1 KL divergence

$$D_{KL}(P\|Q) = \mathop{\mathbb{E}}_{x\sim P}\left[\log\left(\frac{p(x)}{q(x)}\right)\right]$$

2 Expected value

$$\mathbb{E}\left[X\right] = \int x f(x) dx$$

- **Linearity**

$$\mathbb{E}\left[X+Y\right] = \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right]$$
$$\mathbb{E}\left[aX\right] = a\,\mathbb{E}\left[X\right]$$

3 Multivariate Normal Distribution

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k\left|\boldsymbol{\Sigma}\right|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1}$$

- **log likelihood**

$$\begin{aligned}\ln(p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})) &= -\frac{k\ln\left(2\pi\right)}{2} - \frac{\ln\left(\left|\boldsymbol{\Sigma}\right|\right)}{2} - \frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\\ &= -\frac{1}{2}\left[\ln\left((2\pi)^k\left|\boldsymbol{\Sigma}\right|\right) + \left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right]\end{aligned}$$

- **KL divergence**

$$\begin{aligned}D_{KL}\left(\mathcal{N}_a\|\mathcal{N}_b\right) &= \frac{1}{2}\left(\mathrm{Tr}\left(\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\Sigma}_a\right) + \left(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a\right)\boldsymbol{\Sigma}_b^{-1}\left(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a\right) - k + \log\frac{\left|\boldsymbol{\Sigma}_b\right|}{\left|\boldsymbol{\Sigma}_a\right|}\right)\\ &= \frac{1}{2}\left(\left\|\boldsymbol{L}_b\setminus\boldsymbol{L}_a\right\|_F^2 + \left\|\boldsymbol{L}_b\setminus\left(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a\right)\right\|_2^2 - k + 2\left(\log\mathrm{diag}\left(\boldsymbol{L}_b\right)^T\underline{\boldsymbol{1}} - \log\mathrm{diag}\left(\boldsymbol{L}_a\right)^T\underline{\boldsymbol{1}}\right)\right)\end{aligned}$$

- **Affine Transform**

$$\begin{aligned}&\boldsymbol{x}\sim\mathcal{N}\left(\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\right)\\ &\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\\ &\boldsymbol{y}\sim\mathcal{N}\left(\boldsymbol{A}\boldsymbol{\mu}_x+\boldsymbol{b},\boldsymbol{A}\Sigma_x\boldsymbol{A}^T\right)\end{aligned}$$

- **Linear Gaussian systems** Given a linear system:

$$\begin{aligned}p(\boldsymbol{x}) &= \mathcal{N}\left(x\mid\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\right)\\ p(\boldsymbol{y}\mid\boldsymbol{x}) &= \mathcal{N}\left(\boldsymbol{y}\mid\boldsymbol{A}\boldsymbol{x}+b,\boldsymbol{\Sigma}_y\right)\end{aligned}$$

We have the following:

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}\left(x \mid \boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}\right)$$
$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y}\left[\boldsymbol{A}^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}-\boldsymbol{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x\right]$$
$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{A}$$

$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b}, \boldsymbol{\Sigma}_y + \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T\right)$$

- **quadratic relations**
  - $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right] = \mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$
  - $$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{b},\boldsymbol{B})}\left[(\boldsymbol{a}-\boldsymbol{A}\boldsymbol{x})^T\,\Sigma^{-1}\left(\boldsymbol{a}-\boldsymbol{A}\boldsymbol{x}\right)\right] = (\boldsymbol{a}-\boldsymbol{A}\boldsymbol{b})^T\,\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{a}-\boldsymbol{A}\boldsymbol{b}\right) + \mathrm{Tr}\left(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{B}\right)$$

4 Gamma distribution

$x \sim Ga(a,b)$ where a is called the shape and b the rate.

$$Ga(x|a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}\exp\left(-bx\right)$$

- $\mathbb{E}\left[x\right] = \dfrac{a}{b}$
- $\mathbb{E}_{x\sim Ga(a,b)}\left[\ln x\right] = \psi(a) - \ln\left(b\right)$
  where ψ is the polygamma function.

4.1 Inverse gamma

If $x \sim Ga(a,b)$ and $y = \dfrac{1}{x}$, then $y \sim IG(a,b)$

$$IG(y|a,b) = \frac{b^a}{\Gamma(a)}y^{-(a+1)}\exp\left(-b/y\right)$$

4.2 Inverse Wishart (IW)

This distribution is used in Bayesian statistics as the conjugate prior for the covariance matrix of a multivariate normal distribution:

$$\boldsymbol{\Sigma} \sim IW\left(\boldsymbol{S}^{-1}, \upsilon + D + 1\right)$$
$$IW\left(\boldsymbol{\Sigma} \mid \boldsymbol{S}, \upsilon\right) = \frac{1}{\boldsymbol{Z}\left(\boldsymbol{S}, \upsilon\right)}\left|\boldsymbol{\Sigma}\right|^{(\upsilon+D+1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{S}^{-1}\boldsymbol{\Sigma}^{-1}\right)\right)$$
$$Z\left(\boldsymbol{S}, \upsilon\right) = \left|\boldsymbol{S}\right|^{-\upsilon/2}2^{\upsilon D/2}\Gamma_D\upsilon/2$$

where $\boldsymbol{S} \succ 0$