3.1 (a) i $\underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{x \in C_j} \|x-\mu_j\|^2$

$\Rightarrow \underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{x \in C_j} \left(\|x\|^2 + \frac{1}{|C_j|^2}\|\sum_{x' \in C_j} x'\|^2 - \frac{2}{|C_j|} x^T(\sum_{x' \in C_j} x')\right)$

$\Rightarrow \underset{C}{\operatorname{argmin}} \left[\sum_{j=1}^{k} \sum_{x \in C_j} \|x\|^2 + \sum_{j=1}^{k} \sum_{x \in C_j} \frac{1}{|C_j|^2}\|\sum_{x' \in C_j} x'\|^2 - \sum_{j=1}^{k} \sum_{x \in C_j} \frac{2}{|C_j|} x^T(\sum_{x' \in C_j} x')\right]$

$\Rightarrow \underset{C}{\operatorname{argmin}} \left[\sum_{j=1}^{k} \sum_{x \in C_j} \|x\|^2 - \sum_{j=1}^{k} \frac{1}{|C_j|}\|\sum_{x \in C_j} x'\|^2\right]$

$\Rightarrow \underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \left(\sum_{x \in C_j} \|x\|^2 - \frac{1}{|C_j|} \sum_{x, x' \in C_j} x^T x'\right)$

$\underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{x, x' \in C_j} \|x-x'\|^2$

$\Rightarrow \underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{x, x' \in C_j} (\|x\|^2 + \|x'\|^2 - 2x^T x')$

$\Rightarrow \underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \frac{1}{2|C_j|} \left(\sum_{x \in C_j} \sum_{x' \in C_j} \|x\|^2 + \sum_{x \in C_j} \sum_{x' \in C_j} \|x'\|^2 - 2\sum_{x, x' \in C_j} x^T x'\right)$

$\Rightarrow \underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \left(\sum_{x \in C_j} \|x\|^2 - \frac{1}{|C_j|} \sum_{x, x' \in C_j} x^T x'\right)$

So the k-means clustering problem is equivalent to minimizing the pairwise squared deviation between points in the same cluster.

---

ii Based on (i), $\underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{x \in C_j} \|x-\mu_j\|^2 \Rightarrow \underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \left(\sum_{x \in C_j} \|x\|^2 - \frac{1}{|C_j|} \sum_{x, x' \in C_j} x^T x'\right)$

For a given data set, $\sum_{j=1}^{k} \sum_{x \in C_j} \|x\|^2$ is a constant, so the problem equivalent to

$\underset{C}{\operatorname{argmax}} \sum_{j=1}^{k} \sum_{x, x' \in C_j} \frac{1}{|C_j|} x^T x'$, and define $|C|$ is the size of data X.

$\underset{C}{\operatorname{argmax}} \sum_{i=1}^{k} \sum_{j=1}^{k} |C_i||C_j| \|\mu_i - \mu_j\|^2$

$\Rightarrow \underset{C}{\operatorname{argmax}} \sum_{i=1}^{k} \sum_{j=1}^{k} |C_i||C_j| (\|\mu_i\|^2 + \|\mu_j\|^2 - 2\mu_i^T \mu_j)$

$\Rightarrow \underset{C}{\operatorname{argmax}} \left[2\sum_{i=1}^{k} \sum_{j=1}^{k} |C_i||C_j| \|\mu_i\|^2 - 2\sum_{i=1}^{k} \sum_{j=1}^{k} |C_i||C_j| \mu_i^T \mu_j\right]$

$\Rightarrow \underset{C}{\operatorname{argmax}} \left[2|C| \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{x, x' \in C_i} x^T x' - 2\sum_{i=1}^{k} \sum_{j=1}^{k} (\sum_{x \in C_i} x)^T (\sum_{x' \in C_j} x')\right]$

$\Rightarrow \underset{C}{\operatorname{argmax}} 2|C| \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{x, x' \in C_i} x^T x'$

$\Rightarrow \underset{C}{\operatorname{argmax}} \sum_{j=1}^{k} \sum_{x, x' \in C_j} \frac{1}{|C_j|} x^T x'$

And $\sum_{i=1}^{k} \sum_{j=1}^{k} (\sum_{x \in C_i} x)^T (\sum_{x' \in C_j} x')$

$= (\sum_{i=1}^{k} \sum_{x \in C_i} x)^T (\sum_{j=1}^{k} \sum_{x' \in C_j} x')$

$= \|\sum_{x \in C} x\|^2$

which is a constant

So the k-means clustering problem is equivalent to maximizing the between-cluster sum of squares.

(b) i. Using Lloyd's algorithm, there are two steps in each iteration.

① In the relabel step, $c^{(i)}_{new} = \arg\min_j \|x^{(i)} - \mu_j\|^2$.

For any x in $\mathcal{X}$, exists $\|x - \mu_{c^{(i)}_{new}}\|^2 \le \|x - \mu_{c^{(i)}_{old}}\|^2$.

Thus after relabeling, J does not increase.

② In the re-center step, by solve the problem

$\min_{\mu_j} \sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2$, the solution is $\mu_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)}=j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)}=j\}}$

there exists $\sum_{i=1}^{m} \|x^{(i)} - \mu^{new}_{c^{(i)}}\|^2 \le \sum_{i=1}^{m} \|x^{(i)} - \mu^{old}_{c^{(i)}}\|^2$

Because in every sub-step of Lloyd's algorithm's steps, the distortion J does not increase, it can prove that J does not increase in each step of Lloyd's algorithm.

---

ii. The algorithm always converges.

Proof: Based on (i), since every step the distortion decrease, Lloyd's algorithm will never move to a new configuration of clusters assignments unless the new configuration has a lower distortion. Since there are finite number of possible assignments, each with a corresponding unique minimum distortion with regard to $\{\mu_j\}_{j=1}^k$. The Lloyd's algorithm will converge after a finite number of steps, when no changes to the cluster assignment will decrease the distortion, the $\{\mu_j\}_{j=1}^k$ also don't change.

3.2 (a) i. $u^T Cov(x) u = u^T \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T] u$

$= \mathbb{E}[u^T (x - \mathbb{E}[x])(x - \mathbb{E}[x])^T u]$

$= \mathbb{E}[\|(x - \mathbb{E}[x])^T u\|^2] \ge 0$

ii. $tr(Cov(x)) = tr(\mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T])$

$= \mathbb{E}[tr((x - \mathbb{E}[x])(x - \mathbb{E}[x])^T)]$

$= \mathbb{E}[tr((x - \mathbb{E}[x])^T (x - \mathbb{E}[x]))]$

$= \mathbb{E}[(x - \mathbb{E}[x])^T (x - \mathbb{E}[x])]$

$= \mathbb{E}[\|x - \mathbb{E}[x]\|^2]$

(b) Define $X = [x^{(1)} - \hat{\mu}, \dots, x^{(m)} - \hat{\mu}]$, then

$\hat{C} = XX^T$

$rank(\hat{C}) = rank(X)$.

There exists $\sum_{i=1}^{m} (x^{(i)} - \hat{\mu}) = 0$, so the column vectors are not linear independent.

$rank(X) \le \min(d, m-1)$

If $\hat{C}$ is non-singular, $rank(\hat{C}) = d$.

$\min(d, m-1) \ge d$

$m - 1 \ge d$

$m \ge d + 1$

So the minimum value of m is $(d+1)$ such that $\hat{C}$ is non-singular.

3.3 (a)

$$\underset{u:u^Tu=1}{\operatorname{argmin}} \|x^{(i)} - (x^{(i)T}u)u\|^2$$

$$= \underset{u:u^Tu=1}{\operatorname{argmin}} \left(\|x^{(i)}\|^2 - 2(x^{(i)T}u)u^Tx^{(i)} - (u^Tx^{(i)})^2 u^Tu\right)$$

$$= \underset{u:u^Tu=1}{\operatorname{argmin}} \left(\|x^{(i)}\|^2 - 2(u^Tx^{(i)})^2 + (u^Tx^{(i)})^2\right)$$

$$= \underset{u:u^Tu=1}{\operatorname{argmin}} \left(\|x^{(i)}\|^2 - (u^Tx^{(i)})^2\right)$$

$$\Rightarrow \underset{u:u^Tu}{\operatorname{argmax}} (x^{(i)T}u)^2$$

(b) Based on (a).

$$u^* = \underset{u:u^Tu=1}{\operatorname{argmax}} \frac{1}{m}\sum_{i=1}^{m}(x^{(i)T}u)^2$$

$$= \underset{u:u^Tu=1}{\operatorname{argmax}} u^T\left(\frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)T}\right)u$$

$$= \underset{u:u^Tu=1}{\operatorname{argmax}} u^T\Sigma u$$

Generalized Lagrange function of this problem is

$$L(u) = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)T}u)^2 - \beta(u^Tu-1)$$

$$\frac{\partial L}{\partial u} = 2\Sigma u - 2\beta u = 0 \Rightarrow \Sigma u^* = \beta u^*$$

So u^* is an eigenvector of Σ and β is the corresponding eigenvalue.

$$\frac{1}{m}\sum_{i=1}^{m}(x^{(i)T}u^*)^2 = u^{*T}\Sigma u^* = \beta u^{*T}u^* = \beta$$

Thus, minimizing the residual of projections is equivalent to finding the eigenvector with the largest eigenvalue.

(c).

$$\min_{u_1,\cdots,u_k:u_i^Tu_j=\delta_{ij}} \frac{1}{m}\sum_{i=1}^{m}\|x^{(i)} - \sum_{j=1}^{k}(x^{(i)T}u_j)u_j\|^2$$

$$= \min_{u_{k+1},\cdots,u_n:u_i^Tu_j=\delta_{ij}} \frac{1}{m}\sum_{i=1}^{m}\|\sum_{j=k+1}^{n}(x^{(i)T}u_j)u_j\|^2$$

$$= \min_{u_{k+1},\cdots,u_n:u_i^Tu_j=\delta_{ij}} \sum_{j=k+1}^{n} u_j^T\left(\frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)T}\right)u_j$$

$$= \min_{u_{k+1},\cdots,u_n:u_i^Tu_j=\delta_{ij}} \sum_{j=k+1}^{n} u_j^T\Sigma u_j$$

Generalized Lagrange function of this problem is

$$L(u_{k+1},\cdots,u_n) = \sum_{j=k+1}^{n} u_j\Sigma u_j - \sum_{j=k+1}^{n}\sum_{i=k+1}^{n}\beta_{ji}(u_j^Tu_i - \delta_{ij})$$

Left multiply u_i^T (for any $i \neq j$) in both sides of ①

$$2u_i^T\Sigma u_j - 2\beta_{ji}u_i^Tu_j - \beta_{ji}u_i^Tu_i = 0$$

$$\Rightarrow \beta_{ji} = 0$$

$$\frac{\partial L}{\partial u_j} = 2\Sigma u_j - 2\beta_{jj}u_j - \sum_{i=k+1,i\neq j}^{n}\beta_{ji}u_i = 0 \quad ①$$

Bring $\beta_{ji}=0$ in to ①,

$$\Sigma u_j = \beta_{jj}u_j$$

which means u_j is a eigenvector and β_{jj} is the corresponding eigenvalue.

$$\sum_{j=k+1}^{n} u_j^T\Sigma u_j = \sum_{j=k+1}^{n} u_j^T\Sigma u_j = \sum_{j=k+1}^{n}\beta_{jj}$$

$$\min_{u_1,\cdots,u_k:u_i^Tu_j=\delta_{ij}} \frac{1}{m}\sum_{i=1}^{m}\|x^{(i)} - \sum_{j=1}^{k}(x^{(i)T}u_j)u_j\|^2$$

$$\Rightarrow \min \sum_{j=k+1}^{n}\beta_{jj} = \sum_{i=k+1}^{n}\lambda_i$$

So, if want to minimize the residual of projection, must choose the $(n-k)$ eigenvectors with the smallest $(n-k)$ eigenvalues, which are $\lambda_{k+1},\cdots,\lambda_n$.

$$\min_{u_1,\cdots,u_k;u_i^Tu_j=\delta_{ij}} \frac{1}{m}\sum_{i=1}^{m}\|x^{(i)} - \sum_{j=1}^{k}(x^{(i)T}u_j)u_j\|^2 = \sum_{i=k+1}^{n}\lambda_i$$

3.4

Define $X = [x^{(1)},\cdots,x^{(m)}]^T$, $X \in \mathbb{R}^{m\times d}$, $x^{(i)} \in \mathbb{R}^d$.

The covariance matrix is

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)T} = \frac{1}{m}X^TX.$$

It's eigenvectors are the principal axes and eigenvalues are principle component variances.

At the same time,

$$eig(\Sigma) = eig(X^TX) = eig(XX^T)$$

The matrix XX^T has the same eigenvalues, and it's eigenvectors are principle components.

In the linear kernel senario, $K(x_i,x_j) = x_i^Tx_j$

Define $X' = [\phi(x^{(1)}),\cdots,\phi(x^{(m)})]^T$

$$(X'X'^T)_{ij} = \phi(x^{(i)})^T\phi(x^{(j)}) = K(x_i,x_j) = x_i^Tx_j$$

$$\therefore X'X'^T = XX^T \Rightarrow eig(X'X'^T) = eig(\Sigma)$$

which means the linear kernel PCA has the same eigenvalues and eigenvectors as conventional PCA.

3.5 (a). $A v_i = U \Sigma V^T v_i$

$= (U\Sigma)(V^T v_i)$

$= \begin{bmatrix} \sigma_1 u_1 & \sigma_2 u_2 & \cdots & \sigma_r u_r \end{bmatrix} \begin{bmatrix} v_1^T v_i \\ v_2^T v_i \\ \vdots \\ v_r^T v_i \end{bmatrix}$

$= \begin{bmatrix} \sigma_1 u_1 & \sigma_2 u_2 & \cdots & \sigma_r u_r \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ i \\ \\ r \end{matrix}$

$= \sum_{j=1}^{r} \sigma_j u_j v_j^T v_i$

$= \sigma_i u_i$

$A^T u_i = V \Sigma^T U^T u_i$

$= (V\Sigma^T)(U^T u_i)$

$= \begin{bmatrix} \sigma_1 v_1 & \cdots & \sigma_r v_r \end{bmatrix} \begin{bmatrix} u_1^T u_i \\ u_2^T u_i \\ \vdots \\ u_i^T u_i \\ \vdots \\ u_r^T u_i \end{bmatrix}$

$= \sum_{j=1}^{r} \sigma_j v_j u_j^T u_i$

$= \sigma_i v_i$

(b). $\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2}$

$= \max_x \frac{\|U\Sigma V^T x\|_2}{\|x\|_2}$ ..... because U is orthonormal

$= \max_x \frac{\|\Sigma V^T x\|_2}{\|x\|_2}$

$= \max_x \frac{\|\Sigma V^T x\|_2}{\|V^T x\|_2}$ ..... because V is orthonormal

Define $y = V^T x$

$\|A\|_2 = \max_y \frac{\|\Sigma y\|_2}{\|y\|_2}$

$= \max_y \frac{\sqrt{\sigma_1^2 y_1^2 + \cdots + \sigma_n^2 y_n^2}}{\sqrt{y_1^2 + y_2^2 + \cdots + y_r^2}}$

$= \sigma_1$