# 35 数据分布

Distributions of Sample Data

先验分布、证据因子、后验分布

青春不是芳华年少，而是一种心境；青春不是桃花红颜、盈盈朱唇、轻柔体态，而是积极的心志，丰富的想象，炙热的感情；青春是充满生机、清新盎然的生命源泉。

***Youth is not a time of life; it is a state of mind; it is not a matter of rosy cheeks, red lips and supple knees; it is a matter of the will, a quality of the imagination, a vigor of the emotions; it is the freshness of the deep springs of life.***

—— 塞缪尔·厄尔曼 (Samuel Ullman) | 美国诗人 | 1840 ~ 1924

- `matplotlib.pyplot.contour3D()` 绘制三维等高线图
- `matplotlib.pyplot.contourf()`绘制平面填充等高线
- `matplotlib.pyplot.fill_between()` 区域填充颜色
- `matplotlib.pyplot.plot_wireframe()` 绘制线框图
- `matplotlib.pyplot.scatter()` 绘制散点图
- `numpy.ones_like()` 用来生成和输入矩阵形状相同的全 1 矩阵
- `numpy.outer()` 计算外积，张量积
- `numpy.vstack()` 返回竖直堆叠后的数组
- `scipy.stats.gaussian_kde()` 高斯核密度估计
- `seaborn.kdeplot()` 绘制 KDE 概率密度估计曲线
- `statsmodels.api.nonparametric.KDEUnivariate()` 构造一元 KDE
- `statsmodels.nonparametric.kde.kernel_switch()` 更换核函数
- `statsmodels.nonparametric.kernel_density.KDEMultivariate()` 构造多元 KDE

# 35.1 引入新视角：贝叶斯推断

《编程不难》和本书前文介绍过常见数据分布的可视化方案，特别是利用 Seaborn 各种函数展示鸢尾花数据分布。本章引入一个全新视角，贝叶斯推断 (Bayesian statistical inference)，来观察、分析数据。

贝叶斯统计推断是一种基于贝叶斯定理的统计推断方法。它利用先验知识和观测数据来更新对未知量的信念或概率分布，并计算后验概率。

以下是与贝叶斯统计推断相关的一些关键概念。

贝叶斯定理描述了在已知先验概率和条件概率的情况下，如何计算后验概率。条件概率指在某个条件下某事件发生的概率。

似然概率 (likelihood probability) 是在特定条件下，观测数据出现的概率。如图 1、图 2 所示，给定鸢尾花的分类条件下，估计样本数据分布得到的结果便是似然概率。图 1 采用二元高斯分布估算似然概率，等高线为椭圆。图 2 采用高斯核函数估计似然概率。

本章中，证据因子 (evidence) 描述鸢尾花数据的分布，证据因子根据似然概率计算得到。具体计算方法请大家参考《统计至简》第 18、19 章。

后验概率 (posterior probability) 是在考虑了观测数据后，对未知量的概率分布进行更新得到的概率分布。

# 35.2 高斯分布

图 3、图 4、图 5 所示为利用二元高斯分布估算得到的三个似然概率结果。

图 6 为利用图 3、图 4、图 5 计算得到的证据因子结果。

图 7、图 8、图 9 所示为计算得到的后验概率结果。后验概率常用来分类决策。比如，给定某朵鸢尾花花萼长度为 5 cm、花萼宽度 3 cm 条件下，它最可能是哪一类鸢尾花？回答这个问题就可以用后验概率的具体值。

# 35.3 高斯核密度估计

图 10、图 11、图 12 为利用二元高斯核密度估计估算得到的三个似然概率结果。

图 13 为利用图 10、图 11、图 12 计算得到的证据因子结果。

图 14、图 15、图 16 所示为计算得到的后验概率结果。

本章不会展开讨论这些统计学概念。只给定性描述，不会定量计算。大家可以在《统计至简》一书中找到相关的数学工具具体介绍。

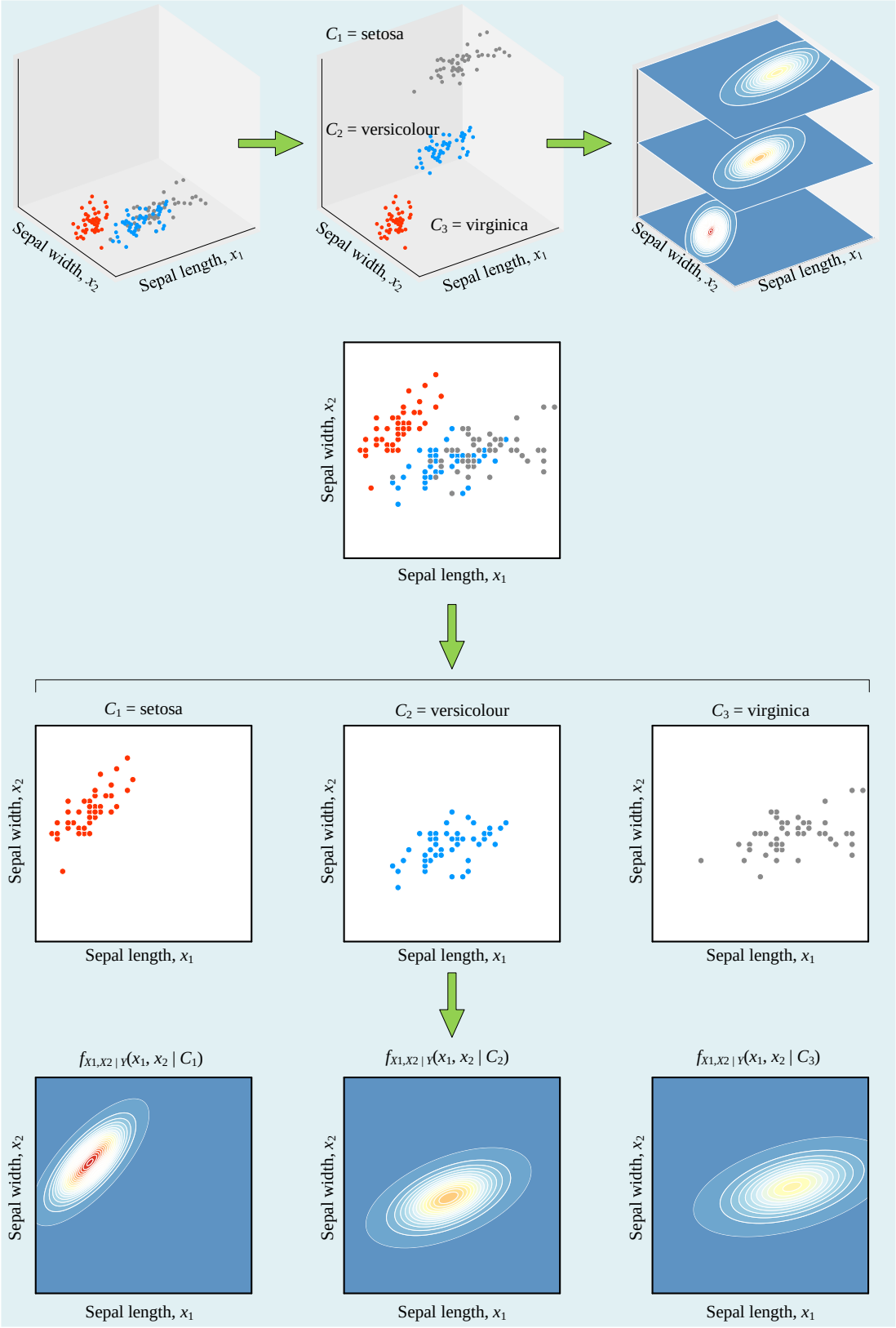


图 1. 似然概率可视化方案，似然概率基于高斯分布

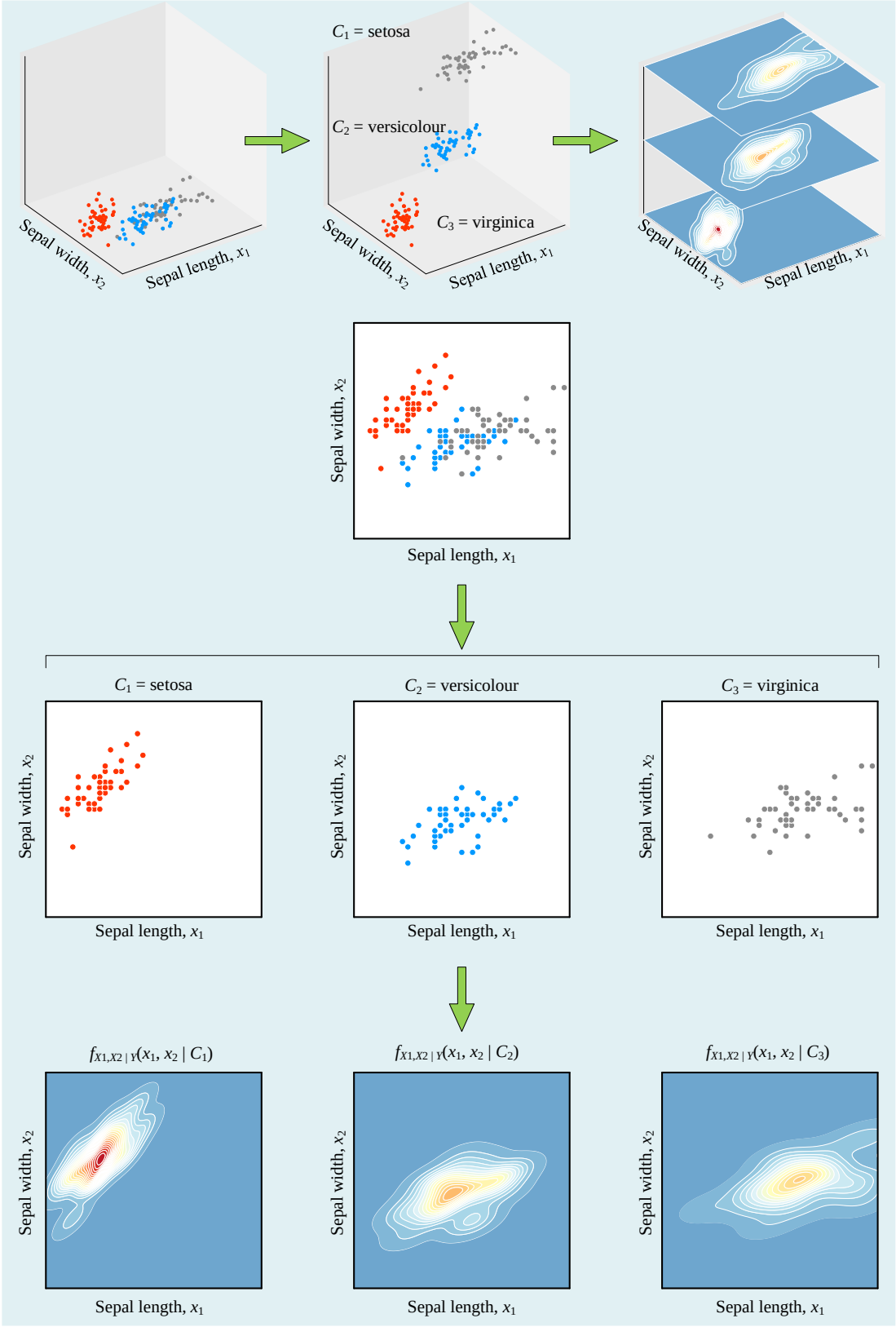


图 2. 似然概率可视化方案，似然概率基于高斯核密度估计

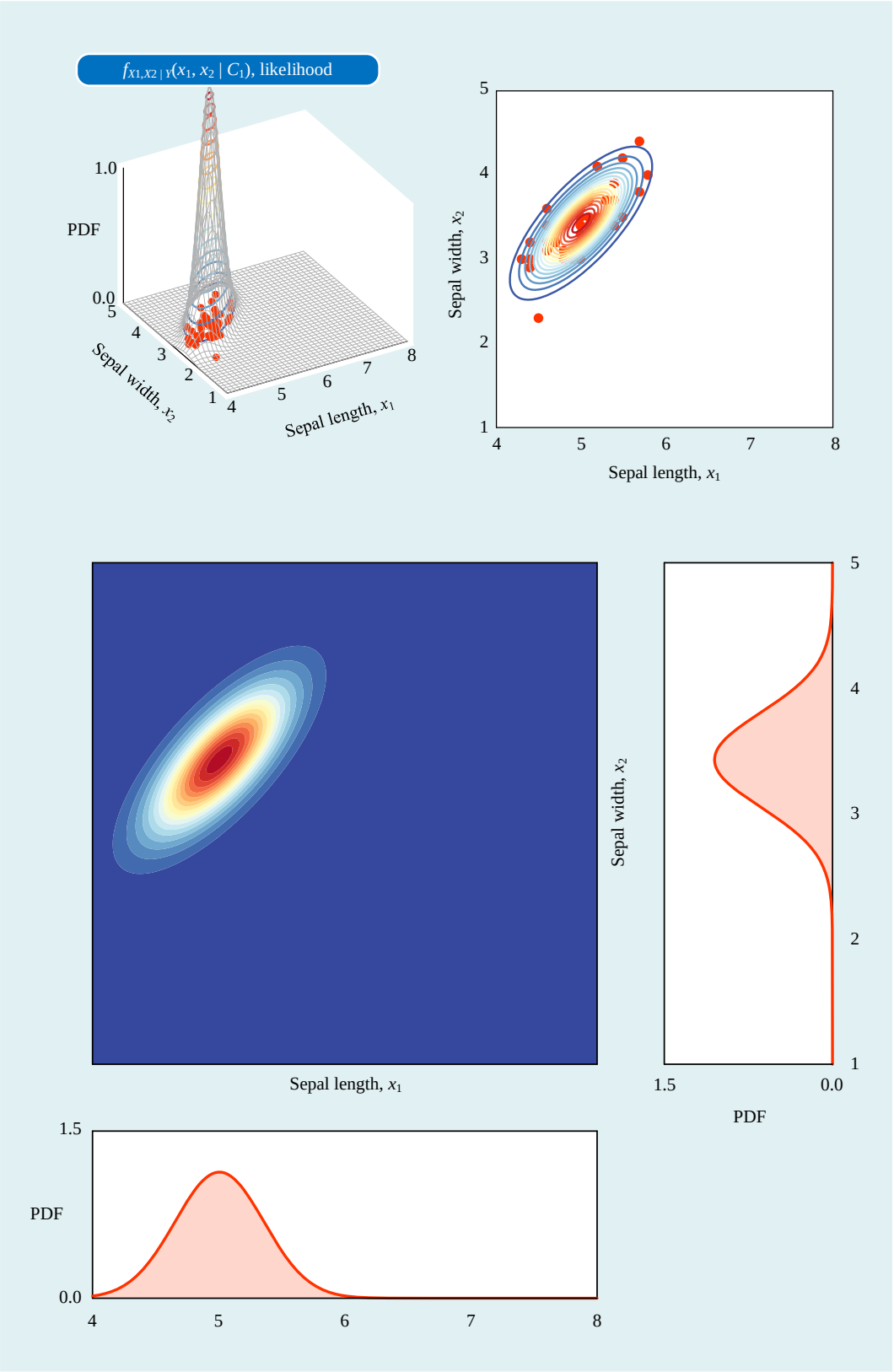


图 3. 似然概率，$f_{X1,X2|Y}(x_1, x_2 \mid C_1)$，似然概率基于高斯分布

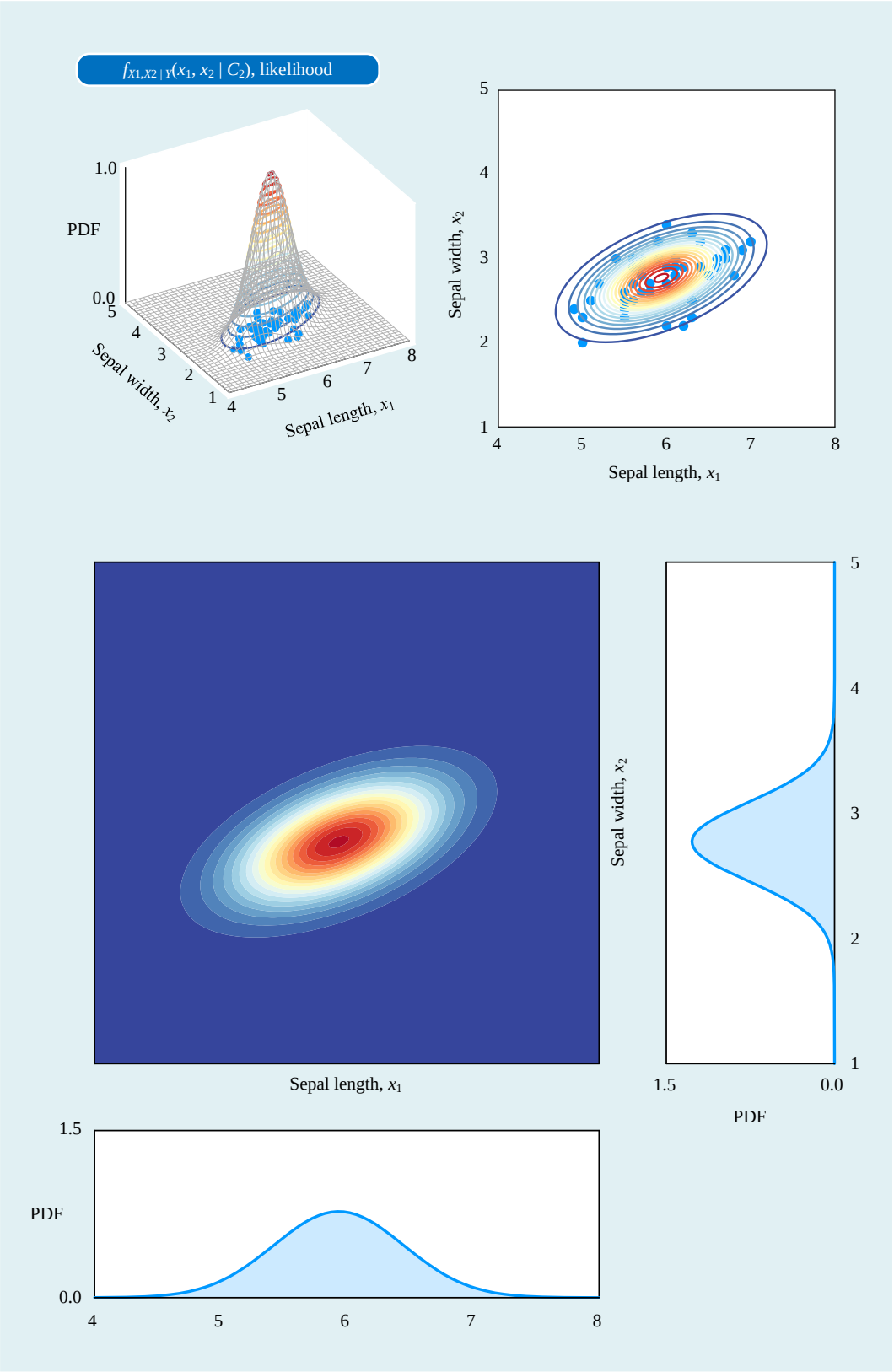


图 4. 似然概率，$f_{X1,X2|Y}(x_1, x_2 \mid C_2)$，似然概率基于高斯分布

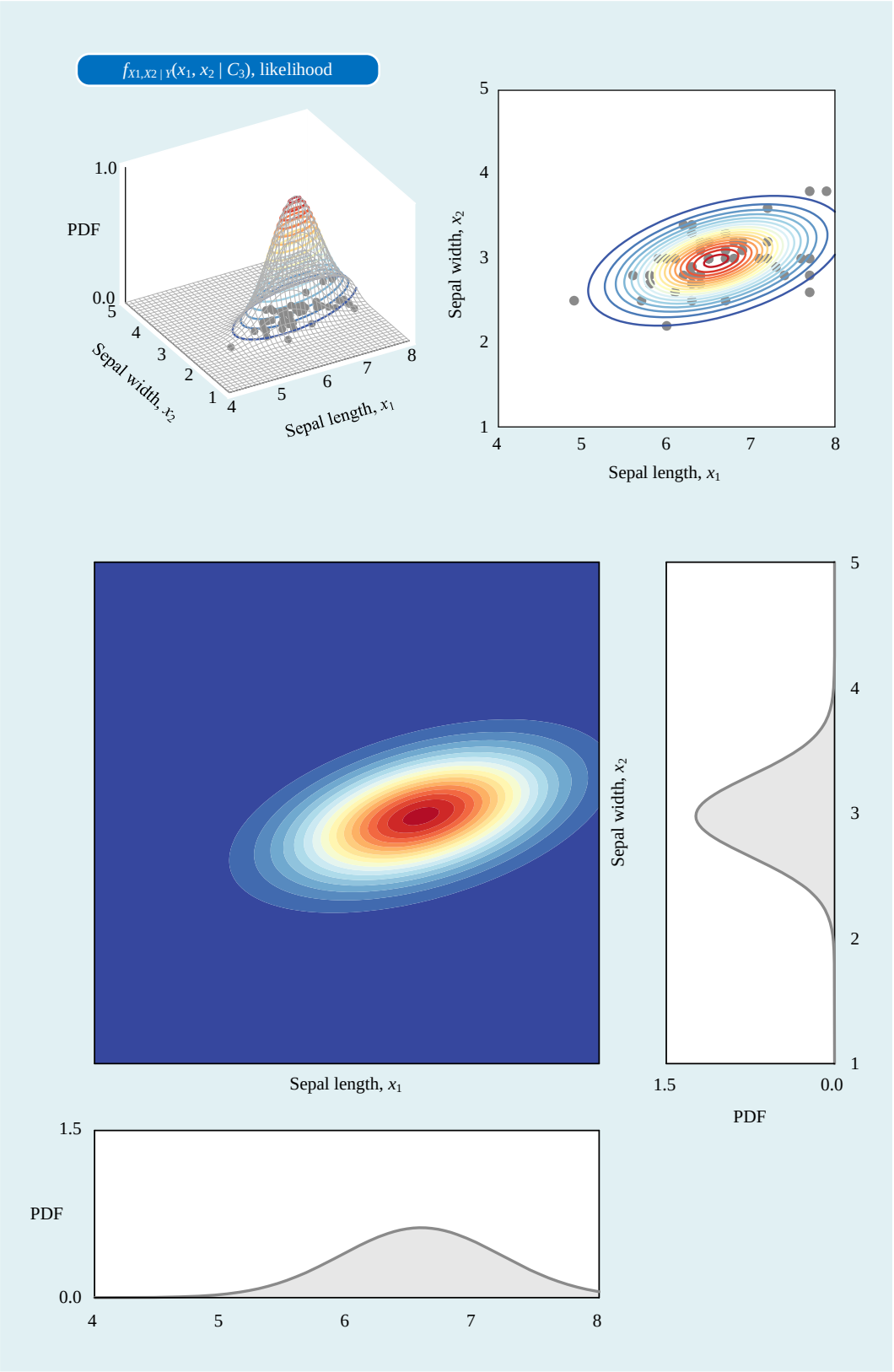


图 5. 似然概率，$f_{X1,X2|Y}(x_1, x_2 \mid C_3)$，似然概率基于高斯分布

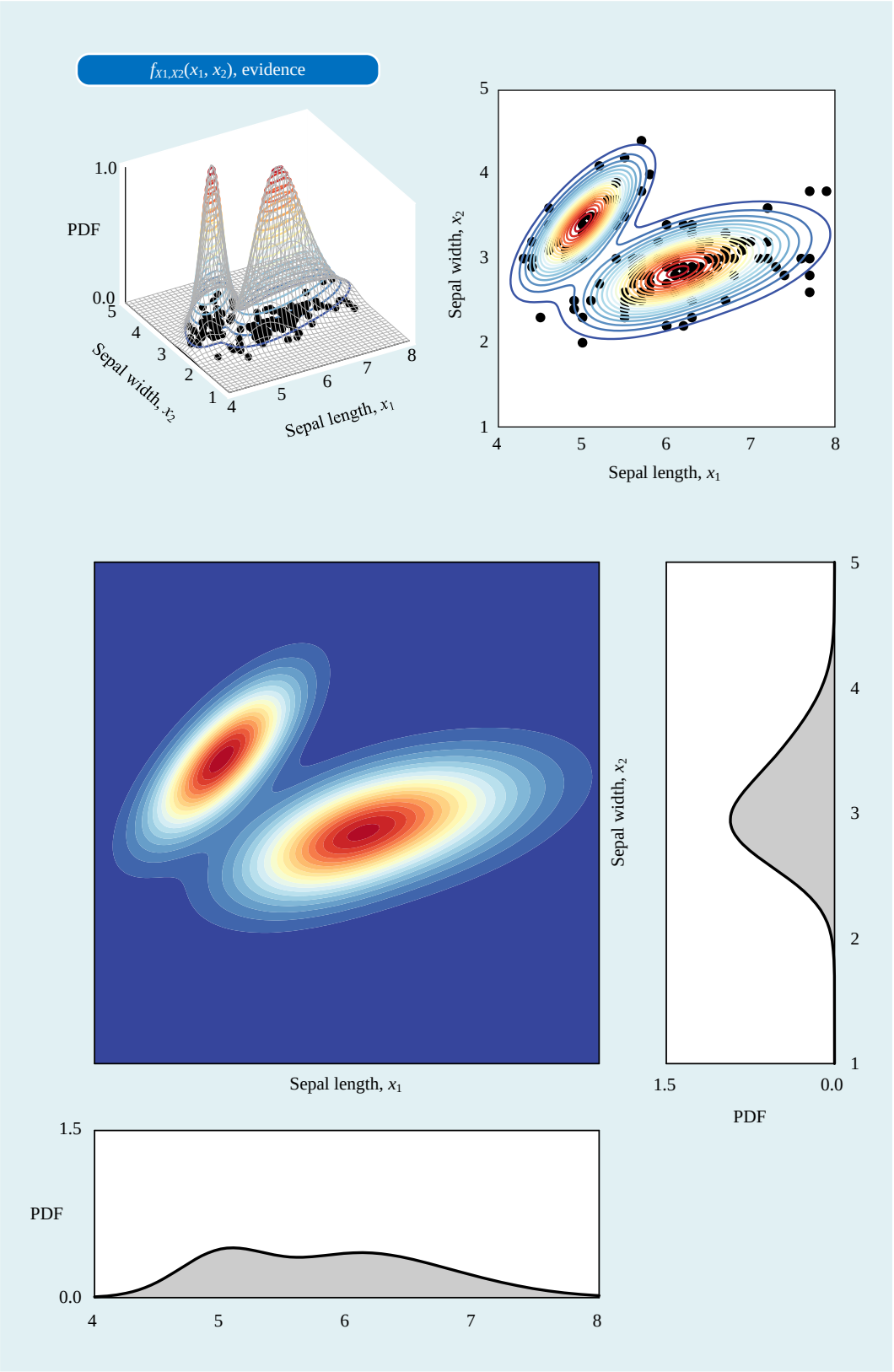


图 6. 证据因子，$f_{X1,X2}(x_1, x_2)$，似然概率基于高斯分布

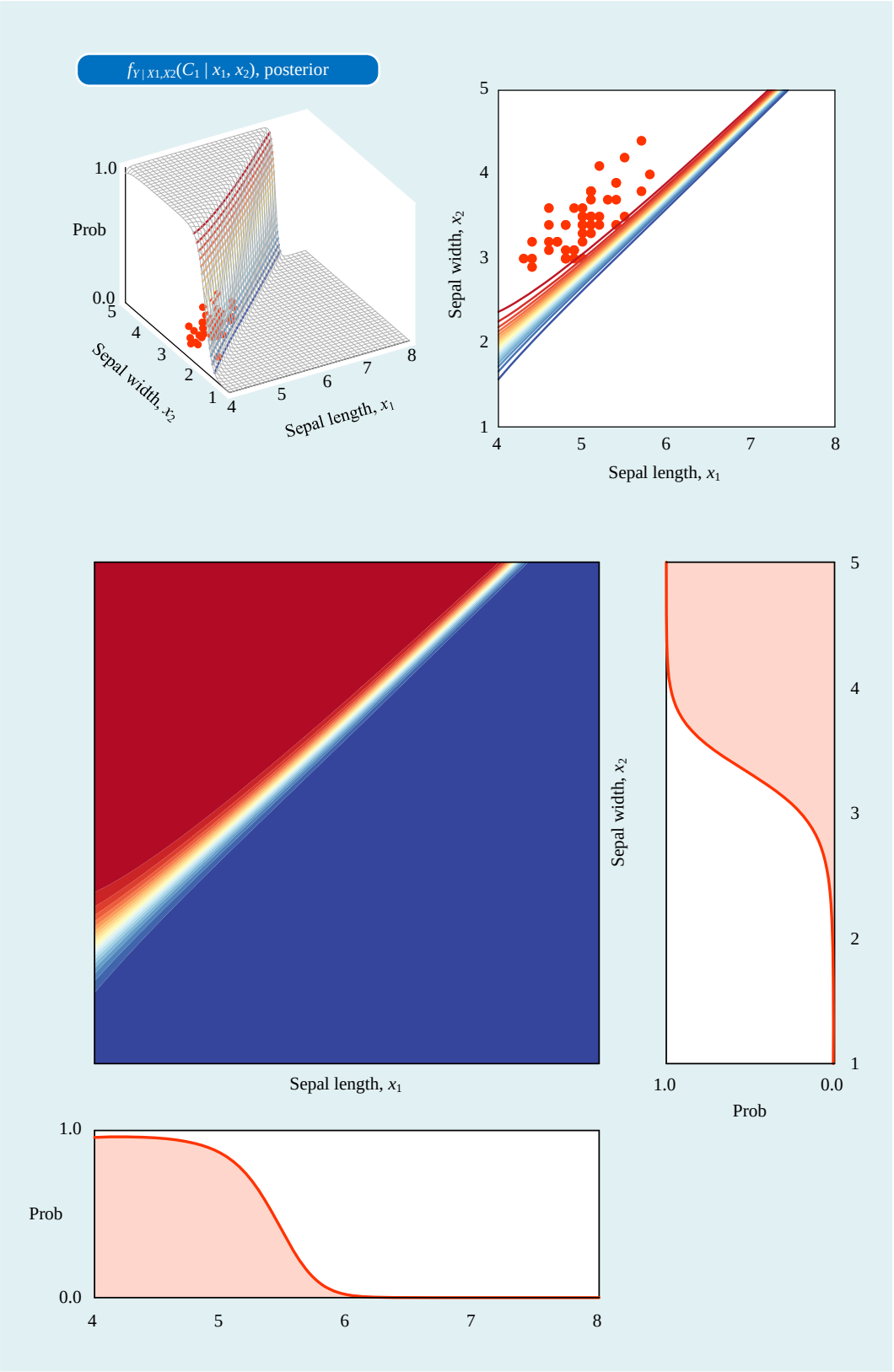


图 7. 后验概率，$f_{Y|X1,X2}(C_1 \mid x_1, x_2)$，似然概率基于高斯分布

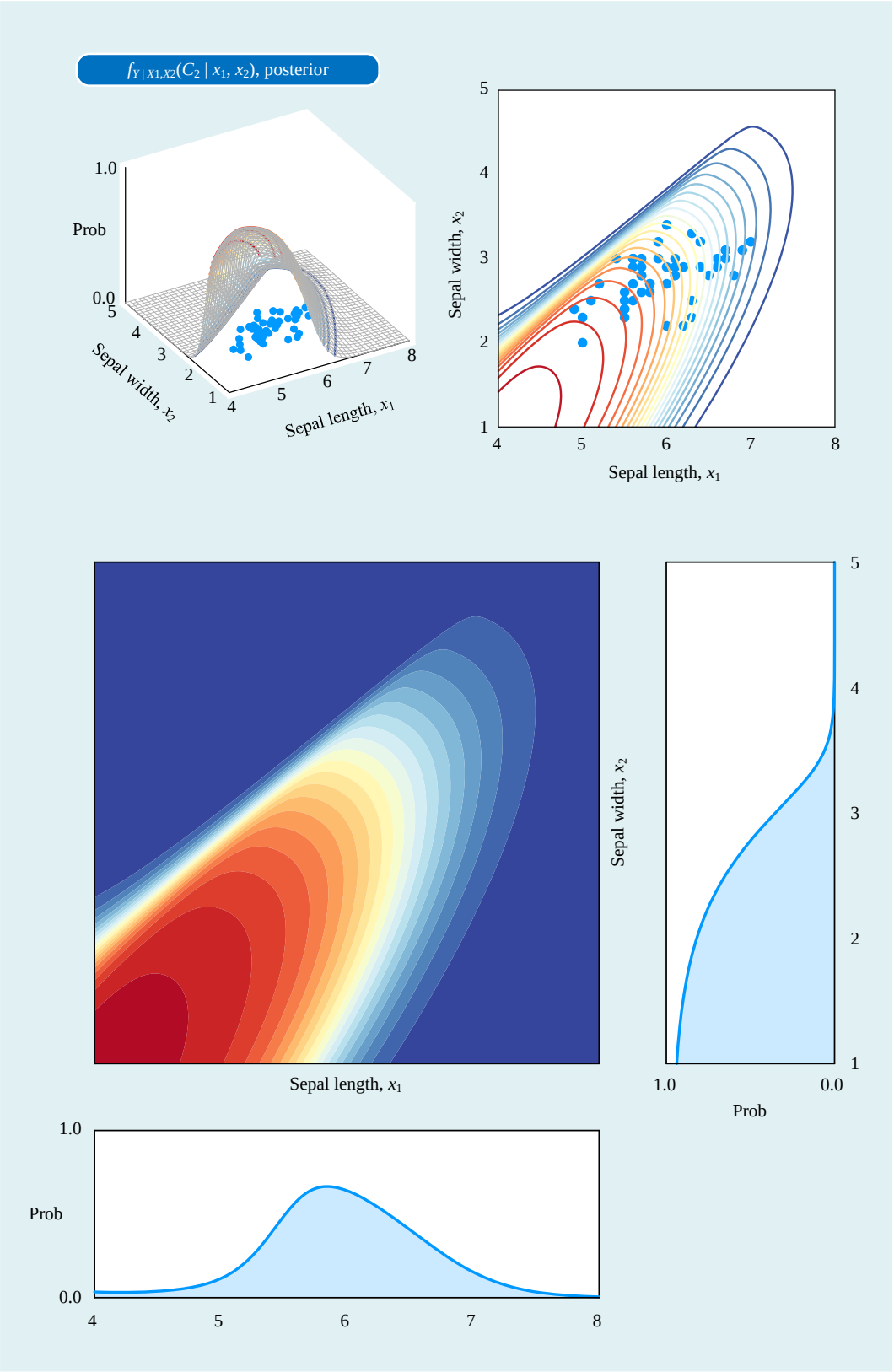


图 8. 后验概率，$f_{Y|X1,X2}(C_2 \mid x_1, x_2)$，似然概率基于高斯分布

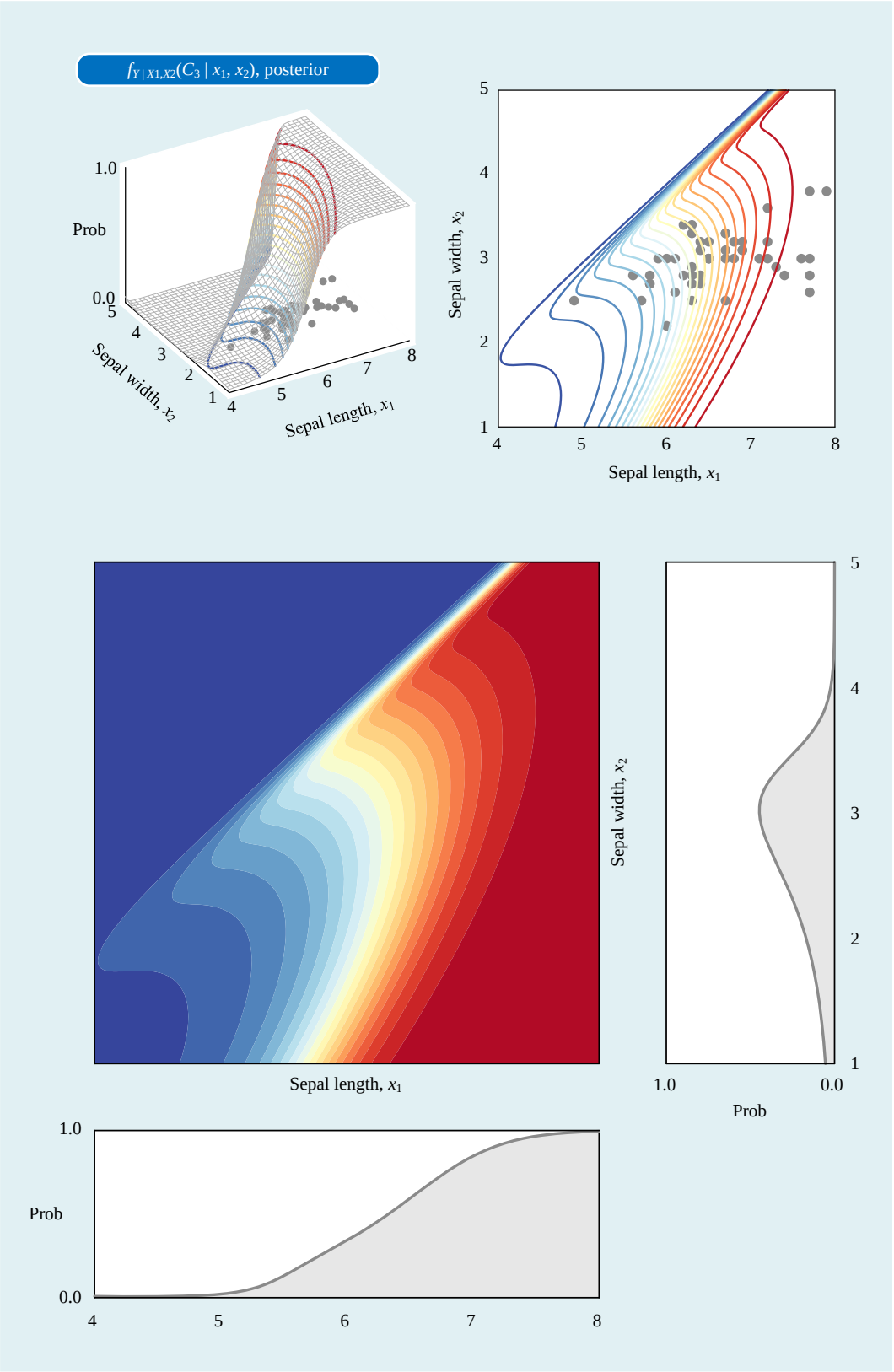


图 9. 后验概率，$f_{Y|X1,X2}(C_3 \mid x_1, x_2)$，似然概率基于高斯分布

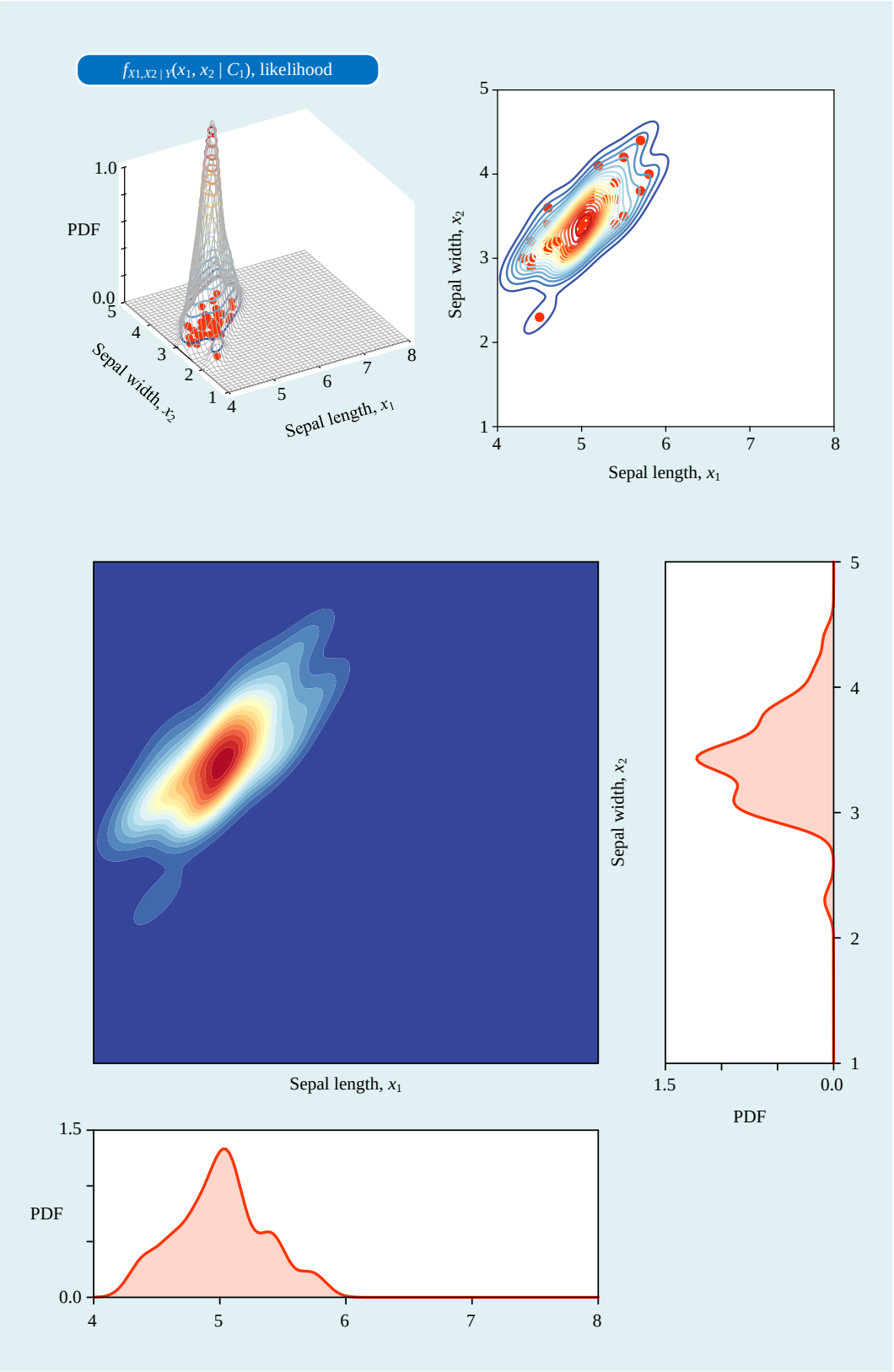


图 10. 似然概率，$f_{X1,X2|Y}(x_1, x_2 \mid C_1)$，似然概率基于高斯核密度估计

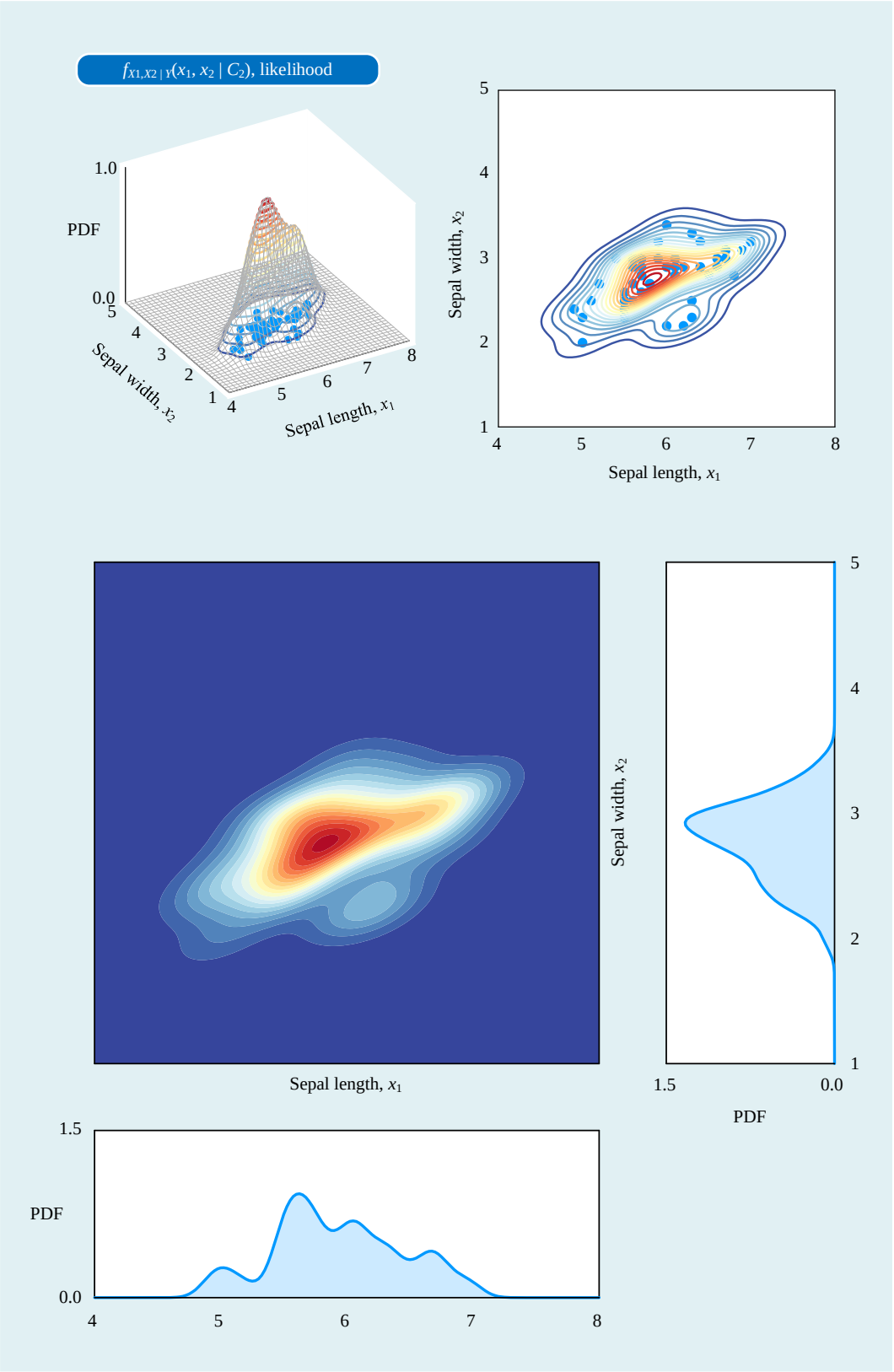


图 11. 似然概率，$f_{X1,X2|Y}(x_1, x_2 \mid C_2)$，似然概率基于高斯核密度估计

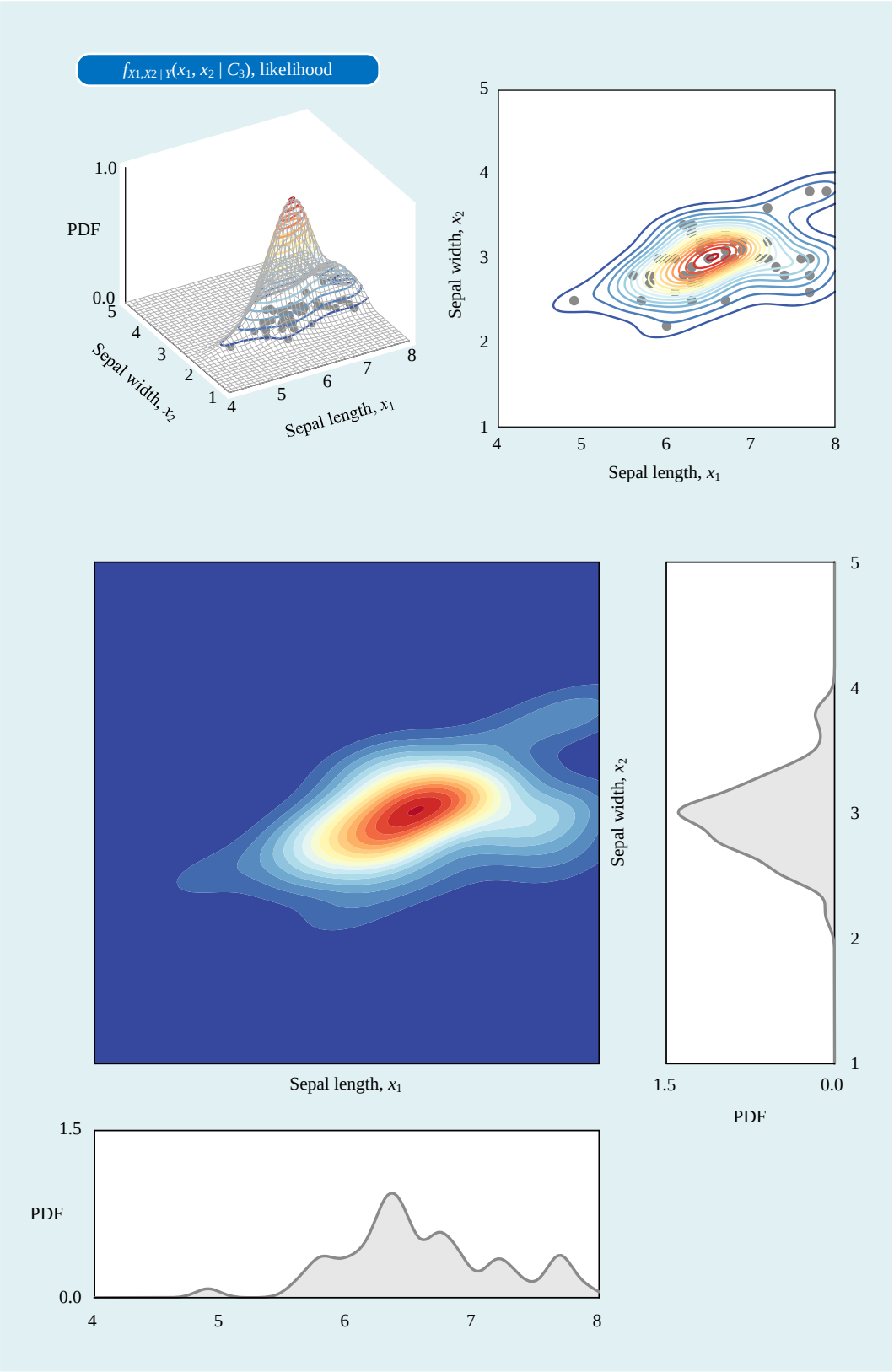


图 12. 似然概率，$f_{X1,X2|Y}(x_1, x_2 \mid C_3)$，似然概率基于高斯核密度估计

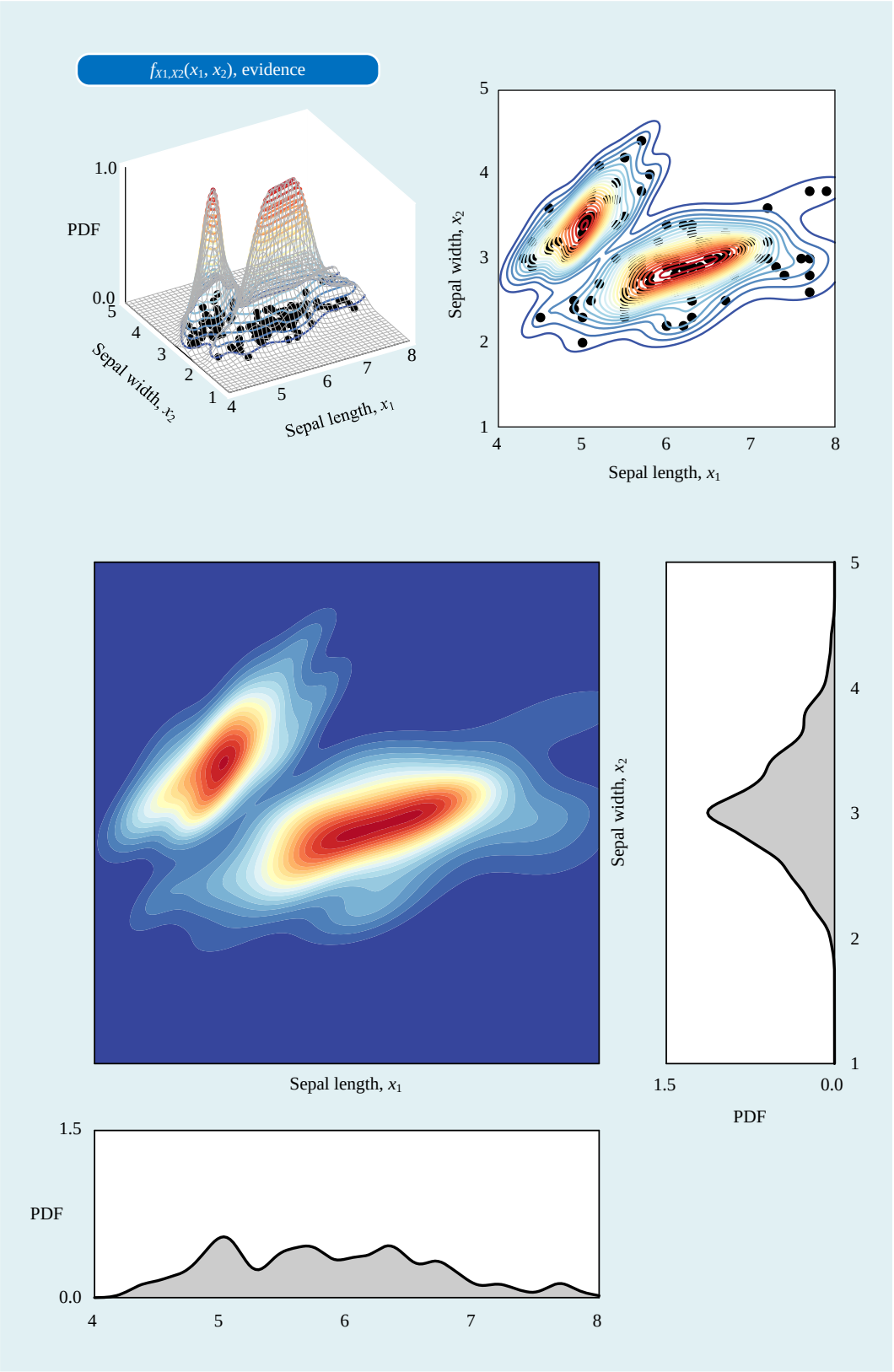


图 13. 证据因子，$f_{X1,X2}(x_1, x_2)$，似然概率基于高斯核密度估计

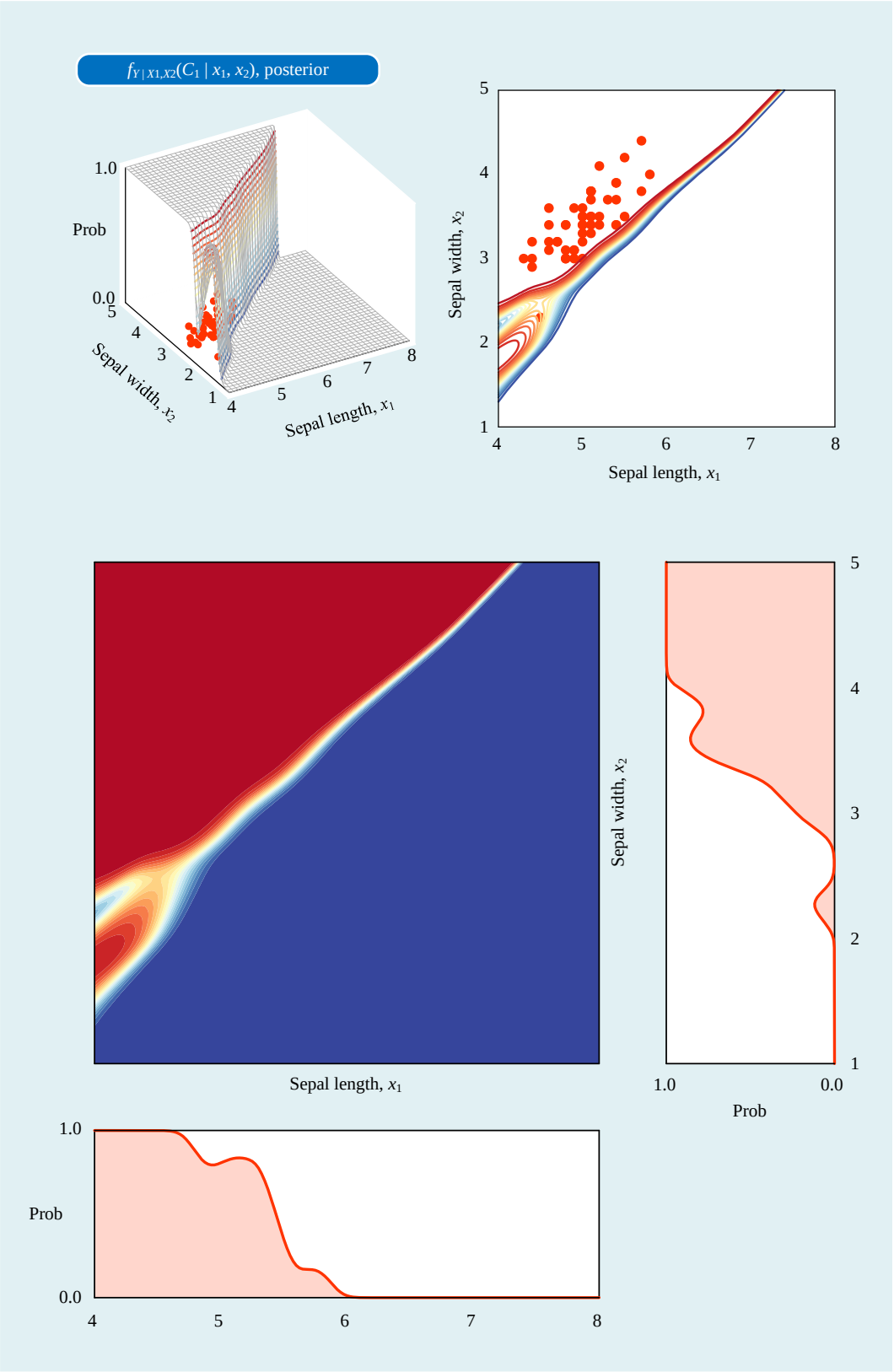


图 14. 后验概率，$f_{Y|X1,X2}(C_1 \mid x_1, x_2)$，似然概率基于高斯核密度估计

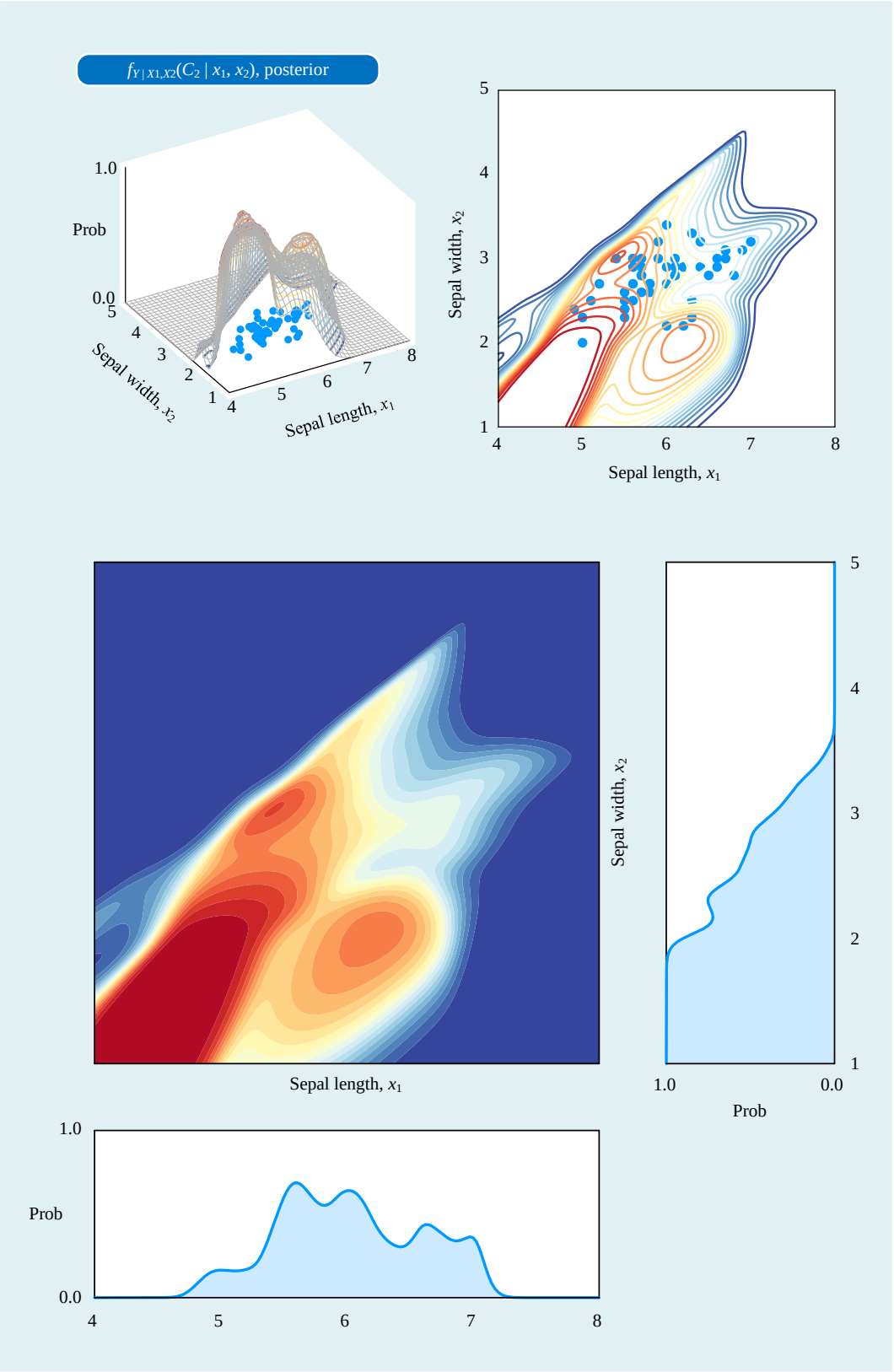


图 15. 后验概率，$f_{Y|X1,X2}(C_2 \mid x_1, x_2)$，似然概率基于高斯核密度估计

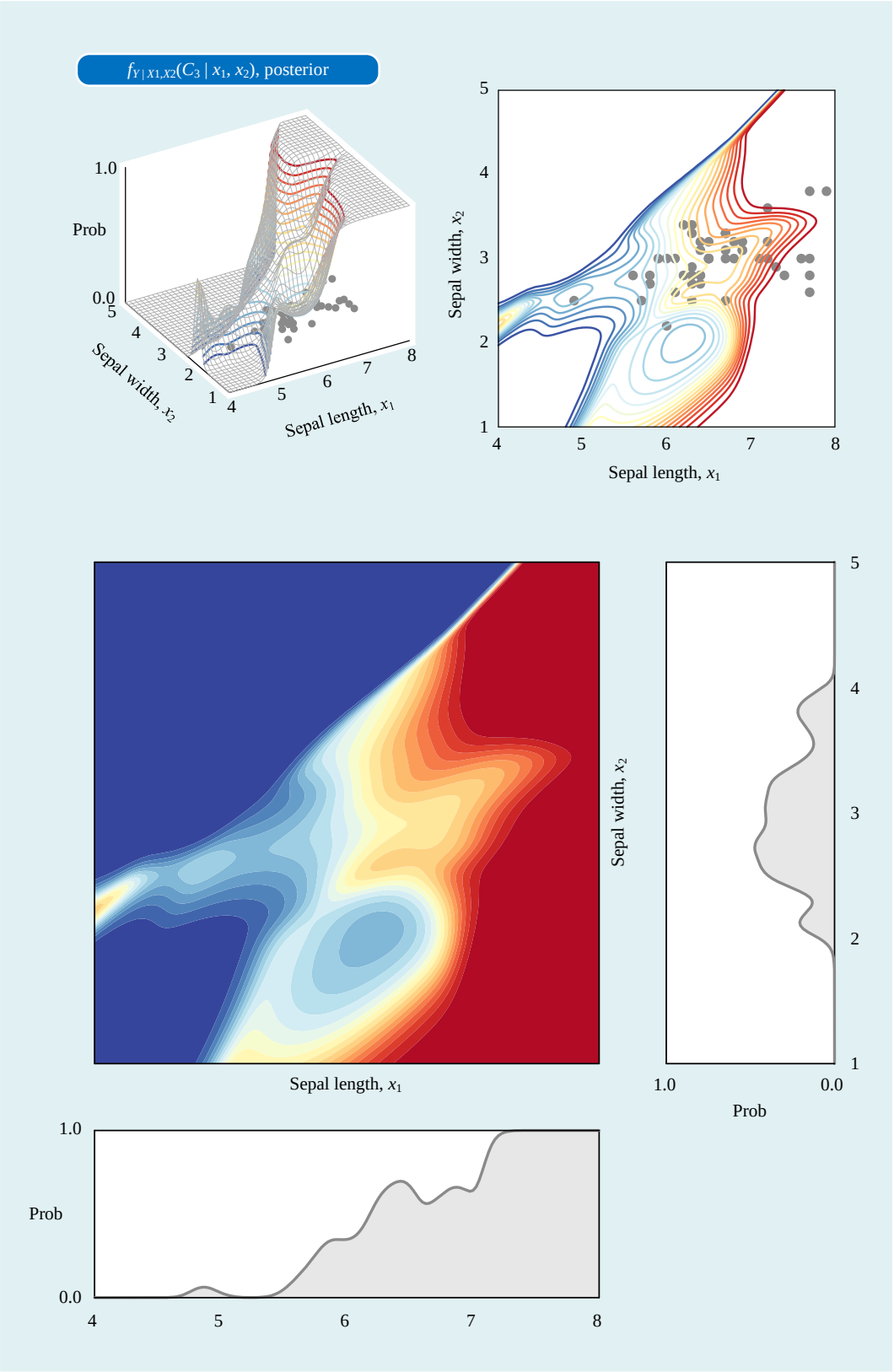


图 16. 后验概率，$f_{Y|X1,X2}(C_3 \mid x_1, x_2)$，似然概率基于高斯核密度估计