数据科学基础 HW1

赵希媛

2025 年 10 月 7 日

1 Problem 1

Consider a mixture of two Gaussian distributions:

$$0.4\mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right)+0.6\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}8.4&2.0\\2.0&1.7\end{bmatrix}\right)$$

1. Compute the marginal distribution for each dimension.
2. Compute the mean, mode and median for each marginal distribution.
3. Compute the mean and mode for the two-dimensional distribution.
4. Plot these distributions and mark the points using Python.

1.1 a

想要求得每个维度的边缘分布，那么我们需要对其他维度进行积分。题中给出的分布为混合高斯分布，因为积分的线性性，我们可以对每一部分分别积分再相加。又由于对于一个高维高斯分布，其边缘分布依旧为高斯分布，且这个一维高斯分布的均值就是原均值向量对应的那个元素，方差就是原协方差矩阵对应的那个对角线上的元素。

因此 x_1 变量的边缘分布为:

$$0.4\cdot\mathcal{N}(10,1)+0.6\cdot\mathcal{N}(0,8.4)$$

x_2 变量的边缘分布为:

$$0.4\cdot\mathcal{N}(2,1)+0.6\cdot\mathcal{N}(0,1.7)$$

1.2 b

对于均值，x_1 的均值为 $0.4\cdot10+0.6\cdot0=4$, x_2 的均值为 $0.4\cdot2+0.6\cdot0=0.8$。

对于众数, 我们要求使概率密度最大的 x 值，即 $\text{mode}(X)=\arg\max_x p(x)$。对于 x_1

$$\text{mode}(x_1)=\arg\max_{x_1}p(x_1)=\arg\max_{x_1}0.4\cdot\frac{\exp(-\frac{(x_1-10)^2}{2})}{\sqrt{2\pi}}+0.6\cdot\frac{\exp(-\frac{{x_1}^2}{16.8})}{\sqrt{2\pi}\sqrt{8.4}}$$

同理对于 x_2，相当于求解

$$\text{mode}(x_2)=\arg\max_{x_2}p(x_2)=\arg\max_{x_2}0.4\cdot\frac{\exp(-\frac{(x_2-2)^2}{2})}{\sqrt{2\pi}}+0.6\cdot\frac{\exp(-\frac{{x_2}^2}{3.4})}{\sqrt{2\pi}\sqrt{1.7}}$$

我们利用 BFGS 方法通过多初始点优化寻找最值，得到 x_1 的众数为 $x_1=9.998395$，x_2 的众数为 $x2=1.330791$。

对于中位数, 我们需要求使满足累积分布函数（CDF）$F(x)=P(X\le x)=0.5$ 的值 x，即 $\text{median}(X)=x,$其中$F(x)=0.5$。

利用二分法，可解得 $\text{median}(x_1)=2.803854,\text{median}(x_2)=0.868645$，并验证可知 x_1 的 CDF 在 $x=2.803854$ 处的值: 0.5000000000，x_2 的 CDF 在 $x=0.868645$ 的值: 0.5000000000，没有问题。

1.3 c

对于均值，混合高维高斯分布各分量的均值就是其边缘分布的均值，因此：

$$mean(X) = 0.4 \begin{bmatrix} 10 \\ 2 \end{bmatrix} + 0.6 \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 0.8 \end{bmatrix}$$

对于众数，混合高维高斯分布的求解相当于

$$\text{mode}(X) = \arg\max_x p(x) = \arg\max_x \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

这里 π_k 为对应权重，μ_k 为对应期望，Σ_k 为对应协方差。

利用相应的 python 代码实现，从多个初始点开始寻找，得到二维分布的众数: $\text{mode}(X) = \begin{bmatrix} 9.99854811 \\ 2.0003569 \end{bmatrix}$。

1.4 d

根据前面得到的数据结果进行可视化：

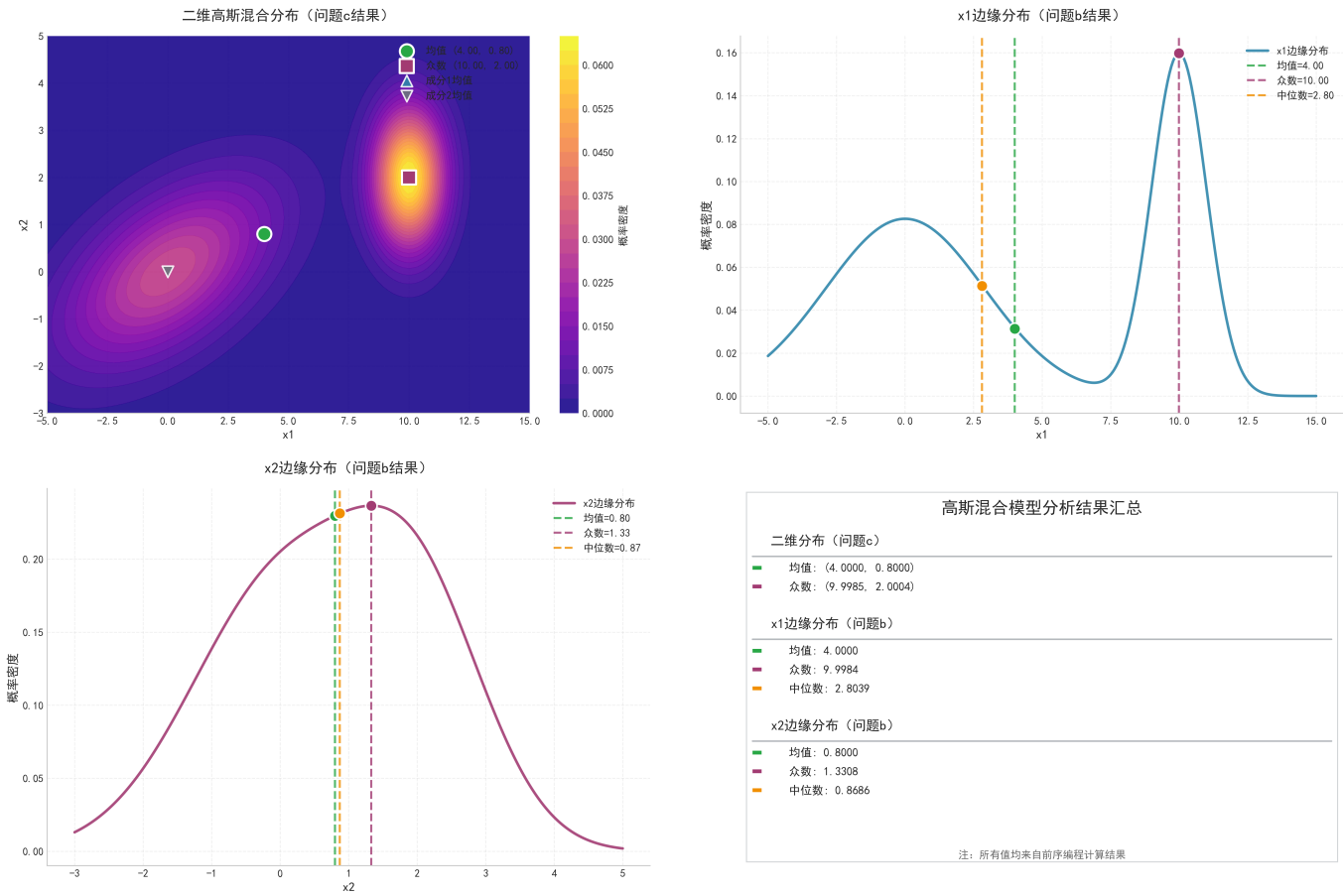

图 1: 可视化结果

2 Problem 2

Consider a Gaussian random variable $x \sim \mathcal{N}(x \mid \mu_x, \Sigma_x)$, where $x \in \mathbb{R}^D$. Furthermore, we have $y = Ax + b + w$ where $y \in \mathbb{R}^E$, $A \in \mathbb{R}^{E \times D}$, $b \in \mathbb{R}^E$, and $w \sim \mathcal{N}(w \mid 0, Q)$ is independent Gaussian noise. "Independent" implies that x and w are independent random variables and that Q is diagonal.

(a) Write down the likelihood $p(y \mid x)$.

(b) The distribution $p(y) = \int p(y \mid x)p(x)dx$ is Gaussian. Compute the mean μ_y and the covariance Σ_y. Derive your result in detail.

2.1 a

似然函数 $p(y \mid x)$ 是条件概率密度函数，相当于是知道 x 取值的条件下 y 的分布。因为 $y = Ax + b + w$，其中 $w \sim \mathcal{N}(w \mid 0, Q)$，那么 $p(y \mid x)$ 就相当于均值为 $Ax + b$，协方差为 Q 的正态分布，即

$$p(y \mid x) \sim \mathcal{N}(Ax + b, Q)$$

2.2 b

由 (a)，

$$p(y \mid x) = \frac{\exp{(y - (Ax + b))^T Q^{-1}(y - (Ax + b))}}{(2\pi)^{\frac{E}{2}}\sqrt{|Q|}}$$

则

$$p(y) = \int p(y \mid x)p(x)dx = \int \frac{\exp{(y - (Ax + b))^T Q^{-1}(y - (Ax + b))}}{(2\pi)^{\frac{E}{2}}\sqrt{|Q|}} \frac{\exp{(x - \mu_x)^T \Sigma_x^{-1}(x - \mu_x)}}{(2\pi)^{\frac{D}{2}}\sqrt{|\Sigma_x|}} dx$$

但我们会发现，直接计算的复杂程度是很大的，那么我们利用 y 也为高斯分布以及高斯分布的简单性质来得到结果。

对于均值 μ_y，由于 $y = Ax + b + w$，则 $\mu_y = E[y] = E[Ax + b + w] = AE[x] + b + E[w] = A\mu_x + b$。

对于协方差 Σ_y，协方差公式为 $\Sigma_y = \mathrm{Cov}[y] = \mathbb{E}\left[(y - \mu_y)(y - \mu_y)^T\right]$。因为 $y - \mu_y = (Ax + b + w) - (A\mu_x + b) = A(x - \mu_x) + w$, 则

$$\Sigma_y = \mathbb{E}\left[(A(x - \mu_x) + w)(A(x - \mu_x) + w)^T\right]$$

整理化简可得：

$$\Sigma_y = \mathbb{E}\left[A(x - \mu_x)(x - \mu_x)^T A^T + A(x - \mu_x)w^T + w(x - \mu_x)^T A^T + ww^T\right]$$

由于 x 和 w 独立，且 $\mathbb{E}[x - \mu_x] = 0$，$\mathbb{E}[w] = 0$，交叉项为零：

$$\mathbb{E}\left[A(x - \mu_x)w^T\right] = A\mathbb{E}\left[(x - \mu_x)\right]\mathbb{E}\left[w^T\right] = 0$$

类似地，其他交叉项也为零。因此：

$$\Sigma_y = A\mathbb{E}\left[(x - \mu_x)(x - \mu_x)^T\right]A^T + \mathbb{E}\left[ww^T\right] = A\Sigma_x A^T + Q$$

其中，$\mathbb{E}\left[(x - \mu_x)(x - \mu_x)^T\right] = \Sigma_x$，且 $\mathbb{E}[ww^T] = Q$。

综上，$\mu_y = A\mu_x + b$，$\Sigma_y = A\Sigma_x A^T + Q$

3 Problem 3

Randomly generate 30 points inside the cube $\left[-\frac{1}{2}, \frac{1}{2}\right]^{100}$ and plot the distance between points and the angle between the vectors from the origin to the points for all pairs of points.

利用 python 代码实现这一过程，可得到如下结果：

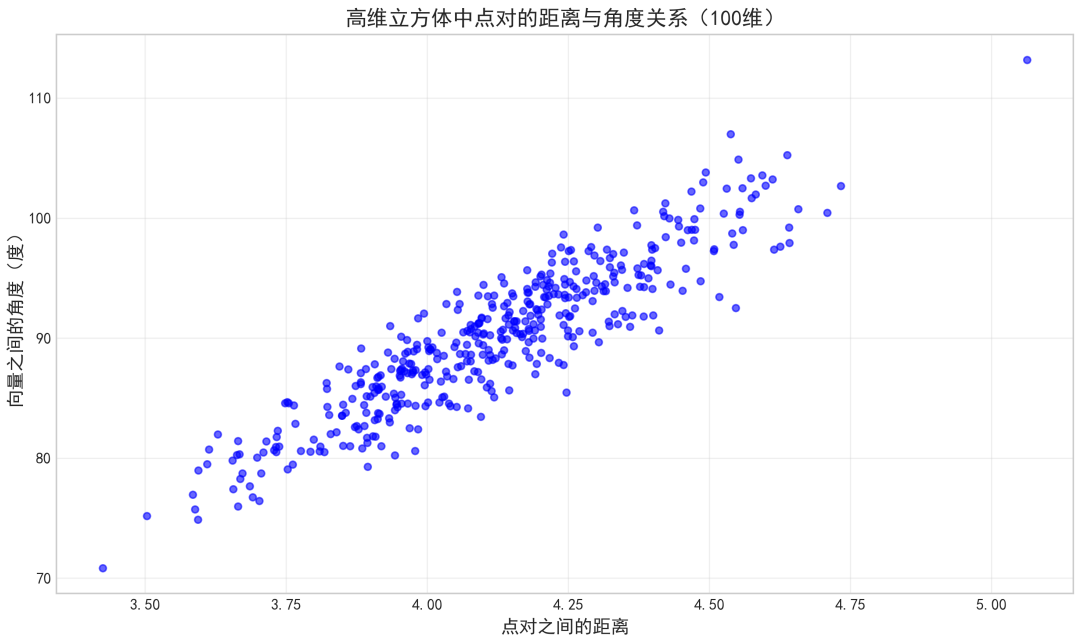

图 2: 高维立方体中点对的距离与角度关系

4 Problem 4

Let G be a d-dimensional spherical Gaussian with variance $\frac{1}{2}$ in each direction, centered at the origin. Derive the expected squared distance to the origin.

根据题意可知，$G \sim \mathcal{N}(0, \frac{1}{2}E)$，其中 E 为单位阵，即若 $X = (X_1, X_2, ..., X_d)$ 为 d 维随机变量，则有 $E[X_i] = 0$，$var(X_i) = \frac{1}{2}$，并且每个分量各自独立。这里我们要求得 $E[||X||^2]$，即 $E[X_1^2 + X_2^2 + ... + X_d^2]$。

$E[X_1^2 + X_2^2 + ... + X_d^2] = E[X_1^2] + E[X_2^2] + ... + E[X_d^2]$，对于任意 X_i，$var(X_i) = E[X_i^2] - E[X_i]^2 = E[X_i^2] = \frac{1}{2}$，故 $E[||X||^2] = E[X_1^2 + X_2^2 + ... + X_d^2] = \frac{d}{2}$。

综上，距原点的期望距离为 $\frac{d}{2}$。

5 Problem 5

How large must ϵ be for 99% of the volume of a 1000-dimensional unit-radius ball to lie in the shell of ϵ-thickness at the surface of the ball?

对于 d 维的半径为 r 的球体，其体积为 $V_d(r) = M(d)r^d$，这里 $M(d)$ 是与维数有关的一个常数。根据题意，我们不妨设球壳厚度为 ϵ，则这个外半径为 1 的球壳的体积为 $V_s = V_1 - V_\epsilon = M(1000)(1 - \epsilon^{1000})$，这里我们要求

$$\frac{V_s}{V_1} = 1 - (1 - \epsilon)^{1000} \geq 0.99$$

即 $(1 - \epsilon)^{1000} \leq 0.01$，两边取以 10 为底的对数，则有

$$1000 \log_{10}(1 - \epsilon) \leq -2$$

则有 $\log_{10}(1 - \epsilon) \leq -\frac{1}{500}$，故 $1 - \epsilon \leq 10^{-\frac{1}{500}}$，即 $\epsilon \geq 1 - 10^{-\frac{1}{500}}$。

综上，球壳厚度 ϵ 只要大于等于 $1 - 10^{-\frac{1}{500}}$（大约为 0.00460517）即可。

6 Problem 6

For what value of d does the volume, $V(d)$, of a d-dimensional unit ball take on its maximum?

(a) Write a recurrence relation for $V(d)$ in terms of $V(d-1)$ by integrating over x_1. Hint: At $x_1 = t$, the $(d-1)$-dimensional volume of the slice is the volume of a $(d-1)$-dimensional sphere of radius $\sqrt{1-t^2}$. Express this in terms of $V(d-1)$ and write down the integral. You need not evaluate the integral.

(b) Verify the formula for $d = 2$ and $d = 3$ by integrating and comparing with $V(2) = \pi$ and $V(3) = \frac{4}{3}\pi$.

6.1 a

我们想像有一个 d 维的单位球，考虑在 $x_1 = t$ 处时，截面为一个半径为 $\sqrt{1-t^2}$ 的 $d-1$ 维的球面，则其体积为 $(\sqrt{1-t^2})^{d-1}V(d-1)$，这里 $V(d-1)$ 为 $d-1$ 维单位球体积，故

$$V(d) = \int_{-1}^{1} (\sqrt{1-t^2})^{d-1} V(d-1)dt = V(d-1)\int_{-1}^{1} (\sqrt{1-t^2})^{d-1} dt$$

令 $t = cos\theta$，则 $\theta \in (0, \pi)$，上式化简为

$$V(d) = -V(d-1)\int_{\pi}^{0} sin\theta^d d\theta = V(d-1)\int_{0}^{\pi} sin\theta^d d\theta$$

由于

$$\int_0^\pi \sin^d\theta\, d\theta = \sqrt{\pi}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)}$$

则

$$\frac{V(d)}{V(d-1)} = \sqrt{\pi}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)} = R(d)$$

接下来考虑 $R(d)$ 的单调性，计算 $\frac{R(d+1)}{R(d)}$：

$$\frac{R(d+1)}{R(d)} = \frac{\sqrt{\pi}\frac{\Gamma\left(\frac{d+2}{2}\right)}{\Gamma\left(\frac{d+1}{2}+1\right)}}{\sqrt{\pi}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)}} = \frac{\Gamma\left(\frac{d+2}{2}\right)\Gamma\left(\frac{d}{2}+1\right)}{\Gamma\left(\frac{d+1}{2}+1\right)\Gamma\left(\frac{d+1}{2}\right)}$$

简化分母和分子：

- $\cdot\,\Gamma\left(\frac{d+1}{2}+1\right) = \Gamma\left(\frac{d+3}{2}\right)$
- $\cdot\,\Gamma\left(\frac{d}{2}+1\right) = \Gamma\left(\frac{d+2}{2}\right)$
- $\cdot\,\Gamma\left(\frac{d+2}{2}\right)$ 在分子中出现两次，因此实际上分子为$\Gamma\left(\frac{d+2}{2}\right)\Gamma\left(\frac{d+2}{2}\right) = \left[\Gamma\left(\frac{d+2}{2}\right)\right]^2$

因此：

$$\frac{R(d+1)}{R(d)} = \frac{\left[\Gamma\left(\frac{d+2}{2}\right)\right]^2}{\Gamma\left(\frac{d+3}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}$$

令 $z = \frac{d+1}{2}$，则：

$$\frac{d+2}{2} = z + \frac{1}{2}$$
$$\frac{d+3}{2} = z + 1$$
$$\frac{d+1}{2} = z$$

代入得：

$$\frac{R(d+1)}{R(d)} = \frac{\left[\Gamma\left(z+\frac{1}{2}\right)\right]^2}{\Gamma(z+1)\Gamma(z)}$$

利用 Γ 函数的递归性质 $\Gamma(z+1)=z\Gamma(z)$，分母变为 $\Gamma(z+1)\Gamma(z)=z[\Gamma(z)]^2$，所以：

$$\frac{R(d+1)}{R(d)}=\frac{\left[\Gamma\left(z+\frac{1}{2}\right)\right]^2}{z[\Gamma(z)]^2}=\frac{1}{z}\left(\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma(z)}\right)^2$$

现在，利用 Γ 函数的一个关键性质：对于 $z>0$，有不等式：

$$\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma(z)}<\sqrt{z}$$

这个不等式源于 Γ 函数的对数凸性（通过柯西-施瓦茨不等式或 Gautschi 不等式可得）。因此：

$$\left(\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma(z)}\right)^2<z$$

代入上式：

$$\frac{R(d+1)}{R(d)}=\frac{1}{z}\left(\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma(z)}\right)^2<\frac{1}{z}\cdot z=1$$

即：

$$\frac{R(d+1)}{R(d)}<1\quad\text{对于所有}\quad d>0$$

故 $R(d)$ 为单调减的函数，又因为 $d=5:R(5)=\frac{16}{15}\approx1.067; d=6:R(6)=\frac{5\pi}{16}\approx0.981$，因此 $d=5$ 时单位球体积最大。（回答大题干中问题）

6.2 b

利用 (a) 中结论，加上 $V(1)=2$，则

$$V(2)=V(1)\int_0^{\pi}sin\theta^2d\theta=2\int_0^{\pi}\frac{1-cos2\theta}{2}d\theta=2(\frac{\pi}{2})=\pi$$

$$V(3)=V(2)\int_0^{\pi}sin\theta^3d\theta=\pi\int_0^{\pi}-sin\theta^2dcos\theta=\pi\int_0^{\pi}(-1+cos\theta^2)dcos\theta=\pi(1+1+\frac{1}{3}(-1-1))=\frac{4}{3}\pi$$

7 Problem 7

Randomly generate 100 points on the surface of a sphere in 3-dimensions and in 100-dimensions. Create a histogram of all distances between the pairs of points in both cases.

利用 python 代码实现，得到结果如下：

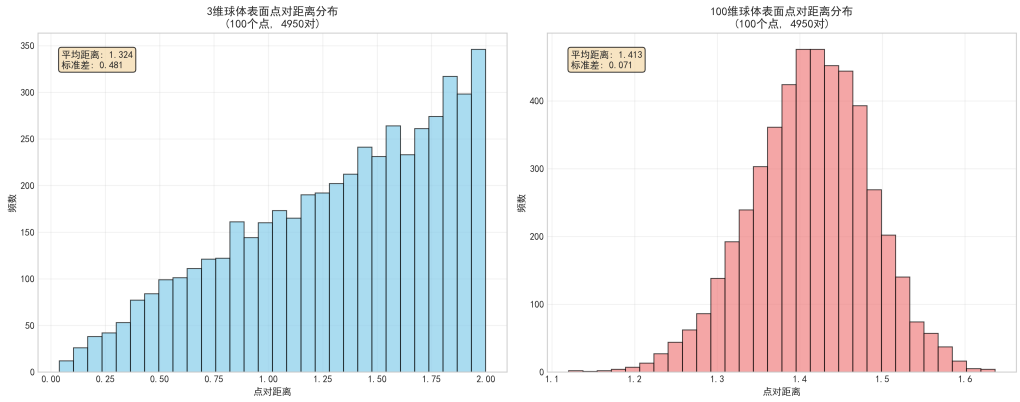

图 3: 多维球表面点距离分布（3 维和 100 维）

通过直方图以及均值和方差的计算会发现平均距离只是略微增大一些，而方差减小了很多，在一定程度上说明了高维空间的一种稀疏性，所以点的分布更加“均匀”。

8 Problem 8

Generate 20 points uniformly at random on a 900-dimensional sphere of radius 30. Calculate the distance between each pair of points. Then, select a method of projection and project the data onto sub-spaces of dimension $k = 100, 50, 10, 5, 4, 3, 2, 1$ and calculate the difference between $\sqrt{k}$ times the original distances and the new pairwise distances. For each value of k what is the maximum difference as a percent of $\sqrt{k}$.

利用 python 代码实现并得到结果如下：

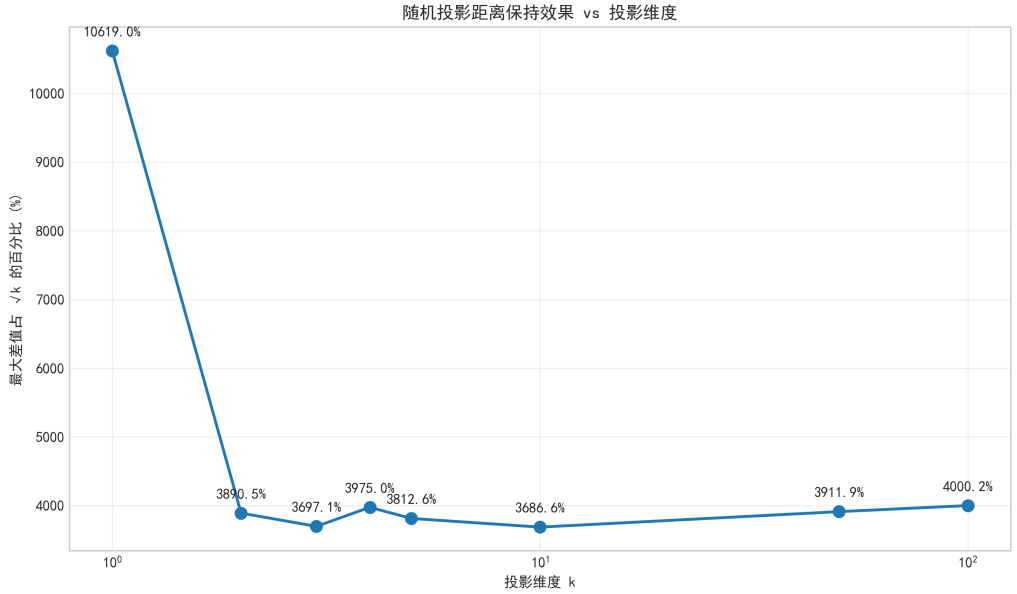

图 4: 高维随机投影

表 1: 详细结果汇总

k	最大差值	占 $\sqrt{k}$ 百分比	原始平均距离	投影平均距离
100	400.0237	4000.24%	42.4791	42.9107
50	276.6129	3911.90%	42.4791	41.6815
10	116.5818	3686.64%	42.4791	36.8148
5	85.2533	3812.64%	42.4791	41.2445
4	79.4995	3974.98%	42.4791	35.1346
3	64.0354	3697.08%	42.4791	40.5509
2	55.0199	3890.50%	42.4791	37.8238
1	106.1897	10618.97%	42.4791	45.2202

这里使用的是随机投影，，这是一种适用于高维数据的降维方法，能近似保持距离关系。通过观察图像和结果汇总会发现随着 k 减小，最大差值的绝对值减小，但占 √k 的百分比可能增加；随机投影在较高维度（k 较大时）能更好地保持距离关系。

9 Problem 9

In d-dimensions there are exactly d unit vectors that are pairwise orthogonal. However, if you wanted a set of vectors that were almost orthogonal you might squeeze in a few more. For example, in 2-dimensions if almost orthogonal meant at least 45 degrees apart, you could fit in three almost orthogonal vectors. Suppose you wanted to find 1000 almost orthogonal vectors in 100 dimensions. Here are two ways you could do it:

(a) Begin with 1,000 orthonormal 1,000-dimensional vectors, and then project them to a random 100-dimensional space.

(b) Generate 1000 100-dimensional random Gaussian vectors. Implement both ideas and compare them to see which does a better job.

利用 python 代码实现并得到结果如下：

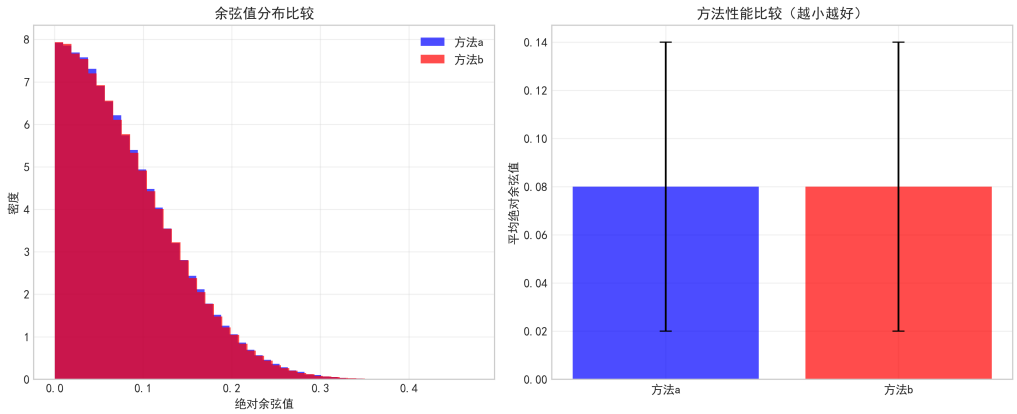

图 5: 高维投影方式比较

为了比较两种方法，我们使用以下指标：

1. 平均绝对余弦值：所有向量对之间夹角余弦的绝对值的平均值（理想值为 0）

$$\text{MeanAbsCosine} = \frac{1}{N}\sum_{i<j}|\cos(\theta_{ij})|$$

其中 N 是向量对的数量，θ_{ij} 是向量 i 和 j 之间的夹角。

2. 余弦值标准差：衡量正交性的一致性

$$\text{CosineStd} = \sqrt{\frac{1}{N}\sum_{i<j}(\cos(\theta_{ij}) - \mu)^2}$$

其中 μ 是余弦值的平均值。

3. 最大余弦值：最差情况下的正交性

$$\text{MaxCosine} = \max_{i<j}|\cos(\theta_{ij})|$$

通过观察图像与数据结果会发现：在高维空间近似正交向量的研究中，方法 b（随机生成法）通常表现良好，因为随机高斯向量在 100 维空间中自然呈现近似正交性，理论预测平均绝对余弦值约为 $1/\sqrt{d} \approx 0.1$（其中 $d = 100$）。相比之下，方法 a（高维投影法）可能略优，因其从严格正交基出发并通过随机投影保持正交性。此外关键发现高维空间存在“维度祝福”现象，即随机向量对夹角自然集中在 90 度附近，这体现了高维几何的反直觉特性；方法比较显示方法 a 更具系统性且有 Johnson-Lindenstrauss 引理保证，而方法 b 更简洁实用；实际应用建议根据需求选择——若需严格理论保证则选方法 a，若优先考虑效率则方法 b 已足够。

10 代码可用性

本文中讨论的所有方法的实现代码可在以下链接获取：
`https://github.com/zoe5xy/Fundamentals-of-Data-Science.git`