如何绘制协方差误差椭圆?

Vincent Spruyt

April 3, 2014

目录

1 序言

在这篇文章中，我将展示如何为 2D 正态分布数据绘制一个误差椭圆，也称为置信椭圆 (confidence ellipse)。误差椭圆表示高斯分布的 iso 轮廓，并允许您可视化 2D 置信区间。图 1 显示了一组 2D 正态分布数据样本的 95% 置信椭圆。该置信椭圆定义了包含 95% 可从基本高斯分布中提取的所有样本的区域。

在下一节中，我们将讨论如何获得不同置信值 (例如 99% 置信区间) 的置信椭圆，并且我们将展示如何使用 Matlab 或 C++ 代码绘制这些椭圆。

2 轴对齐置信椭圆

在推导获得误差椭圆的一般方法之前，让我们先看看椭圆长轴与 X 轴对齐的特殊情况，如图 2 所示：

图 2 说明了椭圆的角度是由数据的协方差决定的。在这种情况下，协方差为零，因此数据不相关，导致轴对齐误差椭圆。

此外，很明显，椭圆轴的大小取决于数据的方差。在我们的例子中，最大的方差在 X 轴方向，而最小的方差在 Y 轴方向。

8.4213	0
0	0.9387

表 1: 图 2 所示数据的协方差矩阵

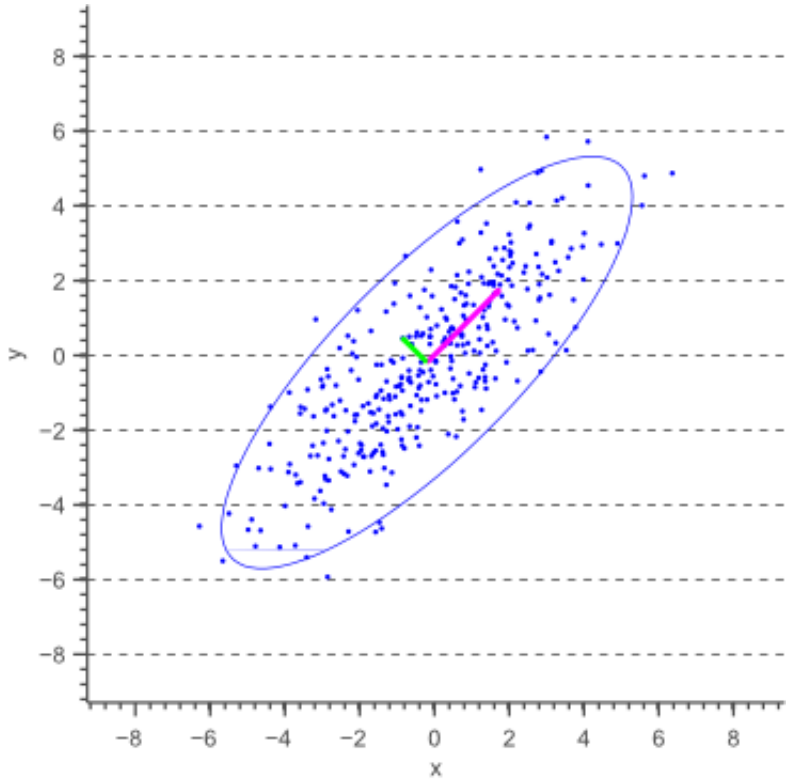

图 1: 正态分布数据的 2D 置信椭圆

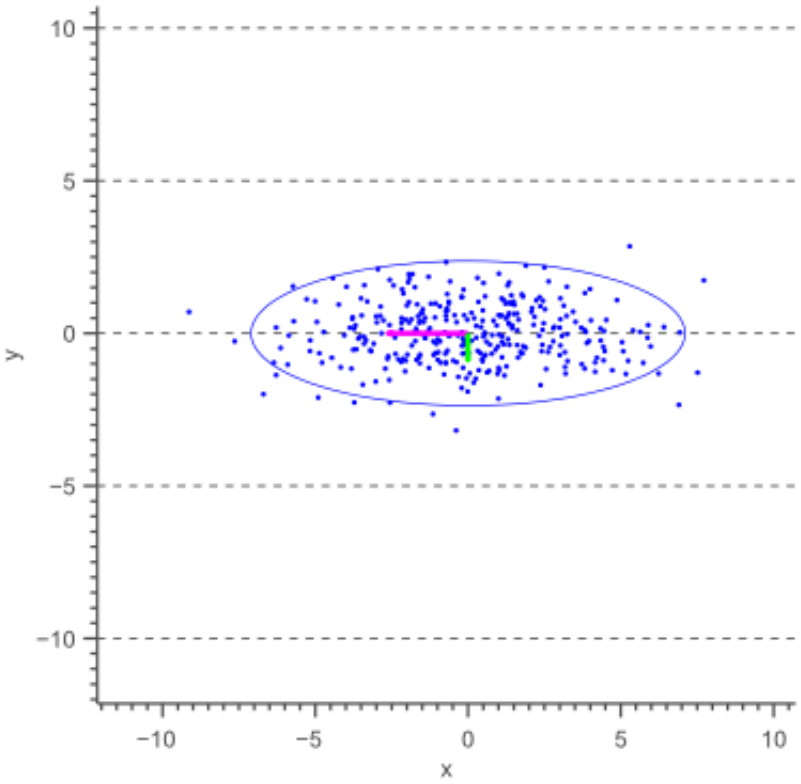

图 2: 不相关的高斯数据的置信椭圆

通常，长轴长度为 $2a$、短轴长度为 $2b$、以原点为中心的轴对齐椭圆方程由以下方程定义：

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1 \tag{1}$$

在我们的例子中，轴的长度由数据的标准偏差 σ_x 和 σ_y 定义，因此误差椭圆的方程变为：

$$\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2 = s \tag{2}$$

其中 s 定义椭圆的比例，可以是任意数字 (例如 $s = 1$)。现在的问题是如何选择 s，使得生成的椭圆的比例表示所选的置信水平 (例如，95% 的置信水平对应于 $s = 5.991$)。

我们的 2D 数据是从一个协方差为零的多元高斯模型中采样的。这意味着 x 值和 y 值也是正态分布的。因此，等式 (2) 的左侧实际上表示独立正态分布数据样本的平方和。已知高斯数据点的平方和根据所谓的卡方分布进行分布。卡方分布定义为“自由度”，表示未知量的数量。在我们的例子中，有 2 个未知数，因此有 2 个自由度。

因此，通过计算卡方似然，我们可以很容易地获得上述总和，因此 s 等于特定值的概率。事实上，由于我们对置信区间感兴趣，我们正在寻找 s 小于或等于特定值的概率，该值可以使用累积卡方分布轻松获得。由于统计学家是懒惰的人，我们通常不尝试计算这种概率，而只是在概率表中查找：`https://people.richland.edu/james/lecture/m170/tbl-chi.html`。

例如，使用此概率表，我们可以很容易地发现，在 2 个自由度的情况下：

$$P(s < 5.991) = 1 - 0.05 = 0.95$$

因此，95% 的置信区间对应于 $s = 5.991$。换句话说，95% 的数据将落在椭圆内，椭圆定义为：

$$\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2 = 5.991 \tag{3}$$

同样，99% 置信区间对应于 $s = 9.210$，90% 置信区间对应于 $s = 4.605$。

因此，图 2 所示的误差椭圆可以绘制为长轴长度等于 $2\sigma_x\sqrt{5.991}$ 且短轴长度等于 $2\sigma_y\sqrt{5.991}$ 的椭圆。

3 任意置信椭圆

在数据不相关的情况下，如存在协方差，则产生的误差椭圆不会与轴对齐。在这种情况下，只有当我们临时定义一个新的坐标系，使椭圆与轴对齐，然后旋转生成的椭圆时，上述段落的推理才成立。

换句话说，虽然我们在前面计算了平行于 x 轴和 y 轴的方差 σ_x 和 σ_y，但我们现在需要计算平行于置信椭圆长轴和短轴的方差。图 1 中的粉红色和绿色箭头说明了需要计算这些差异的方向。

这些方向实际上是数据变化最大的方向，由协方差矩阵定义。协方差矩阵可以看作是对一些原始数据进行线性变换以获得当前观测数据的矩阵。在前一篇关于特征向量和特征值的文章中，我们证明了沿着这种线性变换的方向向量是变换矩阵的特征向量。实际上，图 1 中粉红色和绿色箭头所示的向量是数据协方差矩阵的特征向量，而向量的长度对应于特征值。

因此，特征值表示数据在特征向量方向上的传播。换句话说，特征值表示数据在特征向量方向上的方差。在轴对齐误差椭圆的情况下，即当协方差等于零时，特征值等于协方差矩阵的方差，且

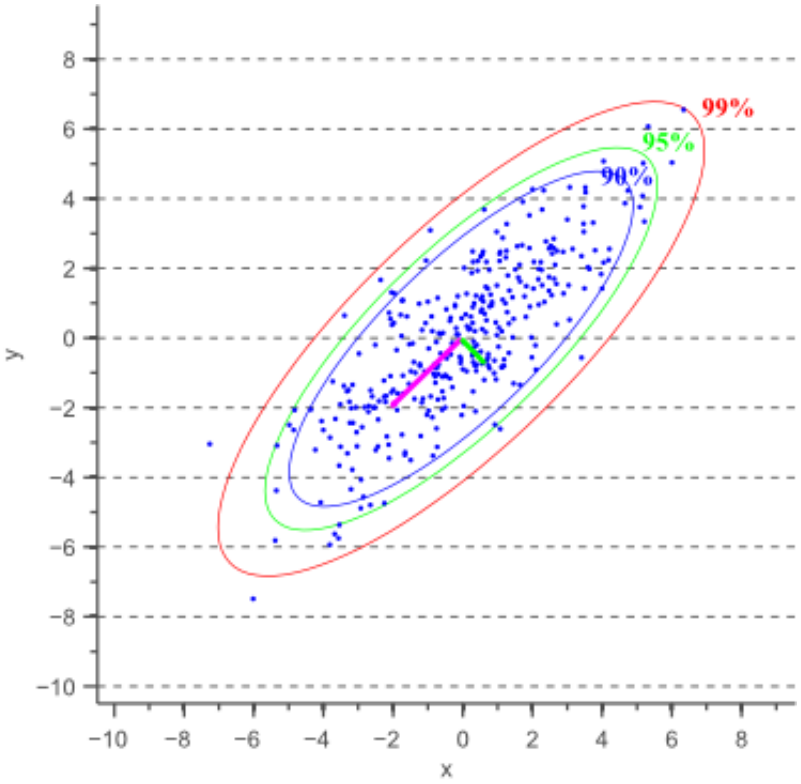

图 3: 正态分布数据的置信椭圆

特征向量等于 x 轴和 y 轴的定义。在任意相关数据的情况下，特征向量表示数据最大扩展的方向，而特征值定义了该扩展的实际大小。

因此，95% 置信椭圆的定义类似于轴对齐情况，长轴长度为 $2\sqrt{5.991\lambda_1}$，短轴长度为 $2\sqrt{5.991\lambda_2}$，其中 λ_1 和 λ_2 表示协方差矩阵的特征值。

为了获得椭圆的方向，我们只需计算最大特征向量相对于 x 轴的角度：

$$\alpha = \arctan \frac{\mathbf{v}_1(y)}{\mathbf{v}_1(x)} \tag{4}$$

其中 $\mathbf{v}_1$ 是对应于最大特征值的协方差矩阵的特征向量。

基于短轴和长轴长度以及长轴和 x 轴之间的角度 α，绘制置信椭圆变得很简单。图 3 显示了几个置信值的误差椭圆。

4 Source Code

- Matlab source code
- C++ source code (uses OpenCV)

5 结论

在这篇文章中，我们展示了如何根据选择的置信值获得 2D 正态分布数据的误差椭圆。这在可视化或分析数据时通常很有用，在以后关于 PCA 的文章中会引起兴趣。

If you’re new to this blog, don’t forget to subscribe, or follow me on twitter!