1 (**Differential entropy, cross-entropy, KL divergence, and mutual information**) Throughout, use natural logarithms (units: nats), so $\log(e) = 1$. Let $X \in \mathbb{R}^n$ be a continuous random vector with density p_X. Assume $\mathbb{E}[X] = \mu$ and $\mathrm{Cov}(X) = \Sigma \succ 0$.

(a) Recall the definitions

$$H(X) := -\int p_X(x)\, \log p_X(x)\, dx, \qquad H(p,q) := -\int p(x)\, \log q(x)\, dx,$$

$$\mathrm{KL}(p \,\|\, q) := \int p(x)\, \log \frac{p(x)}{q(x)}\, dx.$$

Show that

$$H(p,q) = H(p) + \mathrm{KL}(p \,\|\, q).$$

(b) Let (X, Y) have joint density $p_{X,Y}$ and marginals p_X, p_Y. Define mutual information by

$$I(X;Y) := \mathrm{KL}\big(p_{X,Y} \,\|\, p_X p_Y\big).$$

Show the equivalent forms

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X).$$

(c) Let $G \sim \mathcal{N}(\mu, \Sigma)$. Compute $H(G)$.

(d) (Gaussian is the maximum-entropy distribution.) Among all random vectors X with $\mathbb{E}[X] = \mu$ and $\mathrm{Cov}(X) = \Sigma$, show that the Gaussian $\mathcal{N}(\mu, \Sigma)$ has the largest differential entropy. Equivalently, prove that for any such X,

$$H(X) \le \tfrac{1}{2} \log\big((2\pi e)^n \det \Sigma\big),$$

with equality if and only if $X \sim \mathcal{N}(\mu, \Sigma)$.

(e) (Why a constraint is necessary.) Show that differential entropy has no global maximum on $\mathbb{R}^n$: if X has a density and $a > 0$, define $X_a := aX$. Prove

$$H(X_a) = H(X) + n \log a,$$

and conclude that $\sup h(X) = +\infty$ if you do not constrain (for example) the covariance.

(f) Let $P = \mathcal{N}(\mu_0, \Sigma_0)$ and $Q = \mathcal{N}(\mu_1, \Sigma_1)$ on $\mathbb{R}^n$ with $\Sigma_0, \Sigma_1 \succ 0$.

(f1) Derive a closed form for $\mathrm{KL}(P \,\|\, Q)$.

(f2) Using (a), write the cross-entropy $H(P, Q)$ in closed form.

(g) Let $X \sim \mathcal{N}(0, \Sigma_X)$, $Z \sim \mathcal{N}(0, \Sigma_Z)$ independent, and $Y := X + Z$. Compute $I(X;Y)$ in closed form.

(a) $H(p)=-\int p(x)\log p(x)dx$, $KL(p\|q)=\int p(x)\log\frac{p(x)}{q(x)}dx=\int p(x)\left(\log p(x)-\log q(x)\right)dx$

$\Rightarrow KL(p\|q)+H(p)=\int p(x)\log p(x)-\int p(x)\log q(x)-\int p(x)\log p(x)dx=-\int p(x)\log q(x)dx=H(p,q)$

(b)

$I(X;Y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy=\iint p(x,y)\log p(x,y)dxdy-\iint p(x,y)\log p(x)dxdy-\iint p(x,y)\log p(y)dxdy$

$=-H(X,Y)-\int p(x)\log p(x)dx-\int p(y)\log p(y)dy=-H(X,Y)+H(X)+H(Y)$

$H(X,Y)=-\iint p(x,y)\log p(x,y)dxdy=-\iint p(x,y)\log\left(p(x)\cdot p(y|x)\right)dxdy=-\iint p(x,y)\log p(x)dxdy-\iint p(x,y)\log p(y|x)dxdy$

$=-\int p(x)\log p(x)dx-\iint p(x)p(y|x)\log p(y|x)dxdy=-\int p(x)\log p(x)dx-\int p(x)\left(\int p(y|x)\log(y|x)dy\right)dx=H(X)+H(Y|X)=H(Y)+H(X|Y)$

$\Rightarrow I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$

(c)

$G\sim N(\mu,\Sigma)\Rightarrow p(x)=\frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\Rightarrow\log p(G)=-\frac{n}{2}\log(2\pi)-\frac{1}{2}\log(\det\Sigma)-\frac{1}{2}(G-\mu)^T\Sigma^{-1}(G-\mu)$

$\Rightarrow H(G)=-E\log p(G)=\frac{n}{2}\log(2\pi)+\frac{1}{2}\log(\det\Sigma)+\frac{1}{2}E[(G-\mu)^T\Sigma^{-1}(G-\mu)]$

here $E[(G-\mu)^T\Sigma^{-1}(G-\mu)]=E\left[tr\left((G-\mu)^T\Sigma^{-1}(G-\mu)\right)\right]=E\left[tr\left(\Sigma^{-1}(G-\mu)(G-\mu)^T\right)\right]=tr\left(\Sigma^{-1}E[(G-\mu)(G-\mu)^T]\right)=tr(\Sigma^{-1}\Sigma)=n$

$\Rightarrow H(G)=\frac{n}{2}\log(2\pi)+\frac{1}{2}\log(\det\Sigma)+\frac{n}{2}=\frac{1}{2}\log\left((2\pi e)^n\det\Sigma\right)$

(d)

$E[X]=\mu$, $Cov(X)=\Sigma$, $KL(p_X\|p_G)=-H(X)+H(p_X,p_G)\geq 0$, $H(p_X,p_G)=-\int p_X(x)\log p_G(x)dx$

since $\log p_G(x)=C-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Rightarrow$ only depend on μ and Σ of X, so $\int p_X(x)\log p_G(x)dx=\int p_G(x)\log p_G(x)dx$

$\Rightarrow H(p_X,p_G)=H(G)\Rightarrow H(X)\leq H(G)=\frac{1}{2}\log\left((2\pi e)^n\det\Sigma\right)$

(e)

the density of X_a is $p_{X_a}(y)=\frac{1}{a^n}p_X\left(\frac{y}{a}\right)\Rightarrow H(X_a)=-\int p_{X_a}(y)\log p_{X_a}(y)dy=-\int\frac{1}{a^n}p_X\left(\frac{y}{a}\right)\log\left(\frac{1}{a^n}p_X\left(\frac{y}{a}\right)\right)dy$

here $x=\frac{y}{a}$, $dy=a^n dx\Rightarrow H(X_a)=-\int p_X(x)\left(\log p(x)-\log a^n\right)dx=-\int p_X(x)\log p(x)+n\log a\int p_X(x)dx=H(X)+n\log a$

then if $a\to\infty$, $H(X_a)\to+\infty$, $Cov(X_a)=a^2Cov(X)=a^2\Sigma$, so $\sup h(X)=+\infty$ if covariance not constrained

(f)

(f1) $KL(P\|Q)=E_P[\log P-\log Q]=E_P[\log P]-E_P[\log Q]$

$E_P[\log P]=-H(P)=-\frac{n}{2}\log(2\pi)-\frac{1}{2}\log(\det\Sigma_0)-\frac{n}{2}$

$\log Q=-\frac{n}{2}\log(2\pi)-\frac{1}{2}\log(\det\Sigma_1)-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)$

here $E_P[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)]=E_P\left[tr\left(\Sigma_1^{-1}(X-\mu_1)(X-\mu_1)^T\right)\right]=tr\left(\Sigma_1^{-1}E_P\left[(X-\mu_0+\mu_0-\mu_1)(X-\mu_0+\mu_0-\mu_1)^T\right]\right)$, here $E_P[X-\mu_0]=0$

so $E_P[(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)]=tr\left(\Sigma_1^{-1}\left[\Sigma_0+(\mu_0-\mu_1)(\mu_0-\mu_1)^T\right]\right)=tr\left(\Sigma_1^{-1}\Sigma_0\right)+(\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1)$

$\Rightarrow E_P[\log Q]=-\frac{n}{2}\log(2\pi)-\frac{1}{2}\log(\det\Sigma_1)-\frac{1}{2}\left[tr\left(\Sigma_1^{-1}\Sigma_0\right)+(\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1)\right]$

$\Rightarrow KL(P\|Q)=E_P[\log P]-E_P[\log Q]=\frac{1}{2}\left(\log\frac{\det\Sigma_1}{\det\Sigma_0}-n+tr\left(\Sigma_1^{-1}\Sigma_0\right)+(\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1)\right)$

(f2) $H(P,Q)=H(P)+KL(P\|Q)=\frac{n}{2}\log(2\pi)+\frac{1}{2}\log(\det\Sigma_0)+\frac{n}{2}+\frac{1}{2}\left(\log\frac{\det\Sigma_1}{\det\Sigma_0}-n+tr\left(\Sigma_1^{-1}\Sigma_0\right)+(\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1)\right)$

$$= \frac{1}{2}\left(n\log(2\pi) + \log(\det\Sigma_1) + \mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + \mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1)\right)$$

(g) $I(X;Y) = H(Y) - H(Y|X)$, $Y|X=x \sim (x+Z)|X=x \Longrightarrow H(Y|X) = H(Z) = \frac{1}{2}\log((2\pi e)^n \det\Sigma_Z)$

$Y \sim N(0, \Sigma_X + \Sigma_Z) \Rightarrow H(Y) = \frac{1}{2}\log((2\pi e)^n \det(\Sigma_X+\Sigma_Z))$

$\Rightarrow I(X;Y) = \frac{1}{2}\log((2\pi e)^n \det(\Sigma_X+\Sigma_Z)) - \frac{1}{2}\log((2\pi e)^n \det\Sigma_Z) = \frac{1}{2}\log\frac{\det(\Sigma_X+\Sigma_Z)}{\det\Sigma_Z} = \frac{1}{2}\log\det(I_n + \Sigma_Z^{-1}\Sigma_X)$