Chapter 8: Principal Component Analysis in high dimensions:

8.1 V_{j-1} denotes a collection of subspaces of dimension $j-1$.

$$\max_{u \in V^{\perp} \cap S^{d-1}} \langle u, Qu \rangle = \underset{v \in S^{d-1}}{\operatorname{argmax}} \langle u, Qu \rangle$$

$$= \underset{v \in S^{d-1}}{\operatorname{argmax}} \mathbb{E}\left[\langle u, Qu \rangle^2\right]$$

$$= \sum_{j=1}^{n} \gamma_j(Q)(u_j \otimes u_j) \quad \text{where } u_j \otimes u_j = u_j^T \cdot u_j$$

$$\min_{v \in V_{j-1}} \sum_{j=1}^{n} \gamma_j(Q)(u_j \times u_j) = \gamma_j(Q)$$

$$= \min_{v \in V_{j-1}} \max_{u \in V^{\perp} \cap S^{d-1}} \langle u, Qu \rangle$$

$$\gamma_j(Q) = \min_{v \in V_{j-1}} \max_{u \in V^{\perp} \cap S^{d-1}} \langle u, Qu \rangle$$

(Unitarily Invariant)

$$|||M||| = |||VMU||| \text{ when } d_1 \le d_2 ; V_{d_1 \times d_1} ; M_{d_1 \times d_2} ; U_{d_2 \times d_2}$$

8.2

a) i) Frobenium Norm

<u>Method #1:</u>

$$\|VMU\|_F = \sqrt{\left|\sum V_i \sum M_i \sum U_i\right|^2}$$

$$\le \sqrt{\sum |V_i|^2 \sum |M_i|^2 \sum |U_i|^2}$$

$$\le \sqrt{\sum |M_i|^2} \quad \text{when } \sum |V_i|^2 = \sum |U_i|^2 = 1$$

$$\le |||M|||_F$$

<u>Method #2:</u>

$$\|VMU\|_F = \sqrt{\left|\sum V_i \sum M_i \sum U_i\right|^2}$$

$$\le \sqrt{Tr(V_i)^2 \circ Tr(M_i)^2 \circ Tr(U_i)} \le |||M|||_F$$

ii) Nuclear Norm :

$$|||VMU|||_{Nuc} = Tr(\sqrt{VMV}) = \sum_i^{min} \gamma(\sqrt{V \cdot M \cdot V})$$

"This norm describes the minimal form of a matrix"

iii) ℓ_2- operator norm :

$$|||VMU||| = \sqrt{|\sum V_i \sum M_i \sum U_i|^2}$$

$$\leq \sqrt{Tr(v_i)^2 \cdot Tr(M_i)^2 Tr(V_i)^2}$$

$$\leq \sqrt{Tr(M_i)^2}$$ when V_i and also U_i are orthonormal

$$\leq |||M|||_2^2$$

"This is a special case of the ℓ_p-norms."

iv) ℓ_∞-operator norm:

$$|||VMU|||_\infty = \sqrt[\infty]{|\sum V_i \sum M_i \sum U_i|^\infty}$$

$$\neq |||M|||_\infty^2$$

Other Method: $V^T M U = [V_1, \ldots, V_{d_1}] \begin{bmatrix} \gamma_1(M) & & 0 \\ & \ddots & \\ 0 & & \gamma_{d_1 \times d_2}(M) \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_{d_2} \end{bmatrix}$

... uses:

where $V^T V = [V_1, \ldots, V_{d_1}] \begin{bmatrix} V_1 \\ \vdots \\ V_{d_1} \end{bmatrix}$

$= 1$

$U^T U = [u_1, \ldots, U_{d_2}] \begin{bmatrix} u_1 \\ \vdots \\ u_{d_2} \end{bmatrix}$

e.g. the norms work with sums, traces, or matrices.

(Symmetric Gauge Function)

$$\rho(x_1, \ldots, x_{d_1}) = \rho(z_1 x_{\pi(1)}, \ldots, z_{d_1} x_{\pi(d_1)})$$ for all binary strings $z \in \{-1, 1\}^d$

and permutations, π on $\{1, \ldots, d_1\}$

b) $$\rho(|||M|||) = \rho(Tr(\sqrt{m}))$$
$$= \rho\left(\sum_i^{md} \gamma_i(M_i)\right)$$
$$= \rho(\gamma_1(M), \ldots, \gamma_d(m))$$
$$= \rho(\gamma(m)) \in \mathbb{R}^{d_1}$$

c) Frobenius Norm:
$$\rho(|||VMU|||_F) = \rho\left(\sqrt{\left|\sum_i^{d_1 \times d_2} V_i M_i U_i\right|^2}\right) \leq \rho\left(\sqrt{\left|\sum_i^{d_1} V_i \sum_i M_i \sum_i^{d_2} U_i\right|^2}\right) \leq \rho(|||M|||)$$

Nuclear Norm:
$$\rho(|||VMU|||_{Nuc}) = \rho(Tr(\sqrt{VMU})) \leq \rho(Tr(\sqrt{V})\,Tr(\sqrt{m})\,Tr(\sqrt{U})) \leq \rho(|||M|||)$$

l_2-norm:
$$\rho(|||VMU|||_F) = \rho\left(\sqrt{\left|\sum_i^{d_1 \times d_2} V_i M_i U_i\right|^2}\right) \leq \rho\left(\sqrt{\left|\sum_i^{d_1} V_i \sum_i M_i \sum^{d_2} U_i\right|^2}\right) \leq \rho(|||M|||)$$

l_∞-norm:
$$\rho(|||VMU|||_\infty) = \rho\left(\sqrt[\infty]{\left|\sum_i^{d_1 \times d_2} V_i M_i U_i\right|^\infty}\right) \leq \rho\left(\sqrt[\infty]{\left|\sum_i V_i \sum_i M_i \sum_i U_i\right|^\infty}\right) \leq \rho(|||M|||)$$

"The norms example a distance measure to a nearest matrix, a minimal matrix, or a Cartesian matrix"

(Weyl's Inequality)

$$\max_{j=1\ldots d} |\gamma_j(Q) - \gamma_j(R)| \leq |||Q - R|||_2$$

When $Q = R + P$

Q.3 Proof by Example: $Q = R + P$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1.01 \end{bmatrix} + \epsilon \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1.01 \end{bmatrix}$$

Eigenvalues for each matrix are:

$\gamma(Q) = \{ \frac{1}{2}((a+1) \pm \sqrt{(a-1)^2 + 4\epsilon^2})$ where $a = 1.01$

$\gamma(R) = 1, a$

$\gamma(P) = \pm \epsilon$

$$\max_{j=1..2} |\gamma(Q) - \gamma(R)| = |\tfrac{1}{2}((a-1) - \sqrt{(a-1)^2 + 4\epsilon^2}) - a|$$

$$= \tfrac{1}{2}|(1-a) - \sqrt{(a-1)^2 + 4\epsilon^2}|$$

$\leq \epsilon$ when $a > \epsilon$

Q.4. $V \in \mathbb{R}^{d \times r}$

$$\mathbb{E}\|V^T X\|_2^2 = \mathbb{E}\left[\sqrt{\left(\sum_i^{d\times r} V_i X_i\right)^2}^{\,2}\right]$$

$$= \mathbb{E}\left[\left(\sum_i^{d\times r} V_i X_i\right)^2\right]$$

$$= \sum_i^{d\times r} \mathbb{E}\left[(V_i X_i)^2\right]$$

$$= \sum_i^{d\times r} \mathbb{E}\left[\langle V, X\rangle^2\right]$$

Q.5. $\Theta \in S^{d \times d}$

$\theta \in \mathbb{R}^d$; $\{\theta^T\}_{t=0}^{\infty}$; $\theta^{t+1} = \dfrac{Q\theta^t}{\|Q\theta^t\|_2}$

Power Iteration: The largest eigenvector is a good representation for a dataset.

a) Power Iteration example proving extensive eigen vectors:

$$\theta^1 = Q\theta^0 = \begin{bmatrix}2 & -12\\ 1 & -5\end{bmatrix}\begin{bmatrix}1\\ 1\end{bmatrix} = \begin{bmatrix}-10\\ -4\end{bmatrix} \longrightarrow -4\begin{bmatrix}2.50\\ 1.00\end{bmatrix}$$

$$\theta^2 = Q\theta^1 = \begin{bmatrix}2 & -12\\ 1 & -5\end{bmatrix}\begin{bmatrix}-10\\ -4\end{bmatrix} = \begin{bmatrix}28\\ 10\end{bmatrix} \longrightarrow 10\begin{bmatrix}2.80\\ 1.00\end{bmatrix}$$

$$\theta^3 = Q\theta^2 = \begin{bmatrix}2 & -12\\ 1 & -5\end{bmatrix}\begin{bmatrix}28\\ 10\end{bmatrix} = \begin{bmatrix}-64\\ -22\end{bmatrix} \longrightarrow -22\begin{bmatrix}2.90\\ 1.00\end{bmatrix}$$

$$\theta^4 = Q\theta^3 = \begin{bmatrix}2 & -12\\ 1 & -5\end{bmatrix}\begin{bmatrix}-64\\ -22\end{bmatrix} = \begin{bmatrix}136\\ 46\end{bmatrix} \longrightarrow 46\begin{bmatrix}2.96\\ 1.00\end{bmatrix}$$

$$\theta^5 = Q\theta^4 = \begin{bmatrix}2 & -12\\ 1 & -5\end{bmatrix}\begin{bmatrix}136\\ 46\end{bmatrix} = \begin{bmatrix}-280\\ -94\end{bmatrix} \longrightarrow -94\begin{bmatrix}2.98\\ 1.00\end{bmatrix}$$

$$\theta^6 = Q\theta^5 = \begin{bmatrix}2 & -12\\ 1 & -5\end{bmatrix}\begin{bmatrix}-280\\ -94\end{bmatrix} = \begin{bmatrix}568\\ 190\end{bmatrix} \longrightarrow 190\begin{bmatrix}2.99\\ 1.00\end{bmatrix}$$

$$\vdots$$

$$\theta^d = Q\theta^{d-1} = \begin{bmatrix}2 & -12\\ 1 & -5\end{bmatrix}\begin{bmatrix}\vdots\end{bmatrix} = \begin{bmatrix}\vdots\end{bmatrix} \longrightarrow \begin{bmatrix}3.00\\ 1.00\end{bmatrix}$$

The eigenvectors converge to $\theta_n = \theta^*$ as a limit. from $t=0$ to d

Power iteration begins at unit and eigen vectors, rather than eigen values.

b) Convergence never occurs in power iteration with absolute equivalent matrix diagnols, such as:

$$Q = \begin{bmatrix}2 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 1\end{bmatrix} \quad \text{or} \quad Q = \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}$$

$$\lambda = 2, 2, 1 \qquad \lambda = 1, -1$$

In both cases, there are no "largest" absolute eigenvalues.

c) The first eigenvalue derivation is from $\lambda_1 = \dfrac{Q\theta^*\theta^*}{\theta^t \cdot \theta^t}$

$$= \frac{\begin{bmatrix}2 & -12\\ 1 & -5\end{bmatrix}\begin{bmatrix}3.00\\ 1.00\end{bmatrix}\begin{bmatrix}3.00\\ 1.00\end{bmatrix}}{\begin{bmatrix}2.99 & 1.00\end{bmatrix}\begin{bmatrix}2.99\\ 1.00\end{bmatrix}}$$

$$= -2.01$$

Successive eigenvalues appear from subtracting the "largest" eigenvalue from Q:

$$Q_1 = \begin{bmatrix} 2 & -12 \\ 1 & -5 \end{bmatrix}; \quad Q_2 = \begin{bmatrix} 2-\lambda_1 & -12-\lambda_1 \\ 1-\lambda_1 & -5-\lambda_1 \end{bmatrix}$$
$$= \begin{bmatrix} 4 & 10 \\ 3 & 3 \end{bmatrix}$$

Similar steps follow from part a,
$$\theta^1 = Q_2 \theta^0$$
$$\theta^2 = Q_2 \theta^1$$
$$\theta^3 = Q_2 \theta^2$$
$$\vdots$$
$$\theta^d = Q_2 \theta^{d-1}$$
with $\lambda_2 = \dfrac{Q_2 \theta^* \theta^*}{\theta^t \cdot \theta^t}$

0.6 Gaussian Mixture Model:

$$f(x|\theta, \sigma^2 I_d) = \alpha \phi(x|\theta^*, \sigma^2 I_d) + (1-\alpha)\phi(x|-\theta^*, \sigma^2 I_d)$$

where $\phi(x|\theta, \sigma^2 I_d)$ is a Gaussian

$\alpha \in (0,1)$; $\sigma > 0$; $\theta^* =$ mean

a) From corollary 6.20, $\dfrac{|||\hat{\Sigma} - I|||}{\sqrt{n}} \leq 1 + c'\sqrt{\dfrac{d}{n}}$ for Gaussians with $d \times n$ dimensions and $n > d$

$$\frac{|||\hat{\theta} - \theta|||}{\sqrt{n}} \leq 1 + c'\sqrt{\frac{d}{n}} \text{ where } n > c_1 \sigma^2 (1+\sigma^2) d$$

$$|||\hat{\theta} - \theta||| \leq \sqrt{n}\,(1+c')\sqrt{\frac{d}{n}}$$
$$\leq \sqrt{c_1}\,\sigma(1+\sigma^2)(1+c')\sqrt{\frac{d}{n}}$$
$$\leq c_2 \sigma (1+\sigma^2)\sqrt{\frac{d}{n}} \text{ when } c_2 = \left(\sqrt{\frac{n}{d}} + c'\right)\sqrt{c_1}$$

b) A classifier seperates at a boundary, in this case at $\hat{\theta}$.

$$X_i \begin{cases} -1 & \theta < \hat{\theta} \\ 1 & \theta > \hat{\theta} \end{cases}$$

If $X_i = -1$, $4(X_i) = -\theta^*$

else $X_i = 1$, $4(X_i) = \theta^*$

c) A Gaussian mixture model with zero variance has no covariance identity multiple.

$$T = \mathbb{E}[X \otimes X]$$
$$= \theta^* \otimes \theta^d$$
$$\neq \theta^* \otimes \theta^* + \sigma^2 I_d$$

Other mixture models exist without identity multiples, as with Binomials, Gammas and Poissons. Their repetitive moments never contain covariance.

0.7 $\theta^* \in \mathbb{R}^d$; $n = \#$ samples; $\{(X_i, y_i)\}_{i=1}^n = \{y_i + X_i\}_{i=1}^n$ where $y_i = |\langle X_i, \theta^* \rangle|$

$$= \sum_{i=1}^n \theta_i^* X_i + X_i \quad \text{where } X_i = N(0, \sigma^2 I)$$

$$Z = \sum_{i=1}^n \theta_i^* X_i + X_i$$

$$\mathbb{E}[Z] = \mathbb{E}[Z \otimes Z]$$
$$= \mathbb{E}\left[\left(\sum_{i=1}^n \theta_i^* X_i + X_i\right)\left(\sum_{i=1}^n \theta_i^* X_i + X_i\right)\right]$$
$$= \mathbb{E}\left[\sum_{i=1}^n \theta_i^* X_i \cdot \sum_{i=1}^n \theta^* X_i\right] + 2\,\mathbb{E}\left[\left(\sum \theta_i X_i\right) X_i\right] + \mathbb{E}[X_i \circ X_i]$$
$$= \theta_i \otimes \theta_i \cdot \underbrace{\mathbb{E}\left[\sum_{i=1}^n X_i \sum_{i=1}^n X_i\right]}_{\approx 1.004} + \underbrace{2 \cdot \theta\, \mathbb{E}\left[\sum X_i\right] \mathbb{E}[X_i]}_{\text{mean} = 0} + \underbrace{\mathbb{E}[X_i]^2}_{\text{variance}}$$
$$= \theta_i \times \theta_i + \sigma^2 I_d$$

8.8 Equation 8.25a : $\hat{\Theta} = \underset{\|\theta\|_1 = 1}{\arg\min} \{\langle \theta, \hat{\Sigma}\theta \rangle\}$

a) Equation 8.25b does a better job at the relationship between Scotlass's equation and a function/constraint system.

$$\hat{\Theta} = \underset{\|\theta\|_1 = 1}{\arg\max} \langle \theta, \hat{\Sigma}\theta \rangle - \lambda \|\theta\|_1$$

$$= 0$$

	Function	Bound	Type of Function	Number of Minima
Equation	$\lambda = \max_{\theta} \hat{\Sigma}\theta$	$\|\theta\|_1 = 1$	"convex"	1
Function/ Constraint	$\max_{\theta \in S} tr(\hat{\Sigma}\theta)$	$trace(\theta) = 1$	$\sum_{j,k=1}^{d} \lvert\theta_{j,k}\rvert \leq R^2$	$rank(\theta) = 1$

b) A rank constraint for a convex function defines one minimum with rank of one. A function/constraint without rank one satisfies multiple minima and a non-convex type.

8.9 $\hat{\Theta} = \underset{\substack{\theta \in S \\ tr(\theta) = 1}}{\max} \{trace(\hat{\Sigma}\Theta) - \lambda_n \sum_{j,k=1}^{d} \lvert\theta_{j,k}\rvert\}$

$$\hat{U} = \begin{cases} sign(\theta_j, \theta_k) & \text{if } \hat{\theta}_j, \hat{\theta}_k \neq 0 \\ \in [-1, 1] & \text{otherwise} \end{cases}$$

$$\hat{\Theta} = \max_{\theta \in S} \{trace(\hat{\Sigma}\theta) - \lambda_n \sum_{j,k=1}^{d} \lvert\theta_{j,k}\rvert\} = 0$$

$$= \max_{\theta \in S} \{trace(\hat{\Sigma}\theta)^2 - \lambda\}$$

$$= \max_{u \in S} \max_{\theta \in S} \hat{U}\{trace(\hat{\Sigma}\theta)^2 - \lambda\}$$

$$= \max_{u \in S} \hat{U}\{trace(\hat{\Sigma})^2\} \quad \hat{\Theta} = trace(\hat{\Sigma}\theta)^2 = trace(\hat{\Sigma})^2 = 0$$

$$= \theta \dot{x} \theta$$