最优化方法第一次作业

162050127 颜劭铭

目录

1 基础题

1.1 Sherman-Morrison 公式

由逆矩阵的性质：$(A+xy^T)(A+xy^T)^{-1}=I$, 其中 I 指 $n\times n$ 单位矩阵
所以若要验证该公式：

$$\left(A+xy^{\mathrm{T}}\right)^{-1}=A^{-1}-\frac{A^{-1}xy^{\mathrm{T}}A^{-1}}{1+y^{\mathrm{T}}A^{-1}x}$$

可转化为验证：$(A+xy^T)(A+xy^T)^{-1}=I$
验证过程如下：

$$\begin{aligned}
&(A+xy^T)(A+xy^T)^{-1}\\
&=(A+xy^T)(A^{-1}-\frac{A^{-1}xy^{\mathrm{T}}A^{-1}}{1+y^{\mathrm{T}}A^{-1}x})\\
&=AA^{-1}-\frac{AA^{-1}xy^TA^{-1}}{1+y^TA^{-1}x}+xy^TA^{-1}-\frac{xy^TA^{-1}xy^TA^{-1}}{1+y^TA^{-1}x}\\
&=AA^{-1}+xy^TA^{-1}-\frac{AA^{-1}xy^TA^{-1}+xy^TA^{-1}xy^TA^{-1}}{1+y^TA^{-1}x}\\
&=I+xy^TA^{-1}-\frac{xy^TA^{-1}+xy^TA^{-1}xy^TA^{-1}}{1+y^TA^{-1}x}\\
&=I+xy^TA^{-1}-\frac{x(1+y^TA^{-1}x)y^TA^{-1}}{1+y^TA^{-1}x}\\
&=I+xy^TA^{-1}-xy^TA^{-1}\\
&=I
\end{aligned}$$

其中 I 指 $n\times n$ 单位矩阵

1.2 矩阵的子空间

1.2.1 列空间、行空间、零空间、左零空间定义

首先，由于列空间、行空间、零空间、左零空间都是基于矩阵提出的，因此我们应该在某个矩阵的基础上给出四个基本子空间的定义，会使得定义更加清晰。

我们假设对于一个 m×n 的矩阵 A，他蕴含四个基本子空间：列空间、行空间、零空间、左零空间。

列空间，顾名思义，是由矩阵 A 的 n 个 m 维列向量长成，我们用 C(A) 表示。

行空间，由矩阵 A 的 m 个 n 维行向量长成，我们用 R(A) 表示。

由于 A 的行向量就是 A^T 的列向量，因此矩阵 A 的行空间就是 A^T 的列空间，即 $R(A)=C(A^T)$。

零空间，即满足所有 Ax=0 的向量 x 集合，由线性无关和线性相关的知识可以得出，当矩阵 A 的各列都是线性无关，则 x 只有唯一解零向量，当且仅当 A 的各列线性相关，x 才有非零解。根据矩阵乘法的原则，当 A 是 m×n 矩阵，所以 x 一定是 n 维。

左零空间，和零空间相反，零空间是满足所有 $Ax = 0$ 的向量 x 的集合，左零空间是满足所有 $yA = 0$ 的向量 y 的集合，根据矩阵的转置，左零空间也就是满足所有 $A^T y = 0$ 的向量 y 的集合。根据矩阵乘法的原则，当 A 是 m×n 矩阵，A^T 是 n×m 矩阵，所以 y 一定是 m 维。

在文献中，提到了矩阵的维度的概念，综合上述对于四个子空间的描述，我们可以得到以下四个子空间的关系：

1、矩阵 A 的列空间的维度数矩阵 A 的秩，即 r。

2、矩阵 A 的行空间的维度数和列空间的维度数相等，都是 r。

3、矩阵 A 的零空间的维度数是 $n - r$，列空间和零空间的维度数之和为 n。

4、矩阵 A 的左零空间的维度数是 $m - r$，行空间和左零空间的维度数之和为 m。

最后，文献根据正交的定义，给出了四个基本子空间的正交关系：

1、列空间和左零空间正交。

2、行空间和零空间正交。

1.2.2 秩-零化度定理

秩-零化度定理：定义域 V 的维数等于核空间 ker(T) 的维数与值域 ran(T) 的维数之和，即

$$dimV = dimker(T) + rankT$$

假设线性变换 T：$v \to w$ 由 $m \times n$ 阶矩阵 A 表示，因此可得，n=dimv,m=dimw, 列空间 C(A) 和零空间 N（A）分别表示线性变换 T 的值域 ran(T) 和核空间 ker(T)。

所以转换为证明：n=dimN(A)+rankA.

证明过程如下：

$\because$ $m \times n$ 矩阵 A 用 Gauss-Jordan 消元法可化简为分块矩阵 R 的形式：

$$R = \begin{bmatrix} I_r & F \\ 0 & 0 \end{bmatrix}$$

$\therefore$ 矩阵 R 的秩为 r，F 是 $r \times (n - r)$ 阶矩阵。

$\because$ Gauss-Jordan 消元法不改变矩阵 A 的秩和零空间

$\therefore$ rankA=rankR=r,N(A)=N(R)

观察矩阵 R，可得其 $n \times (n - r)$ 的零空间矩阵 P 如下：

$$P = \begin{bmatrix} -F \\ I_{n-r} \end{bmatrix}$$

验证：

$$RP = \begin{bmatrix} I_r & F \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -F \\ I_{n-r} \end{bmatrix} = \begin{bmatrix} -F + F \\ 0 \end{bmatrix} = 0$$

易得：rankP=n-r

$\therefore$ 列向量线性无关

$\therefore$ 只需证 ker(R) 所有向量可由 P 的列向量线性表出，即可得 C(P)=N(R).

假设 $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, 其中 x_1 是 r 维向量，x_2 是 n-r 维向量，使 $Rx = 0$

$\therefore R_x = \begin{bmatrix} I_r & F \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$

$\because x_1 = -Fx_2$

$\therefore x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -Fx_2 \\ x_2 \end{bmatrix} = \begin{bmatrix} -F \\ I_{n-r} \end{bmatrix} x_2 = Px_2$

∴ C(P)=N(R)

即 dimN(A)=dimNR=rankP=n-r

∴ n=dimN(A)+rankA

∴ 秩-零化度定理成立

1.3 矩阵求导

1.3.1 公式 1

$$\begin{aligned} &\frac{\partial}{\partial X} \operatorname{tr}\left(X^T A X\right) \\ =&d(tr(X^TAX)) \\ =&tr(d(X^TAX)) \\ =&tr(d(X^T)AX + X^TAdX) \\ =&tr((dX)^TAX + X^TAdX) \\ =&tr((AX)^TdX + X^TAdX) \\ =&tr(X^TA^TdX) + tr(X^TAdX) \\ =&tr(X^TA^TdX + X^TAdX) \end{aligned}$$

令 $f(X) = tr(X^TAX)$

$\therefore \ (\frac{\partial f(X)}{\partial X})^T = X^T(A^T + A)$

$\therefore \ \frac{\partial f(X)}{\partial X} = (X^T(A^T + A))^T = (A + A^T)X$

$\therefore \ \frac{\partial}{\partial X} \operatorname{tr}\left(X^T A X\right) = \left(A^T + A\right) X$

1.3.2 公式 2

$$\begin{aligned} &\frac{\partial}{\partial X} \operatorname{tr}\left(X^T A Y\right) \\ =&d(tr(X^TAY)) \\ =&tr(d(X^T)AY)) \\ =&tr((dX)^TAY) \\ =&tr(AY(dX)^T) \\ =&tr((AY)^TdX) \end{aligned}$$

同 1.3.1 公式 1:

$\therefore \ \frac{\partial}{\partial X} \operatorname{tr}\ (X^TAY) = AY$

1.3.3 公式 3

$$
\begin{aligned}
&\frac{\partial}{\partial Y}\operatorname{tr}\left(X^{T}AY\right)\\
=&d(tr(X^TAY))\\
=&tr(d(X^TAY))\\
=&tr(X^TAdY)\\
=&tr((A^TX)^TdY)
\end{aligned}
$$

同 1.3.1 公式 1：
$\therefore\ \frac{\partial}{\partial Y}\operatorname{tr}\ (X^TAY)=A^TX$

1.4 多元正态分布的等高面

1.4.1 证明概率密度等高面是椭球面

由题意：$X\sim\mathcal{N}(\mu,\Sigma)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$
因此对于 $f(x;\mu,\Sigma)=a$ 我们可以做以下变形：

$$
\begin{aligned}
\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)=a\\
\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)=(2\pi)^{p/2}|\Sigma|^{1/2}a\\
(x-\mu)^T\Sigma^{-1}(x-\mu)=-2\ln(2\pi)^{p/2}|\Sigma|^{1/2}a
\end{aligned}
$$

根据椭球的定义：形如

$$\left\{x|(x-x_c)^TP^{-1}(x-x_c)\le 1\right\}$$

的集合称为椭球。
因为 $a>0$，所以可以得出 $(x-\mu)^T\Sigma^{-1}(x-\mu)=-2\ln(2\pi)^{p/2}|\Sigma|^{1/2}a$ 是椭球。
$\therefore$ 概率密度等高面 $f(x;\mu,\Sigma)=a$ 是椭球面

1.4.2 求椭圆方程、长轴、短轴

由题意：$X\sim\mathcal{N}(\mu,\Sigma)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$
因此，任给 a>0，我们可以取 $a_0=(2\pi)^{p/2}|\Sigma|^{1/2}$
当 $0<a<\frac{1}{a_0}$ 时，$f(x;\mu,\Sigma)=a\Leftrightarrow(x-\mu)^T\Sigma^{-1}(x-\mu)=b^2$, 其中 $b^2=-2\ln\left[a(2\pi)^{p/2}|\Sigma|^{1/2}\right]=-2\ln\left[aa_0\right]$
$\because \Sigma>0$ 且 $p=2$
$\therefore \Sigma$ 的特征值记为 $\lambda_1\ge\lambda_2>0$
$\therefore \lambda_i$ 对应的特征向量记为 l_1,l_2
$\therefore \Sigma^{-1}$ 的谱分解记为 $\Sigma^{-1}=\sum_{i=1}^{2}\frac{1}{\lambda_i}l_il_i'$
令 $y_i=(x-\mu)'l_i(i=1,2)$
$\therefore$ 椭球面方程为：

$$
\begin{aligned}
(x-\mu)'\Sigma^{-1}(x-\mu) = (x-\mu)'\sum_{i=1}^{2}\frac{1}{\lambda_i}l_i l_i'(x-\mu) = b^2 \\
\Leftrightarrow \quad \sum_{i=1}^{2}\frac{1}{\lambda_i}y_i^2 = b^2 \\
\Leftrightarrow \frac{y_1^2}{\lambda_1 b^2} + \frac{y_2^2}{\lambda_2 b^2} = 1
\end{aligned}
$$

而当 p=2 且 $\Sigma = \sigma^2\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} (\rho > 0)$ 时
我们可以得到：$|\Sigma| = \sigma^4\left(1-\rho^2\right)$
接下来计算 Σ 的特征值：

$$
\begin{aligned}
& |\Sigma - \lambda I_p| \\
=& \begin{vmatrix} \sigma^2-\lambda & \sigma^2\rho \\ \sigma^2\rho & \sigma^2-\lambda \end{vmatrix} \\
=& \left(\sigma^2-\lambda\right)^2 - \sigma^4\rho^2 \\
=& \left(\sigma^2-\lambda-\sigma^2\rho\right)\left(\sigma^2-\lambda+\sigma^2\rho\right) \\
=& 0
\end{aligned}
$$

可得 Σ 的特征值为 $\lambda_1 = \sigma^2(1+\rho), \lambda_2 = \sigma^2(1-\rho)$
通过正交化和归一化得到对应的特征向量为：

$$
l_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \quad l_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}
$$

由刚刚计算得到椭球面方程为：

$$
\frac{y_1^2}{\lambda_1 b^2} + \frac{y_2^2}{\lambda_2 b^2} = 1
$$

其中 $b^2 = -2\ln\left[a(2\pi)|\Sigma|^{1/2}\right] = -2\ln\left[2\pi\sigma^2\sqrt{1-\rho^2 a}\right]$
$\therefore$ 长轴为：$2b\sigma\sqrt{1+\rho}$，短轴为：$2b\sigma\sqrt{1-\rho}$

1.5 多元正态分布的简单性质

1.5.1 计算 $\mathbb{E}[x] = \mu$

由题意得：

$$
p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)
$$

因此，对于 $\mathbb{E}[x]$:

$$
\begin{aligned}
\mathbb{E}[x] &= \int_{\mathbb{R}^n} xp(x \mid \mu, \Sigma)dx \\
&= \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) xdx
\end{aligned}
$$

作变量代换，令 $z = x - \mu$:

$$
\begin{aligned}
\mathbb{E}[x] &= \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}z^T\Sigma^{-1}z\right)(z+\mu)dz \\
&= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \left(\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}z^T\Sigma^{-1}z\right) zdz + \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}z^T\Sigma^{-1}z\right) \mu dz\right)
\end{aligned}
$$

由于 Σ 是正定对称矩阵
$\therefore \Sigma = Q^T\Lambda Q$，$\Sigma^{-1} = Q^T\Lambda^{-1}Q$，$t = Qz$
$\therefore$ 第一项可以解出为奇函数 $\int_{\mathbb{R}^n} zdz$, 由奇函数积分性质可得其积分值为 0
而第二项可以看出其为正态分布概率密度函数在 R^n 中的积分，因此可得最后的结果为:

$$
\int_{\mathbb{R}^n} \mu dz = \mu
$$

$\therefore \mathbb{E}[x] = \int_{\mathbb{R}^n} xp(x \mid \mu, \Sigma)dx = \mu$

1.5.2 计算 $\mathbb{E}\left[xx^T\right] = \Sigma + \mu\mu^T$

根据协方差矩阵的定义可得:

$$
\operatorname{Cov}\left(x, x^T\right) = \mathbb{E}\left[(x-\mu)\left(x^T - \mu^T\right)\right]
$$

$$
\operatorname{Cov}\left(x, x^T\right) = D(X) = \Sigma
$$

所以，对于 $\operatorname{Cov}\left(x, x^T\right)$ 进行展开:

$$
\begin{aligned}
&\operatorname{Cov}\left(x, x^T\right) \\
=&\mathbb{E}\left[(x-\mu)\left(x^T-\mu^T\right)\right] \\
=&\mathbb{E}\left[xx^T\right] - \mathbb{E}\left[x\mu^T\right] - \mathbb{E}\left[\mu x^T\right] + \mathbb{E}\left[\mu\mu^T\right] \\
=&\mathbb{E}\left[xx^T\right] - \mu\mu^T - \mu\mu^T + \mu\mu^T \\
=&\mathbb{E}\left[xx^T\right] - \mu\mu^T
\end{aligned}
$$

结合 $\operatorname{Cov}\left(x, x^T\right) = \Sigma$，可得:

$$
\Sigma = \mathbb{E}\left[xx^T\right] - \mu\mu^T
$$

$\therefore \mathbb{E}\left[xx^T\right] = \Sigma + \mu\mu^T$

1.5.3 证明 $\mathbb{E}\left[x^{\mathrm{T}} A x\right]=\operatorname{tr}(\Sigma A)+\mu^{\mathrm{T}} A \mu$

由矩阵迹的性质：$x^T Ax = tr(x^T Ax) = tr(xx^T A)$
$\because \mathbb{E}[tr(x)] = tr(\mathbb{E}[x])$
$\therefore$ 对 $\mathbb{E}[x^T Ax]$ 进行展开变形：

$$\begin{aligned}
&\mathbb{E}[x^T Ax] \\
=&\mathbb{E}[tr(x^T Ax)] \\
=&\mathbb{E}[tr(xx^T A)] \\
=&tr(\mathbb{E}[xx^T A])
\end{aligned}$$

$\because$ A 为对称矩阵
$\therefore tr(\mathbb{E}[xx^T A]) = tr(A\mathbb{E}[xx^T])$
由 1.5.2 可得:$\mathbb{E}\left[xx^T\right] = \Sigma + \mu\mu^T$
所以对 $\mathbb{E}[x^T Ax]$ 继续进行计算：

$$\begin{aligned}
&\mathbb{E}[x^T Ax] \\
=&tr(A\mathbb{E}[xx^T]) \\
=&tr(A(\Sigma + \mu\mu^T)) \\
=&tr(A\Sigma) + tr(A\mu\mu^T) \\
=&tr(A\Sigma) + tr(\mu^T A\mu) \\
=&tr(A\Sigma) + \mu^T A\mu
\end{aligned}$$

$\therefore \mathbb{E}\left[x^{\mathrm{T}} A x\right]=\operatorname{tr}(\Sigma A)+\mu^{\mathrm{T}} A \mu$

1.5.4 证明 $\mathbb{E}\left(X^T AX\right) = \sigma^2(p-1)$

由 1.5.3 可得：

$$\mathbb{E}\left[x^{\mathrm{T}} A x\right]=\operatorname{tr}(\Sigma A)+\mu^{\mathrm{T}} A \mu$$

当 $\mu = a\mathbf{1}_p, \quad A = I_p - \frac{1}{p}\mathbf{1}_{p\times p}, \quad \Sigma = \sigma^2 I_p$ 时，对 $\mathbb{E}\left[x^{\mathrm{T}} A x\right]$ 进行展开：

$$\begin{aligned}
&\mathbb{E}[x^T Ax] \\
=&\operatorname{tr}(\Sigma A)+\mu^{\mathrm{T}} A \mu \\
=&tr(\sigma^2 I_p(I_p - \frac{1}{p}\mathbf{1}_{p\times p})) + (a\mathbf{1}_p)^T(I_p - \frac{1}{p}\mathbf{1}_{p\times p})(a\mathbf{1}_p)
\end{aligned}$$

由于 $\mathbf{1}_p$ 表示所有分量均为 1 的列向量，$\mathbf{1}_{p\times p}$ 表示所有元素均为 1 的矩阵
因此，对上式中后一项继续计算：

$$
\begin{aligned}
&(a\mathbf{1}_p)^T(I_p - \frac{1}{p}\mathbf{1}_{p\times p})(a\mathbf{1}_p) \\
=&a\mathbf{1}_p^T(I_p - \frac{1}{p}\mathbf{1}_{p\times p})(a\mathbf{1}_p) \\
=&(a\mathbf{1}_p^T - a\frac{1}{p} \times p\mathbf{1}_p^T)(a\mathbf{1}_p) \\
=&(a\mathbf{1}_p^T - a\mathbf{1}_p^T)(a\mathbf{1}_p) \\
=&0
\end{aligned}
$$

根据迹运算的性质，对前一项继续计算：

$$
\begin{aligned}
&tr(\sigma^2 I_p(I_p - \frac{1}{p}\mathbf{1}_{p\times p})) \\
=&tr(\sigma^2 I_p - \frac{\sigma^2}{p}\mathbf{1}_{p\times p}) \\
=&\sigma^2 tr(I_p) - \sigma^2 \\
=&\sigma^2 p - \sigma^2 \\
=&\sigma^2(p-1)
\end{aligned}
$$

$\therefore \mathbb{E}\left(X^TAX\right) = \sigma^2(p-1)$

1.6 多元正态分布的极大似然估计

1.6.1 推导 μ 和 Σ 的极大似然估计

题目给出了 p 元正态分布如下：

$$
N_p(X \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)
$$

所以对于数据 $\{X_i\}$，数据的极大似然函数为：

$$
\begin{aligned}
\mathcal{L} = \prod_{i=1}^{n} p\left(X_i \mid \theta\right) = \prod_{i=1}^{n} N_p(X \mid \mu, \Sigma) \\
= \frac{1}{(2\pi)^{n^2/2}|\Sigma|^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(X_i-\mu\right)^T\Sigma^{-1}\left(X_i-\mu\right)\right)
\end{aligned}
$$

两边取偏导可得：

$$
\ln\mathcal{L} = const - \frac{n}{2}\ln det\Sigma - -\frac{1}{2}\sum_{i=1}^{n}\left(X_i-\mu\right)^T\Sigma^{-1}\left(x_i-\mu\right)
$$

对 μ 求偏导可得：

$$
\frac{\partial \ln\mathcal{L}}{\partial\mu} = -\Sigma^{-1}\sum_{i=1}^{n}\left(X_i-\mu\right)
$$

令梯度 $\frac{\partial \ln \mathcal{L}}{\partial \mu}=0$，可得：

$$\mu=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}$$

已知：

$$\frac{\partial \ln \det \Sigma}{\partial \Sigma}=\Sigma^{-1}$$

$$d\Sigma^{-1}=\Sigma^{-1}d\Sigma\Sigma^{-1}$$

对 Σ 求偏导得：

$$\frac{\partial \ln \mathcal{L}}{\partial \Sigma}=\Sigma^{-1}-\Sigma^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)^T\Sigma^{-1}$$

令梯度 $\frac{\partial \ln \mathcal{L}}{\partial \Sigma}=0$，可得：

$$\Sigma=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)^T=A$$

1.6.2 证明 $\overline{X}\sim N\left(\mu,\frac{1}{n}\Sigma\right)$

要证明 $\overline{X}\sim N\left(\mu,\frac{1}{n}\Sigma\right)$，即证明 $\overline{X}$ 服从正态分布，且 μ 和 $\frac{1}{n}\Sigma$ 分别为该正态分布的期望和方差

首先由于正态分布的线性性，且 $X\sim N\left(\mu,\Sigma\right)$，$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_{(t)}$，则易得 $\overline{X}$ 依然服从正态分布

接下来证明 μ 和 $\frac{1}{n}\Sigma$ 分别为该正态分布的期望和方差：

$$\begin{aligned}&\mathbb{E}\left(\overline{X}\right)=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\\=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i\right)\\=&\frac{1}{n}\sum_{i=1}^{n}\mu\\=&\mu\end{aligned}$$

$$\begin{aligned}&\mathbb{D}\left(\overline{X}\right)=\mathbb{D}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\\=&\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}\left(X_i\right)\\=&\frac{1}{n^2}\sum_{i=1}^{n}\Sigma\\=&\frac{1}{n}\Sigma\end{aligned}$$

$\therefore \overline{X}\sim N\left(\mu,\frac{1}{n}\Sigma\right)$

1.6.3 证明 $\overline{X}$ 是 μ 的无偏估计

根据题目中给出的无偏估计定义：当样本满足 $\mathbb{E}(\hat{\theta})=\theta$，我们就称 $\hat{\theta}$ 是 θ 的无偏估计
根据 1.6.2 中给出的：

$$\begin{aligned}\mathbb{E}\left(\overline{X}\right)=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\\ =&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i\right)\\ =&\frac{1}{n}\sum_{i=1}^{n}\mu\\ =&\mu\end{aligned}$$

$\therefore\ \overline{X}$ 是 μ 的无偏估计

1.7 范数的严格凸性

证明. ($\Rightarrow$):
已知 $\|x+y\|=\|x\|+\|y\|(\forall x\neq\theta,y\neq\theta)$
要证明 $x=cy$，可转化为证明 $\frac{x}{\|x\|}=\frac{y}{\|y\|}$
$\because$ 单位向量 $\frac{x}{\|x\|}=\frac{y}{\|y\|}=1$ 且 $\|x\|+\|y\|>0$
$\therefore \|x\|+\|y\|\neq 0$，且 $\frac{\|x\|}{\|x+y\|}+\frac{\|y\|}{\|x+y\|}=1$
假设 $\frac{x}{\|x\|}\neq\frac{y}{\|y\|}$
由范数的严格凸性，可得：

$$\left(\frac{\|x\|}{\|x+y\|}\right)\left(\frac{\|x\|}{\|x\|}\right)+\left(\frac{\|y\|}{\|x+y\|}\right)\left(\frac{\|y\|}{\|y\|}\right)<1$$

即 $\|\frac{x+y}{\|x+y\|}\|<1$，与已知条件矛盾
$\therefore \frac{x}{\|x\|}=\frac{y}{\|y\|}$，即 $x=\left(\frac{\|x\|}{\|y\|}\right)y$
$\therefore x=cy$

证明. ($\Leftarrow$):
已知 $x=cy$，要证 $\|x+y\|=\|x\|+\|y\|(\forall x\neq\theta,y\neq\theta)$，即证范数是严格凸的
设 $x\neq y$ 且 $\|x\|=\|y\|=1$，α，$\beta\in(0,1)$ 且 $\alpha+\beta=1$
$\therefore$ 根据不等式性质可得：

$$\|\alpha x+\beta y\|\leqslant\|\alpha x\|+\|\beta y\|=\alpha\|x\|+\beta\|y\|=\alpha+\beta=1$$

设 $\|\alpha x+\beta y\|=1$
$\because x=cy$
$\therefore \alpha x=c\left(\beta y\right)$
$\therefore \|\alpha x\|=\|c\left(\beta y\right)\|$
$\because \|x\|=\|y\|=1$

$\therefore \alpha = c\beta$
$\therefore x = y$，与 $x \neq y$ 矛盾
$\therefore \|\alpha x + \beta y\| < 1$，即范数是严格凸的。

1.8 证明：凸函数的局部极小值必然是全空间的极小值

设 x_1，x_2 都是局部极小点，并且 $x_1 \neq x_2$
假设 $\varphi(x_1) > \varphi(x_2)$
$\because$ 不等式 $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)(\forall 0 \leq \lambda \leq 1)$ 成立
$\therefore \varphi(x_1) > (1-\lambda)\varphi(x_1) + (\lambda)\varphi(x_2) \geqslant \varphi(\lambda x_1 + (1-\lambda)x_2)$
令 $y_\lambda = (1-\lambda)x_1 + \lambda x_2$
当 $\lambda \to 0$ 的时候，$\lim_{\lambda\to 0} = x_1$
$\therefore \varphi(x_1) > \varphi(y_\lambda)$，与 x_1 为局部极小点矛盾
$\therefore \varphi(x_1) \geqslant \varphi(x_2)$
同理可得：$\varphi(x_1) \leqslant \varphi(x_2)$
$\therefore \varphi(x_1) = \varphi(x_2)$
$\therefore$ 凸函数的局部极小值必然是全空间的极小值

1.9 投影的最优性条件

1.9.1 计算数据 $\boldsymbol{x}_i$ 在 $\boldsymbol{L}$ 的投影所对应的点 $\tilde{\boldsymbol{x}}$

由题意得：定义 $\tilde{\boldsymbol{x}}_i = \arg\min\limits_{y\in L}\|\boldsymbol{x}_i - \boldsymbol{y}\|$ 为 $\boldsymbol{x}$ 在 $\boldsymbol{L}$ 上的投影，以 $\boldsymbol{w}$ 为方向向量，过点 $\overline{\boldsymbol{x}}$ 的直线为 $\boldsymbol{L} = \{\boldsymbol{y} \in \mathbb{R}^n \mid \boldsymbol{y} = \overline{\boldsymbol{x}} + t\boldsymbol{w}, t \in \mathbb{R}\}$

因此要求 $\boldsymbol{x}$，可转化为求 $\min\limits_t \|\boldsymbol{x_i} - \overline{\boldsymbol{x}} - t\boldsymbol{w}\|_2$

$$\begin{aligned}&\|\boldsymbol{x_i} - \overline{\boldsymbol{x}} - t\boldsymbol{w}\|_2^2\\ =&t^2 - 2t\boldsymbol{w}^T(\boldsymbol{x_i} - \overline{\boldsymbol{x}}) + \|\boldsymbol{x_i} - \overline{\boldsymbol{x}}\|_2^2\\ =&\left(t - \boldsymbol{w}^T(\boldsymbol{x_i} - \overline{\boldsymbol{x}})\right)^2 + constant\end{aligned}$$

$\therefore$ 最优解 $t^* = \boldsymbol{w}^T(\boldsymbol{x_i} - \overline{\boldsymbol{x}})$
$\because \boldsymbol{w}$ 不是单位向量
$\therefore\ t^* = \frac{\boldsymbol{w}^T(\boldsymbol{x_i} - \overline{\boldsymbol{x}})}{\boldsymbol{w}^T\boldsymbol{w}}$
$\therefore \boldsymbol{x_i}$ 在 $\boldsymbol{L}$ 上的投影所对应的点:

$$\tilde{\boldsymbol{x}} = \overline{\boldsymbol{x}} + t^*\boldsymbol{w} = \overline{\boldsymbol{x}} + \frac{\boldsymbol{w}^T(\boldsymbol{x_i} - \overline{\boldsymbol{x}})\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}}$$

1.9.2 对于目标函数 $\min\limits_w J(\boldsymbol{w})$ 求最优点满足的最优性条件

由 1.9.1 可得：

$$\tilde{\boldsymbol{x}} = \overline{\boldsymbol{x}} + t^*\boldsymbol{w} = \overline{\boldsymbol{x}} + \frac{\boldsymbol{w}^T(\boldsymbol{x_i} - \overline{\boldsymbol{x}})\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}}$$

$\because \|\boldsymbol{w}\| = 1$，即 $\boldsymbol{w}$ 是单位向量

$\therefore \min\limits_{\boldsymbol{w}} J(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|^2 = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x_i} - \overline{\boldsymbol{x}} + \boldsymbol{w}^T(\boldsymbol{x_i} - \overline{\boldsymbol{x}})\boldsymbol{w}\|$
令 $\boldsymbol{y_i} = \boldsymbol{x_i} - \overline{\boldsymbol{x}}$
即 $J(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} = \|\boldsymbol{y_i} - \boldsymbol{w}^T\boldsymbol{y_i}\boldsymbol{w}\|$
所以由范数的计算性质:

$$\begin{aligned}J(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} &= \|\boldsymbol{y_i} - \boldsymbol{w}^T\boldsymbol{y_i}\boldsymbol{w}\|\\ =&\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{y_i} - \boldsymbol{w}^T\boldsymbol{y_i}\boldsymbol{w}\right)^T\left(\boldsymbol{y_i} - \boldsymbol{w}^T\boldsymbol{y_i}\boldsymbol{w}\right)\\ =&\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{y_i}^T\boldsymbol{y_i} - \boldsymbol{y_i}^T\boldsymbol{w}^T\boldsymbol{y_i}\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{y_i}^T\boldsymbol{w}\boldsymbol{y_i} + \boldsymbol{w}^T\boldsymbol{y_i}^T\boldsymbol{y_i}\boldsymbol{w}\right)\end{aligned}$$

$\because \|\boldsymbol{w}\| = 1$
$\therefore \boldsymbol{y_i}^T\boldsymbol{w}^T\boldsymbol{y_i}\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{y_i}^T\boldsymbol{w}\boldsymbol{y_i} = \boldsymbol{w}^T\boldsymbol{y_i}^T\boldsymbol{y_i}\boldsymbol{w}$
因此，对 $J(\boldsymbol{w})$ 进行进一步的化简:

$$\begin{aligned}J(\boldsymbol{w}) &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{y_i}^T\boldsymbol{y_i} - \boldsymbol{w}^T\boldsymbol{y_i}^T\boldsymbol{y_i}\boldsymbol{w}\right)\\ =&\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y_i}^T\boldsymbol{y_i} - \boldsymbol{w}^T\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y_i}^T\boldsymbol{y_i}\right)\boldsymbol{w}\end{aligned}$$

记 $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y_i}^T\boldsymbol{y_i}$ 为矩阵 Y
由瑞利商的定义和性质:

$$\lambda_{\min}(\mathbf{A}) = \min_{\boldsymbol{x}^T\boldsymbol{x}=1}\boldsymbol{x}^T\mathbf{A}\boldsymbol{x} = \min_{\boldsymbol{x}\neq 0}\frac{\boldsymbol{x}^T\mathbf{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}$$

可得该最优点的最优性条件为矩阵 Y 的最小特征值

1.9.3 对于目标函数 $\max\limits_{w} J(\boldsymbol{w})$ 求最优点满足的最优性条件

由 1.9.1 可得:

$$\tilde{\boldsymbol{x}} = \overline{\boldsymbol{x}} + t^*\boldsymbol{w} = \overline{\boldsymbol{x}} + \frac{\boldsymbol{w}^T(\boldsymbol{x_i} - \overline{\boldsymbol{x}})\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}}$$

$\because \|\boldsymbol{w}\| = 1$，即 $\boldsymbol{w}$ 是单位向量
$\therefore \boldsymbol{w}^T\boldsymbol{w} = 1$
已知目标函数:

$$\max_{\boldsymbol{w}} K(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}(t_i^*)^2$$

对目标函数进行展开并化简:

$$
\begin{aligned}
&\max_{\boldsymbol{w}} K(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\boldsymbol{w}^T\left(\boldsymbol{x_i}-\overline{\boldsymbol{x}}\right)\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}}\right)^2 \\
=&\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{w}^T\left(\boldsymbol{x_i}-\overline{\boldsymbol{x}}\right)\boldsymbol{w}\right)^2 \\
=&\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{w}^T\left(\boldsymbol{x_i}-\overline{\boldsymbol{x}}\right)\boldsymbol{w}\right)^T\left(\boldsymbol{w}^T\left(\boldsymbol{x_i}-\overline{\boldsymbol{x}}\right)\boldsymbol{w}\right) \\
=&\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{w}^T\left(\boldsymbol{x_i}-\overline{\boldsymbol{x}}\right)\left(\boldsymbol{x_i}-\overline{\boldsymbol{x}}\right)^T\boldsymbol{w} \\
=&\boldsymbol{w}^T\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x_i}-\overline{\boldsymbol{x}}\right)\left(\boldsymbol{x_i}-\overline{\boldsymbol{x}}\right)^T\boldsymbol{w}
\end{aligned}
$$

记 $\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x_i}-\overline{\boldsymbol{x}}\right)\left(\boldsymbol{x_i}-\overline{\boldsymbol{x}}\right)^T$ 为矩阵 X
由瑞利商的定义和性质:

$$
\lambda_{\max}(\mathbf{A}) = \max_{\boldsymbol{x}^T\boldsymbol{x}=1}\boldsymbol{x}^T\mathbf{A}\boldsymbol{x} = \max_{\boldsymbol{x}\neq 0}\frac{\boldsymbol{x}^T\mathbf{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}
$$

可得该最优点的最优性条件为矩阵 X 的最大特征值

1.10 KL 散度的简单介绍

1.10.1 验证对称性

已知:

$$
\mathrm{D_{KL}}(p\|q) = \int_{\mathbb{R}^n} p(x)(\log p(x) - \log q(x))dx
$$

$$
\mathrm{D_{KL}}(q\|p) = \int_{\mathbb{R}^n} q(x)(\log q(x) - \log p(x))dx
$$

所以:

$$
\begin{aligned}
&\mathrm{D_{KL}}(p\|q) - \mathrm{D_{KL}}(q\|p) \\
=&\int_{\mathbb{R}^n}\left[p(x)(\log p(x)-\log q(x)) - q(x)(\log q(x)-\log p(x))\right]dx \\
=&\int_{\mathbb{R}^n}\left[(p(x)+q(x))\left(\log p(x)-\log q(x)\right)\right]dx
\end{aligned}
$$

$\because \log p(x) - \log q(x)$ 和 $p(x)+q(x)$ 都不一定为 0
$\therefore \mathrm{D_{KL}}(p\|q)$ 和 $\mathrm{D_{KL}}(q\|p)$ 在 $\forall p, q$ 的条件下不一定相等
$\therefore$ KL 散度不一定满足对称性。

1.10.2 验证非负性

根据对数的运算性质：
$\mathrm{D_{KL}}(p\|q)=\int_{\mathbb{R}^n}p(x)(\log p(x)-\log q(x))dx=\int_{\mathbb{R}^n}p(x)(-\ln\frac{q(x)}{p(x)})dx$
令 $f(x)=-\ln x$，令 $g(x)=\frac{q(x)}{p(x)}$
由 $f(x)$ 为凸函数，应用 Jensen 不等式：

$$\int_{\mathbb{R}^n}p(x)f(g(x))dx\geqslant f(\int_{\mathbb{R}^n}p(x)\frac{q(x)}{p(x)}dx)=f(\int_{\mathbb{R}^n}q(x)dx)$$

由 q 为概率分布可得：$\int_{\mathbb{R}^n}q(x)dx=1$
$\therefore \int_{\mathbb{R}^n}p(x)f(g(x))dx\geqslant f(1)=-\ln 1=0$，即 $\mathrm{D_{KL}}(p\|q)\geqslant 0$
$\therefore$ KL 散度有非负性。

2 编程题

2.1 Gauss-Jordan 消元法

带帐号名的 AC 截图：

Figure 1: 带帐号名的 AC 截图.

源代码：

```
#include<cstdio>
#include<cmath>

const double EPS=1E-8;
double B[100][100];
int n;

int main(){
    scanf("%d",&n);
    for (int i=0;i<n;i++)
    {
        for (int j=0;j<n;j++)
            scanf("%lf",&B[i][j]);
        scanf("%lf",&B[i][n]);
    }
    for (int i=0;i<n;i++)
    {
        int pivot=i;
        for (int j=i;j<n;j++)
            if (fabs(B[j][i]-B[pivot][i])<=EPS)//寻找主元
                pivot=j;
        for (int j=0;j<=n;j++)//将主元所在行换到第一行
        {
            double t=B[i][j];
            B[i][j]=B[pivot][j];
            B[pivot][j]=t;
        }
        if (fabs(B[i][i])<=EPS){//如果该位置系数等于零，则0x=a，一定无解
            printf("No Solution\n");
            return 0;
        }
```

Figure 2: 源代码 1

```
        for (int j=i+1;j<=n;j++) //将主元所在行的所有元素全部除于主元的值
            B[i][j]/=B[i][i];
        for (int j=0;j<n;j++)//对于其他行进行化简
            if (i!=j)
                for (int k=i+1;k<=n;k++)
                    B[j][k]-=B[j][i]*B[i][k];
    }
    for (int i=0;i<n;i++)
        printf("%.2lf\n",B[i][n]);
    return 0;
}
```

Figure 3: 源代码 2

2.2 凸包的简单应用

2.2.1 伪代码

Algorithm 1: Magic Rabit

Data: 三瓶药水 x,y,z 中材料 a 含量 a_1, a_2, a_3，材料 b 含量 b_1, b_2, b_3，合成药水中材料 a 含量 m，材料 b 含量 n

Result: 是否能配出含有 mg 材料 a，ng 材料 b 的药水

首先根据题意，我们可以得出以下两个方程：

$a_1x + a_2y + a_3(1-x-y) = m = (x+y+1-x-y)m$

$b_1x + b_2y + b_3(1-x-y) = n = (x+y+1-x-y)n$

经过化简可得：

$(a_1-m)x + (a_2-m)y + (a_3-m)(1-x-y) = 0$

$(b_1-n)x + (b_2-n)y + (b_3-n)(1-x-y) = 0$

根据式子的形式，我们可以借助空间几何内积点乘和外积叉乘的定义进行解题：

设 $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}$ 向量，其中：

$\boldsymbol{a} = (a_1-m, a_2-m, a_3-m)^T$ $\boldsymbol{b} = (b_1-n, b_2-n, b_3-n)^T$ $\boldsymbol{q} = (x, y, 1-x-y)^T$

因此，我们将两个式子看成：$<\boldsymbol{a}, \boldsymbol{q}>= \boldsymbol{a}^T\boldsymbol{q} = 0 < \boldsymbol{b}, \boldsymbol{q}>= \boldsymbol{b}^T\boldsymbol{q} = 0$

我们可以得出向量 $<\boldsymbol{a}$ 和向量 $\boldsymbol{q}>$ 正交，向量 $\boldsymbol{b}$ 和向量 $\boldsymbol{q}$ 正交，所以我们可以得到向量 $\boldsymbol{q}$ 垂直于向量 $<\boldsymbol{a}$ 和向量 $\boldsymbol{b}$ 形成的平面。

那么根据外积的定义，我们可以得到：

$$\boldsymbol{a} \times \boldsymbol{b} = \begin{bmatrix} [3]x & y & 1-x-y \\ a_1-m & a_2-m & a_3-m \\ b_1-n & b_2-n & b_3-n \end{bmatrix} = ((a_2-m)(b_3-n)-(b_2-n)(a_3-m))x - ((a_1-m)(b_3-n)-(b_1-n)(a_3-m))y + ((a_1-m)(b_2-n)-(b_1-n)(a_2-m))(1-x-y)$$

为了方便表示：

$num1 \leftarrow (a_2-m)(b_3-n)-(b_2-n)(a_3-m)$

$num2 \leftarrow -(a_1-m)(b_3-n)+(b_1-n)(a_3-m)$

$num3 \leftarrow (a_1-m)(b_2-n)-(b_1-n)(a_2-m)$

为了使药水含量有意义，我们必须保证 num1,num2,num3 都 >0 或者都 <0。

for $i \leftarrow 1$ **to** *count* **do**

- $num1 \leftarrow (a_2-m)(b_3-n)-(b_2-n)(a_3-m)$;
- $num2 \leftarrow -(a_1-m)(b_3-n)+(b_1-n)(a_3-m)$;
- $num3 \leftarrow (a_1-m)(b_2-n)-(b_1-n)(a_2-m)$;
- **if** $num1 \geqslant 0$ *and* $num2 \geqslant 0$ *and* $num3 \geqslant 0$ **then**
  - YES;
- **else if** $num1 \leqslant 0$ *and* $num2 \leqslant 0$ *and* $num3 \leqslant 0$ **then**
  - YES;
- **else** NO;

end

2.2.2 C++ 源代码

```cpp
#include <iostream>
using namespace std;
bool func(int a[4], int b[4], int m, int n)
{
    int num1 = (a[2] - m) * (b[3] - n) - (a[3] - m) * (b[2] - n);
    int num2 = -(a[1] - m) * (b[3] - n) + (a[3] - m) * (b[1] - n);
    int num3 = (a[1] - m) * (b[2] - n) - (a[2] - m) * (b[1] - n);
    if (num1 >= 0 && num2 >= 0 && num3 >= 0)
        return true;
    if (num1 <= 0 && num2 <= 0 && num3 <= 0)
        return true;

    return false;
}

int main()
{
    int a[4], b[4];

    for (int i = 1; i <= 3; i++)
        cin >> a[i] >> b[i];

    int m, n, count;
    cin >> count;
```

Figure 4: 源代码 1

```cpp
int main()
{
    int a[4], b[4];

    for (int i = 1; i <= 3; i++)
        cin >> a[i] >> b[i];

    int m, n, count;
    cin >> count;

    for (int i = 0; i < count; i++)
    {
        cin >> m >> n;

        if (func(a, b, m, n))
            cout << "YES" << endl;
        else
            cout << "NO" << endl;
    }

    return 0;
}
```

Figure 5: 源代码 2