第十一讲

王伟文　暨南大学

1 线性代数回顾

定理 1.1: 对称实矩阵谱分解

任意 n 阶实矩阵 $\boldsymbol{A}$ 可以表示为

$$\boldsymbol{A} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T,$$

其中 $(\lambda_i, \boldsymbol{u}_i)$ 为矩阵 $\boldsymbol{A}$ 的特征值及相应的正交化单位特征向量, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

令 $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r] \in \mathbb{R}^{n\times r}$, $\Lambda = \mathtt{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$, 则

$$\boldsymbol{A} = \boldsymbol{U}\Lambda\boldsymbol{U}^T.$$

从优化的视角看特征值

$$\begin{aligned}
\lambda_1 &= \langle \boldsymbol{A}\boldsymbol{u}_1, \boldsymbol{u}_1\rangle = \max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{u}_1 = \arg\max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \\
\lambda_2 &= \langle \boldsymbol{A}\boldsymbol{u}_2, \boldsymbol{u}_2\rangle = \max_{\substack{\boldsymbol{x}\perp\boldsymbol{u}_1 \\ \|\boldsymbol{x}\|=1}} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{u}_2 = \arg\max_{\substack{\boldsymbol{x}\perp\boldsymbol{u}_1 \\ \|\boldsymbol{x}\|=1}} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \\
\lambda_3 &= \langle \boldsymbol{A}\boldsymbol{u}_3, \boldsymbol{u}_3\rangle = \max_{\substack{\boldsymbol{x}\perp\{\boldsymbol{u}_1,\boldsymbol{u}_2\} \\ \|\boldsymbol{x}\|=1}} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{u}_3 = \arg\max_{\substack{\boldsymbol{x}\perp\{\boldsymbol{u}_1,\boldsymbol{u}_2\} \\ \|\boldsymbol{x}\|=1}} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \\
&\cdots \\
\lambda_n &= \langle \boldsymbol{A}\boldsymbol{u}_n, \boldsymbol{u}_n\rangle = \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{u}_n = \arg\min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}
\end{aligned}$$

命题 1.1

对任意 n-阶对称矩阵 $\boldsymbol{A}$, 有

$$\begin{aligned}
\lambda_1 &= \langle \boldsymbol{A}\boldsymbol{u}_1, \boldsymbol{u}_1\rangle = \max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{u}_1 = \arg\max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \\
\lambda_n &= \langle \boldsymbol{A}\boldsymbol{u}_n, \boldsymbol{u}_n\rangle = \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{u}_n = \arg\min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}
\end{aligned}$$

证明. 由定理 1.1 知, 矩阵 $\boldsymbol{A}$ 可表示谱分解的形式 $\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, 故

$$
\begin{aligned}
\max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} &= \max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{x} \\
&= \max_{\|\boldsymbol{y}\|^2=1} \boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{y} = \max_{\|\boldsymbol{y}\|^2=1} \sum_{i=1}^n y_i^2\lambda_i \quad (\boldsymbol{y}=\boldsymbol{U}^T\boldsymbol{x}) \\
&= \max_{\xi\in\Delta_n} \sum_{i=1}^n \xi_i\lambda_i = \lambda_1 \quad (y_i^2=\xi)
\end{aligned}
$$

余下结论显然成立. □

定理 1.2: 奇异值分解 (Singular Value Decomposition)

$\forall \boldsymbol{A}\in\mathbb{R}^{m\times n}$, 可以表示为

$$\boldsymbol{A}=\sum_{i=1}^r \sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T$$

其中 $\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_r>0$ 为奇异值, $\boldsymbol{u}_i$ 为左奇异向量, $\boldsymbol{v}_i$ 为右奇异向量, 满足

$$\sigma_i=\sqrt{\lambda_i(\boldsymbol{A}\boldsymbol{A}^T)}=\sqrt{\lambda_i(\boldsymbol{A}^T\boldsymbol{A})};$$

$$\{\boldsymbol{u}_i\}\text{为矩阵}\boldsymbol{A}\boldsymbol{A}^T\text{的正交化特征向量};$$

$$\{\boldsymbol{v}_i\}\text{为矩阵}\boldsymbol{A}^T\boldsymbol{A}\text{的正交化特征向量}.$$

且 $r=\texttt{rank}(\boldsymbol{A})$.

奇异值分解的矩阵形式

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T,$$

其中 $\boldsymbol{U}\in\mathbb{R}^{m\times n}$, $\boldsymbol{V}\in\mathbb{R}^{n\times n}$ 为正交矩阵, 分别为矩阵 $\boldsymbol{A}\boldsymbol{A}^T$ 和 $\boldsymbol{A}^T\boldsymbol{A}$ 的正交化特征向量构成的列矩阵, $\boldsymbol{\Sigma}\in\mathbb{R}^{m\times n}$ 的对角元为奇异值 $(\sigma_1,\sigma_2,\ldots,\sigma_r)$, 其他位置为 0.

紧凑形式

$$\boldsymbol{A}=\boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r^T,$$

其中 $\boldsymbol{U}_r\in\mathbb{R}^{m\times r}$ 和 $\boldsymbol{V}_r\in\mathbb{R}^{n\times r}$ 分别为 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的前 r 列, 满足 $\boldsymbol{U}_r^T\boldsymbol{U}_r=\boldsymbol{V}_r^T\boldsymbol{V}_r=\boldsymbol{I}_r$, $\boldsymbol{\Sigma}_r=\texttt{diag}(\sigma_1,\sigma_2,\ldots,\sigma_r)$.

从优化的视角看奇异值

$$
\begin{aligned}
\sigma_1 &= \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|_2, \quad \boldsymbol{v}_1 = \arg\max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|_2 \\
\sigma_2 &= \max_{\substack{\boldsymbol{x}\perp\boldsymbol{v}_1 \\ \|\boldsymbol{x}\|=1}} \|\boldsymbol{A}\boldsymbol{x}\|_2, \quad \boldsymbol{v}_2 = \arg\max_{\substack{\boldsymbol{x}\perp\boldsymbol{v}_1 \\ \|\boldsymbol{x}\|=1}} \|\boldsymbol{A}\boldsymbol{x}\|_2 \\
\sigma_3 &= \max_{\substack{\boldsymbol{x}\perp\{\boldsymbol{v}_1,\boldsymbol{v}_2\} \\ \|\boldsymbol{x}\|=1}} \|\boldsymbol{A}\boldsymbol{x}\|_2, \quad \boldsymbol{v}_3 = \arg\max_{\substack{\boldsymbol{x}\perp\{\boldsymbol{v}_1,\boldsymbol{v}_2\} \\ \|\boldsymbol{x}\|=1}} \|\boldsymbol{A}\boldsymbol{x}\|_2 \\
&\dots \\
\sigma_n &= \min_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|_2, \quad \boldsymbol{v}_n = \arg\min_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|_2
\end{aligned}
$$

命题 1.2

设 $\boldsymbol{A}$ 为 n-阶对称实矩阵，则有

$$
\sigma_1(\boldsymbol{A}) = \max_{\|\boldsymbol{x}\|=1} |\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}| = \max_{i\in[n]} |\lambda_i(\boldsymbol{A})|.
$$

2 矩阵范数

考虑矩阵 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, Frobenius 范数

$$
\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}
$$

矩阵内积

$$
\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathtt{trace}(\boldsymbol{B}^T\boldsymbol{A})
$$

$$
\|\boldsymbol{A}\|_F^2 = \langle \boldsymbol{A}, \boldsymbol{A} \rangle
$$

命题 2.1: $\|\cdot\|_F$ 的正交不变性

对任意正交矩阵 $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ 或 $\boldsymbol{V} \in \mathbb{R}^{n\times n}$，有

$$
\|\boldsymbol{U}\boldsymbol{A}\|_F = \|\boldsymbol{A}\boldsymbol{V}\|_F = \|\boldsymbol{A}\|_F
$$

证明. $\|\boldsymbol{U}\boldsymbol{A}\|_F = \sqrt{\langle \boldsymbol{U}\boldsymbol{A}, \boldsymbol{U}\boldsymbol{A} \rangle} = \sqrt{\mathtt{trace}(\boldsymbol{A}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{A})} = \sqrt{\mathtt{trace}(\boldsymbol{A}^T\boldsymbol{A})} = \|\boldsymbol{A}\|_F$. $\square$

命题 2.2

$\|\boldsymbol{A}\|_F^2 = \sum \sigma_i^2(\boldsymbol{A})$.

证明. 将矩阵 $\boldsymbol{A}$ 表示为 SVD 分解的形式, $\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$.

$$\|\boldsymbol{A}\|_F^2=\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\|\overset{(i)}{=}\|\boldsymbol{\Sigma}\|_F^2=\mathtt{trace}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma})=\sum\sigma_i^2(\boldsymbol{A}),$$

其中 (i) 由 Frobenius 范数的正交不变性得到. □

定义 2.1: 矩阵谱范数

$$\|\boldsymbol{A}\|=\max_{\|\boldsymbol{x}\|_2=1}\|\boldsymbol{A}\boldsymbol{x}\|_2=\max_{\|\boldsymbol{x}\|_2\le 1}\|\boldsymbol{A}\boldsymbol{x}\|_2=\max_{\boldsymbol{x}}\frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$$

关于矩阵谱范数有

(a) 对任意正交矩阵 $\boldsymbol{U}\in\mathbb{R}^{m\times m}$ 或 $\boldsymbol{V}\in\mathbb{R}^{n\times n}$, $\|\boldsymbol{U}\boldsymbol{A}\|=\|\boldsymbol{A}\boldsymbol{V}\|=\|\boldsymbol{A}\|$;

(b) 若矩阵 $\boldsymbol{A}$ 可逆, $\|\boldsymbol{A}^{-1}\|=\frac{1}{\sigma_n(\boldsymbol{A})}$

(c) $\|\boldsymbol{A}+\boldsymbol{B}\|\le\|\boldsymbol{A}\|+\|\boldsymbol{B}\|$

(d) $\|\boldsymbol{A}\boldsymbol{B}\|\le\|\boldsymbol{A}\|\|\boldsymbol{B}\|$ $\quad(\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|_2\le\|\boldsymbol{A}\|\|\boldsymbol{B}\boldsymbol{x}\|_2\le\|\boldsymbol{A}\|\|\boldsymbol{B}\|)$

(e) $\|\boldsymbol{A}\|\le\|\boldsymbol{A}\|_F\le\sqrt{r}\|\boldsymbol{A}\|$ $\quad(\|\boldsymbol{A}\|=\sigma_1(\boldsymbol{A})\le\sqrt{\sum_{i=1}^r\sigma_i^2(\boldsymbol{A})}=\|\boldsymbol{A}\|_F\le\sqrt{r\sigma_1^2(\boldsymbol{A})}=\sqrt{r}\|\boldsymbol{A}\|)$

这里 $r=\mathtt{rank}(\boldsymbol{A})$.

(f) $\|\boldsymbol{A}\|=\max_{\|\boldsymbol{x}\|=1\text{且}\|\boldsymbol{y}\|=1}\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x}$

$(\|\boldsymbol{z}\|_2=\max_{\|\boldsymbol{y}\|_2=1}\boldsymbol{y}^T\boldsymbol{z},\ \max_{\|\boldsymbol{x}\|=1\text{且}\|\boldsymbol{y}\|=1}\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x}=\max_{\|\boldsymbol{x}\|=1}\max_{\|\boldsymbol{y}\|_2=1}\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x}=\max_{\|\boldsymbol{x}\|_2=1}\|\boldsymbol{A}\boldsymbol{x}\|_2=\sigma_1(\boldsymbol{A})=\|\boldsymbol{A}\|)$

3 随机向量

设 X 为一个 d-维随机向量, 其均值为 $\mathbb{E}X=\boldsymbol{\mu}$.

协方差矩阵为

$$\mathtt{cov}(X)=\mathbb{E}(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^T=\mathbb{E}[XX^T]-\boldsymbol{\mu}\boldsymbol{\mu}^T$$

- $(\mathtt{cov}(X))_{i,j}=\mathbb{E}(X_i-\boldsymbol{\mu}_i)(X_j-\boldsymbol{\mu}_j)=\mathtt{cov}(X_i,X_j)$
- $(\mathtt{cov}(X))_{i,i}=\mathtt{var}(X_i)$

命题 3.1

$\mathtt{cov}(X)$ 是一个正半定矩阵.

4 主成分分析 (Principal Component Analysis, PCA)

设 $X \in \mathbb{R}^d$ 是以 0 为均值的随机向量, 协方法为 $\texttt{cov}(X) = \boldsymbol{\Sigma}$. 要找一个单位向量将 X 投影到 1 维空间中, 并使得方差最大, 即

$$\max_{\|\boldsymbol{v}\|_2=1} \mathbb{E}[(\langle X, \boldsymbol{v}\rangle)^2] = \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}$$

由矩阵特征值性质知, 最优的投影变换为 $\boldsymbol{\Sigma}$ 的最大特征对应的单位化特征向量 $\boldsymbol{v}_1(\boldsymbol{\Sigma})$, 此时

$$\lambda_1(\boldsymbol{\Sigma}) = \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}.$$

接下来寻找第二个"最佳投影方向", 满足

$$\max_{\substack{\boldsymbol{v}\perp\boldsymbol{v}_1 \\ \|\boldsymbol{v}\|_2=1}} \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}.$$

得到最优的投影变换为 $\boldsymbol{\Sigma}$ 的最二大特征对应的单位化特征向量 $\boldsymbol{v}_2(\boldsymbol{\Sigma})$, 此时

$$\lambda_2(\boldsymbol{\Sigma}) = \max_{\substack{\boldsymbol{v}\perp\boldsymbol{v}_1 \\ \|\boldsymbol{v}\|_2=1}} \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}.$$

依次类推, 协方差 $\boldsymbol{\Sigma}$ 的前 k 个特征值对应的单位正交化特征向量构成主成分分析的"主成分", 即随机变量 X 投影到 k-为空间的线性变换.

称比率

$$\frac{\sum_{i=1}^{k} \lambda_i(\boldsymbol{\Sigma})}{\sum_{j=1}^{d} \lambda_j(\boldsymbol{\Sigma})}$$

为这 k 个主成分的累积解释度.

然而在实践中, 随机变量 $X \in \mathbb{R}^d$ 的协方差往往是未知的. 给定 n 个来自与 X 同一总体的样本 $X_1, X_2, \ldots, X_n$, 样本协方差为

$$\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T.$$

此时, 主成分分析建立在 $\boldsymbol{\Sigma}_n$ 上.

一个自然的问题是, $PCA(\boldsymbol{\Sigma}_n)$ 在多大程度上近似 $PCA(\boldsymbol{\Sigma})$, 即能否确保

$$\lambda_i(\boldsymbol{\Sigma}_n) \approx \lambda_i(\boldsymbol{\Sigma}) \quad 且 \quad \boldsymbol{v}_i(\boldsymbol{\Sigma}_n) \approx \boldsymbol{v}_i(\boldsymbol{\Sigma}),$$

相应的样本量应该是多少?

我们首先来看看能否在谱范数的度量下控制 $\mathbf{\Sigma}$ 与 $\mathbf{\Sigma}_n$ 的差别. 由谱范数性质

$$\begin{aligned}\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\| &= \max_{\boldsymbol{v}\in\mathcal{S}^{d-1}} \left|\boldsymbol{v}^T(\mathbf{\Sigma}_n - \mathbf{\Sigma})\boldsymbol{v}\right| \\ &= \max_{\boldsymbol{v}\in\mathcal{S}^{d-1}} \left|\boldsymbol{v}^T\mathbf{\Sigma}_n\boldsymbol{v} - \boldsymbol{v}^T\mathbf{\Sigma}\boldsymbol{v}\right| \\ &= \max_{\boldsymbol{v}\in\mathcal{S}^{d-1}} \left|\frac{1}{n}\sum_{i=1}^n \langle X_i, \boldsymbol{v}\rangle^2 - \mathbb{E}[\langle X, \boldsymbol{v}\rangle^2]\right|\end{aligned}$$

其中 $\mathcal{S}^{d-1} \triangleq \{\boldsymbol{x} \in \mathbb{R}^d : \|x\|_2 = 1\}$

定义随机变量

$$Z(\boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^n \langle X_i, \boldsymbol{v}\rangle^2 - \mathbb{E}[\langle X, \boldsymbol{v}\rangle^2],$$

此时

$$\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\| = \max_{\boldsymbol{v}\in\mathcal{S}^{d-1}} |Z(\boldsymbol{v})|$$

命题 4.1: 谱范数与球面上的 ϵ-覆盖

对于矩阵 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, 设 $\mathcal{N}$ 为单位球面 $\mathcal{S}^{n-1}$ 的一个 ϵ-覆盖, 则

$$\|\boldsymbol{A}\| \le \frac{1}{1-\epsilon}\max_{\boldsymbol{x}\in\mathcal{N}} \|\boldsymbol{A}\boldsymbol{x}\|$$

证明. 由 ϵ-覆盖定义, $\forall \boldsymbol{v} \in \mathcal{S}^{n-1}$, $\exists \boldsymbol{x} \in \mathcal{N}$, 满足

$$\|\boldsymbol{v} - \boldsymbol{x}\|_2 \le \epsilon$$

由三角不等式及谱范数定义

$$\|\boldsymbol{A}\boldsymbol{v}\|_2 - \|\boldsymbol{A}\boldsymbol{x}\|_2 \le \|\boldsymbol{A}\boldsymbol{v} - \boldsymbol{A}\boldsymbol{x}\|_2 \le \|\boldsymbol{A}\|\|\boldsymbol{v} - \boldsymbol{x}\|_2 \le \|\boldsymbol{A}\|\epsilon$$

即对任意 $\boldsymbol{v} \in \mathcal{S}^{d-1}$, 存在 $\boldsymbol{x} \in \mathcal{N}$, 使得

$$\|\boldsymbol{A}\boldsymbol{v}\|_2 \le \|\boldsymbol{A}\|\epsilon + \|\boldsymbol{A}\boldsymbol{x}\|_2$$

因此

$$\max_{\boldsymbol{v}\in\mathcal{S}^{n-1}} \|\boldsymbol{A}\boldsymbol{v}\|_2 \le \|\boldsymbol{A}\|\epsilon + \max_{\boldsymbol{x}\in\mathcal{N}} \|\boldsymbol{A}\boldsymbol{x}\|_2$$

即

$$\|\boldsymbol{A}\| \le \epsilon\|\boldsymbol{A}\| + \max_{\boldsymbol{x}\in\mathcal{N}} \|\boldsymbol{A}\boldsymbol{x}\|_2$$

移项得到

$$\|\boldsymbol{A}\| \le \frac{1}{1-\epsilon}\max_{\boldsymbol{x}\in\mathcal{N}} \|\boldsymbol{A}\boldsymbol{x}\|$$

□

命题 4.2: 单位球面的最小覆盖数

$\forall \epsilon > 0$, 在 $\|\cdot\|_2$ 度量下有

$$\mathcal{N}(\epsilon; \mathcal{S}^{d-1}, \|\cdot\|_2) \le \left(\frac{2}{\epsilon} + 1\right)^d$$

4.1 协方差一致收敛性分析

定理 4.1: Bernstein 不等式

设 $X_1, X_2, \cdots, X_N$ 为相互独立、均值为 $\mathbb{E}[X_i] = \mu_i$ 的 sub-exponential 随机变量, 相应的参数为 (ν_i, α_i), 则有

$$\mathbb{P}\left[\sum_{i=1}^{N}(X_i - \mu_i) \ge t\right] \le \exp\left(-\frac{1}{2}\min\left(\frac{t^2}{\tilde{\nu}}, \frac{t}{\tilde{\alpha}}\right)\right) \quad \forall t > 0,$$

其中 $\tilde{\nu}^2 = \sum_{i=1}^{N} \tilde{\nu}_i^2$, $\tilde{\alpha} = \max_{i=1,\dots,N} \alpha_i$.

特别地, 若 X_i 独立同分布, Bernstein 不等式可改写为

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}[X] \ge \delta\right] \le \exp\left(-\frac{n}{2}\min\left(\frac{\delta^2}{\nu}, \frac{\delta}{\alpha}\right)\right) \quad \forall t > 0,$$

定理 4.2: 协方差估计

设 $X_i \sim N(\mathbf{0}, \mathbf{\Sigma})$, 若 $n \ge \frac{16d\log 3}{\delta^2}$ 且 $0 < \delta \le \frac{1}{2}$, 则

$$\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\| \le \delta\|\mathbf{\Sigma}\|$$

至少以概率 $1 - 2\exp(-\frac{n\delta^2}{8})$ 成立.

证明. 不失一般性, 设 $\|\mathbf{\Sigma}\| = 1$.

取任一固定 $\boldsymbol{v} \in \mathcal{S}^{d-1}$, 则 $\langle X, \boldsymbol{v}\rangle$ 是以零为均值的 Gaussian 随机变量, 其方差为

$$\sigma^2 = \mathbb{E}[\langle X, \boldsymbol{v}\rangle^2] = \mathbb{E}[\boldsymbol{v}^T X X^T \boldsymbol{v}] = \boldsymbol{v}^T \mathbf{\Sigma} \boldsymbol{v} \le \|\mathbf{\Sigma}\| = 1$$

故 $\langle X, \boldsymbol{v}\rangle/\sqrt{\sigma} \sim N(0,1)$ 且 $\langle X, \boldsymbol{v}\rangle^2/\sigma$ 是以 $(2,4)$ 为参数的 sub-exponential 随机变量.

因为 $\sigma \le 1$, 故

$$\frac{1}{n}\left|\sum_{i=1}^{n}\frac{\langle X_i, \boldsymbol{v}\rangle^2}{\sigma} - \frac{\mathbb{E}[\langle X, \boldsymbol{v}\rangle^2]}{\sigma}\right| \ge \frac{1}{n}\left|\sum_{i=1}^{n}\langle X_i, \boldsymbol{v}\rangle^2 - \mathbb{E}[\langle X, \boldsymbol{v}\rangle^2]\right|$$

因此有

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^{n}\langle X_i,\boldsymbol{v}\rangle^2-\mathbb{E}[\langle X,\boldsymbol{v}\rangle^2]\right|\geq\delta\right]\leq\mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^{n}\frac{\langle X_i,\boldsymbol{v}\rangle^2}{\sigma}-\frac{\mathbb{E}[\langle X,\boldsymbol{v}\rangle^2]}{\sigma}\right|\right]$$
$$\leq 2\exp\left(-\frac{n}{2}\min\left(\frac{\delta^2}{2},\frac{\delta}{4}\right)\right)\quad\forall\delta>0.$$

由联合界，由

$$\begin{aligned}\mathbb{P}\left[\|\boldsymbol{\Sigma}_n-\boldsymbol{\Sigma}\|\geq\delta\right]&=\mathbb{P}\left[\max_{\boldsymbol{v}\in\mathcal{S}^{d-1}}|Z(\boldsymbol{v})|\geq\delta\right]\\&\leq\bigcup_{\boldsymbol{v}\in\mathcal{S}^{d-1}}\mathbb{P}\left[|Z(\boldsymbol{v})|\geq\delta\right]\\&\leq\left(\frac{2}{\epsilon}+1\right)^d\cdot 2\exp\left(-\frac{n}{2}\min\left(\frac{\delta^2}{2},\frac{\delta}{4}\right)\right)\quad\forall\epsilon>0,\delta>0.\end{aligned}$$

设 δ 充分小，$\epsilon=1$，若 $n\geq\frac{16d\log 3}{\delta^2}$

$$\mathbb{P}\left[\|\boldsymbol{\Sigma}_n-\boldsymbol{\Sigma}\|\geq\delta\right]\leq 2\exp(-\frac{n\delta^2}{8}).$$

此时

$$\|\boldsymbol{\Sigma}_n-\boldsymbol{\Sigma}\|\leq\delta$$

至少以概率 $1-2\exp(-\frac{n\delta^2}{8})$ 成立. □

4.2 扰动理论

定理 4.3: Weyl 不等式

任意 n-阶级对称矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{T}$，有

$$\max_i|\lambda_i(\boldsymbol{S})-\lambda_i(\boldsymbol{T})|\leq\|\boldsymbol{S}-\boldsymbol{T}\|$$

定理 4.4: Davis-Kahan 不等式

设有任意 n-阶级对称矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{T}$，$\lambda_i(\boldsymbol{S})$ 为 $\boldsymbol{S}$ 的特征值从大到小排列次序为 i 的特征值，记 $\lambda_i(\boldsymbol{S})$ 与余下特征值的间隙为

$$\Delta_i\triangleq\min_{j:j\neq i}|\lambda_i(\boldsymbol{S})-\lambda_j(\boldsymbol{S})|>0,$$

则存在 $\theta\in\{-1,+1\}$，使得 $\boldsymbol{S}$ 和 $\boldsymbol{T}$ 中次序为 i 的特征值对应的单位化特征向量 $\boldsymbol{v}_i(\boldsymbol{S})$ 和 $\boldsymbol{v}_i(\boldsymbol{T})$ 满足

$$\|\boldsymbol{v}_i(\boldsymbol{S})-\theta\boldsymbol{v}_i(\boldsymbol{T})\|_2\leq\frac{2\sqrt{2}\|\boldsymbol{S}-\boldsymbol{T}\|}{\Delta_i}$$

总结：PCA 有较好的拟合效果要求

- $n=\mathcal{O}(d)$;

- $\mathbf{\Sigma}_n$ 有较好分离度的特征值间隙.