4.1 $\mathbb{E}[g^2(Y)] - \mathbb{E}[X^2]$

$$= \sum_{y\in Y} P_Y(y) g^2(y) - \sum_{x\in X} P_X(x) x^2$$

$$= \sum_{y\in Y} P_Y(y)(\mathbb{E}[X|Y=y])^2 - \sum_{x\in X}\sum_{y\in Y} P_{X|Y}(x|y) x^2$$

$$= \sum_{y\in Y} P_Y(y)(\mathbb{E}[X|Y=y])^2 - \sum_{y\in Y}\sum_{x\in X} P_{X|Y}(x|y) x^2$$

$$= \sum_{y\in Y} P_Y(y)\left(\mathbb{E}[X|Y=y]^2 - \sum_{x\in X} P_{X|Y}(x|y) x^2\right)$$

$$= \sum_{y\in Y} P_Y(y)\left(\mathbb{E}[X|Y=y]^2 - \mathbb{E}[X^2|Y=y]\right)$$

$$= \sum_{y\in Y} P_Y(y)\cdot(-\mathrm{Var}[X|Y=y])$$

$$\leq 0$$

$$\therefore \mathbb{E}[g^2(Y)] \leq \mathbb{E}[X^2]$$

4.2 (a). $\phi(x) = f(x)\phi_1$

$$\Rightarrow \phi_1 = \frac{\phi(x)}{f(x)}$$

$$= \text{constant}$$

$$= 1(x)$$

(b). $\phi(x) = f(x)\sqrt{P_X(x)}$

$$\Rightarrow f(x) = \frac{\phi(x)}{\sqrt{P_X(x)}}$$

$$\mathbb{E}[f^2(x)] = \sum_{x\in X} P_X(x) f^2(x)$$

$$= \sum_{x\in X} P_X(x)\frac{\phi^2(x)}{P_X(x)}$$

$$= \sum_{x\in X}\phi^2(x)$$

$$= \|\phi\|^2$$

(c). $\mathbb{E}[f_1(x) f_2(x)] = \sum_{x\in X}\frac{P_X(x)}{\sqrt{P_X(x)}\sqrt{P_X(x)}}\phi_1(x)\phi_2(x)$

$$= \sum_{x\in X}\phi_1(x)\phi_2(x)$$

$$= \langle\phi_1, \phi_2\rangle$$

4.3 (a). Because S_1, S_2 are independent,

$$P(S_1, S_2) = \frac{1}{2\pi}\exp\left(-\frac{S_1^2+S_2^2}{2}\right)$$

(b). Based on (a), define $S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$

$$P(S_1, S_2) = \frac{1}{2\pi}\exp\left(-\frac{S^TS}{2}\right)$$

Define $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = AS$

$$P(X_1, X_2) = P_S(A^{-1}X)|A^{-1}|$$

$$= \frac{1}{2\pi}\exp\left(-\frac{X^T(A^{-1})^TA^{-1}X}{2}\right)|A^T|$$

$$= \frac{1}{2\pi}\exp\left(-\frac{X^T(AA^{-1})X}{2}\right)$$

$$= \frac{1}{2\pi}\exp\left(-\frac{X_1^2+X_2^2}{2}\right).$$

$$= P(S_1, S_2)$$

which means the joint density is the same for any orthogonal mixing matrix. In the case of Gaussian variables, ICA can only determine the mixing matrix up to an orthogonal ~~the~~ transformation. So the Gaussian variables are forbidden.

4.4. (a)

E-step:

$$P_{\theta^{(t)}}(Z=z|X=x_i) = \frac{P_{\theta^{(t)}}(X=x_i|Z=z)\,P_{\theta^{(t)}}(Z=z)}{P_{\theta^{(t)}}(X=x_i)}$$

$$= \frac{\phi_z(t)\exp\left(-\frac{1}{2}(x_i-\mu_z^{(t)})^T\Sigma_z^{-1}(x_i-\mu_z^{(t)})\right)}{(2\pi)^{\frac{d}{2}}|\Sigma_z|^{\frac{1}{2}}P_{\theta^{(t)}}(X=x_i)}$$

$$= \frac{\phi_z(t)\exp\left(-\frac{1}{2}(x_i-\mu_z^{(t)})^T\cancel{\Sigma_z}(x_i-\mu_z^{(t)})\right)}{(2\pi)^{\frac{d}{2}}P_{\theta^{(t)}}(X=x_i)}$$

$P_{\theta^{(t)}}(X=x_i)$ is a normalization factor which ensures that $\sum_{z\in Z}P_{\theta^{(t)}}[Z=z|X=x_i]$ sums to 1.

M step:

$$\theta^{(t+1)} = \arg\max_\theta \sum_{i=1}^{m}\sum_{z\in Z}P_{\theta^{(t)}}(Z=z|X=x_i)\log\left(P_\theta[X=x_i, Z=z]\right)$$

$$= \arg\max_\theta \sum_{i=1}^{m}\sum_{z\in Z}P_{\theta^{(t)}}(Z=z|X=x_i)\log\left(P_\theta[X=x_i|Z=z]\,P_\theta(Z=z)\right)$$

$$= \arg\max_\theta \sum_{i=1}^{m}\sum_{z\in Z}P_{\theta^{(t)}}(Z=z|X=x_i)\log\left(\phi_z\cdot\exp\left(-\frac{1}{2}(x_i-\mu_z)^T(x_i-\mu_z)\right)\right)$$

$$= \arg\max_\theta \sum_{i=1}^{m}\sum_{z\in Z}P_{\theta^{(t)}}(Z=z|X=x_i)\left(\log\phi_z - \frac{1}{2}\|x_i-\mu_z\|^2\right)$$

(b). $J(\theta)=\sum_{i=1}^{m}\sum_{z\in Z} P_{\theta^{(t)}}(Z=z|X=x_i)\left(\log\phi_z-\frac{1}{2}\|x_i-\mu_z\|^2\right)$.

$$\frac{\partial J(\theta)}{\partial \mu_z}=\sum_{i=1}^{m} P_{\theta^{(t)}}(Z=z|X=x_i)(x_i-\mu_z)$$

$$=0$$

$$\therefore\ \mu_z^{(t+1)}=\frac{\sum_{i=1}^{m} P_\theta^{(t)}[Z=z|X=x_i]\,x_i}{\sum_{i=1}^{m} P_{\theta}^{(t)}[Z=z|X=x_i]}$$

To compute $\phi_z^{(t+1)}$,

$$\max_\theta J(\theta).$$

$$\text{s.t. } \sum_{z\in Z}\phi_z=1.$$

using Lagrange method.

$$L(\theta)=J(\theta)-c\left(\sum_{z\in Z}\phi_z-1\right)$$

$$\frac{\partial L}{\partial \phi_z}=\frac{\sum_{i=1}^{m} P_{\theta^{(t)}}(Z=z|X=x_i)}{\phi_z}-c=0$$

$$\Rightarrow\ \phi_z=\frac{\sum_{i=1}^{m} P_{\theta^{(t+1)}}(Z=z|X=x_i)}{c}.$$

$$\sum_{z\in Z}\phi_z=1\ \Rightarrow\ c=\sum_{i=1}^{m}\sum_{z\in Z}P_{\theta^{(t)}}(Z=z|X=x_i).$$

$$\therefore\ \phi_z^{(t+1)}=\frac{\sum_{i=1}^{m} P_\theta^{(t)}(Z=z|X=x_i)}{\sum_{z\in Z}\sum_{i=1}^{m} P_{\theta^{(t)}}(Z=z|X=x_i)}$$

Compared with K-means, they both have 2 steps. First step is to assign and second step is to update parameters. But in the first step, mixed Gaussian assign "softly" instead of just assign x to the closest cluster. In the second step, the form of the updated centroid is quite similar, but K-means do not need to compute p(z).

4.5 Define $V_1(B), V_2(B), \dots, V_n(B)$ are the corresponding eigenvectors of B, and they are orthonormal.

For any $V\in R^n$, can find $V=\sum_{i=1}^{n} x_i V_i(B)$

$$\frac{V^TBV}{\|V\|^2}=\frac{(x_1V_1(B)+\cdots+V_n(B)x_n)^T(x_1\lambda_1(B)V_1(B)+\cdots+x_n\lambda_n(B)V_n(B))}{x_1^2+x_2^2+\cdots+x_n^2}$$

$$=\frac{x_1^2\lambda_1(B)+\cdots+x_n^2\lambda_n(B)}{x_1^2+\cdots+x_n^2}$$

$$\lambda_n(B)=\frac{\lambda_n(B)(x_1^2+\cdots+x_n^2)}{x_1^2+\cdots+x_n^2}\le\frac{V^TBV}{\|V\|^2}\le\frac{\lambda_1(B)(x_1^2+\cdots+x_n^2)}{x_1^2+\cdots+x_n^2}=\lambda_1(B)$$

$$\lambda_k(A+B)=\min_U\left\{\max_V\left\{\frac{V^T(A+B)V}{\|V\|^2}\,\middle|\,V\in U \text{ and } V\neq 0\right\}\,\middle|\,\dim(U)=n+1-k\right\}$$

$$=\min_U\left\{\max_V\left\{\frac{V^TAV}{\|V\|^2}+\frac{V^TBV}{\|V\|^2}\,\middle|\,V\in U \text{ and } V\neq 0\right\}\,\middle|\,\dim(U)=n+1-k\right\}$$

$$\le\min_U\left\{\max_V\left\{\frac{V^TAV}{\|V\|^2}+\lambda_1(B)\,\middle|\,V\in U \text{ and } V\neq 0\right\}\,\middle|\,\dim(U)=n+1-k\right\}$$

$$=\lambda_k(A)+\lambda_1(B).$$

Also,

$$\cancel{\lambda_k(A+B}\ \lambda_k(A)=\lambda_k(A+B+(-B))$$

$$\le\lambda_k(A+B)+\lambda_1(-B)$$

$$=\lambda_k(A+B)-\lambda_n(B)$$

$$\therefore\ \lambda_k(A+B)\ge\lambda_k(A)+\lambda_n(B).$$

Finally, $\lambda_k(A)+\lambda_n(B)\le\lambda_k(A+B)\le\lambda_k(A)+\lambda_1(B)$