1 Assigning Probabilities

Given some constraints on the uncertainty of a variable, the question that arises is what probability distribution is reasonable to be assigned to represent the uncertainty in the variable. The best approach is to choose the probability distribution that corresponds to the larger uncertainty. This choices arises from the concept of maximum information entropy. The maximum information entropy is often used to assign the prior probability distribution for uncertain variables.

1.1 Information Entropy

The concept of **information entropy** has been introduced by Shannon to describe the information content of an event. The information entropy is a unique scalar measure of the uncertainty in a variable (or random variable). The information entropy for a probability distribution $p\underline{\theta}$ is defined by

$$I(p) = \mathbb{E}_{\underline{\theta}}[-\log[p(\underline{\theta})]] = -\int_{\Theta} p(\underline{\theta}) \log[p(\underline{\theta})] \, \mathrm{d}\underline{\theta}, \tag{1}$$

where Θ is the domain of definition of the parameter space or the support of the PDF of the parameters in $\underline{\theta}$. $I(p)$ gives a measure of the uncertainty in $\underline{\theta}$. The more uncertainty in the value of $\underline{\theta}$, the higher the information entropy.

1.2 Information entropy for the univariate normal (Gaussian) distribution

The information entropy for a univariate normal (Gaussian) distribution, $\theta \sim N(\mu, \sigma^2)$, given by

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(\theta - \mu)^2\right] \tag{2}$$

is

$$I(p) = \mathbb{E}_{\theta}[-\log[p(\theta)]] = -\int_{-\infty}^{\infty} p(\theta) \left[-\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(\theta - \mu)^2\right] \, \mathrm{d}\theta \tag{3}$$

$$= \log(\sqrt{2\pi}\sigma) \underbrace{\int_{-\infty}^{\infty} p(\theta) \, \mathrm{d}\theta}_{=1} + \frac{1}{2\sigma^2} \underbrace{\int_{-\infty}^{\infty} (\theta - \mu)^2 \, p(\theta) \, \mathrm{d}\theta}_{=\sigma^2} \tag{4}$$

$$= \log(\sqrt{2\pi}\sigma) + \frac{1}{2}. \tag{5}$$

So

$$I(p) = \frac{1}{2}[\log(2\pi\sigma^2) + 1] \tag{6}$$

For the Gaussian distribution of a parameter, the information entropy $I(p)$ depends only on the standard deviation σ. The higher the value of the standard deviation σ, the higher the information entropy and the uncertainty in the parameter.

1.3 Information entropy for the multivariate normal (Gaussian) distribution

The multivariate normal (Gaussian) distribution of a set of parameters, $\underline{\theta} \sim N(\underline{\mu}, C)$, is given by

$$p(\underline{\theta}) = \frac{1}{(\sqrt{2\pi})^n |C|^{1/2}} \exp\left[-\frac{1}{2}(\underline{\theta}-\underline{\mu})^T C^{-1}(\underline{\theta}-\underline{\mu})\right]$$

In this case the information entropy takes the form

$$I(p) = \mathbb{E}_{\underline{\theta}}\left[-\log p(\underline{\theta})\right] = -\int p(\underline{\theta}) \log[p(\underline{\theta})]\, \mathrm{d}\underline{\theta} =$$

$$-\int p(\underline{\theta}) \left[-\log\left\{\left(\sqrt{2\pi}\right)^n |C|^{1/2}\right\} - \frac{1}{2}(\underline{\theta}-\underline{\mu})^T C^{-1}(\underline{\theta}-\underline{\mu})\right] \mathrm{d}\underline{\theta} =$$

$$\log\left\{\left(\sqrt{2\pi}\right)^n |C|^{1/2}\right\} \underbrace{\int p(\underline{\theta})\, \mathrm{d}\underline{\theta}}_{1} + \frac{1}{2} \underbrace{\int p(\underline{\theta})(\underline{\theta}-\underline{\mu})^T C^{-1}(\underline{\theta}-\underline{\mu})\, \mathrm{d}\underline{\theta}}_{n} \qquad (7)$$

which means that the information entropy is

$$I(p) = \log\left\{\left(\sqrt{2\pi}\right)^n |C|^{1/2}\right\} + \frac{n}{2} = \frac{1}{2}[\log\{(2\pi)^n |C|\} + n]$$

and depends on the determinant of the covariance matrix.

To prove that the second integral in (7), introduce the new vector $\underline{y}$ by the transformation

$$\underline{\theta} - \underline{\mu} = \Phi\sqrt{\Lambda}\underline{y} \qquad (8)$$

where full matrix Φ and the diagonal matrix Λ contain the eigenvectors and the eigenvalues of the covariance matrix C, i.e. they satisfy

$$C\Phi = \Phi\Lambda \qquad (9)$$

Note that $\Phi\Phi^T = I$, $C = \Phi\Lambda\Phi^T$ and $C^{-1} = \Phi\Lambda^{-1}\Phi^T$. Thus taking expectations in (8) one readily derives that $\underline{0} = \mathbb{E}[\underline{\theta}-\underline{\mu}] = \mathbb{E}[\Phi\sqrt{\Lambda}\underline{y}] = \Phi\sqrt{\Lambda}\mathbb{E}[\underline{y}]$, which yields $\mathbb{E}[\underline{y}] = \sqrt{\Lambda}^T \Phi^T \underline{0} = \underline{0}$, i.e that the new parameter vector $\underline{y}$ has mean zero. The covariance matrix of $\underline{y}$ is obtained by noting that

$$C = \mathbb{E}[(\underline{\theta}-\underline{\mu})(\underline{\theta}-\underline{\mu})^T] = \mathbb{E}[\Phi\sqrt{\Lambda}\underline{y}\underline{y}^T\sqrt{\Lambda}\Phi^T] = \Phi\sqrt{\Lambda}\mathbb{E}[\underline{y}\underline{y}^T]\sqrt{\Lambda}\Phi^T \qquad (10)$$

Multiplying the equation by $\sqrt{\Lambda}^{-1}\Phi^T$ from the left and $\Phi\sqrt{\Lambda}^{-1}$ from the right and using that $C = \Phi\Lambda\Phi^T$ and $\Phi\Phi^T = I$, one derives that

$$\sqrt{\Lambda}^{-1}\Phi^T C\Phi\sqrt{\Lambda}^{-1} = \mathbb{E}[\underline{y}\underline{y}^T] \tag{11}$$

or equivalently

$$\mathbb{E}[\underline{y}\underline{y}^T] = \sqrt{\Lambda}^{-1}\Phi^T\Phi\Lambda\Phi^T\Phi\sqrt{\Lambda}^{-1} = \sqrt{\Lambda}^{-1}\Lambda\sqrt{\Lambda}^{-1} = I \tag{12}$$

Thus the new parameter vector $\underline{y}$ has covariance matrix equal to I. This means that the elements in the parameter vector $\underline{y}$ are standard normal variables following the distribution $y_i \sim N(0,1)$.

The second integral in (7) can thus simplify to

$$\begin{aligned}
\int p(\underline{\theta})(\underline{\theta}-\underline{\mu})^T C^{-1}(\underline{\theta}-\underline{\mu})\,\mathrm{d}\underline{\theta} &= \int p(\underline{y})(\Phi\sqrt{\Lambda}\underline{y})^T C^{-1}(\Phi\sqrt{\Lambda}\underline{y})\,\mathrm{d}\underline{y} && (13)\\
&= \int p(\underline{y})\underline{y}^T\sqrt{\Lambda}(\Phi^T C^{-1}\Phi)\sqrt{\Lambda}\underline{y}\,\mathrm{d}\underline{y} && (14)\\
&= \int p(\underline{y})\underline{y}^T\sqrt{\Lambda}(\Lambda^{-1})\sqrt{\Lambda}\underline{y}\,\mathrm{d}\underline{y} && (15)\\
&= \int \underline{y}^T\underline{y}p(\underline{y})\,\mathrm{d}\underline{y} && (16)\\
&= \int \sum_{i=1}^{n} y_i^2 p(\underline{y})\,\mathrm{d}\underline{y} && (17)\\
&= \sum_{i=1}^{n}\int y_i^2 p(\underline{y})\,\mathrm{d}\underline{y} && (18)\\
&= \sum_{i=1}^{n}\int y_i^2 p(y_1)\dots p(y_n)\,\mathrm{d}y_1\dots y_n && (19)\\
&= \sum_{i=1}^{n}\underbrace{\int y_i^2 p(y_i)\,\mathrm{d}y_i}_{1}\prod_{j\neq i}\underbrace{\int p(y_j)\,\mathrm{d}y_j}_{1} && (20)\\
&= \sum_{i=1}^{n} 1\prod_{j\neq i} 1 = \sum_{i=1}^{n} 1 = n && (21)
\end{aligned}$$

which proves the result in the second integral in (7).

Introduce the matrix $H(\underline{\theta}) := C^{-1}(\underline{\theta})$ and let $\lambda_1, ..., \lambda_n$ be the eigenvalues of $H(\underline{\theta})$ and $\sigma_1^2, ..., \sigma_n^2$ be the eigenvalues of $C(\underline{\theta})$. Then

$$\sigma_i^2 = \frac{1}{\lambda_i} \Rightarrow \sigma_i = \frac{1}{\sqrt{\lambda_i}} \;\; \forall i = 1, ..., n$$

and

$$|C| = \sigma_1^2 \cdot ... \cdot \sigma_n^2 = \frac{1}{\lambda_1} \cdot ... \cdot \frac{1}{\lambda_n}.$$

Now one can rewrite the the information entropy in the form:

$$I(p) = \frac{1}{2}\log\left[(2\pi)^n \frac{1}{\lambda_1}\cdot ... \cdot \frac{1}{\lambda_n}\right] + \frac{n}{2} \tag{22}$$

$$= \frac{1}{2}\log\left[(2\pi)^n \sigma_1^2 \cdot ... \cdot \sigma_n^2\right] + \frac{n}{2} \tag{23}$$

Consider the special case of $n = 2$. The spread of uncertainty is proportional to $|C| = \prod_i \sigma_i^2 = \prod_i \lambda_i^{-1}$ (see Figure 1). This means that the larger the σ_i the larger the $|C|$, the higher the information entropy and the uncertainty in the parameter vector $\underline{\theta}$.

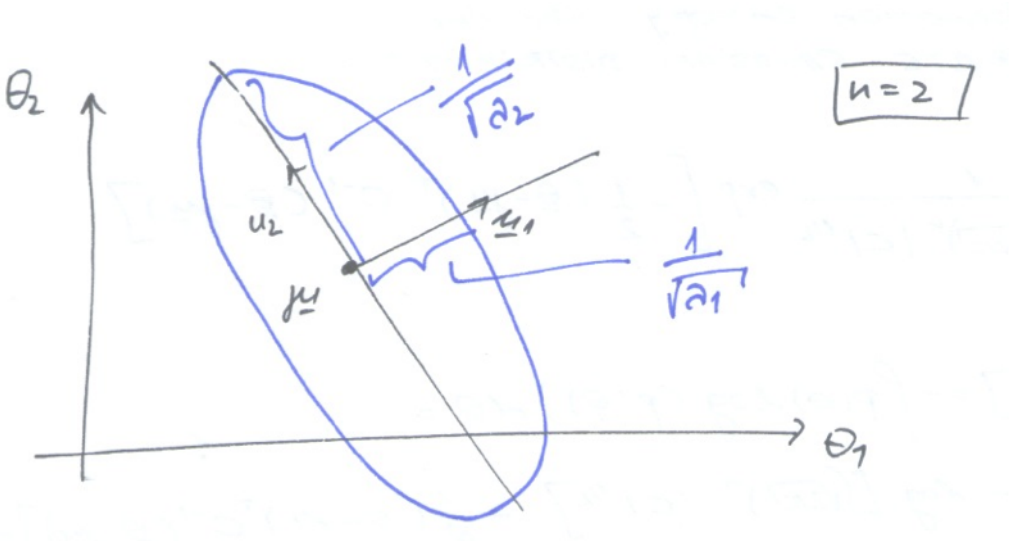

Figure 1: Spread of uncertainty for $n = 2$.

1.4 Principle of maximum information entropy (PMIE)

The PMIE is a theoretical approach for assigning probabilities, such as prior PDFs in Bayesian inference, based on the available information. It states that among all PDFs, the PDF which best represents the current state of knowledge is the one with the largest information entropy.

The *maximum entropy distribution* represents the *least informative* distribution given the constraints (knowledge) about the parameters.

The maximum entropy distribution is found by maximizing the information on entropy with respect to the distribution, given the prescribed constraints (e.g. moments or bounds of distribution).

1.4.1 Example: Maximum entropy distribution given mean and variance

Consider the case when the mean and the variance of an uncertain scalar parameter θ are given. The distribution $p(\theta)$ with the least information or the highest uncertainty is found by maximizing the information entropy

$$I(p) = \mathbb{E}[-\log(p(\theta))] = -\int p(\theta)\log(p(\theta))\,\mathrm{d}\theta \tag{24}$$

subject to constraints related to the mean μ and variance σ^2 of the distribution

$$\int \theta p(\theta)\,\mathrm{d}\theta = \mu, \tag{25}$$

$$\int (\theta-\mu)^2 p(\theta)\,\mathrm{d}\theta = \sigma^2 \tag{26}$$

and the constraint related to the fact that any PDF $p(\theta)$ has to integrate to one

$$\int p(\theta)\,\mathrm{d}\theta = 1 \tag{27}$$

This optimization problem can be solved using calculus of variations. First the constrained optimization problem is transformed to an unconstrained one by introducing Lagrange multipliers $\lambda_1, \lambda_2, \lambda_3$ and the Lagrange function

$$L(p) = I(p) + \lambda_1\left[\int \theta p(\theta)\,\mathrm{d}\theta - \mu\right] + \lambda_2\left[\int (\theta-\mu)^2 p(\theta)\,\mathrm{d}\theta - \sigma^2\right] + \lambda_3\left[\int p(\theta)\,\mathrm{d}\theta - 1\right] \tag{28}$$

Now the problem is reduced to minimizing $L(p)$ with respect to p which is equivalent of fonding the function p which satisfies $\delta L(p) = 0$.

This is a calculus of variation problem for a function $\delta F(p)$ of the form $F(p) = \int f(p,\theta)\,\mathrm{d}\theta$. The variation of the function $F(p)$ is given by

$$\delta F(p) = \int \frac{\partial f(p,\theta)}{\partial p}\delta p\,\mathrm{d}\theta = 0$$

for an arbirtary δp which yields that

$$\frac{\partial f(p,\theta)}{\partial p} = 0 \tag{29}$$

The equation (29) can be solved for p as a function of $\lambda_1, \lambda_2, \lambda_3$. To check that the found p is a true maximum, one has to check that

$$\frac{\partial^2 f(p,\theta)}{\partial p^2} < 0.$$

Starting from (28) and using (24) one has:

$$L(p) = \int \left\{-\log(p(\theta)) + \lambda_1\theta + \lambda_2(\theta-\mu)^2 + \lambda_3\right\} p(\theta)\,\mathrm{d}\theta +$$

$$\int (-\lambda_1\mu - \lambda_2\sigma^2 - \lambda_3)h(\theta)\,\mathrm{d}\theta,$$

where $h(\theta)$ denotes an arbitrary function which integrates to one.

Now $L(p)$ is represented as an integral of the function

$$f(p,\theta) = \left\{-\log(p(\theta)) + \lambda_1\theta + \lambda_2(\theta-\mu)^2 + \lambda_3\right\} p(\theta) -$$

$$(\lambda_1\mu + \lambda_2\sigma^2 + \lambda_3)h(\theta)$$

and the maximum condition is given by

$$0 = \frac{\partial f(p,\theta}{\partial p} = -1 - \log(p(\theta)) + \lambda_1\theta + \lambda_2(\theta-\mu)^2 + \lambda_3.$$

Solving with respect to $p(\theta)$, the desired PDF is

$$p(\theta) = \exp[g(\theta)]$$

where

$$g(\theta) = (\lambda_1 - 2\lambda_2\mu)\theta + \lambda_2(\theta^2 + \mu^2) + \lambda_3 - 1$$

Note that the function $g(\theta)$ is quadratic in θ and can be re-writtem in the form

$$g(\theta) = \lambda_2(\theta - A)^2 + B,$$

where

$$A = \frac{2\lambda_2\mu - \lambda_1}{2\lambda_2},$$

$$B = \lambda_2\mu^2 + \lambda_3 - 1 - \lambda_2\frac{(2\lambda_2\mu - \lambda_1)^2}{4\lambda_2^2} = \lambda_3 - 1 + \lambda_1\mu - \frac{\lambda_1^2}{4\lambda_2^2}.$$

Thus, the PDF $p(\theta)$ is given by

$$p(\theta) = e^{\lambda_2(\theta-A)^2+B} \tag{30}$$

Substituting $p(\theta)$ to the constraint equations, one has

$$\int \theta e^{\lambda_2(\theta-A)^2+B}d\theta = \mu \tag{31}$$

$$\int (\theta-\mu)^2 e^{\lambda_2(\theta-A)^2+B}\,\mathrm{d}\theta = \sigma^2 \tag{32}$$

$$\int e^{\lambda_2(\theta-A)^2+B}\,\mathrm{d}\theta = 1 \tag{33}$$

which, after carrying out the integration analytically (**Exercise**), yield $\lambda_1 = 0$; $\lambda_2 = -\frac{1}{2\sigma^2}$; $\lambda_3 = 1-\ln(\sqrt{2\pi}\sigma)$. Substituting these values to the expressions for A and B gives $A = \mu$ and $B = \lambda_3 - 1 = -\ln(\sqrt{2\pi}\sigma)$. Substituting the derived expressions for λ_2, A and B to (30) for $p(\theta)$ gives

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(\theta-\mu)^2} \tag{34}$$

which is a Normal distribution with mean μ and variance σ^2.

Finally, noting that the second derivative

$$\frac{\partial^2 f(p,\theta)}{\partial p^2} = -\frac{1}{p} < 0, \tag{35}$$

since $p(\theta)$ as a distribution function is always positive, it is clear that the derived PDF $p(\theta)$ maximized $L(p)$ and the information entropy.

1.4.2 Remarks

1. The maximum entropy distribution for a bounded distribution within the interval [a, b] with given mean μ and variance σ^2 is a **truncated Gaussian distribution** given by

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma\left[\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})\right]}e^{-\frac{1}{2\sigma^2}(\theta-\mu)^2} \tag{36}$$

where Φ is the cumulative distribution function of the standard Gaussian distribution with zero mean and unit variance, given by

$$\Phi(\theta) = \frac{1}{\sqrt{2\pi}}\int_{-\theta}^{\theta} e^{-\frac{s^2}{2}}ds = \frac{1}{2} + \frac{1}{2}erf\left(\frac{\theta}{\sqrt{2}}\right) \tag{37}$$

2. The maximum entropy distibution defined within the interval [a, b] is the **uniform distribution**.

3. The maximum entropy distribution given only mean is the **exponential distribution**.