24 Gaussian Process
高斯过程

高斯核协方差矩阵，多元高斯分布

今天，是别人的；我苦苦耕耘的未来，是我的。

The present is theirs; the future, for which I really worked, is mine.

—— 尼古拉·特斯拉 (Nikola Tesla) | 发明家、物理学家 | 1856 ~ 1943

- XXXXX
- XXXXXX
- XXXXXX
- XXXXXX

24.1 高斯过程原理

高斯过程 (Gaussian Process, GP) 是一种概率模型，用于建模连续函数或实数值变量的概率分布。在高斯过程中，任意一组数据点都可以被视为多元高斯分布的样本，该分布的均值和协方差矩阵由先验信息和数据点间的相似度计算而得。通过高斯过程，可以对函数进行预测并对其不确定性进行量化，这使得其在机器学习、优化和贝叶斯推断等领域中被广泛应用。

在使用高斯过程进行预测时，通常使用条件高斯分布来表示先验和后验分布。通过先验分布和数据点的观测，可以计算后验分布，并通过该分布来预测新数据点的值。在高斯过程中，协方差函数或核函数起着重要的作用，它定义了数据点间的相似性，不同的核函数也适用于不同的应用场景。一些常见的核函数包括线性核、多项式核、高斯核、拉普拉斯核等。

本章将以高斯核 (Gaussian kernel) 为例，介绍如何理解高斯过程算法原理。注意，高斯核也叫径向基 (Radial Basis Function kernel, RBF kernel)。

先验

$\boldsymbol{x}_2$ 为一系列需要预测的点，$\boldsymbol{y}_2 = \text{GP}(\boldsymbol{x}_2)$ 对应高斯过程预测结果。

高斯过程的先验为：

$$\boldsymbol{y}_2 \sim N\left(\boldsymbol{\mu}_2, \boldsymbol{K}_{22}\right) \tag{1}$$

其中，$\boldsymbol{\mu}_2$ 为高斯过程的均值 (通常默认为全 0 向量)，$\boldsymbol{K}_{22}$ 为协方差矩阵。之所以写成 $\boldsymbol{K}_{22}$ 这种形式，是因为高斯过程的协方差矩阵通过核函数定义。

在 Scikit-learn 中，高斯核的定义为：

$$\kappa\left(x_i, x_j\right) = \text{cov}\left(y_i, y_j\right) = \exp\left(-\frac{\left(x_i - x_j\right)^2}{2l^2}\right) \tag{2}$$

图 1 所示为 $l = 1$ 时先验协方差矩阵的热图。

为了保证形式上和协方差矩阵一致，图 1 纵轴上下调转。这个协方差矩阵显然是对称矩阵，它的主对角线也是方差 (variance)，非主对角线元素为协方差 (covariance)。

回到 (2)，不难发现 $\kappa\left(x_i, x_j\right)$ 体现的是 y_i 和 y_j 的协方差 (即描述协同运动)，但是 $\kappa\left(x_i, x_j\right)$ 通过 x_i 和 x_j 两个坐标点确定的。更确切地说，如图 2 所示，当 l 一定时，x_i 和 x_j 间距绝对值 ($\Delta x = x_i - x_j$) 越大，$\kappa\left(x_i, x_j\right)$ 越小；反之，Δx 越小，$\kappa\left(x_i, x_j\right)$ 越大。这一点后续将会反复提及。

此外，图 2 还展示了参数 l 对高斯函数的影响。

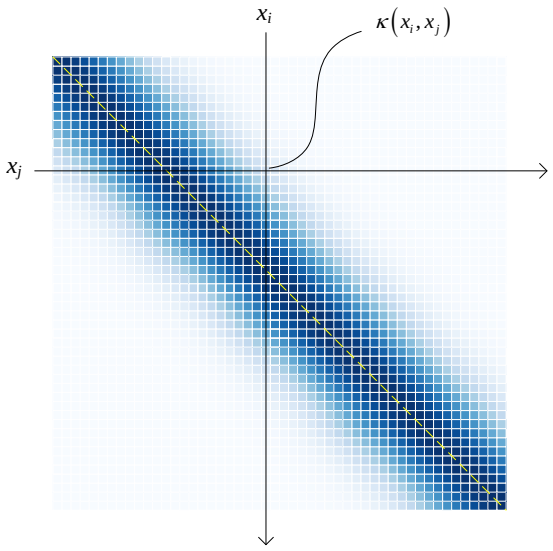


图 1. 高斯过程的先验协方差矩阵，高斯核

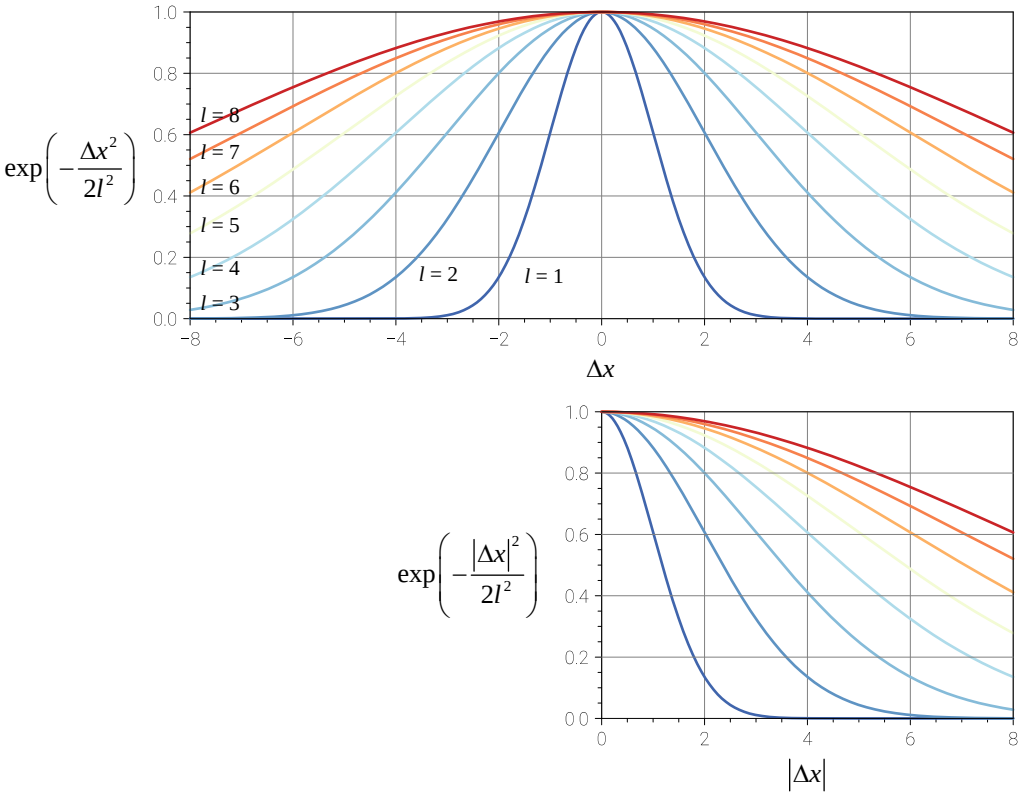


图 2. 高斯核函数受参数 l 的影响

图 3 所示每一条代表一个根据当前先验均值、先验协方差的函数采样。打个比方，在没有引入数据之前，图 3 的曲线可以看成是一捆没有扎紧的丝带，随着微风飘动。

图 3 中的红线为高斯过程的先验均值，本章假设均值为 0。本章接下来要解释为什么图 3 中曲线是这种形式。

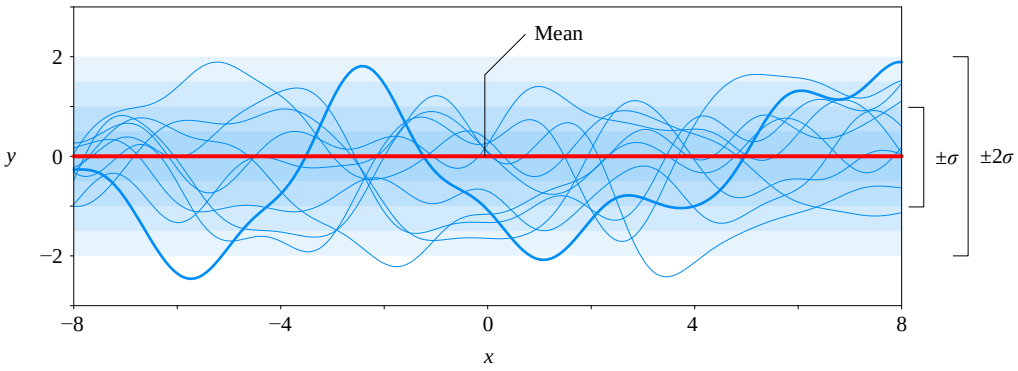


图 3. 高斯过程的采样，高斯核先验，$\sigma = 1$

样本数据

观测到的样本数据为 $(\boldsymbol{x}_1, \boldsymbol{y}_1)$。图 4 给出 5 个样本点，大家很快就会发现这 5 个点相当于扎紧丝带的 5 个节点。下面，我们要用贝叶斯方法来帮我们整合“先验 + 数据”，并计算后验分布。

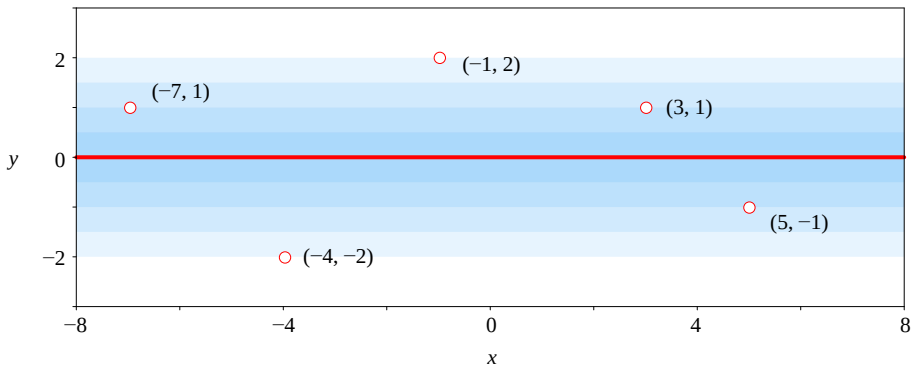


图 4. 给定 5 个样本数据

联合分布

假设样本数据 $\boldsymbol{y}_1$ 和预测值 $\boldsymbol{y}_2$ 服从联合高斯分布：

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}_{11} & \boldsymbol{K}_{12} \\ \boldsymbol{K}_{21} & \boldsymbol{K}_{22} \end{bmatrix} \right) \tag{3}$$

简单来说，高斯过程对应的分布可以看成是无限多个随机变量的联合分布。图 5 中的协方差矩阵来自 $[\boldsymbol{x}_1, \boldsymbol{x}_2]$ 的核函数。本章后文会用实例具体展示如何计算上式中的协方差矩阵 $(\boldsymbol{K}_{11}, \boldsymbol{K}_{22})$ 和互协方差矩阵 $(\boldsymbol{K}_{12}, \boldsymbol{K}_{21})$。

注意，一般假设 $\boldsymbol{\mu}_1$、$\boldsymbol{\mu}_2$ 为全 0 向量。

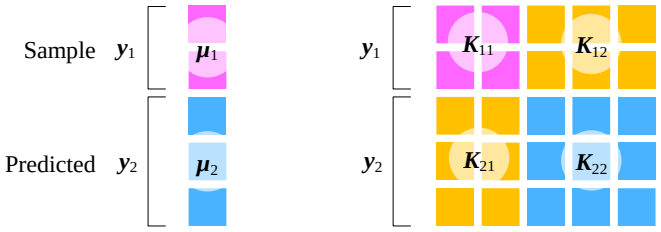


图 5. 样本数据 y_1 和预测值 y_2 服从联合高斯分布

后验分布

根据条件高斯分布，我们可以获得后验分布为：

$$f(\boldsymbol{y}_2 \mid \boldsymbol{y}_1) \sim N\left(\underbrace{\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}(\boldsymbol{y}_1-\boldsymbol{\mu}_1)+\boldsymbol{\mu}_2}_{\text{Expectation}}, \underbrace{\boldsymbol{K}_{22}-\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}\boldsymbol{K}_{12}}_{\text{Covariance matrix}}\right) \tag{4}$$

看到这个式子，特别是条件期望部分，大家是否想到了多元线性回归？

如图 6 所示，在 5 个样本点位置丝带被锁紧，而其余部分丝带仍然舞动。

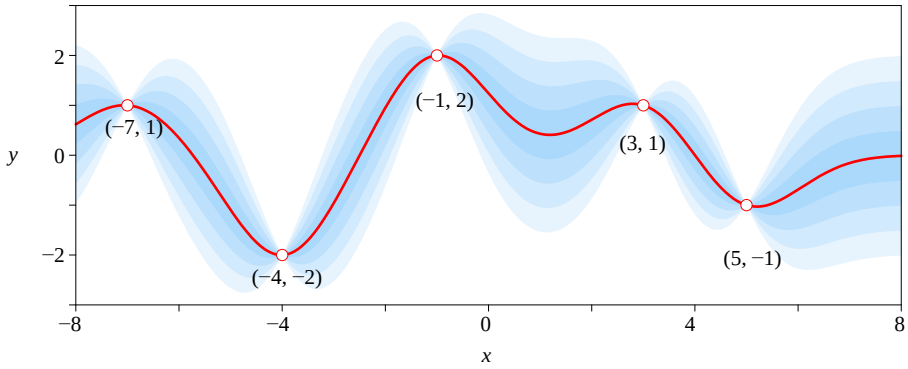


图 6. 高斯过程后验分布的采样函数，高斯核

图 6 中红色曲线对应后验分布的均值：

$$\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}(\boldsymbol{y}_1-\boldsymbol{\mu}_1)+\boldsymbol{\mu}_2 \tag{5}$$

图 6 中带宽对应一系列标准差：

$$\text{sqrt}\left(\text{diag}\left(\boldsymbol{K}_{22}-\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}\boldsymbol{K}_{12}\right)\right) \tag{6}$$

其中，diag() 表示获取对角线元素；sqrt() 代表开平方得到一组标准差序列，代表纵轴位置的不确定性。

如图 7 所示，在高斯过程算法中，贝叶斯定理将先验和数据整合到一起得到后验。看到这里，大家如果还是不理解高斯过程原理，不要紧。下面，我们就用这个例子展开讲解高斯过程中的技术细节。

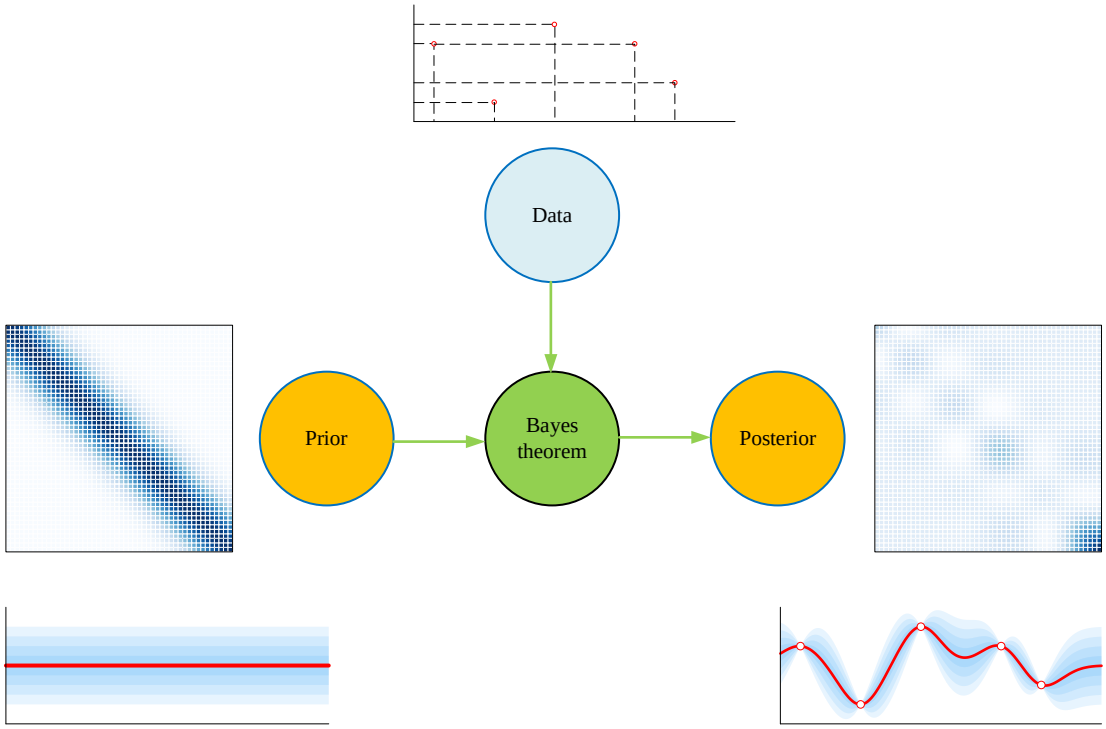


图 7. 高斯过程算法中贝叶斯定理的作用

24.2 协方差矩阵

基于高斯核的协方差矩阵相当是一种“人造”协方差矩阵。当然，这种“人造”协方差矩阵有它的独到之处，下面就近距离观察这个协方差矩阵。

以往，我们看到的协方差矩阵都是有限大小，通常用热图表示。高斯过程算法用到的协方差矩阵实际上是无限大。比如，如果 x 的取值范围为 $[-8, 8]$，在这个区间内满足条件的 x 值有无数个。

如图 8 所示，我们用三维网格图呈现这个无限大的协方差矩阵。和一般的协方差矩阵一致，这个协方差矩阵的主对角线元素为方差，非主对角线元素为协方差。

我们容易发现图 8 这个协方差矩阵特别像是一个二元函数。我们可以固定一个变量，看协方差值随另外一个变量变化。不难发现，图中的每条曲线都是一条高斯函数。

再次强调，x 为预测点，我们关注的是高斯过程预测结果 $y = \mathrm{GP}(x)$ 之间的关系。简单来说，x 提供位置坐标，不同 $y = \mathrm{GP}(x)$ 之间的协同运动用高斯核来描述。

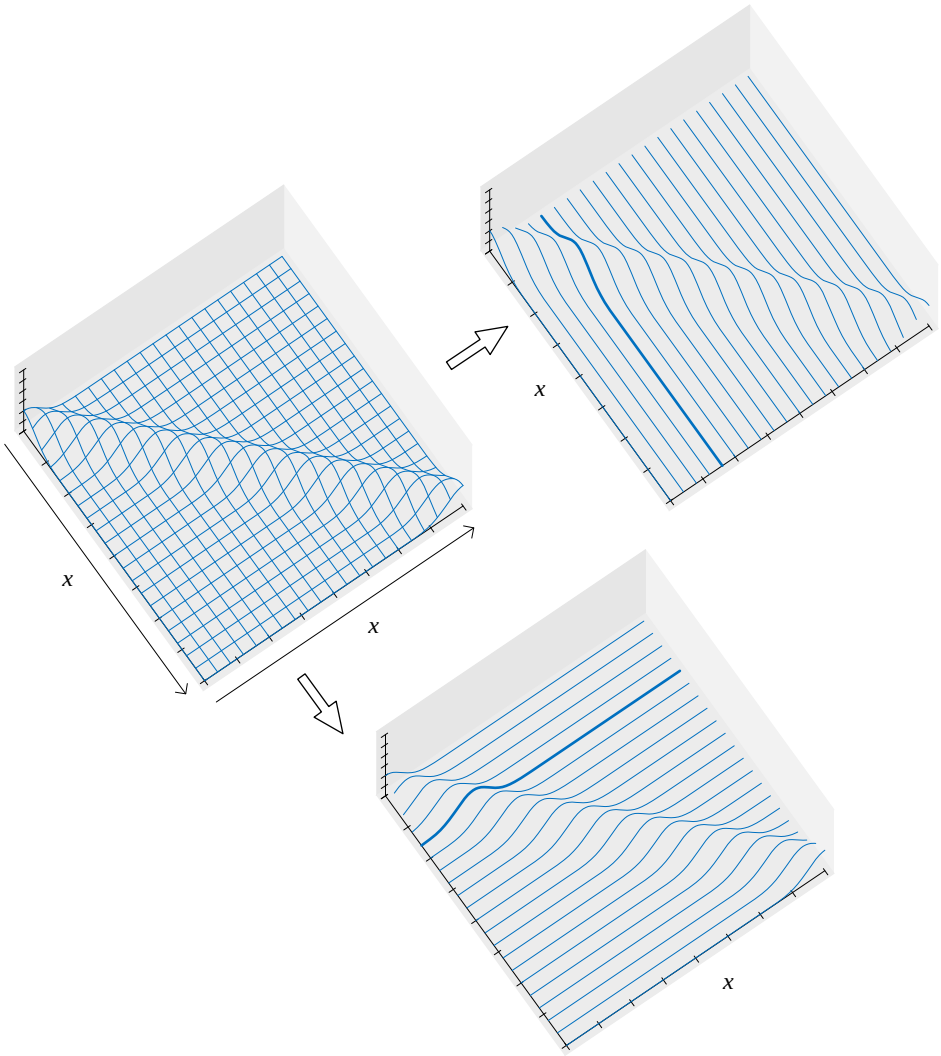


图 8. 无限大的先验协方差矩阵

为了方便可视化，同时为了和我们熟悉的协方差矩阵对照来看，我们选取 $[-8, 8]$ 区间中 50 个点，并绘制如图 9 所示的协方差矩阵热图。下面，我们来观察图 9 中协方差矩阵的每一行。

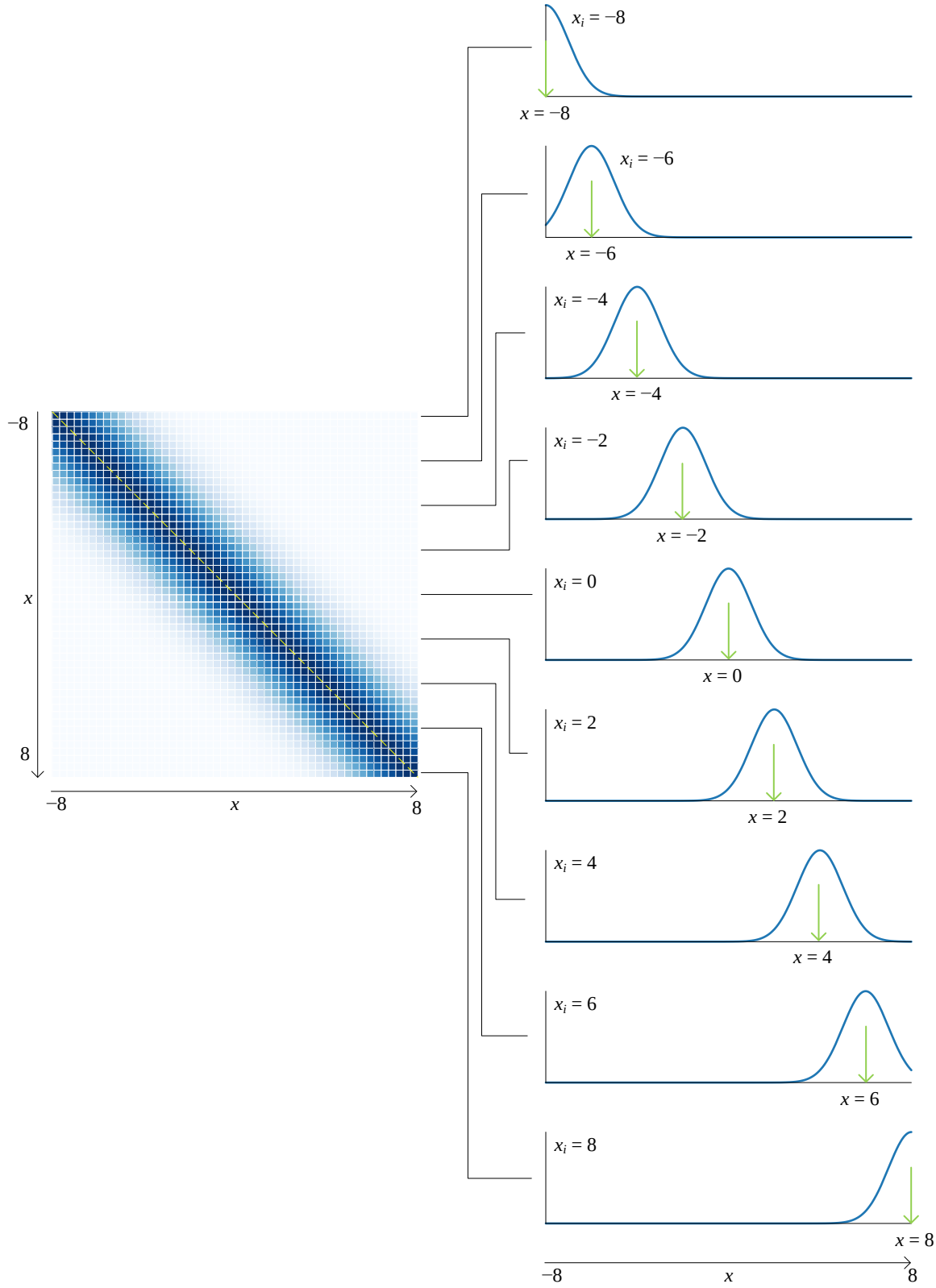


图 9. 观察协方差矩阵的每一行

当确定一个 x_i 取值后，比如 $x_i = 0$，对于区间 $[-8, 8]$ 上任意一点 x，利用高斯核我们都可以计算得到一个协方差值

$$\kappa(x_i, x) = \exp\left(-\frac{(x_i - x)^2}{2l^2}\right) \tag{7}$$

观察这个函数，我们可以发现，函数在 $x = x_i$ 取得最大值。而随着 x 不断远离 x_i，即 $x_i - x$ 绝对值增大，协方差值不断减小，不断靠近 0。

对于 (7)，$x = x_i$ 这一点又恰好是 x_i 位置处的 y_i = GP(x_i) 方差，即 $\text{var}(y_i) = \kappa(x_i, x_i) = 1$。有了这一观察，我们可以发现图 10 中协方差主对角线元素都是 1；也就是说，在给定均值为 0，高斯核为先验函数的条件下，任何一点处 x_i 的 y_i = GP(x_i) 具有相同的“不确定性”。这就是我们可以在图 3 中观察到的，“丝带”任何一点在一定范围内飘动。

学习本章配套代码时，大家就会发现图 3 的每条“丝带”上的纵轴值服从多元高斯分布。

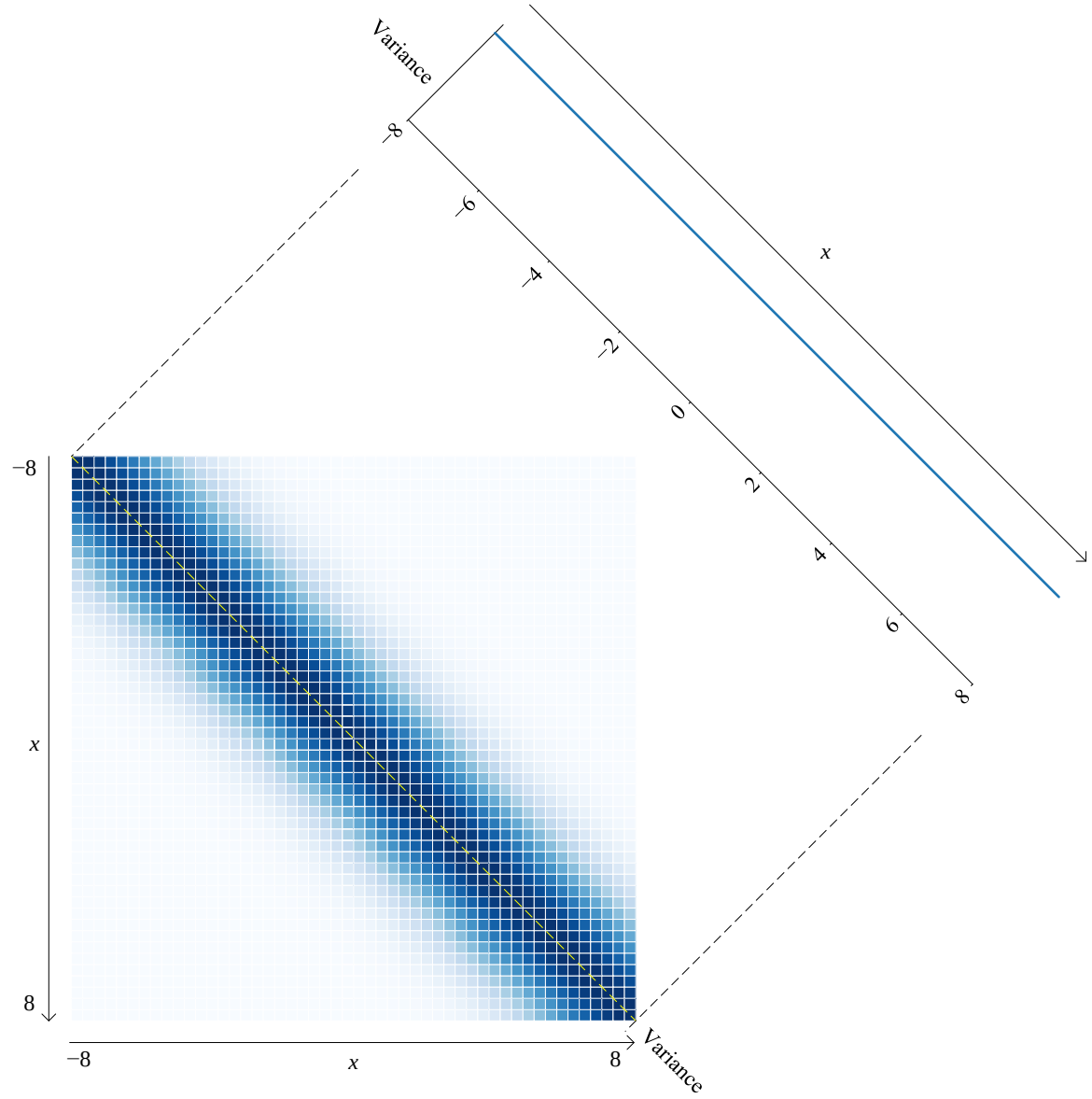


图 10. 高斯核协方差矩阵的方差 (对角元素)

但是观察图 3，我们还发现同一条丝带看上去很“顺滑”，这又是为什么？

想要理解这一点，我们就要关注协方差矩阵中的协方差成分。

在给定 (2) 这种形式的高斯核条件下，对于点 x_i，它和 $x_i + \Delta x$ 的 2 × 2 协方差矩阵可以写成。

$$\begin{bmatrix} 1 & \exp\left(-\frac{\Delta x^2}{2l^2}\right) \\ \exp\left(-\frac{\Delta x^2}{2l^2}\right) & 1 \end{bmatrix} \tag{8}$$

我们已经知道，点 x_i 和点 $x_i + \Delta x$ 的方差都是 1，因此两者的相关性系数为 $\exp\left(-\frac{\Delta x^2}{2l^2}\right)$。

当 l 为定值时，Δx 绝对值越大，即点 $x_i + \Delta x$ 离 x_i 越远，两者的相关性越靠近 0；相反，Δx 绝对值越小，即点 $x_i + \Delta x$ 离 x_i 越近，两者的相关性越靠近 1。

如图 11 所示，当相关性系数 $\exp\left(-\frac{\Delta x^2}{2l^2}\right)$ (非负值) 为不同值时，代表 2 × 2 协方差矩阵 (8) 的椭圆不断变化。

这就解释了图 3 中每一条丝带看上去很顺滑的原因。越靠近丝带的任意一点，相关性越高，也就是说具有更高的协同运动；距离特定点越远，相关性越低，协同运动关系也就越差。

请大家注意，不限定取值范围时，x_i 可以是实轴上任意一点；换个角度来看，x_i 有无数个。

很多其他文献上中高斯核定义为：

$$\kappa\left(x_i, x_j\right) = \sigma^2 \exp\left(-\frac{\left(x_i - x_j\right)^2}{2l^2}\right) \tag{9}$$

上式中先验协方差矩阵中的方差不再是 1，而是 σ^2。

大家可能会问既然这个高斯核协方差矩阵是“人造”的，我们可不可以创造其他形式的协方差矩阵？

答案是肯定的！下一章将介绍高斯过程中常用的其他核函数。

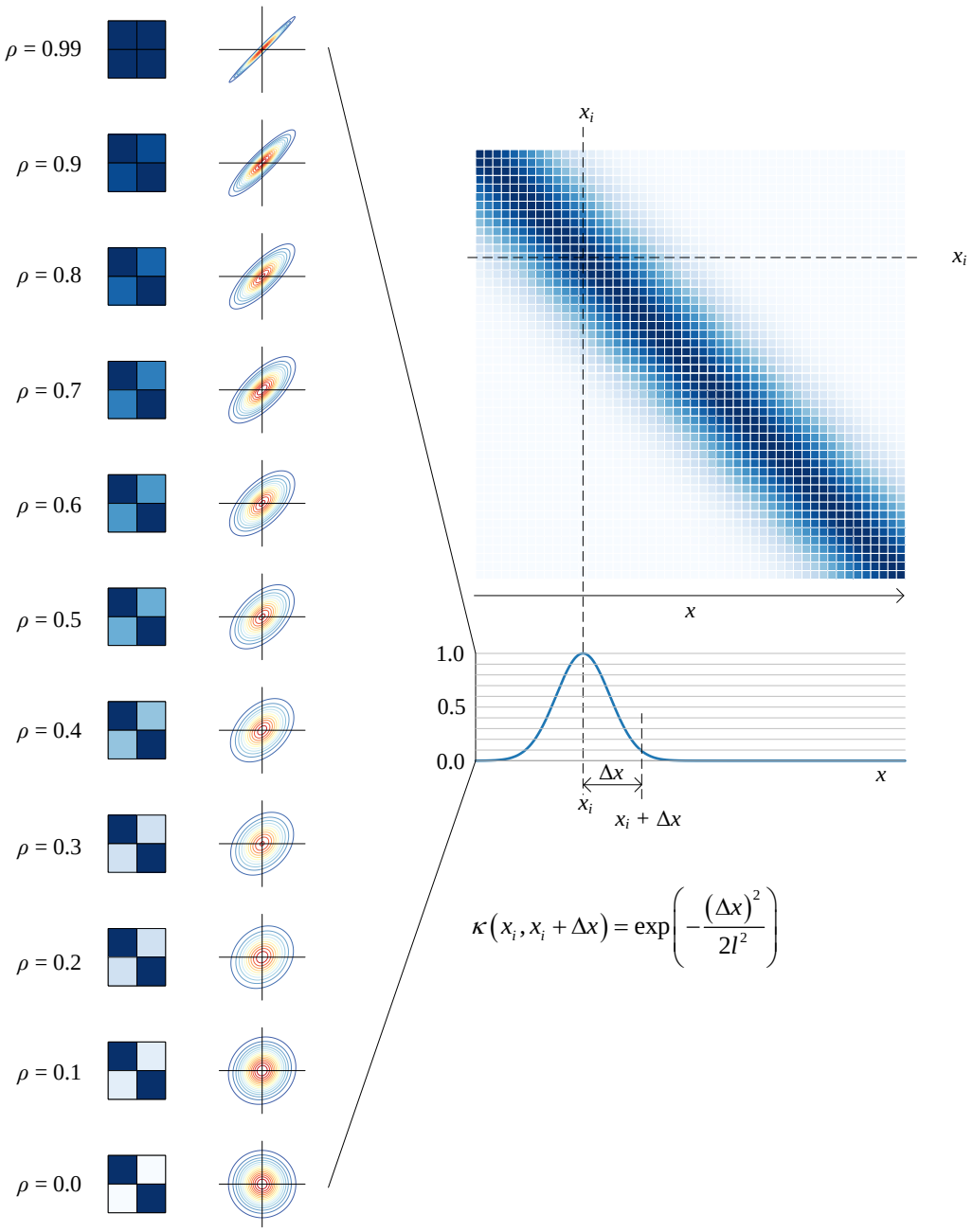


图 11. 高斯核协方差矩阵的协方差 (非对角元素)

24.3 分块协方差矩阵

根据 (3)，我们先将协方差矩阵分块，具体如图 12 所示。其中，$\boldsymbol{K}_{11}$ 是样本数据的协方差矩阵，$\boldsymbol{K}_{22}$ 是先验协方差矩阵。$\boldsymbol{K}_{12}$ 和 $\boldsymbol{K}_{21}$ 都是互协方差矩阵 (cross covariance matrix)，且互为转置。

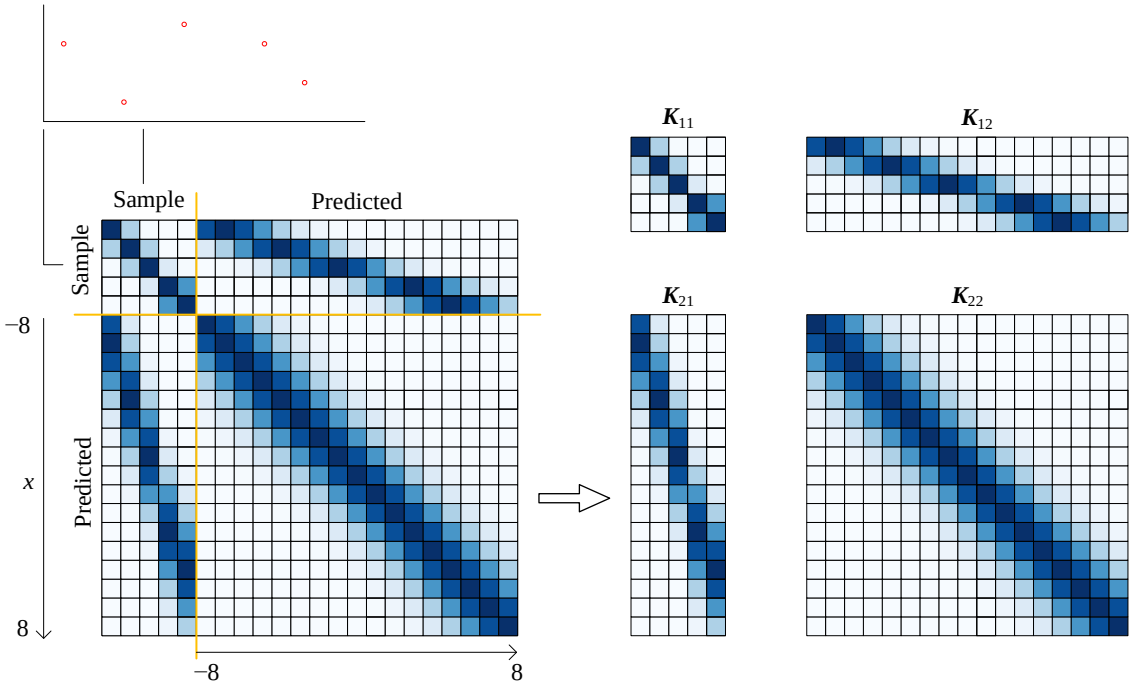


图 12. 协方差矩阵 $\boldsymbol{K}$ 分块

24.4 后验

下面就是利用 (4)，计算条件期望向量和条件协方差矩阵。

图 13 所示为计算条件期望向量 $\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}\left(\boldsymbol{y}_1-\boldsymbol{\mu}_1\right)+\boldsymbol{\mu}_2$ 的过程。

$\boldsymbol{y}_1$ 向量对应图 4 中 5 个红色点的 y 值序列。

默认，$\boldsymbol{\mu}_1$ 向量为全 0 向量，和 $\boldsymbol{y}_1$ 形状相同。

图 13 中协方差矩阵 $\boldsymbol{K}_{11}$ 则图 4 中 5 个红色点的 x 序列 $(\boldsymbol{x}_1)$ 采用 (2) 计算得到。图 14 所示为计算 $\boldsymbol{K}_{11}$ 的示意图。再次强调 $(\boldsymbol{x}_1, \boldsymbol{y}_1)$ 代表样本数据。

$\boldsymbol{K}_{21}$ 的行代表 [−8, 8] 区间上顺序采样的一组数值，即预测点序列 $\boldsymbol{x}_2$；$\boldsymbol{K}_{21}$ 的列代表 5 个红色点的 x 值序列。图 15 所示为计算互协方差矩阵 $\boldsymbol{K}_{21}$ 的示意图。显然，$\boldsymbol{K}_{21}$ 是根据 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 计算得到的。

计算结果 $\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}\left(\boldsymbol{y}_1-\boldsymbol{\mu}_1\right)+\boldsymbol{\mu}_2$ 为列向量，对应中图 6 中红色线 y 值序列。

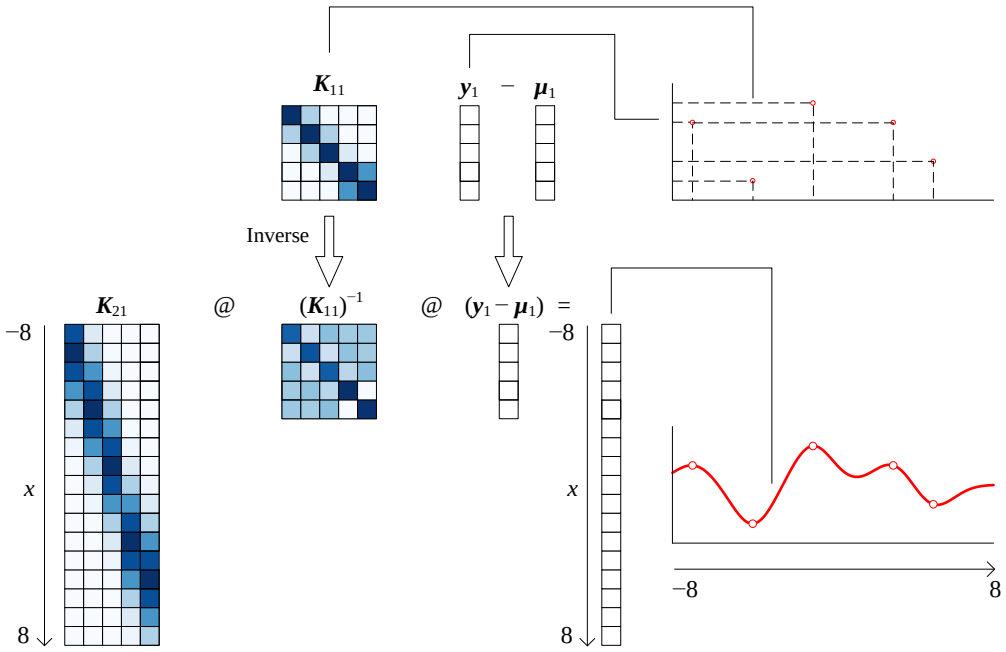


图 13. 计算条件期望

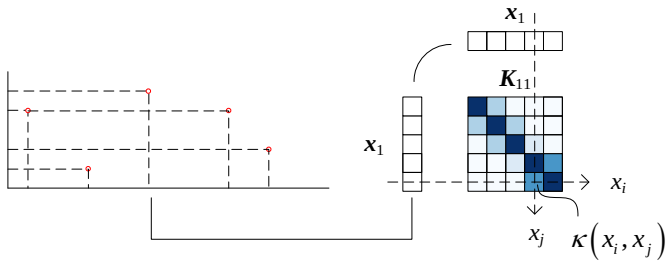


图 14. 计算协方差矩阵 K_{11}

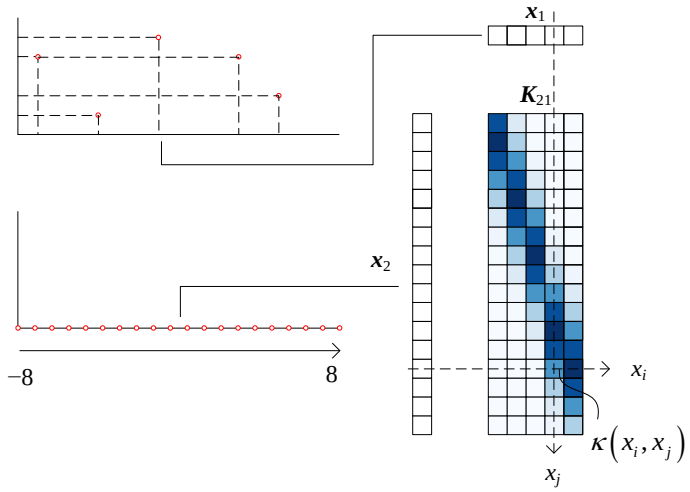


图 15. 计算互协方差矩阵 K_{21}

图 16 所示为计算条件协方差 $\boldsymbol{K}_{22}-\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}\boldsymbol{K}_{12}$ 的过程。比较图 15 和图 17，很容易发现 $\boldsymbol{K}_{21}$ 和 $\boldsymbol{K}_{12}$ 互为转置。图 18 所示为自己算协方差矩阵的示意图 $\boldsymbol{K}_{22}$。

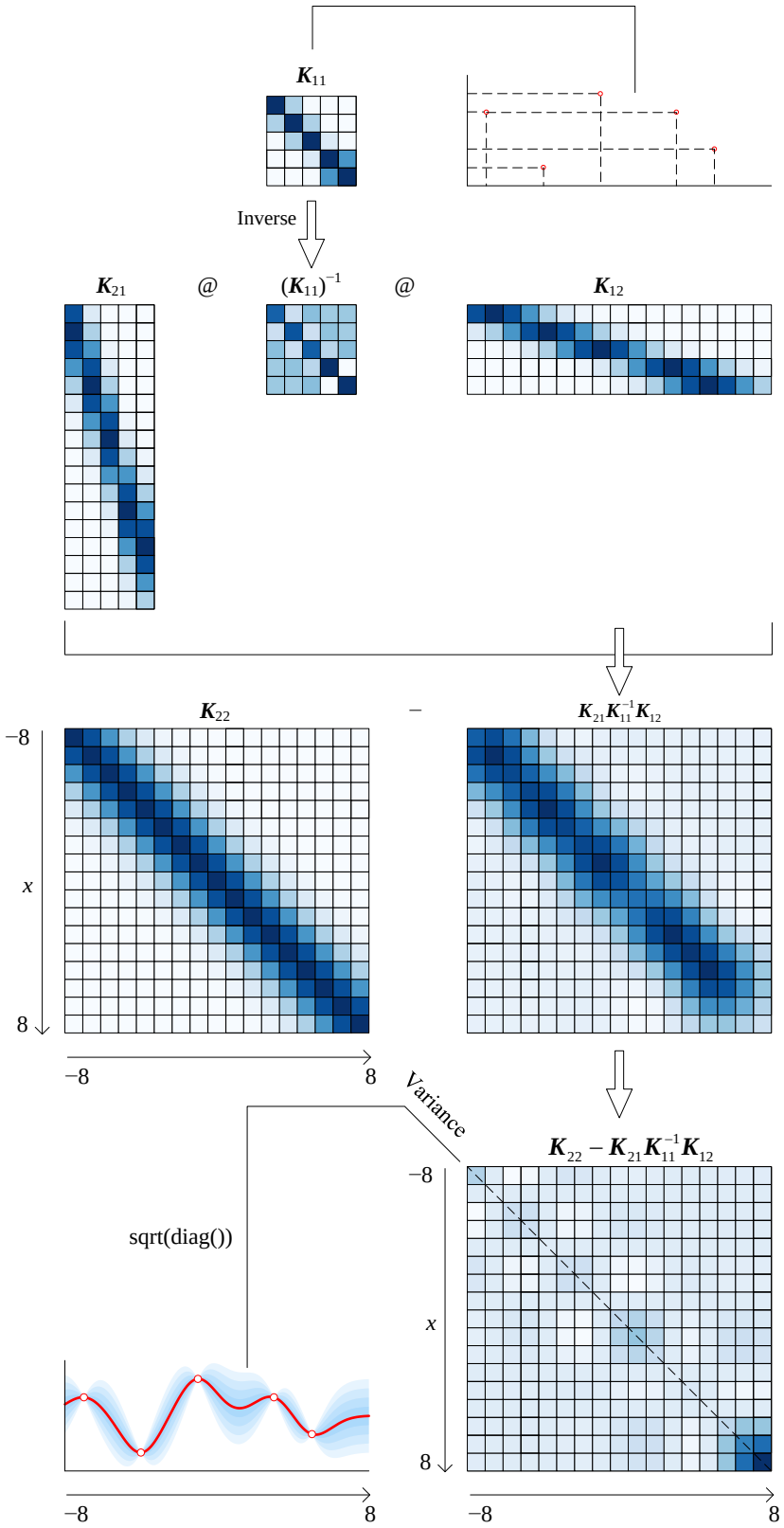


图 16. 计算条件协方差矩阵

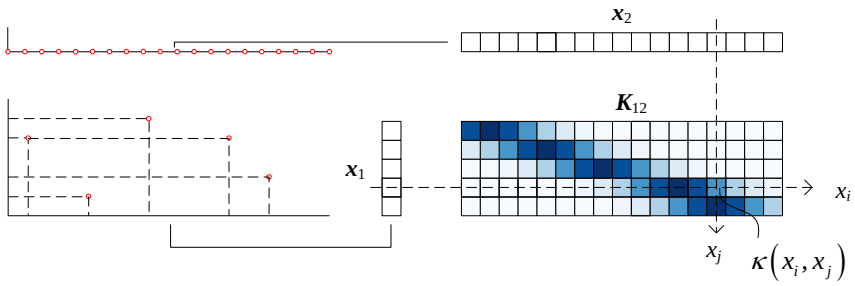


图 17. 计算互协方差矩阵 K_{21}

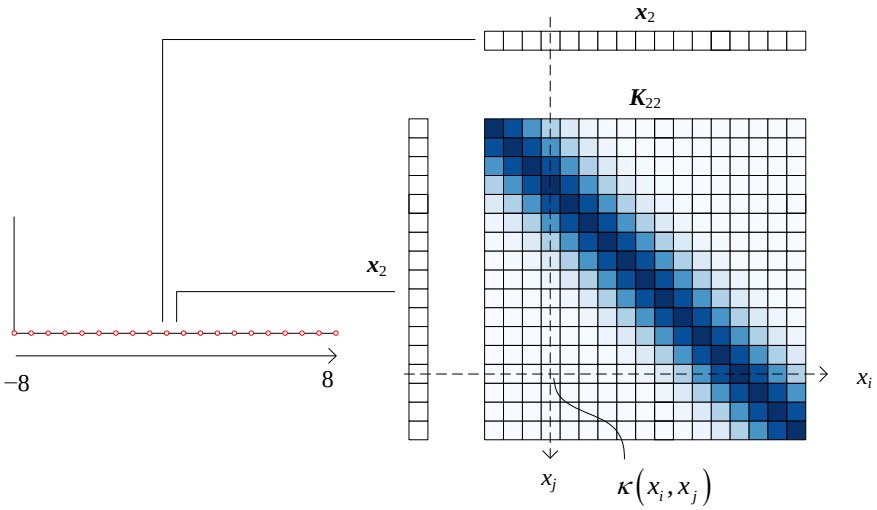


图 18. 计算协方差矩阵 K_{22}

图 19 所示为“无限大”的后验协方差矩阵曲面。和图 8 中先验协方差矩阵相比，我们可以发现在存在样本数据的位置，曲面发生了明显的塌陷。特别是在对角线上。

为了更清楚地看到这一点，我们特别绘制了图 20。后验协方差矩阵 $K_{22} - K_{21}K_{11}^{-1}K_{12}$ 的对角线元素为后验方差，体现了 x_2 不同位置上 y_2 值的不确定性。

图 20 中的标准差在 5 个数据点位置下降到 0；也就是说，不确定性为 0。这就解释了图 6 中 5 个“扎紧”的节点。换个角度来看，这也说明模型样本数据不存在任何“噪音”。倘若“噪音”存在，我们就需要修正 K_{11}。这是下一节要介绍的内容。

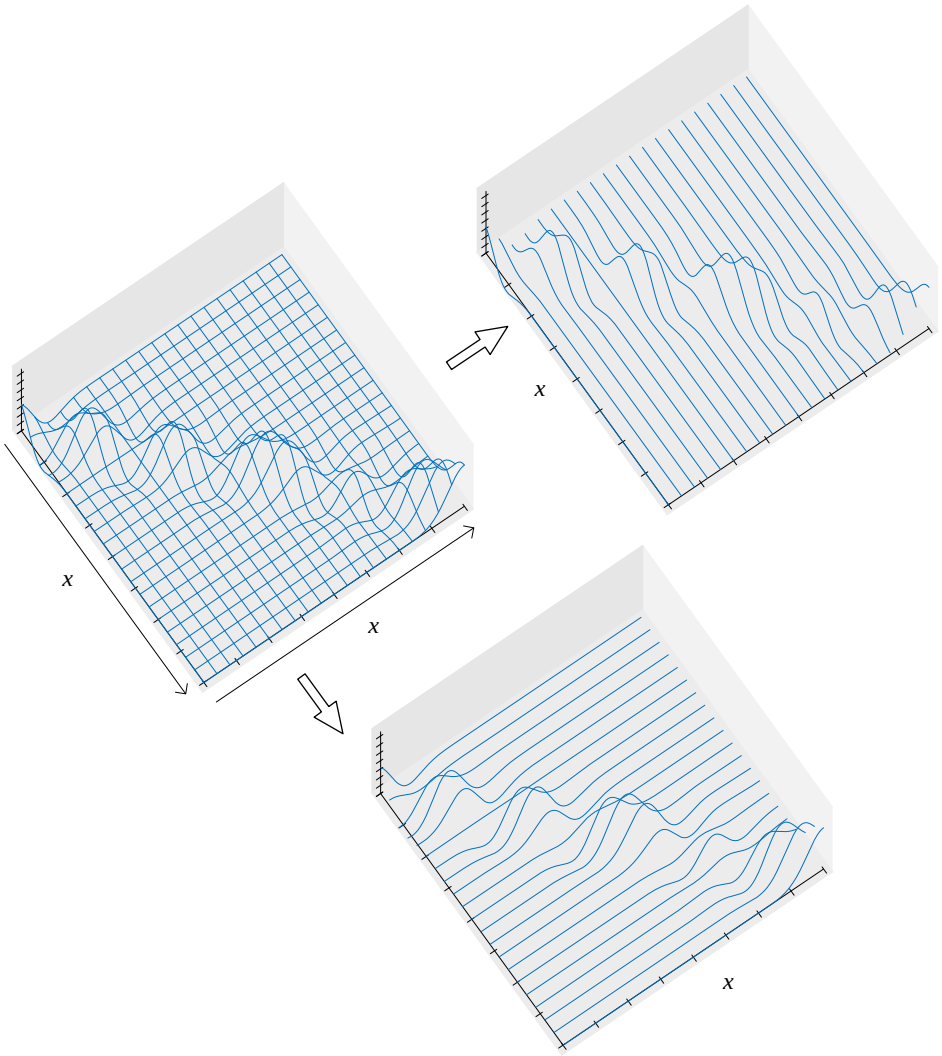


图 19. 无限大的后验协方差矩阵

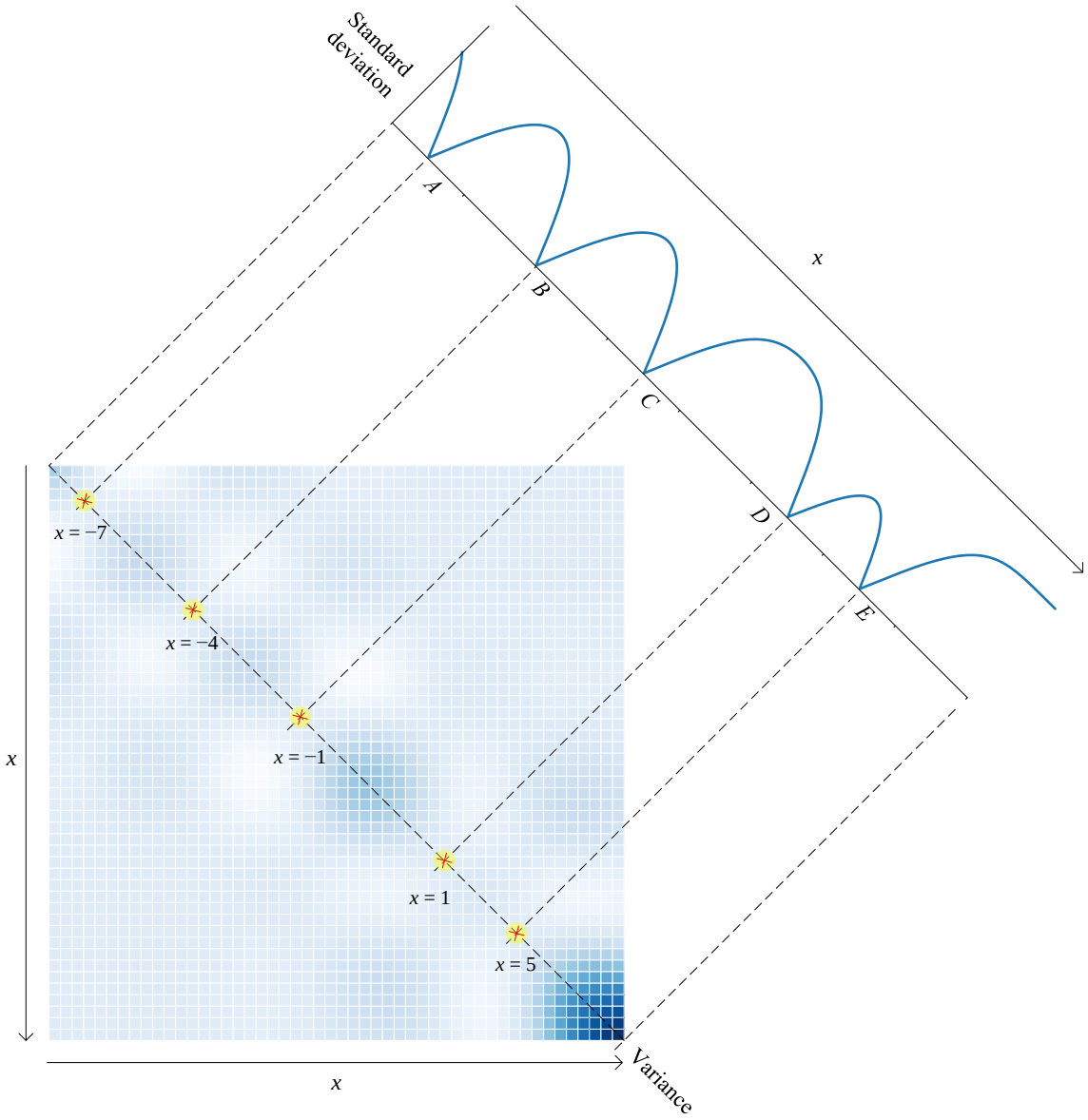


图 20. 后验协方差矩阵对角线元素

24.5 噪音

上一节提到图 20 中给定数据点处后验方差将至 0，这意味着数据不存在噪音。反之，当说数据存在噪音时，意味着观测到的数据可能受到随机误差或不确定性的影响。

在高斯过程中处理带有噪音的数据通常包括在模型中的 $\boldsymbol{K}_{11}$ 中引入噪声项 (如图 21)，以反映实际观测中的不确定性。图 22 所示为不同噪音水平对结果影响。

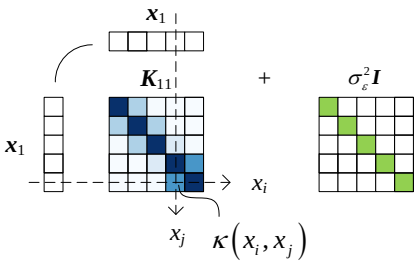


图 21. 在协方差矩阵 K_{11} 加噪音

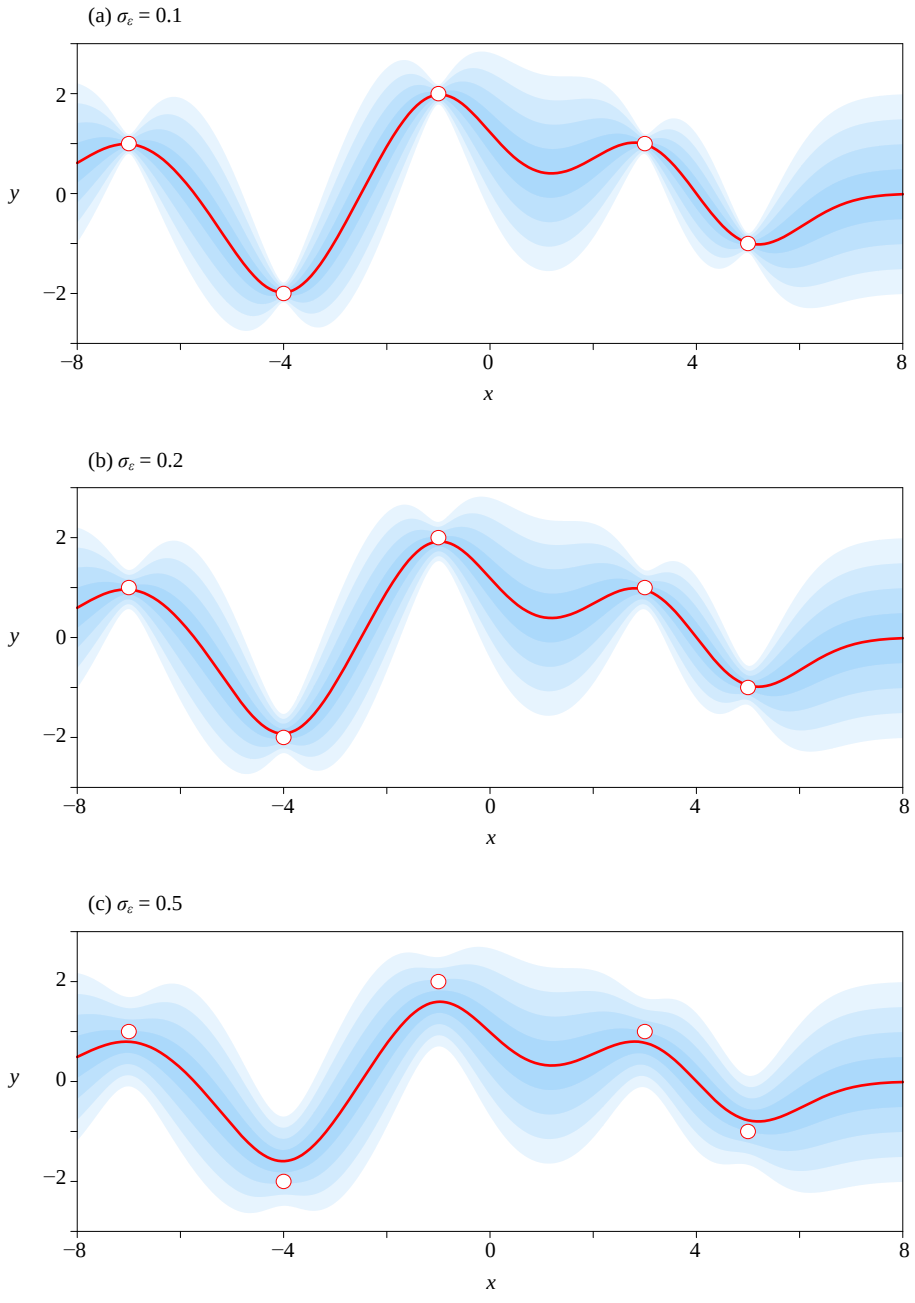
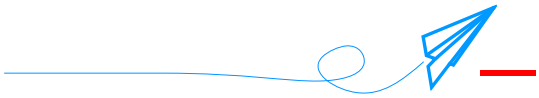


图 22. 在噪音对结果影响



高斯过程可谓高斯分布和贝叶斯定理的完美结合体。本章的关键是理解高斯过程算法原理。希望大家学完这章后，能够掌握如何用高斯核函数构造协方差矩阵，并计算后验分布。下一章将介绍高斯过程中可能用到的更多核函数。

此外，《机器学习》一册还会再介绍高斯过程，我们会用 Scikit-learn 中高斯过程工具完成回归和分类两种不同问题。