Appendix D

Probability Distributions

This appendix archives a number of useful results from texts by Papoulis [44], Lee [33] and Cover [12]. Table 16.1 in Cover (page 486) gives entropies of many distributions not listed here.

D.1 Transforming PDFs

Because probabilities are defined as areas under PDFs when we transform a variable

$$y = f(x) \tag{D.1}$$

we transform the PDF by preserving the areas

$$p(y)|dy| = p(x)|dx| \tag{D.2}$$

where the absolute value is taken because the changes in x or y (dx and dy) may be negative and areas must be positive. Hence

$$p(y) = \frac{p(x)}{|\frac{dy}{dx}|} \tag{D.3}$$

where the derivative is evaluated at $x = f^{-1}(y)$. This means that the function $f(x)$ must be one-to-one and invertible.

If the function is many-to-one then it's inverse will have multiple solutions $x_1, x_2, ..., x_n$ and the PDF is transformed at each of these points (Papoulis' Fundamental Theorem [44], page 93)

$$p(y) = \frac{p(x_1)}{|\frac{dy}{dx_1}|} + \frac{p(x_2)}{|\frac{dy}{dx_2}|} + ... + \frac{p(x_n)}{|\frac{dy}{dx_n}|} \tag{D.4}$$

D.1.1 Mean and Variance

For more on the mean and variance of functions of random variables see Weisberg [64].

Expectation is a *linear operator*. That is

$$E[(a_1 x + a_2 x)] = a_1 E[x] + a_2 E[x] \tag{D.5}$$

Therefore, given the function

$$y = ax \tag{D.6}$$

we can calculate the mean and variance of y as functions of the mean and variance of x.

$$\begin{aligned} E[y] &= aE[x] \\ Var(y) &= a^2 Var(x) \end{aligned} \tag{D.7}$$

If y is a function of many *uncorrelated* variables

$$y = \sum_i a_i x_i \tag{D.8}$$

we can use the results

$$E[y] = \sum_i a_i E[x_i] \tag{D.9}$$

$$Var[y] = \sum_i a_i^2 Var[x_i] \tag{D.10}$$

But if the variables are correlated then

$$Var[y] = \sum_i a_i^2 Var[x_i] + 2 \sum_i \sum_j a_i a_j Var(x_i, x_j) \tag{D.11}$$

where $Var(x_i, x_j)$ denotes the covariance of the random variables x_i and x_j.

Standard Error

As an example, the mean

$$m = \frac{1}{N} \sum_i x_i \tag{D.12}$$

of uncorrelated variables x_i has a variance

$$\begin{aligned} \sigma_m^2 \equiv Var(m) &= \sum_i \frac{1}{N} Var(x_i) \\ &= \frac{\sigma_x^2}{N} \end{aligned} \tag{D.13}$$

where we have used the substitution $a_i = 1/N$ in equation D.10. Hence

$$\sigma_m = \frac{\sigma_x}{\sqrt{N}} \tag{D.14}$$

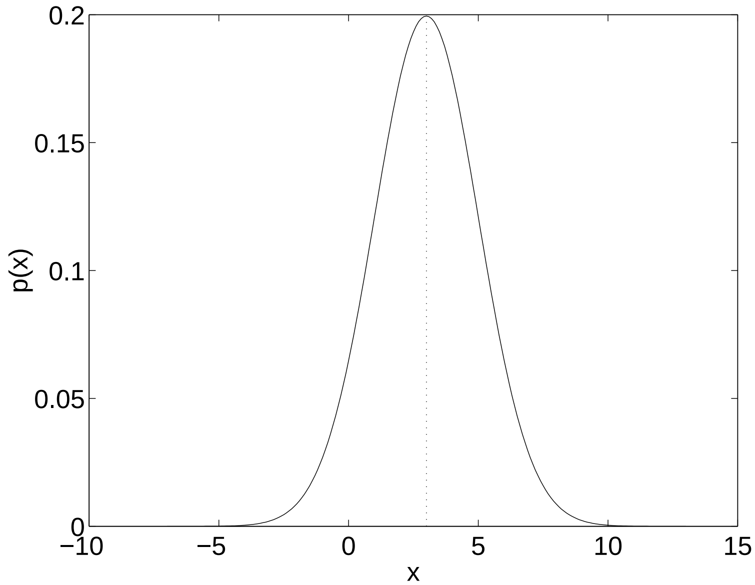

Figure D.1: *The Gaussian Probability Density Function with* $\mu = 3$ *and* $\sigma = 2$.

D.2 Uniform Distribution

The uniform PDF is given by

$$U(x; a, b) = \frac{1}{b-a} \tag{D.15}$$

for $a \leq x \leq b$ and zero otherwise. The mean is $0.5(a+b)$ and variance is $(b-a)^2/12$.

The entropy of a uniform distribution is

$$H(x) = \log(b-a) \tag{D.16}$$

D.3 Gaussian Distribution

The *Normal* or *Gaussian* probability density function, for the case of a single variable, is

$$N(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{D.17}$$

where μ and σ^2 are the mean and variance.

D.3.1 Entropy

The entropy of a Gaussian variable is

$$H(x) = \frac{1}{2}\log\sigma^2 + \frac{1}{2}\log 2\pi + \frac{1}{2} \tag{D.18}$$

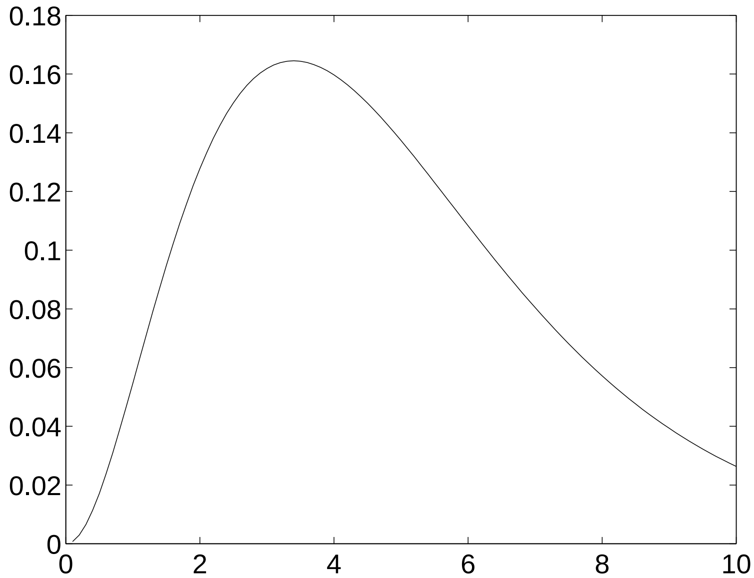

Figure D.2: *The Gamma Density for* $b = 1.6$ *and* $c = 3.125$.

For a given variance, the Gaussian distribution has the highest entropy. For a proof of this see Bishop ([3], page 240).

D.3.2 Relative Entropy

For Normal densities $q(x) = N(x; \mu_q, \sigma_q^2)$ and $p(x) = N(x; \mu_p, \sigma_p^2)$ the KL-divergence is

$$D[q||p] = \frac{1}{2} \log \frac{\sigma_p^2}{\sigma_q^2} + \frac{\mu_q^2 + \mu_p^2 + \sigma_q^2 - 2\mu_q \mu_p}{2\sigma_p^2} - \frac{1}{2} \tag{D.19}$$

D.4 The Gamma distribution

The Gamma density is defined as

$$\Gamma(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b}\right) \tag{D.20}$$

where $\Gamma()$ is the *gamma function* [49]. The mean of a Gamma density is given by bc and the variance by $b^2 c$. Logs of gamma densities can be written as

$$\log \Gamma(x; b, c) = \frac{-x}{b} + (c - 1) \log x + K \tag{D.21}$$

where K is a quantity which does not depend on x; the log of a gamma density comprises a term in x and a term in $\log x$. The Gamma distribution is only defined for positive variables.

D.4.1 Entropy

Using the result for Gamma densities

$$\int p(x) \log x = \Psi(c) + \log b \tag{D.22}$$

where $\Psi()$ is the digamma function [49] the entropy can be derived as

$$H(x) = log\Gamma(c) + c \log b - (c-1)(\Psi(c) + \log b) + c \tag{D.23}$$

D.4.2 Relative Entropy

For Gamma densities $q(x) = \Gamma(\boldsymbol{x}; b_q, c_q)$ and $p(x) = \Gamma(\boldsymbol{x}; b_p, c_p)$ the KL-divergence is

$$\begin{aligned} D[q||p] &= (c_q - 1)\Psi(c_q) - \log b_q - c_q - \log \Gamma(c_q) \\ &+ \log \Gamma(c_p) + c_p \log b_p - (c_p - 1)(\Psi(c_q) + \log b_q) + \frac{b_q c_q}{b_p} \end{aligned} \tag{D.24}$$

D.5 The χ^2-distribution

If $z_1, z_2, ..., z_N$ are independent normally distributed random variables with zero-mean and unit variance then

$$x = \sum_{i=1}^{N} z_i^2 \tag{D.25}$$

has a χ^2-distribution with N degrees of freedom ([33], page 276). This distribution is a special case of the Gamma distribution with $b = 2$ and $c = N/2$. This gives

$$\chi^2(x; N) = \frac{1}{\Gamma(N/2)} \frac{x^{N/2-1}}{2^{N/2}} \exp\left(\frac{-x}{2}\right) \tag{D.26}$$

The mean and variance are N and $2N$. The entropy and relative entropy can be found by substituting the the values $b = 2$ and $c = N/2$ into equations D.23 and D.24. The χ^2 distribution is only defined for positive variables.

If x is a χ^2 variable with N degrees of freedom and

$$y = \sqrt{x} \tag{D.27}$$

then y has a χ-density with N degrees of freedom. For $N = 3$ we have a *Maxwell* density and for $N = 2$ a *Rayleigh* density ([44], page 96).

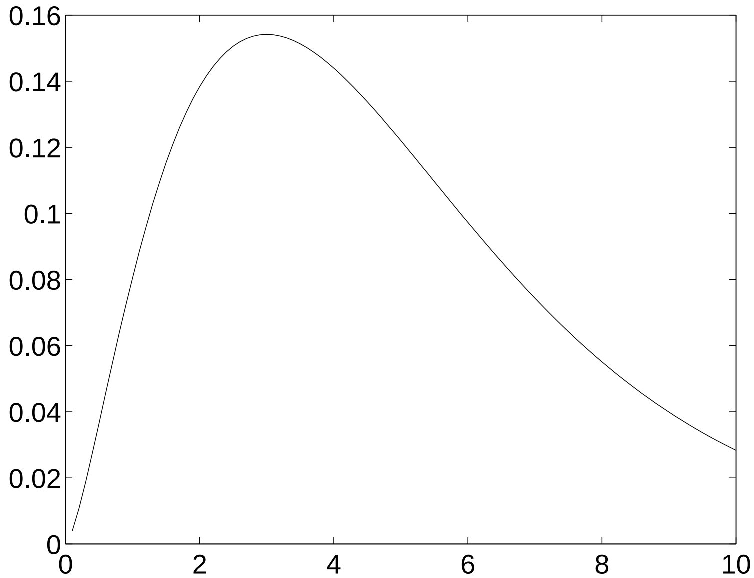

Figure D.3: *The χ^2 Density for $N = 5$ degrees of freedom.*

D.6 The t-distribution

If $z_1, z_2, ..., z_N$ are independent Normally distributed random variables with mean μ and variance σ^2 and m is the sample mean and s is the sample standard deviation then

$$x = \frac{m - \mu}{s/\sqrt{N}} \tag{D.28}$$

has a t-distribution with $N - 1$ degrees of freedom. It is written

$$t(x; D) = \frac{1}{B(D/2, 1/2)} \left(1 + \frac{x^2}{D}\right)^{-(D+1)/2} \tag{D.29}$$

where D is the number of 'degrees of freedom' and

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)} \tag{D.30}$$

is the *beta function*. For $D = 1$ the t-distribution reduces to the standard Cauchy distribution ([33], page 281).

D.7 Generalised Exponential Densities

The 'exponential power' or 'generalised exponential' probability density is defined as

$$p(a) = G(a; R, \beta) = \frac{R\beta^{1/R}}{2\Gamma(1/R)} \exp(-\beta|a|^R) \tag{D.31}$$

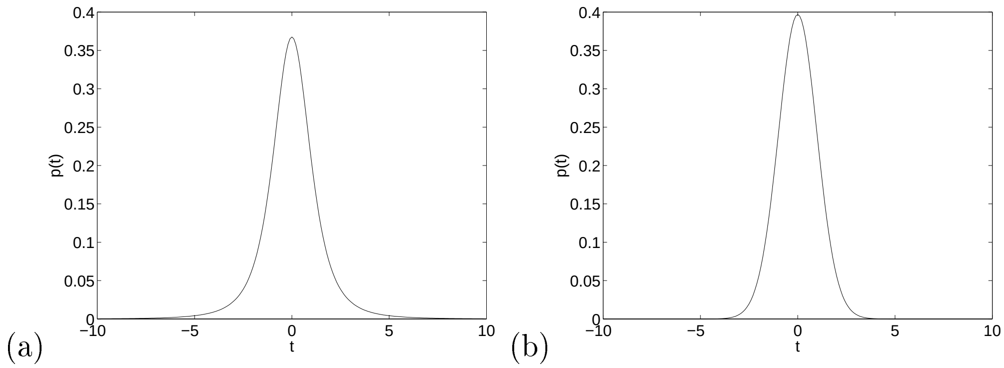

Figure D.4: *The t-distribution with (a) $N = 3$ and (b) $N = 49$ degrees of freedom.*

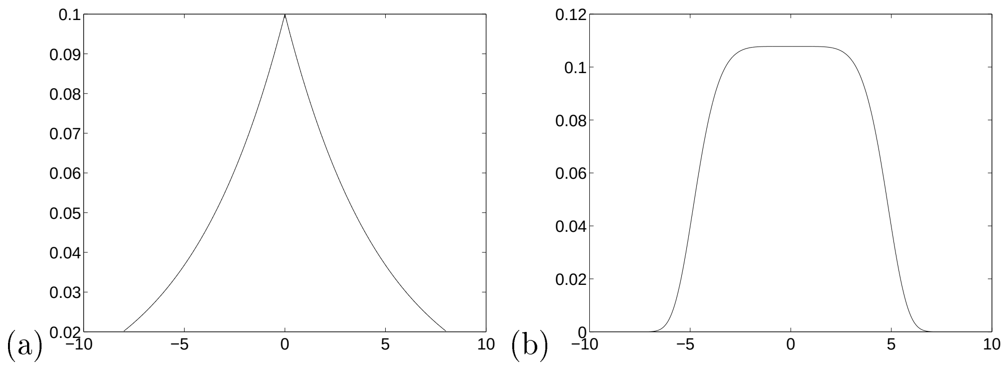

Figure D.5: *The generalised exponential distribution with (a) $R = 1, w = 5$ and (b) $R = 6, w = 5$. The parameter R fixes the weight of the tails and w fixes the width of the distribution. For (a) we have a Laplacian which has positive kurtosis ($k = 3$); heavy tails. For (b) we have a light-tailed distribution with negative kurtosis ($k = -1$).*

where $\Gamma()$ is the gamma function [49], the mean of the distribution is zero [1], the width of the distribution is determined by $1/\beta$ and the weight of its tails is set by R. This gives rise to a Gaussian distribution for $R = 2$, a Laplacian for $R = 1$ and a uniform distribution in the limit $R \to \infty$. The density is equivalently parameterised by a variable w, which defines the width of the distribution, where $w = \beta^{-1/R}$ giving

$$p(a) = \frac{R}{2w\Gamma(1/R)} \exp(-|a/w|^R) \tag{D.32}$$

The variance is

$$V = w^2 \frac{\Gamma(3/R)}{\Gamma(1/R)} \tag{D.33}$$

which for $R = 2$ gives $V = 0.5w^2$. The kurtosis is given by [7]

$$K = \frac{\Gamma(5/R)\Gamma(1/R)}{\Gamma(3/R)^2} - 3 \tag{D.34}$$

where we have subtracted 3 so that a Gaussian has zero kurtosis. Samples may be generated from the density using a rejection method [59].

[1] For non zero mean we simply replace a with $a - \mu$ where μ is the mean.

D.8 PDFs for Time Series

Given a signal $a = f(t)$ which is sampled uniformly over a time period T, its PDF, $p(a)$ can be calculated as follows. Because the signal is uniformly sampled we have $p(t) = 1/T$. The function $f(t)$ acts to transform this density from one over t to to one over a. Hence, using the method for transforming PDFs, we get

$$p(a) = \frac{p(t)}{|\frac{da}{dt}|} \tag{D.35}$$

where $||$ denotes the absolute value and the derivative is evaluated at $t = f^{-1}(x)$.

D.8.1 Sampling

When we convert an analogue signal into a digital one the sampling process can have a crucial effect on the resulting density. If, for example, we attempt to sample uniformly but the sampling frequency is a multiple of the signal frequency we are, in effect, sampling non-uniformly. For true uniform sampling it is necessary that the ratio of the sampling and signal frequencies be irrational.

D.8.2 Sine Wave

For a sine wave, $a = \sin(t)$, we get

$$p(a) = \frac{1}{|cos(t)|} \tag{D.36}$$

where $cos(t)$ is evaluated at $t = \sin^{-1}(a)$. The inverse sine is only defined for $-\pi/2 \leq t \leq \pi/2$ and $p(t)$ is uniform within this. Hence, $p(t) = 1/\pi$. Therefore

$$p(a) = \frac{1}{\pi\sqrt{1-a^2}} \tag{D.37}$$

This density is *multimodal*, having peaks at $+1$ and -1. For a more general sine wave

$$a = R\sin(wt) \tag{D.38}$$

we get $p(t) = w/\pi$

$$p(a) = \frac{1}{\pi\sqrt{1-(a/R)^2}} \tag{D.39}$$

which has peaks at $\pm R$.

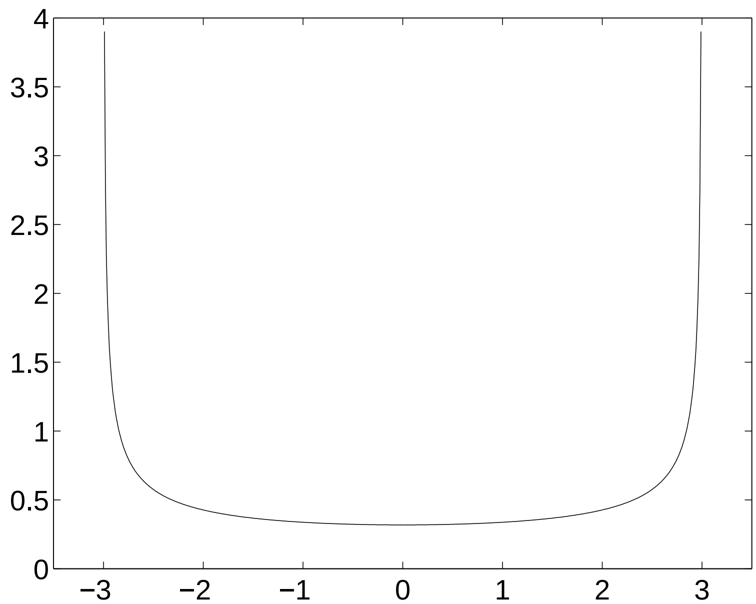

Figure D.6: *The PDF of* $a = R\sin(wt)$ *for* $R = 3$.

Appendix E

Multivariate Probability Distributions

E.1 Transforming PDFs

Just as univariate Probability Density Functions (PDFs) are transformed so as to preserve area so multivariate probability distributions are transformed so as to preserve volume. If

$$y = f(\boldsymbol{x}) \tag{E.1}$$

then this can be achieved from

$$p(\boldsymbol{y}) = \frac{p(\boldsymbol{x})}{abs(|\boldsymbol{J}|)} \tag{E.2}$$

where $abs()$ denotes the absolute value and $||$ the determinant and

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & .. & \frac{\partial y_1}{\partial x_d} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & .. & \frac{\partial y_2}{\partial x_d} \\ .. & .. & .. & .. \\ \frac{\partial y_d}{\partial x_1} & \frac{\partial y_d}{\partial x_2} & .. & \frac{\partial y_d}{\partial x_d} \end{bmatrix} \tag{E.3}$$

is the Jacobian matrix for d-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. The partial derivatives are evaluated at $\boldsymbol{x} = f^{-1}(\boldsymbol{y})$. As the determinant of $\boldsymbol{J}$ measures the volume of the transformation, using it as a normalising term therefore preserves the volume under the PDF as desired. See Papoulis [44] for more details.

E.1.1 Mean and Covariance

For a vector of random variables (Gaussian or otherwise), $\boldsymbol{x}$, with mean $\boldsymbol{\mu}_x$ and covariance $\boldsymbol{\Sigma}_x$ a linear transformation

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x} + \boldsymbol{C} \tag{E.4}$$

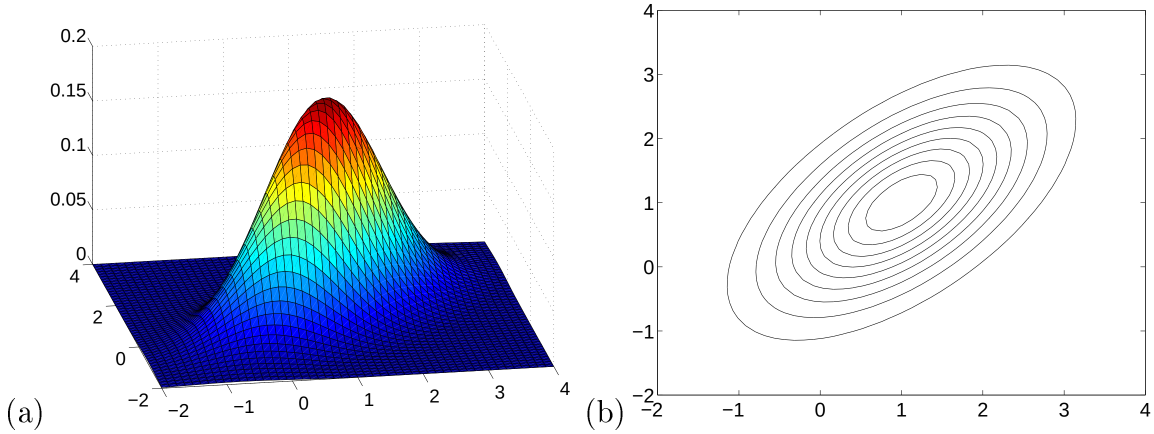

Figure E.1: *(a) 3D-plot and (b) contour plot of Multivariate Gaussian PDF with* $\boldsymbol{\mu} = [1, 1]^T$ *and* $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22} = 1$ *and* $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = 0.6$ *ie. a positive correlation of* $r = 0.6$.

gives rise to a random vector $\boldsymbol{y}$ with mean

$$\boldsymbol{\mu}_y = \boldsymbol{F}\boldsymbol{\mu}_x + \boldsymbol{C} \tag{E.5}$$

and covariance

$$\boldsymbol{\Sigma}_y = \boldsymbol{F}\boldsymbol{\Sigma}_x\boldsymbol{F}^T \tag{E.6}$$

If we generate another random vector, this time from a *different* linear transformation of $\boldsymbol{x}$

$$\boldsymbol{z} = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{D} \tag{E.7}$$

then the covariance *between* the random vectors $\boldsymbol{y}$ and $\boldsymbol{z}$ is given by

$$\boldsymbol{\Sigma}_{y,z} = \boldsymbol{F}\boldsymbol{\Sigma}_x\boldsymbol{G}^T \tag{E.8}$$

The i,jth entry in this matrix is the covariance between y_i and z_j.

E.2 The Multivariate Gaussian

The multivariate normal PDF for d variables is

$$N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{E.9}$$

where the mean $\boldsymbol{\mu}$ is a d-dimensional vector, $\boldsymbol{\Sigma}$ is a $d \times d$ covariance matrix, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

E.2.1 Entropy

The entropy is

$$H(\boldsymbol{x}) = \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{d}{2}\log 2\pi + \frac{d}{2} \tag{E.10}$$

E.2.2 Relative Entropy

For Normal densities $q(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ and $p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ the KL-divergence is

$$D[q||p] = 0.5 \log \frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} + 0.5 Trace(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) + 0.5(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p) - \frac{d}{2} \quad \text{(E.11)}$$

where $|\boldsymbol{\Sigma}_p|$ denotes the determinant of the matrix $\boldsymbol{\Sigma}_p$.

E.3 The Multinomial Distribution

If a random variable x can take one of one m discrete values $x_1, x_2, ..x_m$ and

$$p(x = x_s) = \pi_s \quad \text{(E.12)}$$

then x is said to have a multinomial distribution.

E.4 The Dirichlet Distribution

If $\boldsymbol{\pi} = [\pi_1, \pi_2, ...\pi_m]$ are the parameters of a multinomial distribution then

$$q(\boldsymbol{\pi}) = \Gamma(\lambda_{tot}) \prod_{s=1}^{m} \frac{\pi_s^{\lambda_s - 1}}{\Gamma(\lambda_s)} \quad \text{(E.13)}$$

defines a Dirichlet distribution over these parameters where

$$\lambda_{tot} = \sum_s \lambda_s \quad \text{(E.14)}$$

The mean value of π_s is $\lambda_s / \lambda_{tot}$.

E.4.1 Relative Entropy

For Dirichlet densities $q(\boldsymbol{\pi}) = D(\boldsymbol{\pi}; \boldsymbol{\lambda}_q)$ and $p(\boldsymbol{\pi}) = D(\boldsymbol{\pi}; \boldsymbol{\lambda}_p)$ where the number of states is m and $\boldsymbol{\lambda}_q = [\lambda_q(1), \lambda_q(2), .., \lambda_q(m)]$ and $\boldsymbol{\lambda}_p = [\lambda_p(1), \lambda_p(2), .., \lambda_p(m)]$. the KL-divergence is

$$\begin{aligned} D[q||p] &= \Gamma(\log \lambda_{qtot}) + \sum_{s=1}^{m} (\lambda_q(s) - 1)(\Psi(\lambda_q(s)) - \Psi(\lambda_{qtot}) - \log \Gamma(\lambda_q(s)) \quad \text{(E.15)} \\ &- \Gamma(\log \lambda_{ptot}) + \sum_{s=1}^{m} (\lambda_p(s) - 1)(\Psi(\lambda_q(s)) - \Psi(\lambda_{qtot}) - \log \Gamma(\lambda_p(s)) \end{aligned}$$

where

$$\lambda_{qtot} = \sum_{s=1}^{m} \lambda_q(s) \tag{E.16}$$

$$\lambda_{ptot} = \sum_{s=1}^{m} \lambda_p(s)$$

and $\Psi()$ is the digamma function.