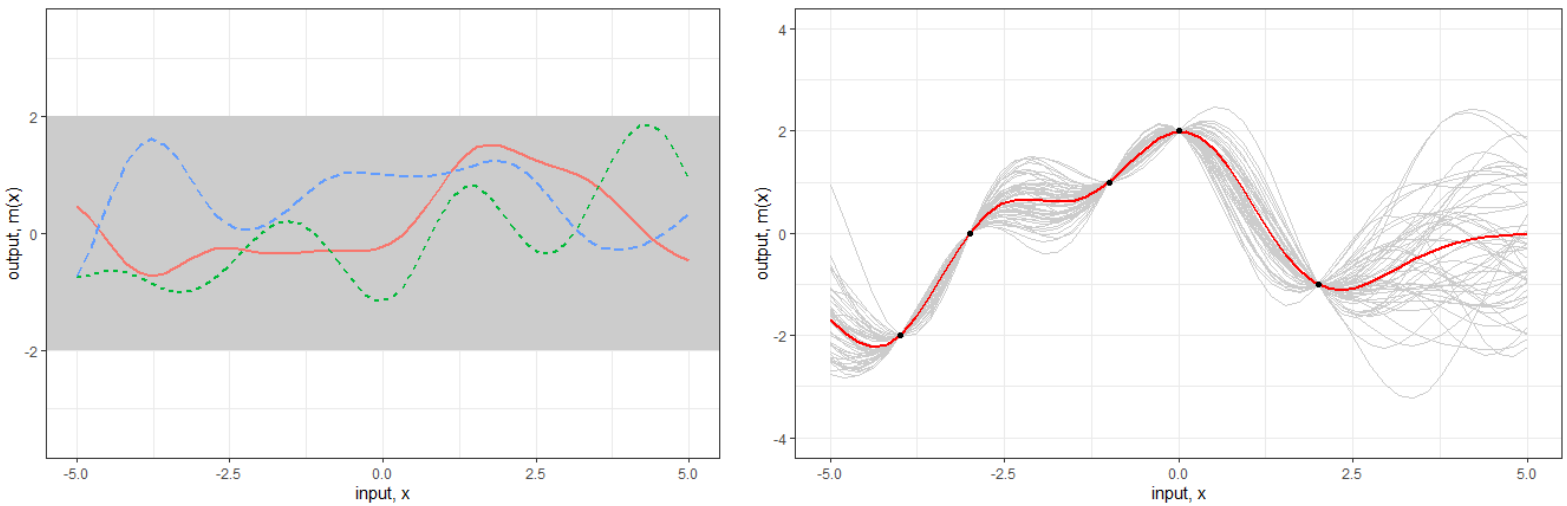
其中 $K = (K_{ij}) = (K(x_i, x_j)).$

- 首先如果没有误差 (噪音), 我们已知 $\{x_i, m(x_i)), i = 1, \ldots, n\}$ 后, 对新的观测点集 $\mathbf{x}_*,$ 及其函数 $m(x_*) = \mathbf{m}_*,$ 因为

$$\begin{bmatrix} \mathbf{m} \\ \mathbf{m}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) & K(\mathbf{x}, \mathbf{x}_*) \\ K(\mathbf{x}_*, \mathbf{x}) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right)$$

所以条件分布

$$\mathbf{m}_*|\mathbf{x}_*,\mathbf{x},\mathbf{m} \sim N\left(K\left(\mathbf{x}_*,\mathbf{x}\right)K(\mathbf{x},\mathbf{x})^{-1}\mathbf{m},\right.$$
$$\left.K\left(\mathbf{x}_*,\mathbf{x}_*\right)-K\left(\mathbf{x}_*,\mathbf{x}\right)K(\mathbf{x},\mathbf{x})^{-1}K\left(\mathbf{x},\mathbf{x}_*\right)\right)$$



左图表示从一个高斯过程先验中随机抽取的三条函数; 右图表示三条从后验分布中抽取的随机函数, 给定五个指定的 (无噪音) 观测点. 两张图中阴影表示逐点的均值加减 2 倍标准差

- 在可加高斯噪音下, 即我们观测到 $y_i = m(x_i) + \epsilon_i$, 其中 $\epsilon_i \sim N(0, \sigma^2)$.
- 注意对数后验分布正比于

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{m}) + \log \pi(\mathbf{m}) &= -\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{m}\|_2^2 - \frac{1}{2}\mathbf{m}^T K^{-1}\mathbf{m} + \text{ const} \\ &= -\frac{1}{2}[(\mathbf{m} - A^{-1}\mathbf{y}/\sigma^2)A(\mathbf{m} - A^{-1}\mathbf{m}/\sigma^2) + \text{ const}\end{aligned}$$

  其中 $A = K^{-1} + 1/\sigma^2 I_n$. 即 $\mathbf{m}|\mathbf{y}, \sigma^2 \sim N(A^{-1}\mathbf{y}/\sigma^2, A^{-1})$.
- 因此估计的回归函数为

$$\begin{aligned}\hat{m}(\mathbf{x}) &= (K^{-1} + 1/\sigma^2 I_n)^{-1}\mathbf{y}/\sigma^2 \\ &= K(K + \sigma^2 I_n)^{-1}\mathbf{y}\end{aligned}$$

- 对一组新观测点集 $\mathbf{x}_*$, 及其函数 $m(x_*) = \mathbf{m}_*$, 则有

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{m}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{x},\mathbf{x}) + \sigma^2 I_n & K(\mathbf{x},\mathbf{x}_*) \\ K(\mathbf{x}_*,\mathbf{x}) & K(\mathbf{x}_*,\mathbf{x}_*) \end{bmatrix}\right)$$

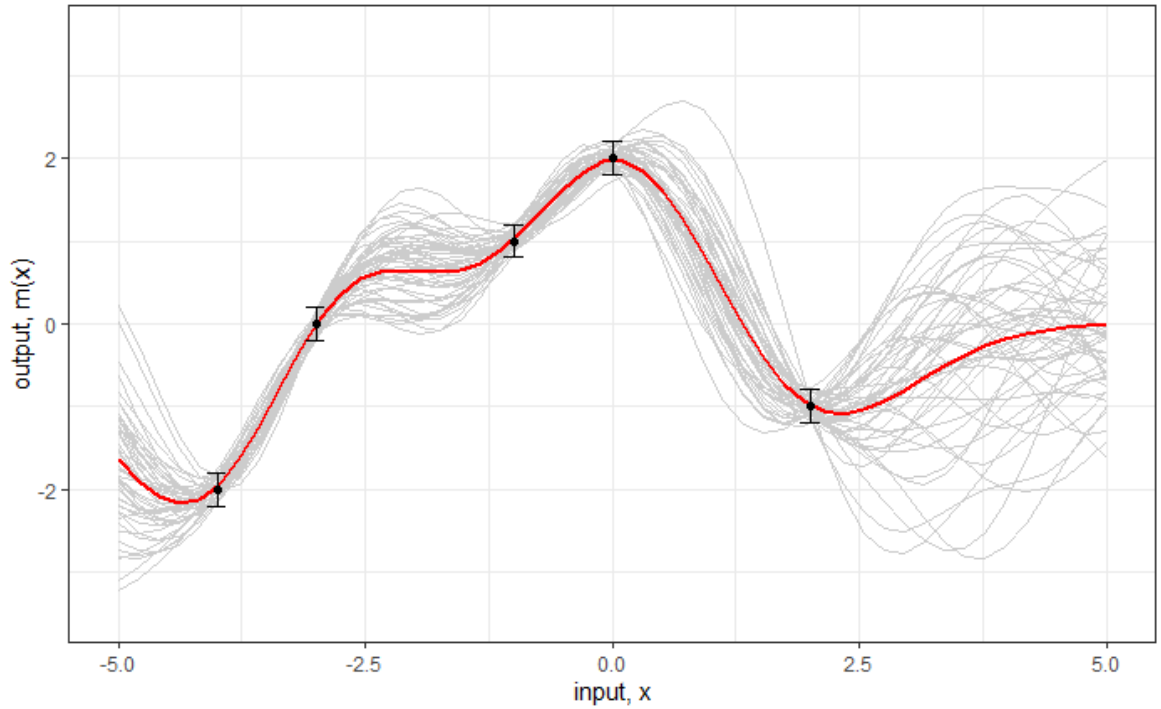
- 因此

$$\mathbf{m}_* | \mathbf{x}, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}\left(\overline{\mathbf{m}}_*, \operatorname{cov}(\mathbf{m}_*)\right)$$

其中

$$\overline{\mathbf{m}}_* = E[\mathbf{m}_* | \mathbf{x}, \mathbf{y}, \mathbf{x}_*] = K(\mathbf{x}_*,\mathbf{x})\left[K(\mathbf{x},\mathbf{x}) + \sigma_n^2 I\right]^{-1}\mathbf{y}$$

$$\operatorname{cov}(\mathbf{m}_*) = K(\mathbf{x}_*,\mathbf{x}_*) - K(\mathbf{x}_*,\mathbf{x})\left[K(\mathbf{x},\mathbf{x}) + \sigma_n^2 I\right]^{-1} K(\mathbf{x},\mathbf{x}_*)$$



观测点带有随机正态误差时候, 给定观测点, 回归函数的随机抽样

- 在变量代换 $m = K\alpha$ 下, 我们有 $\alpha \sim N\left(0, K^{-1}\right)$, 因此

$$\pi(\alpha) = (2\pi)^{-n/2}|K|^{-1/2}\exp\left(-\frac{1}{2}\alpha^T K\alpha\right)$$

- 在可加高斯噪音下, 即我们观测到 $Y_i = m\left(x_i\right) + \epsilon_i$, 其中 $\epsilon_i \sim N\left(0, \sigma^2\right)$. 因此对数后验分布正比于

$$\begin{aligned}\log p(y|m) + \log \pi(m) &= -\frac{1}{2\sigma^2}\|y - K\alpha\|_2^2 - \frac{1}{2}\alpha^T K\alpha + \text{ const} \\ &= -\frac{1}{2}(\alpha - B^{-1}K\mathbf{y}/\sigma^2)'B(\alpha - B^{-1}K\mathbf{y}/\sigma^2) + \text{ const}\end{aligned}$$

其中 $B = K^2/\sigma^2 + K$, 即

$$\alpha|\mathbf{y}, \mathbf{x} \sim N((K + \sigma^2 I_n)^{-1}\mathbf{y}, B^{-1})$$

- 在高斯过程先验下, 什么函数具有比较高的密度? 此先验偏好于 $\alpha' K^{-1}\alpha$ 小的函数. 假设我们考虑 K 的一个特征向量 v, 相

应特征根为 λ, 因此 $Kv = \lambda v$. 则我们有

$$\frac{1}{\lambda} = v^T K^{-1} v$$

因此, 高斯过程先验偏好具有比较大特征根的特性函数, 这对应于光滑函数. 弯曲较多的特征函数对应于较小的特征根.

- 在贝叶斯框架下, MAP 估计对应于 Mercer 核回归, 其对平方误差使用 RKHS 范数 $\|\alpha\|^2$ 进行正则化. 后验均值为

$$\mathbb{E}(\alpha|Y) = \left(K + \sigma^2 I\right)^{-1} \mathbf{y}$$

因此

$$\widehat{m} = \mathbb{E}(m|Y) = K\left(K + \sigma^2 I\right)^{-1} \mathbf{y}$$

即 $\widehat{m}$ 就是一个线性光滑器, 其非常类似于频率下的核回归. 不同的是, 我们需要选择核函数 $K(x, y)$.

- 对一个新观测点 $y_{n+1} = m(x_{n+1}) + \epsilon_{n+1}$, 为了计算其预测分布, 我们注意到 $(y_1, \ldots, y_n) \sim N\left(0, \left(K + \sigma^2 I\right)\right)$. 令 k 为向量

$$k = \left(K(x_1, x_{n+1}), \ldots, K(x_n, x_{n+1})\right)$$

则 $(y_1, \ldots, y_{n+1})$ 服从联合正态分布, 协方差矩阵为

$$\begin{pmatrix} K + \sigma^2 I & k \\ k^T & K(x_{n+1}, x_{n+1}) + \sigma^2 \end{pmatrix}$$

因此, Y_{n+1} 的条件分布为

$$y_{n+1} | \mathbf{y}, \mathbf{x}, \sigma^2 \sim N\left(k^T \left(K + \sigma^2 I\right)^{-1} y, K(x_{n+1}, x_{n+1}) + \sigma^2 - k^T \left(K + \sigma^2 I\right)^{-1} k\right)$$

包含确定的基函数

- 假设高斯过程均值为 0 是常见的, 但不是必须的.
- 使用一个**固定已知**均值函数 $m_0(x)$ 时候, 前面的方法很容易适用于该场合: 只需对 $m(x) - m_0(x)$ 假设零均值高斯过程即可. 此时

$$m(x) \sim \mathcal{GP}\left(m_0(x), K\left(\mathrm{x}, \mathrm{x}'\right)\right)$$

  因此预测均值为

$$\overline{\mathbf{m}}_* = \mathbf{m}\left(\mathbf{x}_*\right) + K\left(\mathbf{x}_*, \mathbf{x}\right) K_y^{-1}(\mathbf{y} - \mathbf{m}_0(\mathbf{x}))$$

  其中 $K_y = K + \sigma^2 I_n$.
- 但是实际中指定一个已知固定的均值函数是非常困难的. 很多场合指定一个基函数比较方便:

$$g(\mathbf{x}) = m(\mathbf{x}) + \mathbf{h}(\mathbf{x})'\beta, \quad \text{其中} \quad m(\mathbf{x}) \sim \mathcal{GP}\left(0, k\left(\mathbf{x}, \mathbf{x}'\right)\right)$$

其中参数 β 需要由数据进行估计, h 为一组固定已知的基函数.

- 如果假设 $\beta \sim N(b, B)$, 则可以得到

$$g(\mathbf{x}) \sim \mathcal{GP}\left(\mathrm{h}(\mathbf{x})'\mathbf{b}, K\left(\mathbf{x}, \mathbf{x}'\right) + \mathrm{h}(\mathbf{x})' B \mathrm{h}\left(\mathbf{x}'\right)\right)$$

- 此时预测均值和协方差函数为

$$\begin{aligned}
\overline{\mathbf{g}}\left(\mathbf{x}_*\right) &= H_*'\overline{\beta} + K_*' K_y^{-1}\left(\mathbf{y} - H'\overline{\beta}\right) = \bar{\mathbf{f}}(X_*) + R'\overline{\beta} \\
cov\left(\mathbf{g}_*\right) &= cov\left(\mathbf{f}_*\right) + R'\left(B^{-1} + H K_y^{-1} H'\right)^{-1} R
\end{aligned}$$

其中 H 由 h 在所有训练集上的值组成. H_* 在测试集上的值组成. $\overline{\beta} = \left(B^{-1} + H K_y^{-1} H'\right)^{-1}\left(H K_y^{-1}\mathbf{y} + B^{-1}\mathbf{b}\right)$, $R = H_* - H K_y^{-1} K_*$