模型和超参的选取

黄政宇
北京大学北京国际数学研究中心
北京大学国际机器学习研究中心

模型和超参的选取

- 高斯过程

$$f(x) \sim GP(h(x)\beta, \kappa(x, x'))$$

- 核函数的选取

  - 平方指数类核函数
  - Matern类核函数
  - 具有紧致支集的分片多项式核函数
  - 混合核函数

  ……

- 超参的选取

  - 核函数大小系数
  - 特征长度尺度

  ……

模型和超参的选取

- 平方指数(squared exponential)类核函数

$$\kappa(x,x') = \sigma_f^2 \exp\left(-\frac{1}{2}(x-x')^T M (x-x')\right) + \sigma_n^2 \delta_{pq}$$

$$f_* \,|X_*, X, y \sim \mathcal{N}(K(X_*,X)(K(X,X)+\sigma_n^2 I)^{-1} y,$$
$$K(X_*,X_*) - K(X_*,X)(K(X,X)+\sigma_n^2 I)^{-1} K(X,X_*)$$

参数包含：

σ_f^2：不确定性

σ_n^2：正则化强求

M是对称正定矩阵，常用的形式包括

$$M_1 = l^{-2} I$$
$$M_2 = \mathrm{diag}(l)^{-2} \qquad (l \in R^N)$$
$$M_3 = \Lambda\Lambda^T + \mathrm{diag}(l)^{-2} \; (l \in R^N, \Lambda \in R^{N\times k})$$

模型和超参的选取

- 平方指数(squared exponential)类核函数

$\kappa(x, x')$定义x和x'之间的“距离”

$M_1 = l^{-2}I$　　各向同性

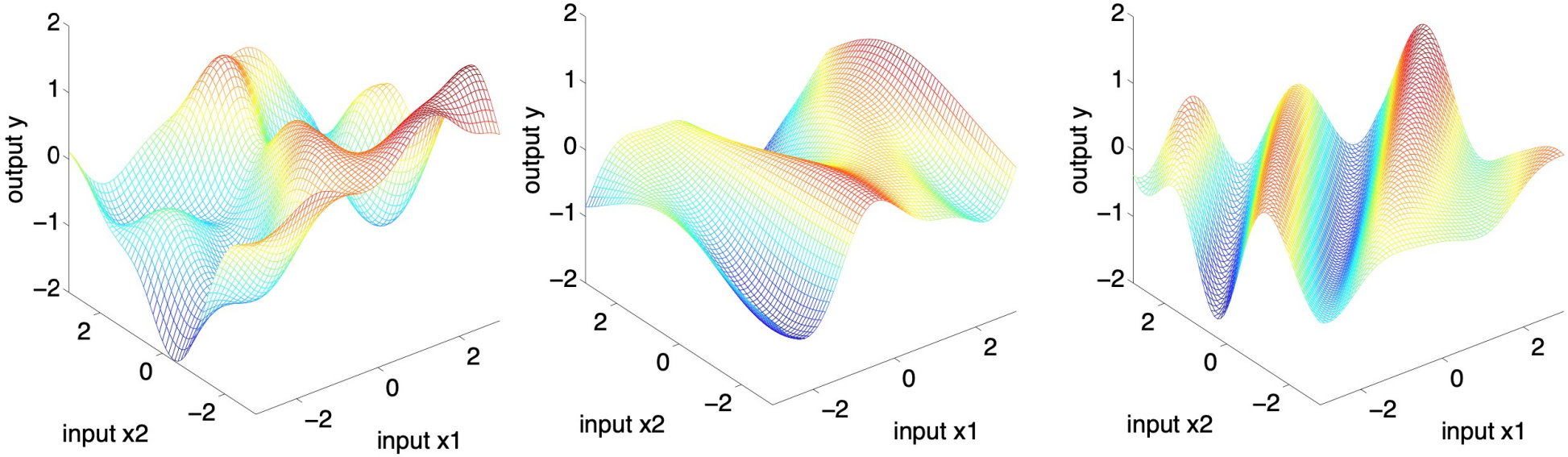
$M_2 = \mathrm{diag}(l)^{-2}$　　各向异性，$l_1 \cdots\cdots l_N$是各个方向的特征长度尺度

$M_3 = \Lambda\Lambda^T + \mathrm{diag}(l)^{-2}$ $(l \in R^N, \Lambda \in R^{N\times k})$
对于高维数据集，Λ矩阵的k列可能在输入空间中标识出几个具有特别高“相关性”的方向，而它们的长度给出了沿这些方向的逆特征长度尺度。

模型和超参的选取

- 平方指数(squared exponential)类核函数

a) $M_1 \quad l = 1$
b) $M_2 \quad l = (1,3)^T$
c) $M_3 \quad \Lambda = (1,-1)^T, l = (6,6)^T$

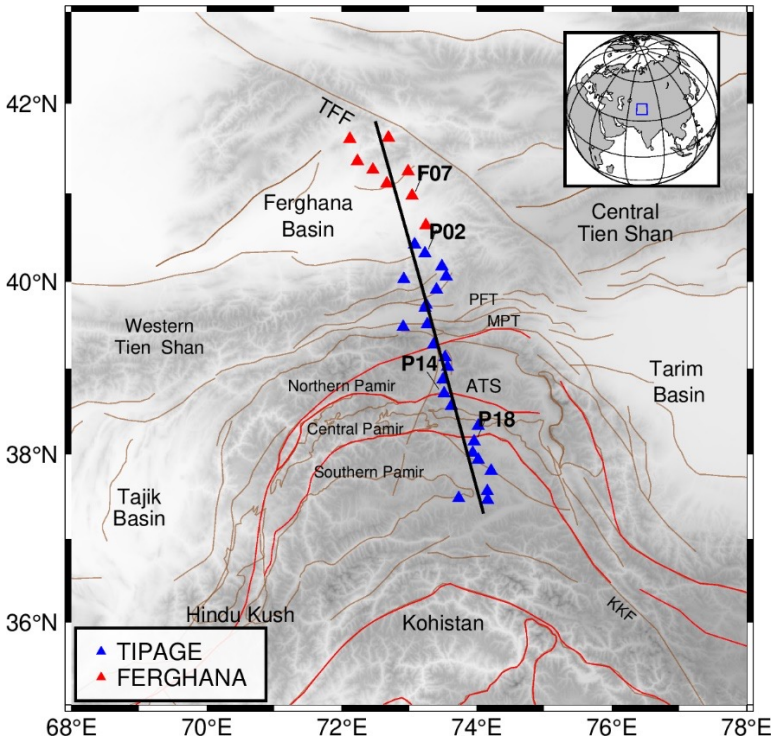


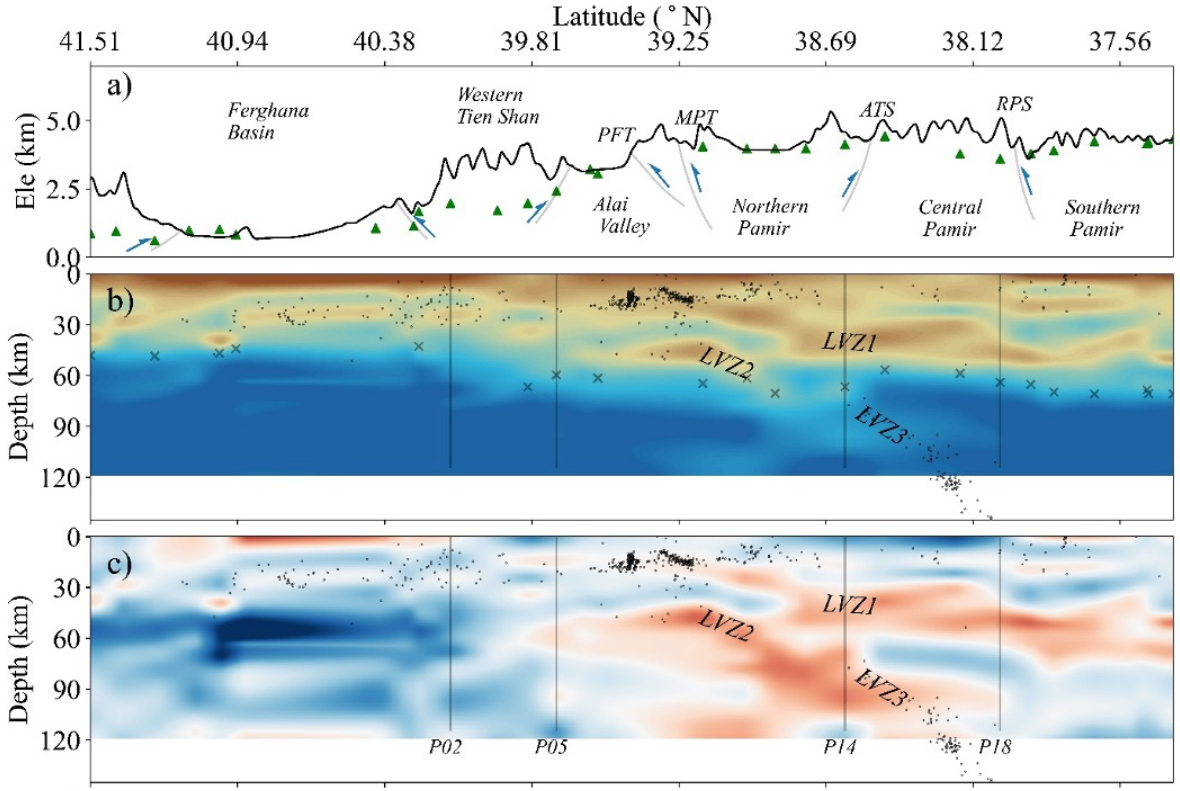
(a) (b) (c)

模型和超参的选取

➢ 底层反演

$$y = \mathcal{G}(\theta) + \eta \qquad \eta \sim \mathcal{N}(0, \Sigma_\eta)$$





模型和超参的选取

- 目标

基于一组训练数据，对协方差函数的形式和参数作出推断，或者说，对数据中的关系进行推断。
发展一种系统且实用的模型选择方法，能够比较两种（或多种）有不同参数值、协方差函数的方法，或将高斯过程模型与任何其他类型的模型进行比较

- 基本准则

- 计算给定数据下模型的概率
- 估计泛化误差
- 估计泛化误差的上界

贝叶斯模型选择

- 模型$\mathcal{H}_i$
  - 不同次数线性回归模型
  - 不同核函数高斯过程回归
  ……
- 超参数θ
  - 核函数参数
  - 线性回归正则项参数
  ……
- 模型参数w
  - 线性回归系数
  - 神经网络参数
  ……

贝叶斯模型选择

- 模型参数估计

$$\rho(w|y,X,\theta,\mathcal{H}_i)=\frac{\rho(y|X,w,\theta,\mathcal{H}_i)\rho_0(w|\theta,\mathcal{H}_i)}{\rho(y|X,\theta,\mathcal{H}_i)}$$

$$\rho(y|X,\theta,\mathcal{H}_i)=\int\rho(y|X,w,\theta,\mathcal{H}_i)\rho_0(w|\theta,\mathcal{H}_i)dw$$

$\rho(y|X,w,\theta,\mathcal{H}_i)$：似然函数

$\rho_0(w|\theta,\mathcal{H}_i)$：参数先验分布

$\rho(y|X,\theta,\mathcal{H}_i)$：边缘似然函数或者证据（evidence）

贝叶斯模型选择

- 超参数估计

$$\rho(\theta|y,X,\mathcal{H}_i)=\frac{\rho(y|X,\theta,\mathcal{H}_i)\rho_0(\theta|\mathcal{H}_i)}{\rho(y|X,\mathcal{H}_i)}$$

$$\rho(y|X,\mathcal{H}_i)=\int\rho(y|X,\theta,\mathcal{H}_i)\rho_0(\theta|\mathcal{H}_i)d\theta$$

$\rho_0(\theta|\mathcal{H}_i)$：超参数的先验分布

- 模型选择

$$\rho(\mathcal{H}_i|y,X)=\frac{\rho(y|X,\mathcal{H}_i)\rho_0(\mathcal{H}_i)}{\rho(y|X)}\propto\rho(y|X,\mathcal{H}_i)$$

$$\rho_0(\mathcal{H}_i)\propto 1$$

贝叶斯模型选择

- 例子

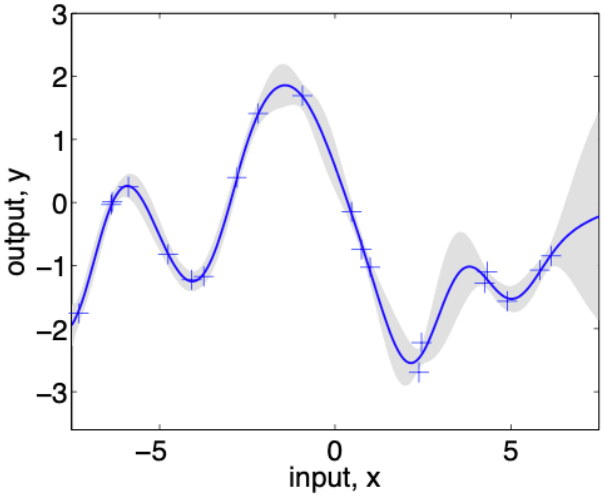
考虑一维高斯过程，期望函数$m(x)=0$，核函数为

$$\kappa_y(x_p,x_q)=\sigma_f^2\exp\left(-\frac{1}{2l^2}\left(x_p-x_q\right)^2\right)+\sigma^2\delta_{pq}$$

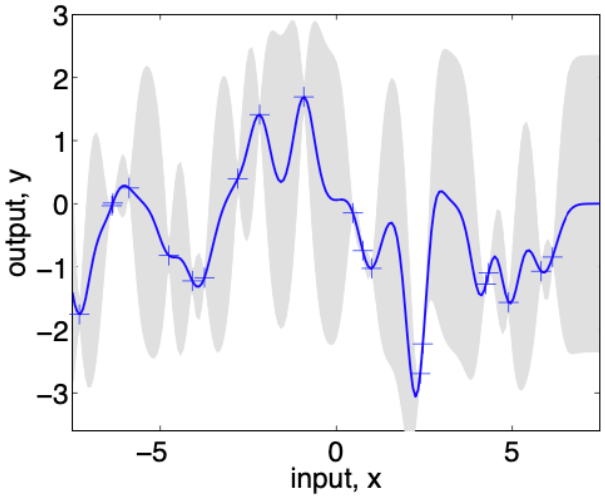
用$(l,\sigma_f,\sigma)=(1.0,1.0,0.1)$生成数据点$x\sim U[-8,8]$，进行超参估计

高斯过程

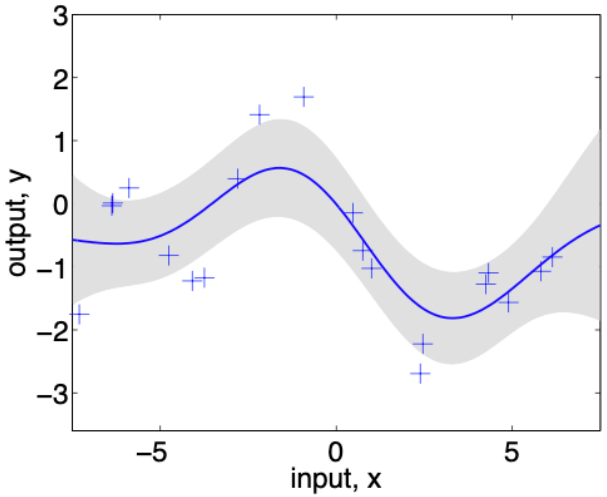
➢ 练习



(a), $\ell = 1$



(b), $\ell = 0.3$

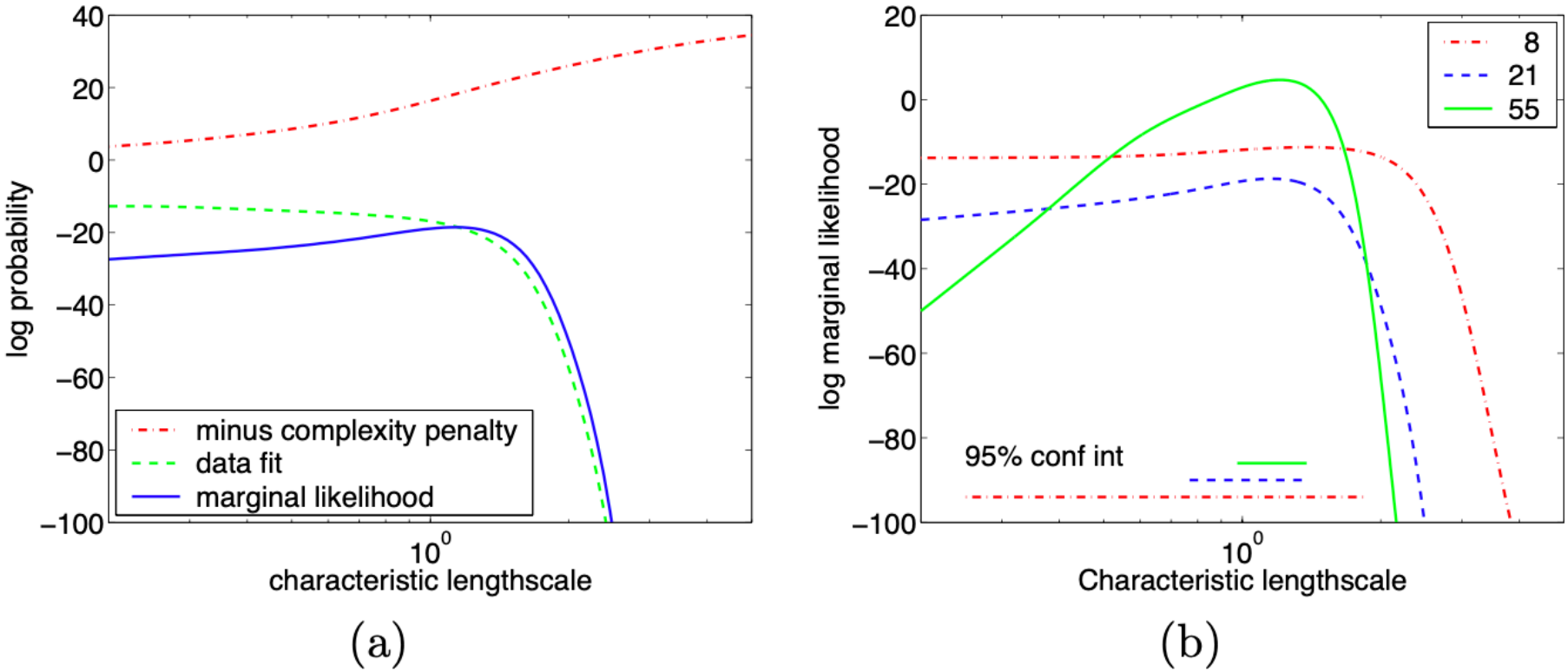


(c), $\ell = 3$

贝叶斯模型选择

➢ 模型参数估计

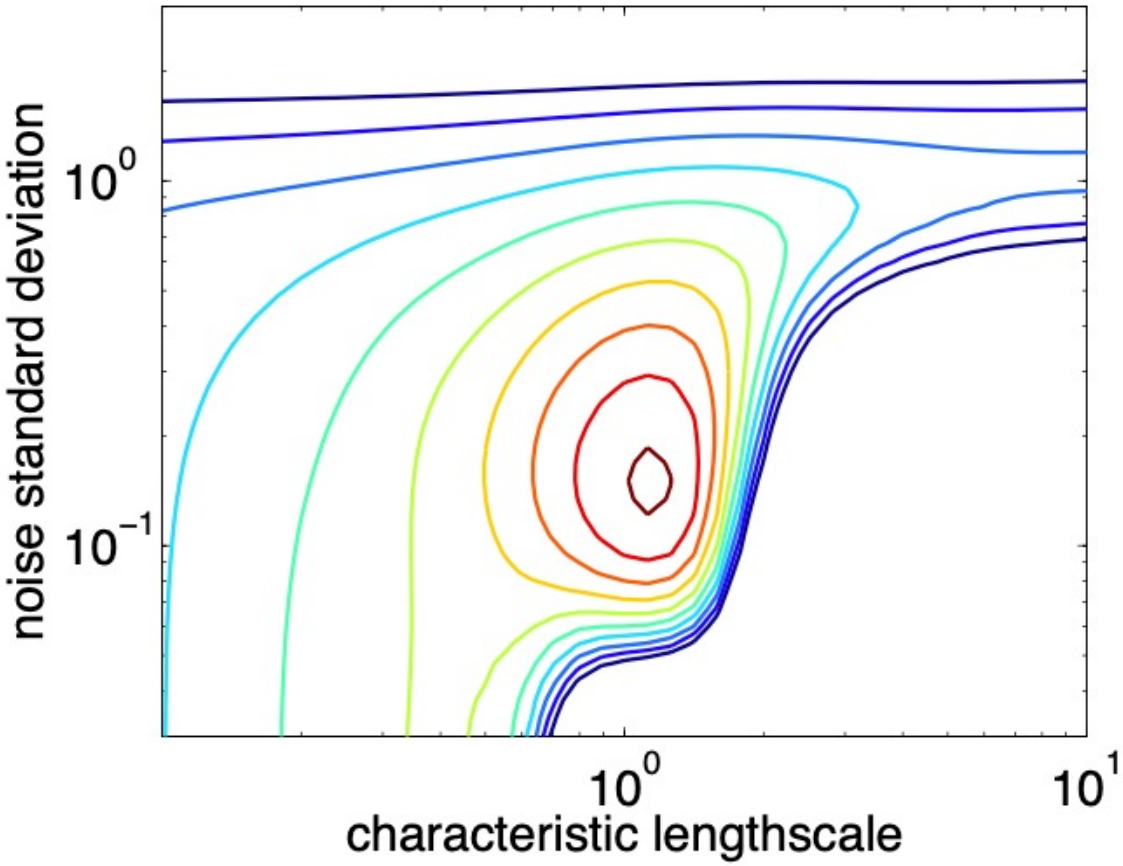
训练集数据量



a)各种对数函数b)对数边际似然函数和特征长度尺度的函数关系。

贝叶斯模型选择

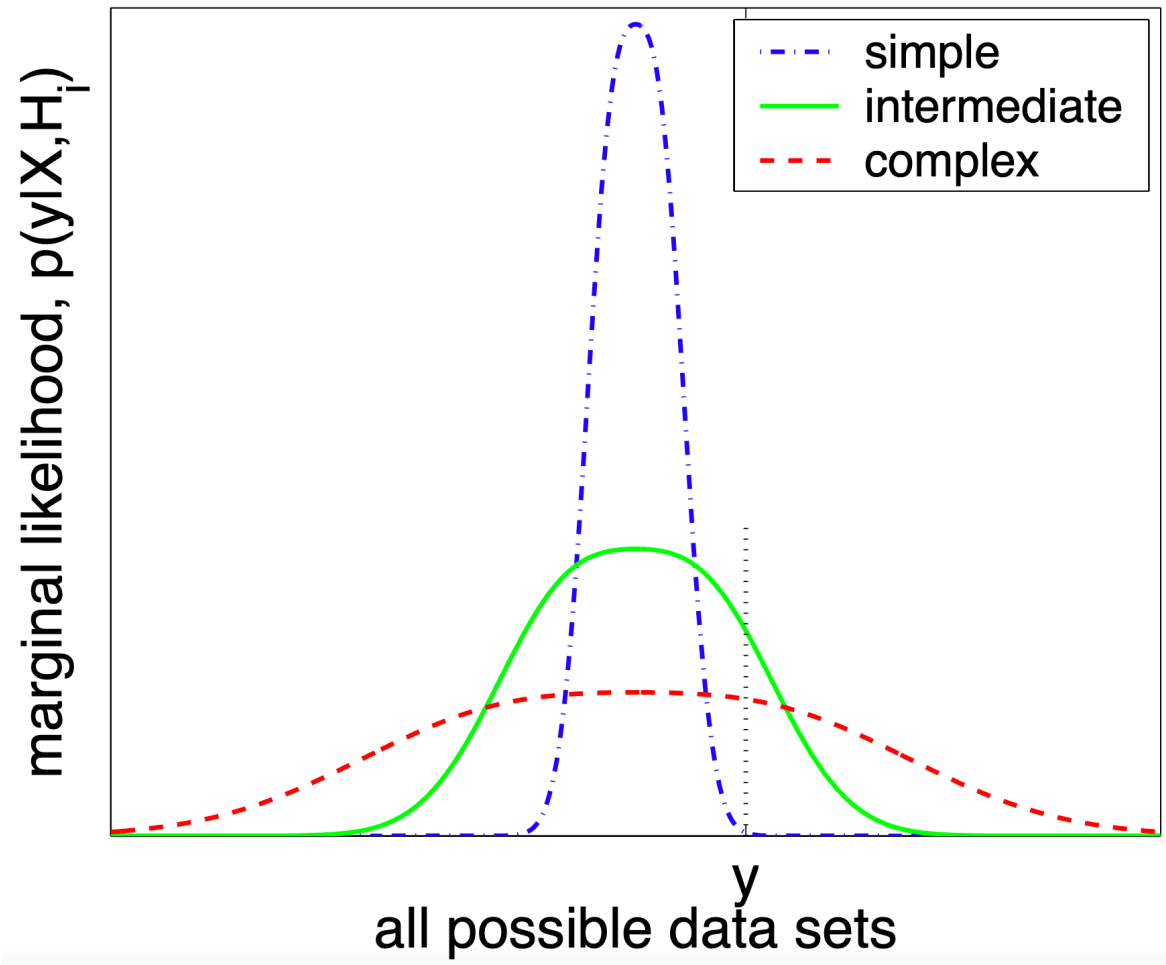
- 模型参数估计



等高线图显示了对数边际似然函数作为特征长度尺度和噪声水平的函数关系。

贝叶斯模型选择

- 奥卡姆剃刀原理(Occam’s razor)



对于三种不同复杂度模型的边际似然函数 $p(y|X,\mathcal{H}_i)$ 。横轴理想化地表示了所有可能的数据集X,y 。

简单模型可能是一个线性模型，而复杂模型可能是一个大型神经网络。威廉·奥卡姆，他提出的原理是：“Entities should not be multiplied unnecessarily”(简单有效原理)。

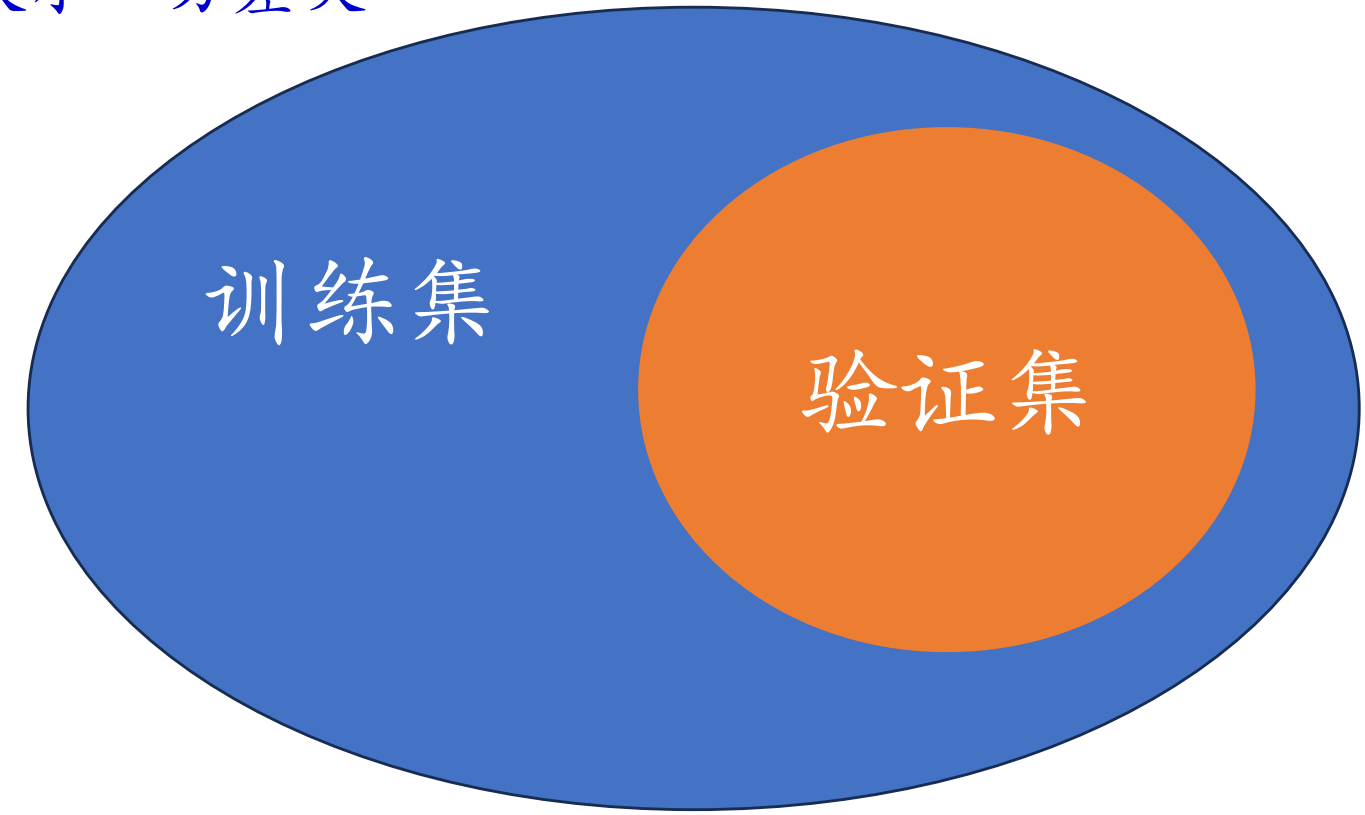
交叉验证

- 数据集

  验证集：估计泛化误差（generalized error）检验模型

  验证集太大：训练集小，浪费数据

  验证集太小：方差大



交叉验证

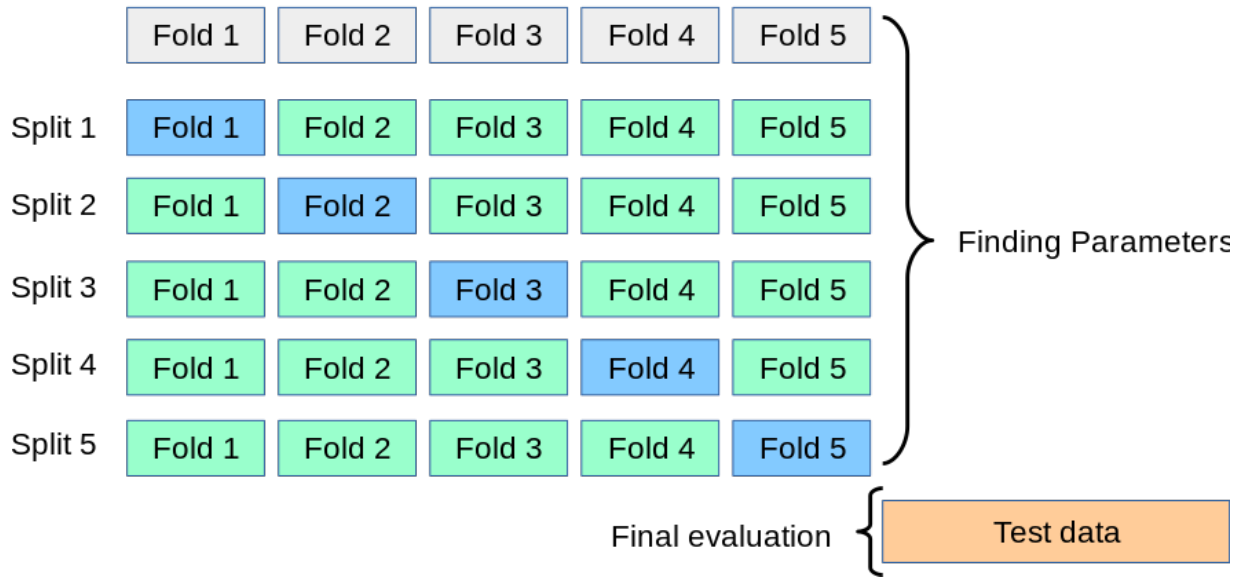
- k折（k-fold） 交叉验证

  把数据分为k分，用k-1分作为训练集，1分作为验证集

  重复k 次，计算泛化误差$\hat{e}_i$

$$e = \frac{1}{k}\sum_i \hat{e}_i$$

一般$3 \le k \le 10$



- 留一（leave-one） 交叉验证

  $k = n$

贝叶斯模型选择

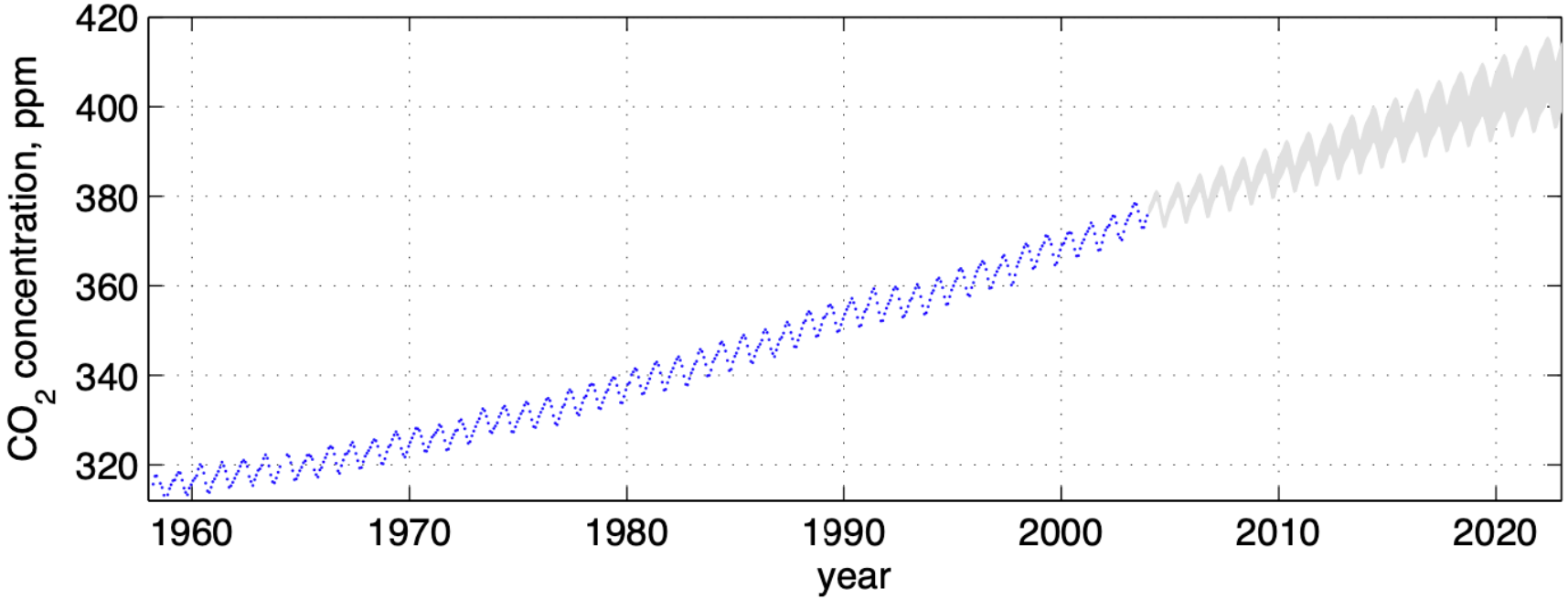
- 例子

考虑一维高斯过程，期望函数$m(x)=0$，核函数为

$$\kappa_y(x_p,x_q)=\sigma_f^2\exp\left(-\frac{1}{2l^2}\left(x_p-x_q\right)^2\right)+\sigma^2\delta_{pq}$$

用$(l,\sigma_f,\sigma)=(1.0,1.0,0.1)$生成数据点$x\sim U[-8,8]$，进行超参估计。使用留一交叉验证进行超参估计。

大气二氧化碳预测



这组数据包含了从1958年至2003年间（其中有一些缺失值）在夏威夷莫纳罗亚观测站采集的现场空气样本得出的月平均大气二氧化碳浓度。

大气二氧化碳预测

长期平滑上升趋势

$$k_1(x,x') = \theta_1^2 \exp(-\frac{(x-x')^2}{2\theta_2^2})$$

局部的季节性趋势

$$k_2(x,x') = \theta_3^2 \exp(-\frac{(x-x')^2}{2\theta_4^2} - \frac{2\sin^2(\pi(x-x'))}{\theta_5^2})$$

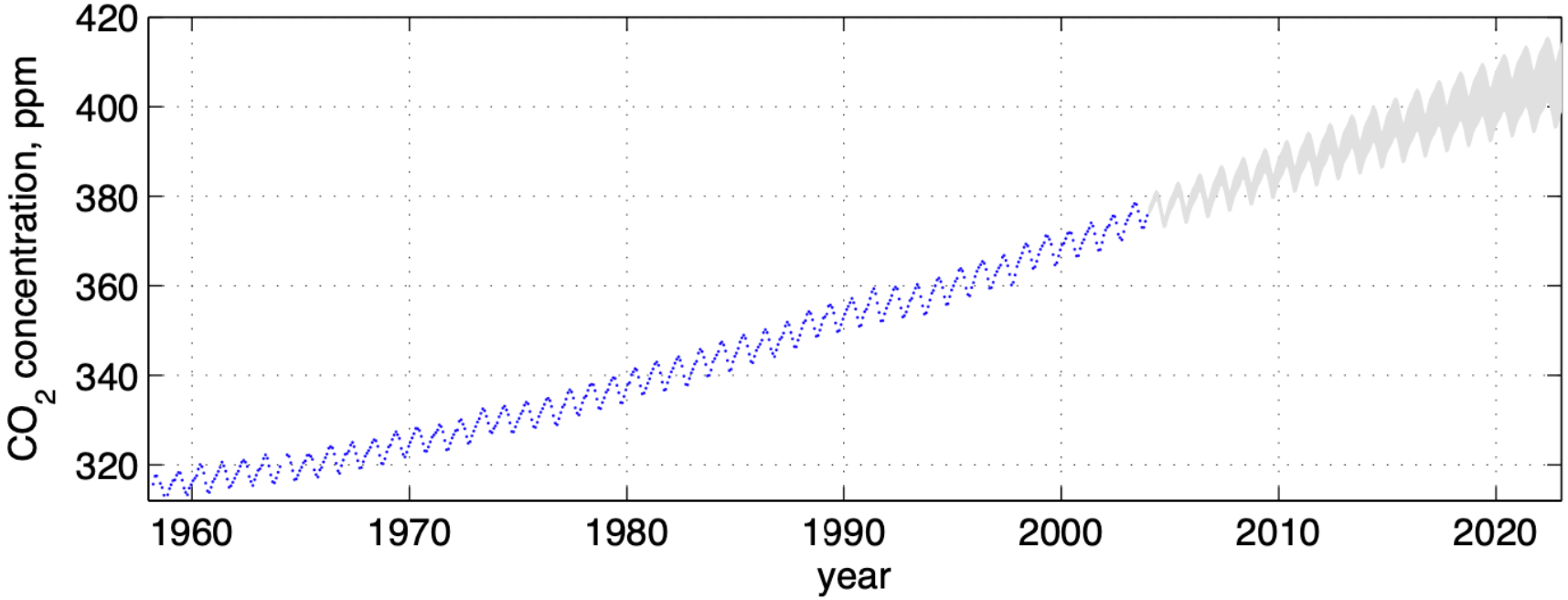
较小的中尺度趋势

$$k_3(x,x') = \theta_6^2 \left(1 + \frac{(x-x')^2}{2\theta_8\theta_7^2}\right)^{-\theta_8}$$

噪声模型

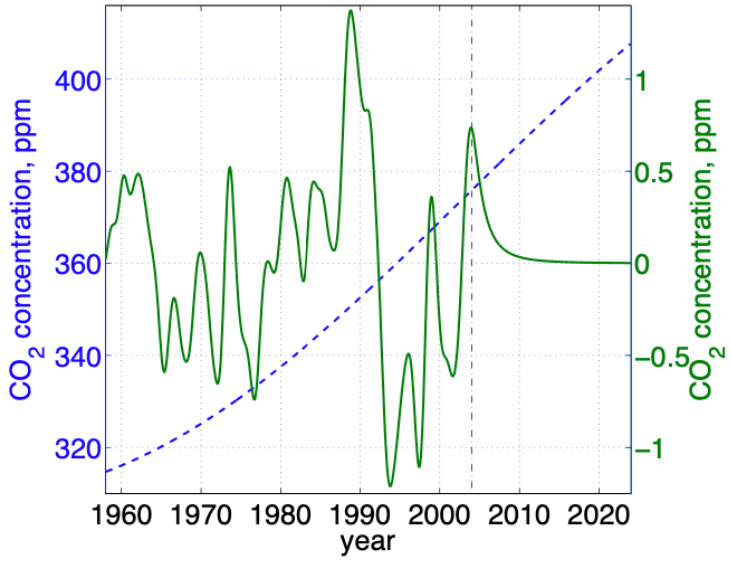
$$k_4(x,x') = \theta_9^2 \exp(-\frac{(x-x')^2}{2\theta_{10}^2}) + \theta_{11}^2 \delta_{xx'}$$

大气二氧化碳预测

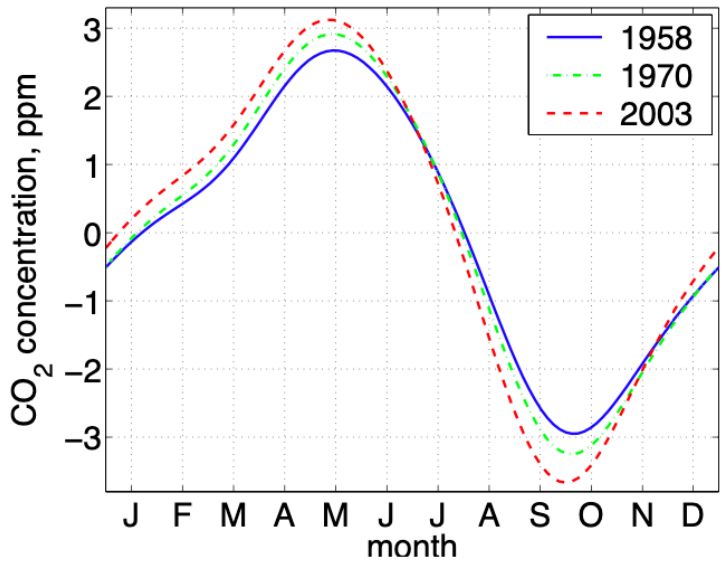


$$k(x,x') = k_1(x,x') + k_2(x,x') + k_3(x,x') + k_4(x,x')$$

大气二氧化碳预测



(a)



(b)

$$k(x,x') = k_1(x,x') + k_2(x,x') + k_3(x,x') + k_4(x,x')$$

参考文献

- 参考文献

C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, Chapter 5