Kullback–Leibler divergence between multivariate Gaussians

The Kullback–Leibler divergence from distribution $p(x)$ to $q(x)$ is:

$$\mathrm{KL}(p, q) = \mathbb{E}_p\{\log(p) - \log(q)\} = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\,\mathrm{d}x$$

If $p = \mathcal{N}(\mu, \Sigma)$ and $q = \mathcal{N}(\nu, \Lambda)$ are multivariate Gaussian distributions, then

$$\begin{aligned}\log p &= -\frac{D}{2}\log\tau - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^T\Sigma(x-\mu) \\ \log q &= -\frac{D}{2}\log\tau - \frac{1}{2}\log|\Lambda| - \frac{2}{2}(x-\nu)^T\Lambda(x-\nu)\end{aligned}$$

and so

$$\mathrm{KL}(p, q) = \frac{1}{2}\log\frac{|\Sigma|}{|\Lambda|} - \frac{1}{2}\mathbb{E}_p\{(x-\mu)^T\Sigma^{-1}(x-\mu)\} + \frac{1}{2}\mathbb{E}_q\{(x-\nu)^T\Lambda^{-1}(x-\nu)\}$$

Since the terms are scalars, they are equal to their trace, allowing us to take the covariance matrices out of the expectation value:

$$\mathrm{KL}(p, q) = \frac{1}{2}\log\frac{|\Sigma|}{|\Lambda|} - \frac{1}{2}\,\mathrm{tr}\left[\mathbb{E}_p\{(x-\mu)(x-\mu)^T\}\Sigma^{-1}\right] + \frac{1}{2}\,\mathrm{tr}\left[\mathbb{E}_q\{(x-\nu)(x-\nu)^T\}\Lambda^{-1}\right]$$

In more detail each term is rewritten as follows:

$$\begin{aligned}&\mathrm{tr}\left[\mathbb{E}_p\{(x-\mu)^T\Sigma^{-1}(x-\mu)\}\right] \\ &= \mathbb{E}_p\{\mathrm{tr}[(x-\mu)^T\Sigma^{-1}(x-\mu)]\} \\ &= \mathbb{E}_p\{\mathrm{tr}[(x-\mu)(x-\mu)^T\Sigma^{-1}]\} \\ &= \mathrm{tr}\left[\mathbb{E}_p\{(x-\mu)(x-\mu)^T\Sigma^{-1}\}\right] \\ &= \mathrm{tr}\left[\mathbb{E}_p\{(x-\mu)(x-\mu)^T\}\Sigma^{-1}\right]\end{aligned}$$

Since $\mathbb{E}_p\{(x-\mu)(x-\mu)^T\} = \Sigma$ by definition, the first term becomes $\mathrm{tr}[\Sigma\Sigma^{-1}] = D$, while the second can be rewritten as

$$\begin{aligned}&\mathbb{E}_p\{(x-\nu)(x-\nu)^T\} \\ &= \mathbb{E}_p\{((x-\mu) - (\nu-\mu))((x-\mu) - (\nu+\mu))^T\} \\ &= \mathbb{E}_p\{(x-\mu)(x-\mu)^T\} - (x-\mu)(\nu-\mu) - (\nu-\mu)(x-\mu) + (\nu-\mu)(\nu-\mu)^T)) \\ &= \Sigma + (\mu-\nu)(\mu-\nu)^T\end{aligned}$$

Pulling this together,

$$\begin{aligned}\mathrm{KL}(p, q) &= \frac{1}{2}\log\frac{|\Lambda|}{|\Sigma|} - \frac{D}{2} + \frac{1}{2}\,\mathrm{tr}[(\Sigma + (\mu-\nu)(\mu-\nu)^T)]\Lambda^{-1}) \\ &= \frac{1}{2}\log\frac{|\Lambda|}{|\Sigma|} - \frac{D}{2} + \frac{1}{2}\,\mathrm{tr}[\Sigma\Lambda^{-1}] + \mathrm{tr}[(\mu-\nu)^T\Lambda^{-1}(\mu-\nu)]\end{aligned}$$

Univariate case

For $D = 1$, this result becomes:

$$\mathrm{KL}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\nu, \rho^2)) = \log\frac{\rho}{\sigma} + \frac{\sigma^2 + (\mu-\nu)^2}{2\rho^2} - \frac{1}{2}$$