PAC - Bayes

Question 1

Consider two multivariate Gaussian distributions over $\mathbb{R}^r$ - $N(\mu_0, \Sigma_0)$ and $N(\mu_1, \Sigma_1)$ where Σ_0 and Σ_1 are non-singular (positive definite).
Then it holds that:

$$KL(N(\mu_0,\Sigma_0)\,||\,N(\mu_1,\Sigma_1)) = \frac{1}{2}\Big[tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - r + \ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)\Big]$$

Solution

First, let's sign the two distributions as:

$P = N(\mu_0, \Sigma_0)$ and $Q = N(\mu_1, \Sigma_1)$.

By definition: $KL(P||Q) = E_P\left[\ln\left(\frac{P}{Q}\right)\right]$.

Remember that the density function of a multivariate Gaussian distribution $N(\mu, \Sigma)$ is:

$$p(x) = \frac{1}{(2\pi)^{\frac{r}{2}}(\det\Sigma)^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

$\Rightarrow$

$$\frac{P(x)}{Q(x)} = P(x)\cdot\frac{1}{Q(x)} = \frac{1}{(2\pi)^{\frac{r}{2}}(\det\Sigma_0)^{\frac{1}{2}}}\cdot e^{-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)} \cdot \frac{(2\pi)^{\frac{r}{2}}(\det\Sigma_1)^{\frac{1}{2}}}{e^{-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}}$$

$$= \frac{(\det\Sigma_1)^{\frac{1}{2}}}{(\det\Sigma_0)^{\frac{1}{2}}}\cdot e^{-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) - \left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)}$$

$$= \left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)^{\frac{1}{2}}\cdot e^{\frac{1}{2}\cdot\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right]}$$

Thus

$$KL(P\|Q) = \mathbb{E}_P\left[\ln\left(\frac{P}{Q}\right)\right]$$

$$= \mathbb{E}_P\left[\ln\left(\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)^{\frac{1}{2}} e^{\frac{1}{2}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)-(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right]}\right)\right]$$

$$[\ln(ab)=\ln(a)+\ln(b)] = \mathbb{E}_P\left[\ln\left[\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)^{\frac{1}{2}}\right] + \ln\left[e^{\frac{1}{2}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)-(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right]}\right]\right]$$

$$= \mathbb{E}_P\left[\frac{1}{2}\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + \frac{1}{2}\cdot\left((x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)-(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)\right]$$

linearity of $\mathbb{E}(\cdot)$ $\searrow$

$$= \mathbb{E}_P\left[\frac{1}{2}\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)\right] + \mathbb{E}_P\left[\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right]$$
$$- \mathbb{E}_P\left[\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right]$$

$$= \frac{1}{2}\cdot\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + \mathbb{E}_P\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right] - \mathbb{E}_P\left[(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right]\right]$$

<u>claim</u>

For each $x \in \mathbb{R}^r$, $A \in \mathbb{R}^{r,r}$: $\quad x^TAx = tr(Axx^T)$

<u>proof</u>

$$x^TAx = (x_1 \cdots x_r)\begin{pmatrix} \\ Ax \\ \\ \end{pmatrix} = (x_1 \cdots x_r)\begin{pmatrix}(Ax)_1 \\ \vdots \\ (Ax)_r\end{pmatrix} = (x_1 \cdots x_r)\begin{pmatrix}\sum_{j=1}^{r} a_{1j}x_j \\ \vdots \\ \sum_{j=1}^{r} a_{rj}x_j\end{pmatrix} =$$

$$= \sum_{j=1}^{r} a_{1j}x_1x_j + \cdots + \sum_{j=1}^{r} a_{rj}x_rx_j =$$

$$= \sum_{j=1}^{r} a_{1j}(xx^T)_{j1} + \cdots + \sum_{j=1}^{n} a_{rj}(xx^T)_{jr} =$$

$$= (Axx^T)_{11} + \cdots + (Axx^T)_{rr} = tr(Axx^T)$$

$\square$

Continue ~~nex~~ next page...

Thus:

$$KL(P||Q) = \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + \mathbb{E}_P[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)] - \mathbb{E}_P[(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)]\right]$$

$$= \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + \mathbb{E}_P[tr(\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T)] - \mathbb{E}_P[tr(\Sigma_0^{-1}(x-\mu_0)(x-\mu_0)^T)]\right]$$

$tr(\mathbb{E}[A]) = \mathbb{E}[tr(A)]$ by linearity of $\mathbb{E}[\cdot]$

$$\to = \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + tr(\mathbb{E}_P[\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T]) - tr(\mathbb{E}_P[\Sigma_0^{-1}(x-\mu_0)(x-\mu_0)^T])\right]$$

$$= \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + tr(\Sigma_1^{-1}\mathbb{E}_P[xx^T - \mu_1 x^T - x\mu_1^T + \mu_1\mu_1^T]) - tr(\Sigma_0^{-1}\mathbb{E}_P[(x-\mu_0)(x-\mu_0)^T])\right] ⊜$$

By definition, $P = N(\mu_0, \Sigma_0)$ thus

$\mathbb{E}_P[x] = \mu_0$ and $\mathbb{E}_P[(x-\mu_0)(x-\mu_0)^T] = \Sigma_0$

and $\mathbb{E}_P[xx^T] = \Sigma_0 + \mu_0\mu_0^T$, so:

$$⊜ \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + tr(\Sigma_1^{-1}(\Sigma_0 + \mu_0\mu_0^T - \mu_1\mu_0^T - \mu_0\mu_1^T + \mu_1\mu_1^T)) - tr(\Sigma_0^{-1}\Sigma_0)\right]$$

$$= \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + tr(\Sigma_1^{-1}\Sigma_0) + tr(\Sigma_1^{-1}(\mu_0\mu_0^T - \mu_1\mu_0^T - \mu_0\mu_1^T + \mu_1\mu_1^T)) - tr(I_r)\right]$$

$x^TAx = Axx^T$ by claim we proved

$$\leftarrow = \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) - r + tr(\Sigma_1^{-1}\Sigma_0) + tr(\mu_0^T\Sigma_1^{-1}\mu_0 - \mu_0^T\Sigma_1^{-1}\mu_1 - \mu_1^T\Sigma_1^{-1}\mu_0 + \mu_1^T\Sigma_1^{-1}\mu_1)\right]$$

$$= \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) - r + tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0)\right]$$

$\Rightarrow$

$$KL(N(\mu_0,\Sigma_0)||N(\mu_1,\Sigma_1)) = \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) - r + tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0)\right]$$

as required.

$\square$