14 Functions of Random Variables 随机变量的函数

从几何视角探讨随机变量的线性变换

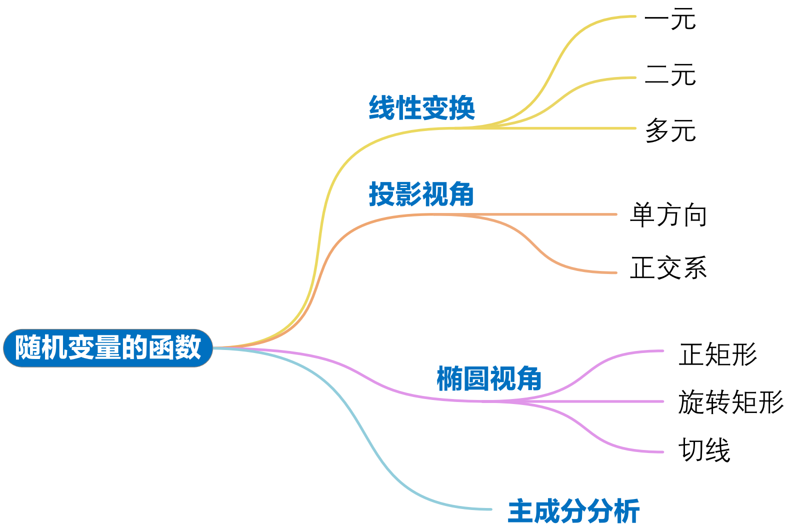
自然的一般规律在大多数情况下不是直接的感知对象。

The general laws of Nature are not, for the most part, immediate objects of perception.

—— 乔治·布尔 (George Boole) | 英格兰数学家和哲学家 | 1815 ~ 1864

- `numpy.cov()` 计算协方差矩阵
- `numpy.linalg.eig()` 特征值分解
- `numpy.linalg.svd()` 奇异值分解
- `sklearn.decomposition.PCA()` 主成分分析函数
- `seaborn.heatmap()` 绘制热图
- `seaborn.kdeplot()` 绘制 KDE 核概率密度估计曲线
- `seaborn.pairplot()` 绘制成对分析图

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

14.1 随机变量的函数：以鸢尾花为例

我们在本书第 3、4 章聊过色子点数的“花式玩法”，比如点数之和、点数平均值、点数之差、点数平方、点数之商等等。这些“花式玩法”都可以叫做随机变量的函数。

随机变量的函数可以分为两类：线性变换 (linear transformation)、非线性变换 (nonlinear transformation)。线性变换是本章的核心内容。

比如，点数之和 $(X_1 + X_2)$、点数之差 $(X_1 - X_2)$、点数平均值 $((X_1 + X_2)/2)$ 等都是线性变换。此外，去均值 $(X_1 - \mathrm{E}(X_1))$、标准化 $((X_1 - \mathrm{E}(X_1))/\mathrm{std}(X_1))$ 也都是常见的随机变量的线性变换。

线性变换之外的随机变量变换都统称为非线性变换，比如平方 (X_1^2)、平方求和 ($\sum_j X_j^2$)、乘积 (X_1X_2)、比例 (X_1/X_2)、倒数 $(1/X_1)$、对数变换 $(\ln X_1)$ 等等。此外，本书第 9 章介绍的经验分布函数 ECDF 也是常用的非线性变换，ECDF 将原始数据转化成 (0, 1) 区间之内的分位值。

注意，经过转换后的随机变量，其分布类型、期望、方差等都可能会发生变化。

从数据角度来看，以上变换又叫数据转化 (data transformation)，这是《数据有道》一册的话题。

以鸢尾花数据为例

鸢尾花数据的前 4 列特征分别为花萼长度 (X_1)、花萼宽度 (X_2)、花瓣长度 (X_3)、花瓣宽度 (X_4)。假如在一个有关鸢尾花的研究中，为了进一步挖掘鸢尾花数据中可能存在的量化关系，我们需要分析如下几个指标：

- 花萼长度去均值，即 $X_1 - \mathrm{E}(X_1)$；
- 花萼宽度去均值，即 $X_2 - \mathrm{E}(X_2)$；
- 花萼长度、宽度之和，即 $X_1 + X_2$；
- 花萼长度、宽度之差，即 $X_1 - X_2$；
- 花萼长度、宽度乘积，即 X_1X_2；
- 花萼长度、宽度比例，即 X_1/X_2。

图 1 所示为经过上述转换后得到的鸢尾花新特征之间的成对特征散点图。这些新特征之间的成对关系中，有些展现出明显的线性关系，有些特征更方便判别鸢尾花分类，有些特征展现出更好的“正态性”，有些则更容易发现“离群值”。

请大家利用成对特征图分析更多鸢尾花特征的随机变量函数。此外，请大家依照同样的方法分析花瓣长度、宽度数据，并且交叉分析花萼、花瓣量化关系。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

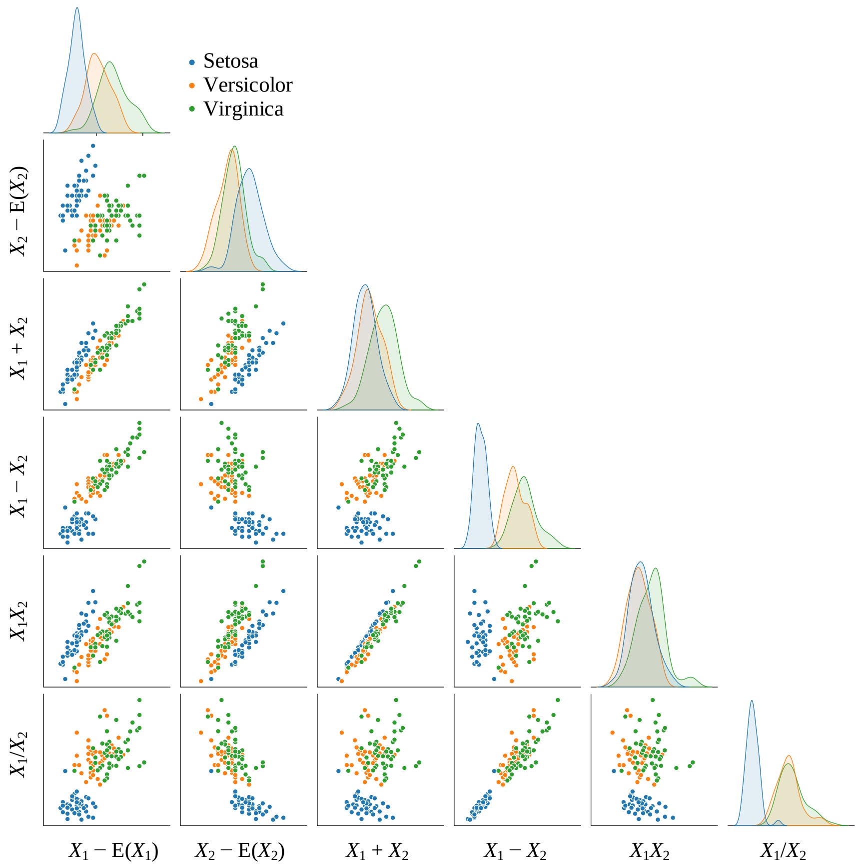


图 1. 鸢尾花花萼长度、宽度特征完成转换后的成对特征散点图

14.2 线性变换：投影视角

《矩阵力量》第 25 章介绍过随机变量的线性变换，我们部分内容"抄"过来。本章后文会用鸢尾花数据展开讲解。

一元随机变量

如果 X 为一个随机变量，对 X 进行函数变换，可以得到其他的随机变量 Y:

$$Y = h(X) \tag{1}$$

特别地，如果 $h()$ 为线性函数，则 X 到 Y 进行的就是线性变换，比如：

$$Y = h(X) = aX + b \tag{2}$$

其中，a 和 b 为常数。这相当于几何中的缩放、平移两步操作。在线性代数中，上式相当于仿射变换 (affine transformation)。展开来说，在线性代数中，仿射变换是指一类在二维或三维欧几里得

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

空间中的变换，可以描述为一种线性变换和一个平移向量的组合。与仿射变换不同，线性变换仅由一个矩阵表示，它可以用来缩放、旋转、镜像、剪切一个图形，但不能进行平移操作。

(2) 中，Y 的期望和 X 的期望之间关系：

$$\mathrm{E}(Y)=a\,\mathrm{E}(X)+b \tag{3}$$

(2) 中，Y 和 X 方差之间关系：

$$\operatorname{var}(Y)=\operatorname{var}(aX+b)=a^2\operatorname{var}(X) \tag{4}$$

二元随机变量

如果 Y 和二元随机变量 (X_1, X_2) 存在如下关系：

$$Y=aX_1+bX_2 \tag{5}$$

(5) 可以写成：

$$Y=\begin{bmatrix}a & b\end{bmatrix}\begin{bmatrix}X_1\\X_2\end{bmatrix} \tag{6}$$

Y 和二元随机变量 (X_1, X_2) 期望之间存在如下关系：

$$\mathrm{E}(Y)=\mathrm{E}(aX_1+bX_2)=a\,\mathrm{E}(X_1)+b\,\mathrm{E}(X_2) \tag{7}$$

(7) 可以写成如下矩阵运算形式：

$$\mathrm{E}(Y)=\begin{bmatrix}a & b\end{bmatrix}\begin{bmatrix}\mathrm{E}(X_1)\\\mathrm{E}(X_2)\end{bmatrix} \tag{8}$$

Y 和二元随机变量 (X_1, X_2) 方差、协方差存在如下关系：

$$\operatorname{var}(Y)=\operatorname{var}(aX_1+bX_2)=a^2\operatorname{var}(X_1)+b^2\operatorname{var}(X_2)+2ab\operatorname{cov}(X_1,X_2) \tag{9}$$

(9) 可以写成：

$$\operatorname{var}(Y)=\begin{bmatrix}a & b\end{bmatrix}\underbrace{\begin{bmatrix}\operatorname{var}(X_1) & \operatorname{cov}(X_1,X_2)\\\operatorname{cov}(X_1,X_2) & \operatorname{var}(X_2)\end{bmatrix}}_{\boldsymbol{\Sigma}}\begin{bmatrix}a\\b\end{bmatrix} \tag{10}$$

相信大家已经在上式中看到了如下协方差矩阵：

$$\boldsymbol{\Sigma}=\begin{bmatrix}\operatorname{var}(X_1) & \operatorname{cov}(X_1,X_2)\\\operatorname{cov}(X_1,X_2) & \operatorname{var}(X_2)\end{bmatrix} \tag{11}$$

也就是说，(10) 可以写成：

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$\operatorname{var}(Y)=\begin{bmatrix}a & b\end{bmatrix}\boldsymbol{\Sigma}\begin{bmatrix}a\\ b\end{bmatrix} \tag{12}$$

D 维随机变量：朝单一方向投影

如果随机向量 $\boldsymbol{\chi} = [X_1, X_2, \ldots, X_D]^{\mathrm{T}}$ 服从 $N(\boldsymbol{\mu}_{\chi}, \boldsymbol{\Sigma}_{\chi})$，$\boldsymbol{\chi}$ 在单位向量 $\boldsymbol{v}$ 方向上投影得到 Y:

$$Y=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\chi}=\boldsymbol{v}^{\mathrm{T}}\begin{bmatrix}X_1\\ X_2\\ \vdots\\ X_D\end{bmatrix} \tag{13}$$

Y 的期望 $\mathrm{E}(Y)$ 为:

$$\mathrm{E}(Y)=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\mu}_{\chi}=\boldsymbol{v}^{\mathrm{T}}\begin{bmatrix}\mathrm{E}(X_1)\\ \mathrm{E}(X_2)\\ \vdots\\ \mathrm{E}(X_D)\end{bmatrix} \tag{14}$$

Y 的方差 $\mathrm{E}(Y)$ 为:

$$\operatorname{var}(Y)=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}_{\chi}\boldsymbol{v} \tag{15}$$

D 维随机变量：朝正交系投影

$\boldsymbol{\chi} = [X_1, X_2, \ldots, X_D]^{\mathrm{T}}$ 服从 $N(\boldsymbol{\mu}_{\chi}, \boldsymbol{\Sigma}_{\chi})$，$\boldsymbol{\chi}$ 在规范正交系 $\boldsymbol{V}$ 投影得到 $\boldsymbol{\gamma} = [Y_1, Y_2, \ldots, Y_D]^{\mathrm{T}}$:

$$\boldsymbol{\gamma}=\begin{bmatrix}Y_1\\ Y_2\\ \vdots\\ Y_D\end{bmatrix}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\chi}=\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\\ \boldsymbol{v}_2^{\mathrm{T}}\\ \vdots\\ \boldsymbol{v}_D^{\mathrm{T}}\end{bmatrix}\boldsymbol{\chi}=\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\chi}\\ \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\chi}\\ \vdots\\ \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\chi}\end{bmatrix} \tag{16}$$

$\boldsymbol{\gamma}$ 的期望 (质心) $\mathrm{E}(\boldsymbol{\gamma})$ 为:

$$\mathrm{E}(\boldsymbol{\gamma})=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\mu}_{\chi}=\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\\ \boldsymbol{v}_2^{\mathrm{T}}\\ \vdots\\ \boldsymbol{v}_D^{\mathrm{T}}\end{bmatrix}\boldsymbol{\mu}_{\chi}=\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\mu}_{\chi}\\ \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\mu}_{\chi}\\ \vdots\\ \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\mu}_{\chi}\end{bmatrix} \tag{17}$$

$\boldsymbol{\gamma}$ 的协方差矩阵 $\operatorname{var}(\boldsymbol{\gamma})$ 为:

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$\operatorname{var}(\gamma) = V^{\mathrm{T}} \Sigma_{\chi} V = \begin{bmatrix} v_1^{\mathrm{T}} \\ v_2^{\mathrm{T}} \\ \vdots \\ v_D^{\mathrm{T}} \end{bmatrix} \Sigma_{\chi} \begin{bmatrix} v_1 & v_2 & \cdots & v_D \end{bmatrix} = \begin{bmatrix} v_1^{\mathrm{T}} \Sigma_{\chi} v_1 & v_1^{\mathrm{T}} \Sigma_{\chi} v_2 & \cdots & v_1^{\mathrm{T}} \Sigma_{\chi} v_D \\ v_2^{\mathrm{T}} \Sigma_{\chi} v_1 & v_2^{\mathrm{T}} \Sigma_{\chi} v_2 & \cdots & v_2^{\mathrm{T}} \Sigma_{\chi} v_D \\ \vdots & \vdots & \ddots & \vdots \\ v_D^{\mathrm{T}} \Sigma_{\chi} v_1 & v_D^{\mathrm{T}} \Sigma_{\chi} v_2 & \cdots & v_D^{\mathrm{T}} \Sigma_{\chi} v_D \end{bmatrix} \tag{18}$$

上式还告诉我们，$v_i^{\mathrm{T}}\chi$ 和 $v_j^{\mathrm{T}}\chi$ 的协方差为：

$$\operatorname{cov}\left(v_i^{\mathrm{T}}\chi, v_j^{\mathrm{T}}\chi\right) = v_i^{\mathrm{T}} \Sigma_{\chi} v_j \tag{19}$$

14.3 单方向投影：鸢尾花两特征为例

本节以鸢尾花数据花萼长度、花萼宽度两特征为例讲解线性变换。我们首先看两个最简单的例子，将数据分别投影到横轴、纵轴。然后再看更一般的情况。

投影到 x 轴

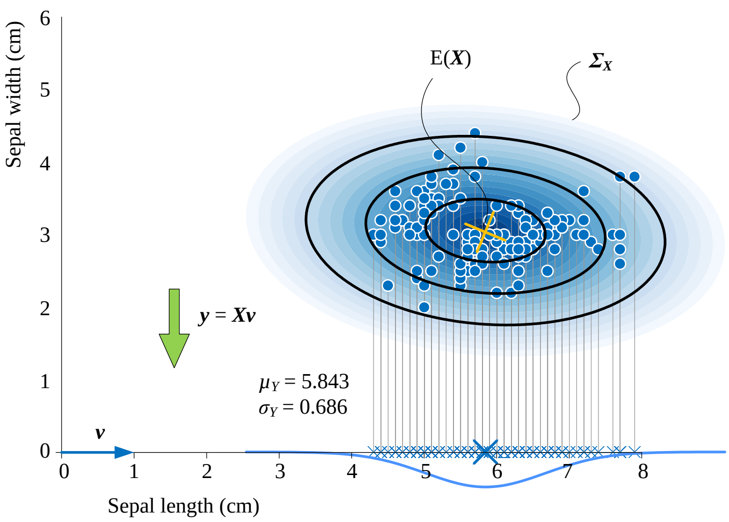
鸢尾花数据矩阵为 $X = [x_1, x_2]$，对应随机变量为 $\chi = [X_1, X_2]^{\mathrm{T}}$。

如图 2 所示，将 X 投影到横轴，即：

$$y = Xv = X\begin{bmatrix} 1 \\ 0 \end{bmatrix} = x_1 \tag{20}$$

从随机变量角度来看上述运算，即：

$$Y = v^{\mathrm{T}}\chi = \begin{bmatrix} 1 \\ 0 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = X_1 \tag{21}$$



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 2. 逆时针 0 度，$\boldsymbol{X}$ 向 $\boldsymbol{v}$ 投影

$\boldsymbol{X}$ 的质心为：

$$\mathrm{E}(\boldsymbol{X})=\begin{bmatrix}5.8433 & 3.0573\end{bmatrix} \tag{22}$$

由此计算得到图 2 中 $\boldsymbol{y}$ 的质心为：

$$\mathrm{E}(\boldsymbol{y})=\mathrm{E}(\boldsymbol{X})\boldsymbol{v}=\begin{bmatrix}5.8433 & 3.0573\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}=5.8433 \tag{23}$$

$\boldsymbol{X}$ 的协方差矩阵为：

$$\boldsymbol{\Sigma}_{\boldsymbol{X}}=\begin{bmatrix}0.6856 & -0.0424\\-0.0424 & 0.1899\end{bmatrix} \tag{24}$$

由此计算得到图 2 中 $\boldsymbol{y}$ 的方差为：

$$\operatorname{var}(\boldsymbol{y})=\boldsymbol{v}^{\mathrm{T}}\operatorname{var}(\boldsymbol{X})\boldsymbol{v}=\begin{bmatrix}1\\0\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}0.6856 & -0.0424\\-0.0424 & 0.1899\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}=0.6856 \tag{25}$$

注意，图 2 中椭圆代表马氏距离。三个黑色旋转椭圆分别代表马氏距离为 1、2、3。将图 2 中三个椭圆也投影到横轴上，大家会发现得到的三条线段分别代表 $\mu_1\pm\sigma_1$、$\mu_1\pm2\sigma_1$、$\mu_1\pm3\sigma_1$。这绝不是几何上的巧合，本章后续会展开讲解。

投影到 $\boldsymbol{y}$ 轴

如图 3 所示，将 $\boldsymbol{X}$ 投影到纵轴，即：

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{v}=\boldsymbol{X}\begin{bmatrix}0\\1\end{bmatrix}=\boldsymbol{x}_2 \tag{26}$$

从随机变量角度来看上述运算，即：

$$Y=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\chi}=\begin{bmatrix}1\\1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}X_1\\X_2\end{bmatrix}=X_2 \tag{27}$$

计算图 3 中 $\boldsymbol{y}$ 的质心为：

$$\mathrm{E}(\boldsymbol{y})=\mathrm{E}(\boldsymbol{X})\boldsymbol{v}=\begin{bmatrix}5.8433 & 3.0573\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix}=3.0573 \tag{28}$$

计算得到图 3 中 $\boldsymbol{y}$ 的方差为：

$$\operatorname{var}(\boldsymbol{y})=\boldsymbol{v}^{\mathrm{T}}\operatorname{var}(\boldsymbol{X})\boldsymbol{v}=\begin{bmatrix}0\\1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}0.6856 & -0.0424\\-0.0424 & 0.1899\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix}=0.1899 \tag{29}$$

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

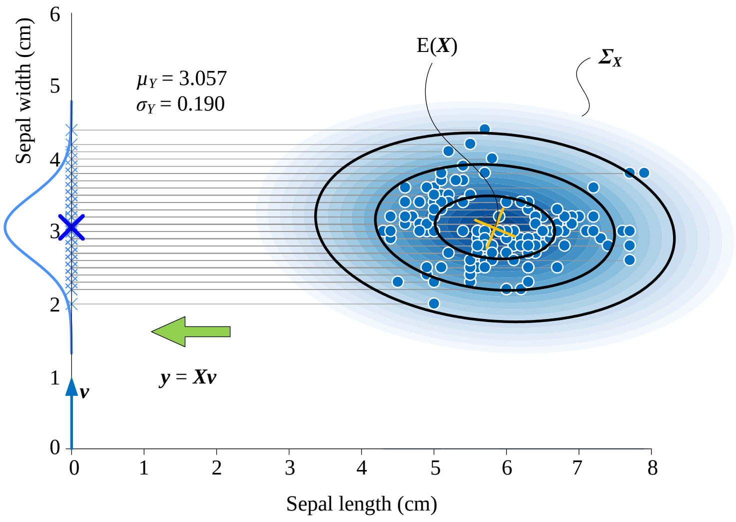


图 3. 逆时针 90 度，$\boldsymbol{X}$ 向 $\boldsymbol{v}$ 投影

其他情况

图 4 ~ 图 7 所示为其他四个投影场景，请大家自己分析。

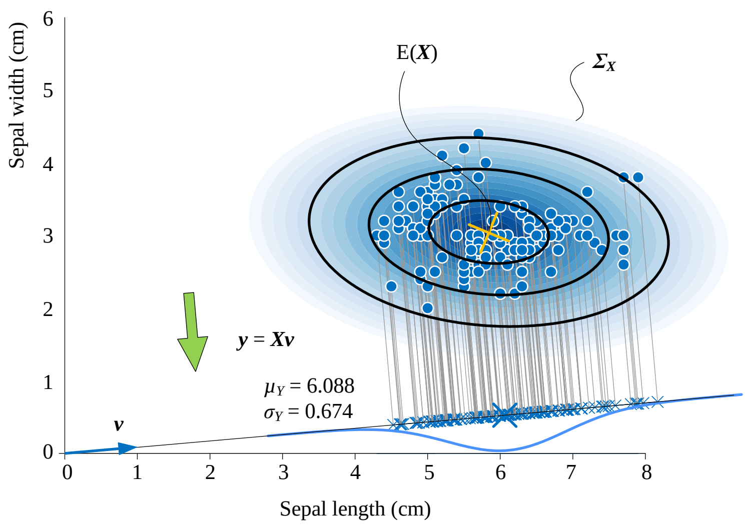


图 4. 逆时针 5 度，$\boldsymbol{X}$ 向 $\boldsymbol{v}$ 投影

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

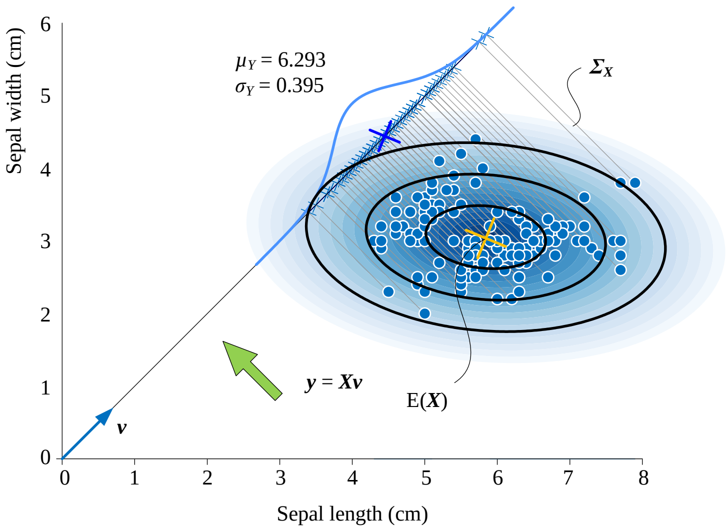


图 5. 逆时针 45 度，X 向 v 投影

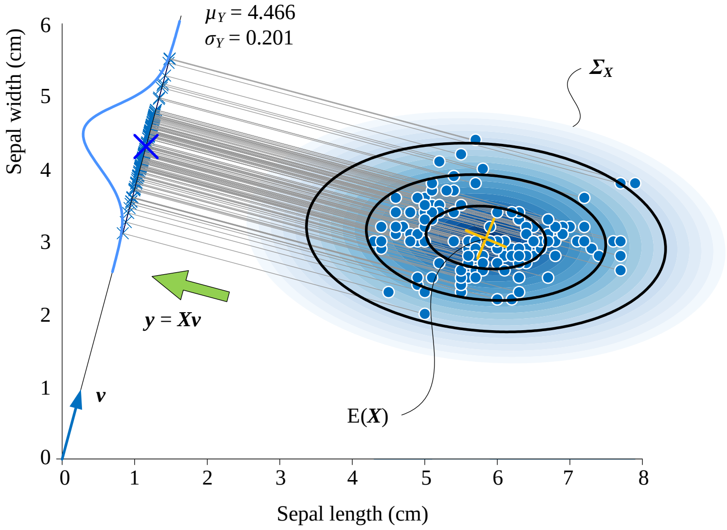


图 6. 逆时针 75 度，X 向 v 投影

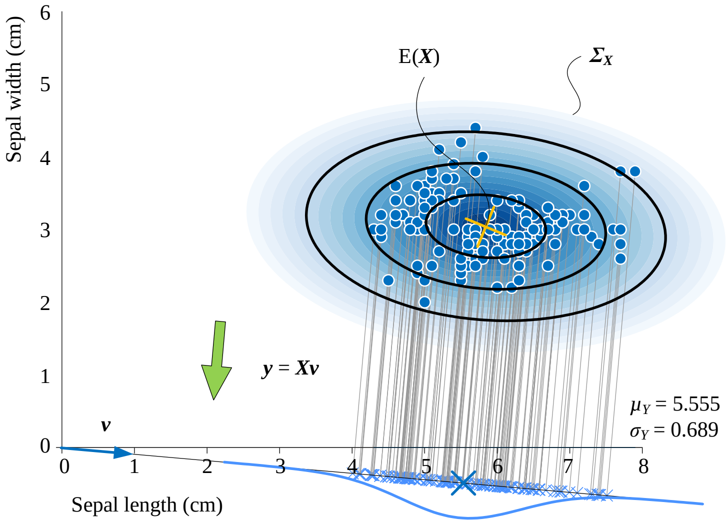


图 7. 逆时针−5 度，X 向 v 投影

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

代码 Bk5_Ch14_01.py 绘制图 2 ~ 图 7。

14.4 正交系投影：鸢尾花两特征为例

正交系

给定正交系 $\boldsymbol{V}$:

$$\boldsymbol{V} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{30}$$

如图 8 所示，数据 $\boldsymbol{X}$ 可以投影到正交系 $\boldsymbol{V}$ 中得到数据 $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{V} \tag{31}$$

展开上式得到：

$$\begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 \end{bmatrix} = \boldsymbol{X}\begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}\boldsymbol{v}_1 & \boldsymbol{X}\boldsymbol{v}_2 \end{bmatrix} \tag{32}$$

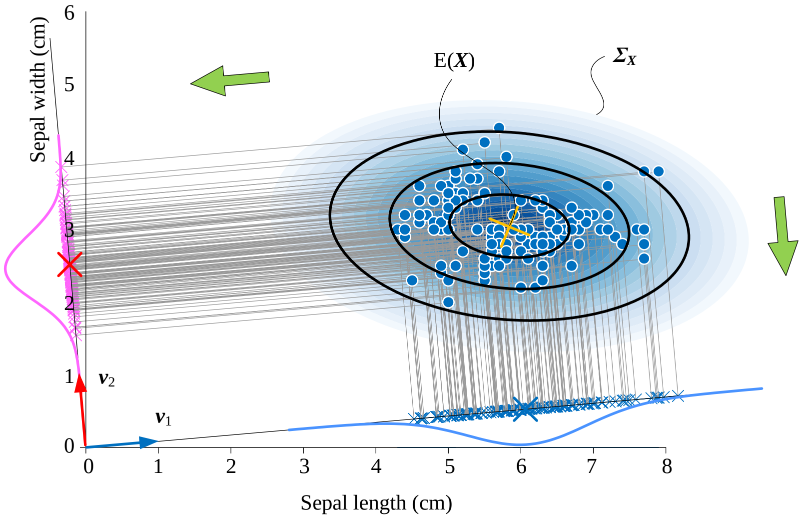


图 8. $\boldsymbol{X}$ 向正交系 $\boldsymbol{V}$ 投影

随机变量为 $\boldsymbol{\chi} = [X_1, X_2]^{\mathrm{T}}$ 投影到 $\boldsymbol{V}$ 得到 $\boldsymbol{\gamma} = [Y_1, Y_2]^{\mathrm{T}}$:

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$\gamma = V^{\mathrm{T}} \chi \tag{33}$$

展开上式得到：

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = V^{\mathrm{T}} \chi = \begin{bmatrix} v_1^{\mathrm{T}} \\ v_2^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} v_1^{\mathrm{T}} \chi \\ v_2^{\mathrm{T}} \chi \end{bmatrix} \tag{34}$$

注意比较 (31) 和 (33) 的转置关系。

向第一方向投影

先考虑 X 向 v_1 投影：

$$v_1 = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \tag{35}$$

将数据 X 投影到 v_1 得到：

$$y_1 = X v_1 \tag{36}$$

类似地，将 $\chi = [X_1, X_2]^{\mathrm{T}}$ 投影到 v_1 得到 Y_1：

$$Y_1 = \begin{bmatrix} X_1 & X_2 \end{bmatrix} v_1 = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} = \cos\theta X_1 + \sin\theta X_2 \tag{37}$$

Y_1 的质心为：

$$\begin{aligned} \mathrm{E}(Y_1) = \mathrm{E}(X) v_1 &= \begin{bmatrix} 5.8433 & 3.0573 \end{bmatrix} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \\ &\approx 3.0573 \times \sin(\theta) + 5.8433 \times \cos(\theta) \end{aligned} \tag{38}$$

Y_1 的方差为：

$$\begin{aligned} \operatorname{var}(Y_1) = v_1^{\mathrm{T}} \Sigma_X v_1 &= \begin{bmatrix} \cos\theta & \sin\theta \end{bmatrix} \begin{bmatrix} 0.6856 & -0.0424 \\ -0.0424 & 0.1899 \end{bmatrix} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \\ &\approx -0.0424 \times \sin(2\theta) + 0.2478 \times \cos(2\theta) + 0.4378 \end{aligned} \tag{39}$$

向第二方向投影

同理，给定 v_2：

$$v_2 = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} \tag{40}$$

将数据 X 投影到 v_2 得到：

$$y_2 = X v_2 \tag{41}$$

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

将 $\boldsymbol{\chi} = [X_1, X_2]^{\mathrm{T}}$ 投影到 $\boldsymbol{v}_2$ 得到 Y_2:

$$Y_2 = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \boldsymbol{v}_2 = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} = -\sin\theta X_1 + \cos\theta X_2 \tag{42}$$

Y_2 的质心为:

$$\begin{aligned} \mu_{Y2} &= \mathrm{E}(\boldsymbol{X})\boldsymbol{v}_2 = \begin{bmatrix} 5.8433 & 3.0573 \end{bmatrix} \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} \\ &\approx -5.8433 \times \sin(\theta) + 3.0573 \times \cos(\theta) \end{aligned} \tag{43}$$

Y_2 的方差为:

$$\begin{aligned} \operatorname{var}(Y_2) &= \boldsymbol{v}_2^{\mathrm{T}} \boldsymbol{\Sigma}_X \boldsymbol{v}_2 = \begin{bmatrix} -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 0.6856 & -0.0424 \\ -0.0424 & 0.1899 \end{bmatrix} \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} \\ &\approx 0.0424 \times \sin(2\theta) - 0.2478 \times \cos(2\theta) + 0.4378 \end{aligned} \tag{44}$$

协方差

Y_1 和 Y_2 的协方差为:

$$\begin{aligned} \operatorname{cov}(Y_1, Y_2) &= \boldsymbol{v}_1^{\mathrm{T}} \boldsymbol{\Sigma}_X \boldsymbol{v}_2 = \begin{bmatrix} \cos\theta & \sin\theta \end{bmatrix} \begin{bmatrix} 0.6856 & -0.0424 \\ -0.0424 & 0.1899 \end{bmatrix} \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} \\ &\approx 0.2478 \times \sin(2\theta) + 0.0424 \times \cos(2\theta) \end{aligned} \tag{45}$$

利用如下三角函数关系:

$$\begin{aligned} f(\theta) &= a\sin(\theta) + b\cos(\theta) \\ &= \sqrt{a^2+b^2}\left(\frac{a}{\sqrt{a^2+b^2}}\sin(v) + \frac{b}{\sqrt{a^2+b^2}}\cos(\theta) \right) \\ &= \sqrt{a^2+b^2}\big(\sin(\theta)\cos(\phi) + \cos(\theta)\sin(\phi)\big) \\ &= A\sin(\theta+\phi) \end{aligned} \tag{46}$$

其中,

$$\begin{aligned} \phi &= \arctan\left(\frac{b}{a}\right) \\ A &= \sqrt{a^2+b^2} \end{aligned} \tag{47}$$

我们可以进一步整理 (38)、(39)、(43)、(44)、(45)，这部分推导交给大家完成。

如图 9 所示，期望、方差、协方差随 θ 变化。请大家特别注意 Y_1 和 Y_2 的方差之和为定值，即:

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$\begin{aligned}\operatorname{var}(Y_1)+\operatorname{var}(Y_2) &\approx -0.0424\times\sin(2\theta)+0.2478\times\cos(2\theta)+0.4378+\\ &\quad 0.0424\times\sin(2\theta)-0.2478\times\cos(2\theta)+0.4378\\ &\approx 0.8756\end{aligned} \tag{48}$$

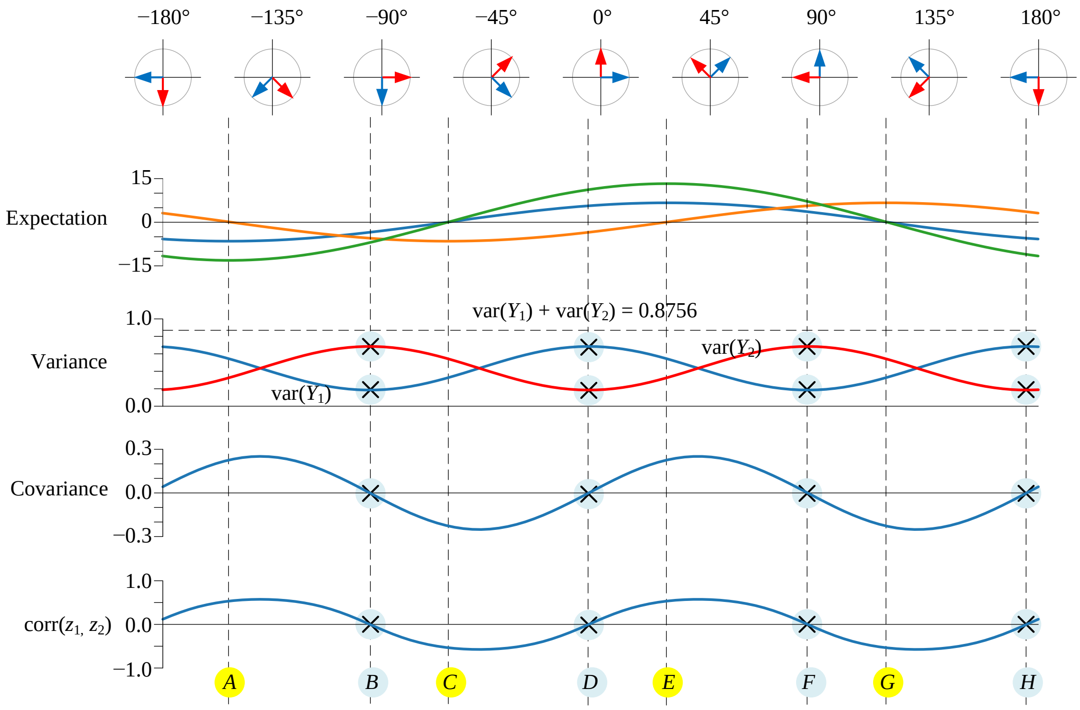


图 9. $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 各种量化关系随 θ 变化

协方差矩阵

$\boldsymbol{\gamma} = [Y_1, Y_2]^{\mathrm{T}}$ 的协方差矩阵 $\boldsymbol{\Sigma}_\gamma$ 为：

$$\operatorname{var}(\boldsymbol{\gamma}) = \boldsymbol{\Sigma}_\gamma = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}_X\boldsymbol{V} \tag{49}$$

图 10 所示为当 θ 取不同值时，协方差矩阵 $\boldsymbol{\Sigma}_\gamma$ 的三种不同可视化方案的变化情况。

特别地，如图 10 (b) 所示，当 θ 约为−4.85 度时，协方差矩阵 $\boldsymbol{\Sigma}_\gamma$ 为对角方阵。这意味着 Y_1 和 Y_2 的相关性系数为 0。

在图 9 中，我们可以发现，当 θ 约为−4.85 度时，$\operatorname{var}(Y_1)$ 取得最大值，$\operatorname{var}(Y_2)$ 取得最小值。如图 11 所示为数据矩阵在这个正交坐标系中投影的结果。这一点对于本章后续要讲解的主成分分析非常重要。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

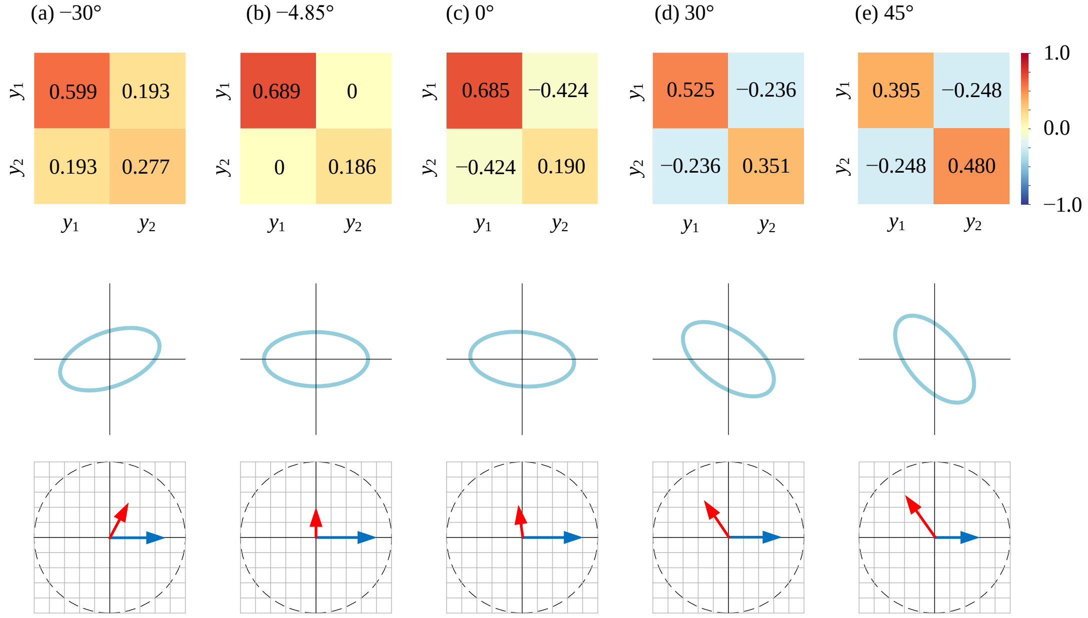


图 10. 协方差矩阵的可视化

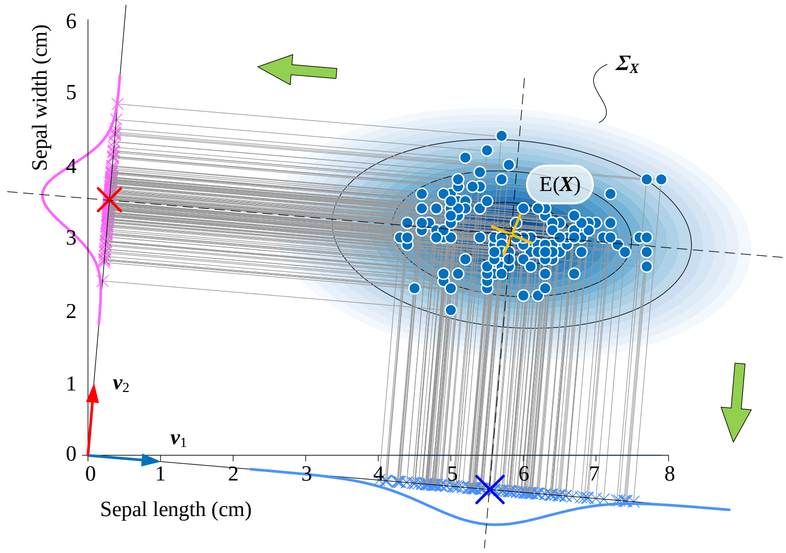


图 11. X 向正交系 V 投影，-4.8575 度

14.5 以椭圆投影为视角看线性变换

本节将从椭圆投影视角理解随机变量的线性转换。

“正”矩形

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

如图 12 所示，三个“正”矩形的四条边分别和马氏距离为 1、2、3 的椭圆相切。其中，和马氏距离为 1 的矩形相切的矩形的长、宽分别为 $2\sigma_1$、$2\sigma_2$。

上一章提到过，图 12 中这个大矩形的面积为 $4\sigma_1\sigma_2$，其对角线长度为矩形对角线长度为 $2\sqrt{\sigma_1^2+\sigma_2^2}$。

上一章特别强调图 12 中阴影区域对应的 1/4 矩形。这个 1/4 矩形的面积为 $\sigma_1\sigma_2$，1/4 矩形对角线长度为 $\sqrt{\sigma_1^2+\sigma_2^2}$，这个值是其协方差迹的平方根 $\sqrt{\sigma_1^2+\sigma_2^2}=\sqrt{\operatorname{tr}\left(\boldsymbol{\Sigma}_{2\times 2}\right)}$。

根据本章有关随机变量线性变换内容，如图 12 所示，这三个矩形“长边”所在位置分别对应 $\mu_1 \pm \sigma_1$、$\mu_1 \pm 2\sigma_1$、$\mu_1 \pm 3\sigma_1$。“宽边”所在位置分别对应 $\mu_2 \pm \sigma_2$、$\mu_2 \pm 2\sigma_2$、$\mu_2 \pm 3\sigma_2$。这并不是巧合，本节后续将用数学工具加以证明。

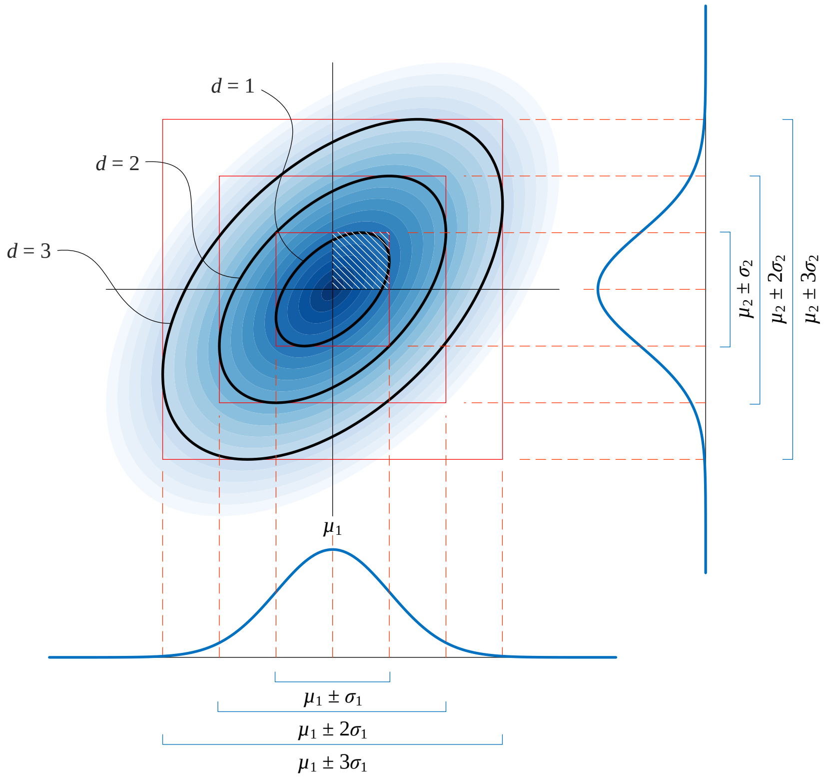


图 12. 和马氏距离椭圆相切的“正”矩形

“主轴”矩形

如图 13 所示，和马氏距离为 1 的椭圆相切的矩形有无数个。观察这些矩形，大家能够发现它们的顶点位于正圆之上。这意味这这些矩形的对角线长度相同，都是 $2\sqrt{\sigma_1^2+\sigma_2^2}$。想要证明这个观察，需要用到矩阵迹的性质，证明留给大家自行完成。

除了图 12 中的“正”矩形之外，还有一个“旋转”矩形特别值得我们关注。这就是图 14 所示的“主轴”矩形。之所以叫“主轴”矩形，是因为这个矩形的四条边平行于椭圆的两条主轴 (长轴、短轴)。

而特征值分解协方差矩阵就是获得椭圆主轴方向、长轴长度、短轴长度的数学工具。请大家根据上一章内容自行分析图 14 中和马氏距离为 1 椭圆相切的“主轴”矩形的几何特征。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

请大家回忆协方差特征值分解得到的特征值和投影获得的两个分布的方差、标准差关系。

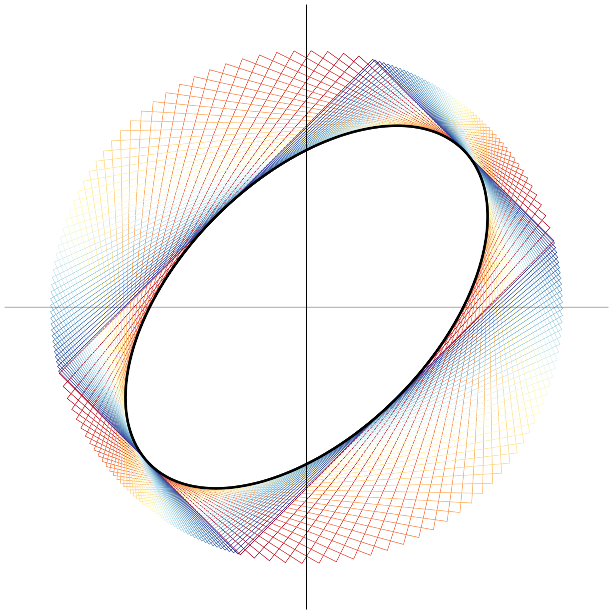


图 13. 和马氏距离椭圆相切的一组“旋转”矩形

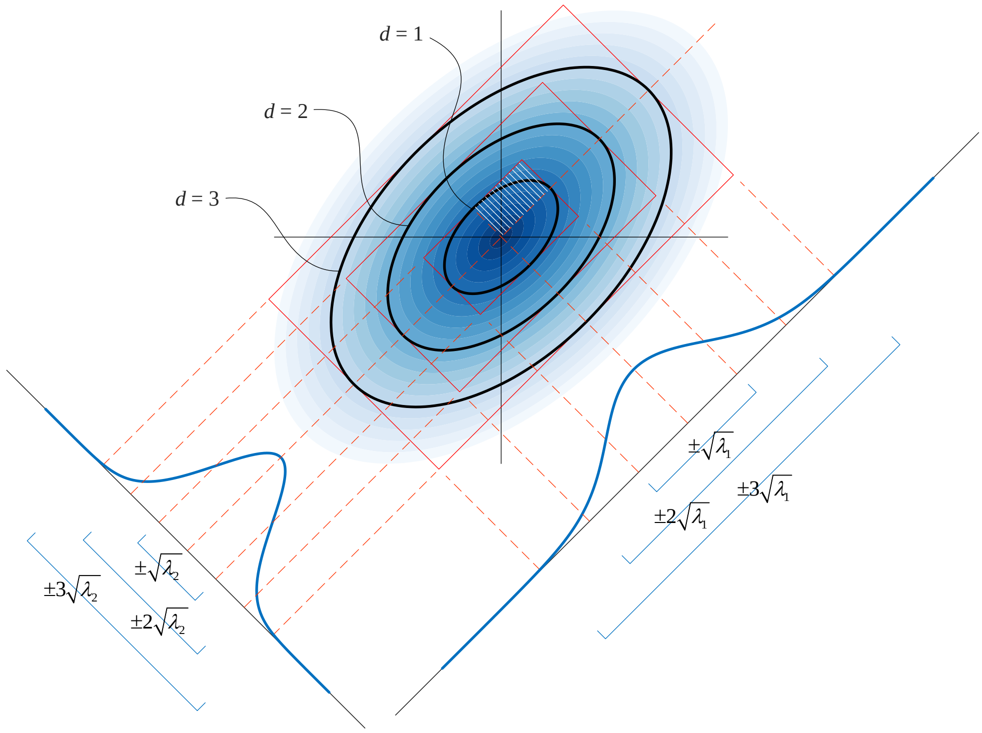


图 14. 和马氏距离椭圆相切的“主轴”矩形

椭圆切线

大家可能好奇如何绘制图 13 这组旋转矩形。如图 15 所示，首先，计算椭圆圆心 $\boldsymbol{\mu}$ 和椭圆上任意一点 $\boldsymbol{p}$ 切线的距离 h。$2h$ 就是矩形一条边长度。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

而切线的梯度向量 $\boldsymbol{n}$ 可以用来定位矩形的旋转角度。然后，根据矩形的对角线长度为 $2\sqrt{\sigma_1^2+\sigma_2^2}$，我们便得到矩形另外一条边的长度。

问题来了，如何计算距离 h 和梯度向量 $\boldsymbol{n}$？

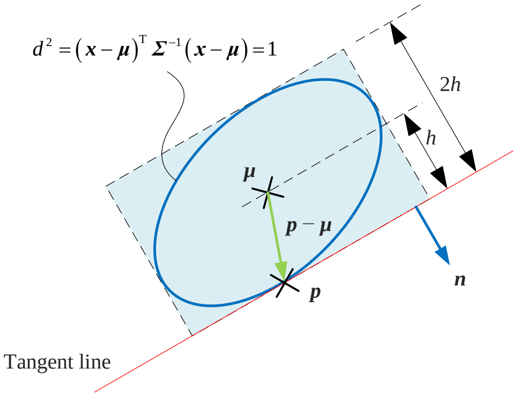


图 15. 计算马氏椭圆上任意一点切线原理

我们在《矩阵力量》第 20 章介绍过如何求解椭圆切线。图 15 中椭圆的解析式为：

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})-1=0 \tag{50}$$

$\boldsymbol{p}$ 在椭圆上，如下等式成立：

$$(\boldsymbol{p}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{p}-\boldsymbol{\mu})-1=0 \tag{51}$$

定义如下函数 $f(\boldsymbol{x})$：

$$f(\boldsymbol{x})=(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})-1=0 \tag{52}$$

$f(\boldsymbol{x})$ 对 $\boldsymbol{x}$ 求偏导便得到梯度向量 $\boldsymbol{n}$：

$$\boldsymbol{n}=\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}=2\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{53}$$

上式用到了《矩阵力量》第 17 章的多元微分。也就是说，图 13 中椭圆上 $\boldsymbol{p}$ 点处切线的法向量为：

$$\boldsymbol{n}=2\boldsymbol{\Sigma}^{-1}(\boldsymbol{p}-\boldsymbol{\mu}) \tag{54}$$

切点 $\boldsymbol{p}$ 和椭圆圆心 $\boldsymbol{\mu}$ 的距离向量 $\boldsymbol{p}-\boldsymbol{\mu}$，对应图 13 中的绿色箭头。而距离 h 就是向量 $\boldsymbol{p}-\boldsymbol{\mu}$ 在梯度向量 $\boldsymbol{n}$ 上的标量投影：

$$h=\frac{\boldsymbol{n}^{\mathrm{T}}(\boldsymbol{p}-\boldsymbol{\mu})}{\|\boldsymbol{n}\|} \tag{55}$$

有了以上推导，请大家自行编写代码绘制图 14。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

《可视之美》一册曾详细讲解过这段代码。

14.6 主成分分析：换个视角看数据

下面以鸢尾花数据作为原始数据，从随机变量的线性变换角度理解主成分分析。

首先将鸢尾花花萼长度、花萼宽度数据中心化，即获得 $\boldsymbol{X}_c = \boldsymbol{X} - \mathrm{E}(\boldsymbol{X})$。图 16 所示为中心化数据的散点图。将数据投影到角度为逆时针 30 度的正交系 $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2]$ 中。如前文所述，数据投影到正交系中好比在 $\boldsymbol{V}$ 中观察数据，如图 17 所示。在 $\boldsymbol{V}$ 中，我们看到代表协方差矩阵的椭圆发生了明显旋转。在 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 方向上，我们可以求得投影数据的分布情况。

图 18 ~ 图 23 所示为其他 3 组投影角度。请大家格外注意图 22 和图 23，这就是前文说的最优化角度。这两幅图中的 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 分别为第一、第二主成分方向。

本章仅仅从随机变量的线性函数角度介绍主成分分析，本书第 25 章将深入介绍主成分分析。此外，《矩阵力量》第 25 章介绍过，特征值分解协方差矩阵仅仅是主成分分析六条基本技术路径之一，《数据有道》还会介绍其他路径，并做区分。

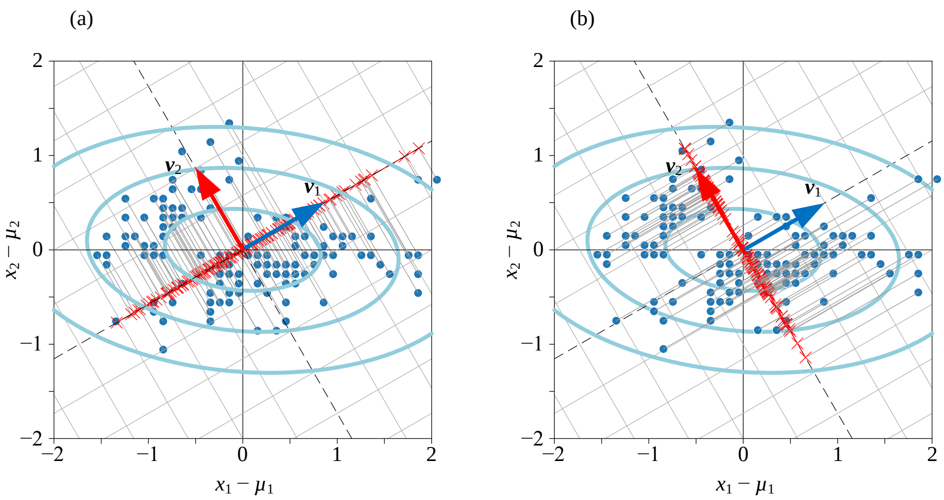


图 16. 正交系，逆时针 30 度

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

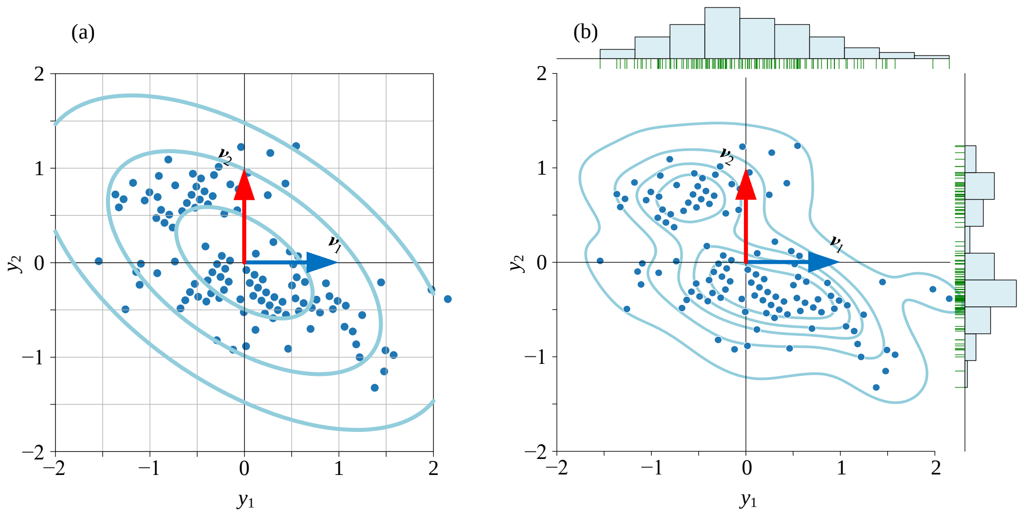


图 17. 数据顺时针旋转 30 度

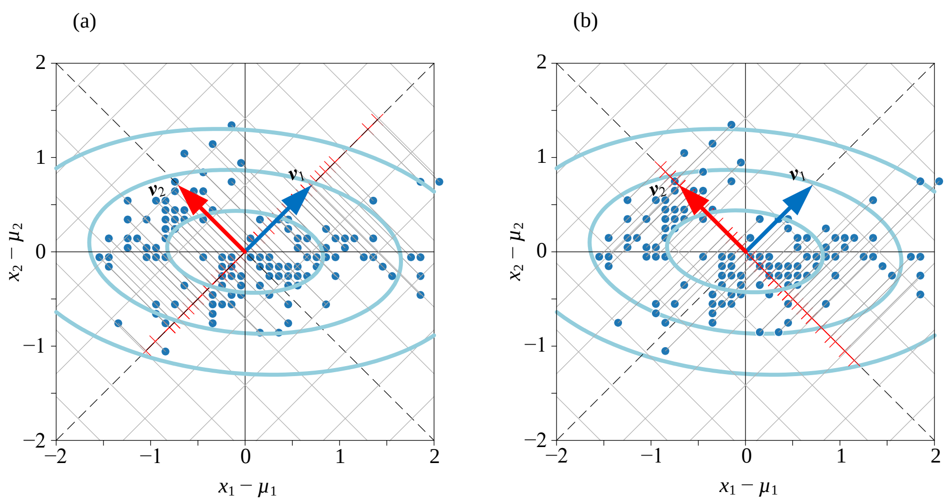


图 18. 正交系，逆时针 45 度

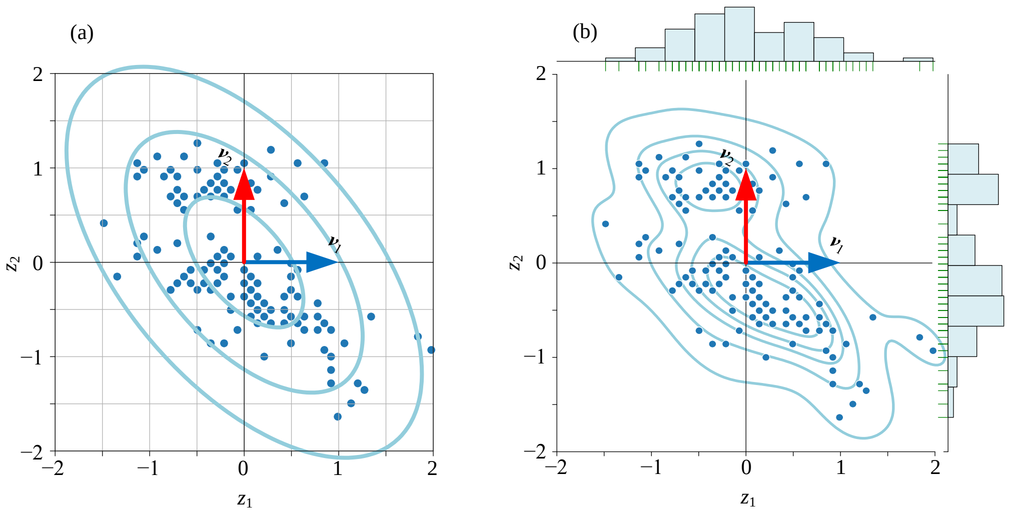


图 19. 数据顺时针旋转 45 度

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

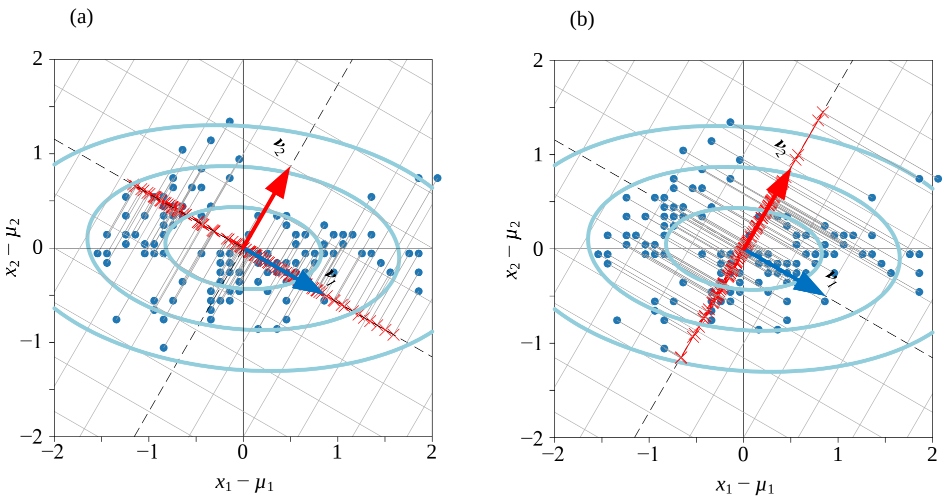


图 20. 正交系，逆时针−30 度

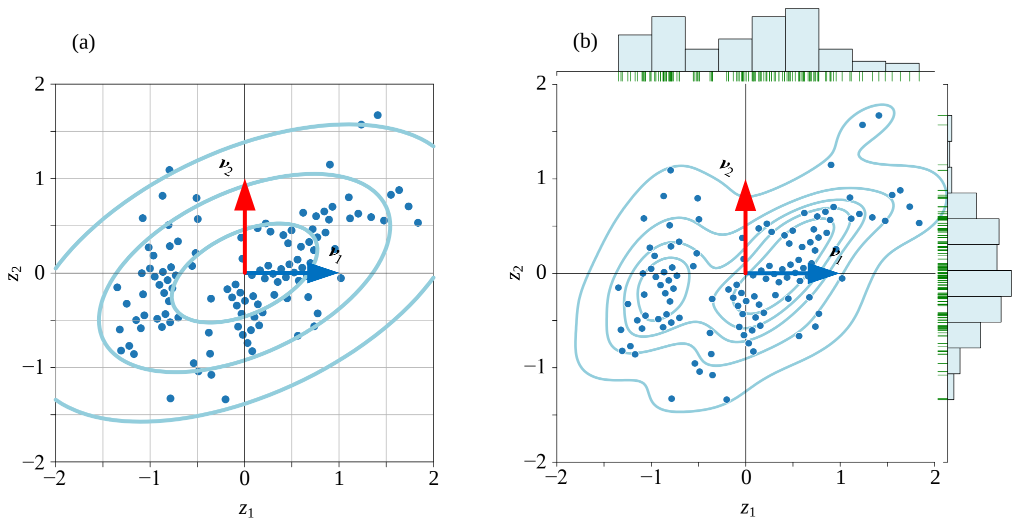


图 21. 数据顺时针旋转−30 度

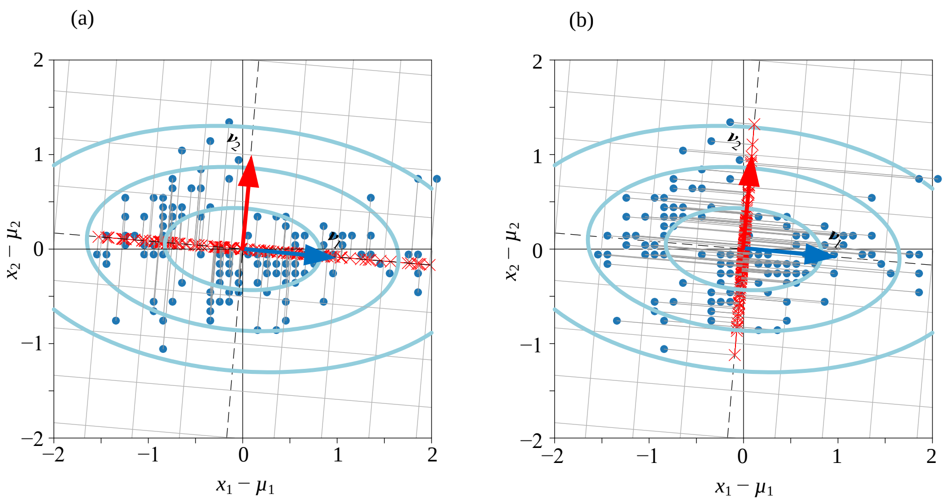


图 22. 正交系，顺时针−4.85 度

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

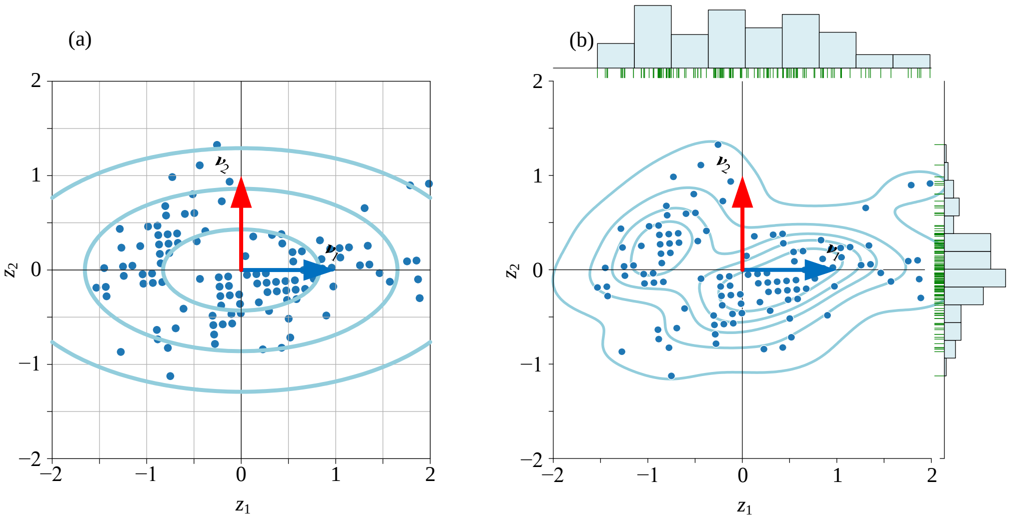


图 23. 数据逆时针旋转−4.85 度

随机变量的函数是指一个或多个随机变量组成的函数，其值也是一个随机变量。它们可以用来描述随机变量之间的关系或随机事件的性质。随机变量的函数在概率论和统计学中都有广泛的应用，例如用于建立概率模型、描述随机事件的分布和性质、进行概率推断和预测等。这一章，我们特别关注的是随机变量的线性变换。这是指将一个随机变量通过一个线性函数转化为另一个随机变量的过程，相当于线性代数中的仿射变换。

随机变量的线性变换在统计学和概率论中经常被用来描述随机变量之间的关系，例如线性回归模型、协方差矩阵和主成分分析等。通过线性变换，可以将随机变量从原始空间中转换到一个新的空间，从而发现不同随机变量之间的联系和规律。

请大家特别重视通过投影、椭圆视角理解随机变量的线性变换。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：https://github.com/Visualize-ML
本书配套微课视频均发布在 B 站——生姜 DrGinger：https://space.bilibili.com/513194466
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com